COVID Needles in Social Media Haystacks: Identifying Cross-Language and Longitudinal Changes in Pandemic-Related Discussion Topics

Sarah Marvi[1], Valerie Novak[1], Michelle Morrison[1], Ruthanna Gordon[1], Tess Wood[1],
Sarah Oates[1], Anton Rytting[1], Kelly Jones[1], Shawn Janzen[1], and Mike Maxwell[1]

[1] University of Maryland, College Park MD 20742, USA
smarvi@terpmail.umd.edu; rgordon@arlis.umd.edu

Abstract. An "infodemic" of malinformation – misleading and harmful messaging – has spread alongside the COVID-19 pandemic and potentially interfered with uptake of public health measures. We use topic modeling to examine U.S. English- and Spanish-language Twitter discussion before and after the initial vaccine rollout, examining differences in vaccine-related messaging across time periods and language communities. We find that misinformation changes substantially across time periods with minimal continuity. English and Spanish misinformation reflects shared themes and concerns, but often plays out with different details (e.g., ostensible negative side effects of vaccines).

Keywords: Social Media, COVID-19, Topic Modeling.

1 Introduction

The COVID-19 pandemic has been paralleled--and likely exacerbated--by an "infodemic" of rumor, shifting guidance, and misleading information. Content and spread of COVID-related misinformation have been well-documented [1,2,3,4], as has the role of deliberate disinformation from a range of actors [5,6,7,8,9,10]. Most work has focused on single-language datasets. However, language plays an important role in access to accurate information and exposure to ideas and opinions, particularly when institutional information is shared primarily in a dominant language. It is therefore important to understand how discussions around the pandemic, including mitigation behaviors (e.g., masking, vaccine acceptance), vary across language communities.

As an initial effort toward this type of comparison, we carried out topic modeling on US English-language and Spanish-language subsets of a social media dataset built around COVID-focused keywords. Spanish is the second most commonly spoken language in the US [11]; while multilingual health information is increasing in availability, it remains more difficult to access such information in languages other than English [12]. Given findings that disinformation flourishes when gaps in credible information are more extensive, misleading online messages have the potential to exacerbate health disparities between language communities. Conversely, another possibility is that high-output disinformation sources focus primarily on dominant languages, thus creating more misleading content for English-language communities.

Comparing both across languages and across two time period samples, we examine the following questions:

1. How does discussion in US Twitter change before and after vaccine rollout?
2. How do English and Spanish vaccine discussions in US Twitter differ?
3. What trends in misinformation took off during this period, and do they differ across languages?

This work is part of a larger pilot project examining attitudes and health behaviors related to COVID and the COVID vaccine in different language communities in the US, and sets a foundation for more extensive and rigorous exploration of cross-language COVID discussions planned for the coming year.

2 Methods

We carried out analyses on a subset of Twitter data acquired from Zignal Labs. The initial dataset begins in January 2020 and continues through the present (June 2021), based on a multi-language search of a wide range of social and traditional media venues using COVID-related keywords. We examined English-language and Spanish-language US data, sampling from March-September 2020 in English and December 2020 - February 2021 in English and Spanish (sampling every 5th day from the English-language dataset to compensate for the larger amount of data).

Topic modeling was performed by adapting code from Kapadia (2020)[13]. Tweets were preprocessed by removing punctuation, removing twitter URL short links, and making all letters lowercase. Duplicates were removed after cleaning the text. After preprocessing the data, there were a total of 767,623 tweets. Next stopwords were removed using the nltk package. In addition to the premade set of stopwords, the stopwords list was extended to include ‘from’, ‘subject’, ‘re’, ‘edu’, and ‘use’. Tokens were lemmatized using the spaCy package with the “en_core_web_sm” pipeline. The topic model with the highest coherence score was selected for analysis. These coherence values were calculated using the gensim LDA multicore model with 1 pass. Using parameter settings of alpha = 0.01, beta = 0.9 topic number = 6, resulted in the final model used which has a coherence score of 0.355.

2.1 Topic Modeling: Spanish

For the Spanish dataset, keywords included several prominent pharmaceutical and biotech companies that also appeared in the English dataset, resulting in a final keyword list of ‘vacuna’, ‘vacunas’, vacunacion’, ‘Pfizer’, ‘Moderna’, and ‘BioNTech’. We also expanded the search settings to include tweet subtypes ‘quote’, ‘reply’, and ‘mention’. With these modifications, the final compiled dataset included about 50,000 total tweets during the timeframe of December 3, 2020 to February 9, 2021.

The process for topic modeling the Spanish Twitter data is the same as for the English data. An additional step for data cleaning was added; usernames that were in the text were removed due to the inclusion of "mentions" For removing stopwords from the text, the nltk package was used to create a list of Spanish stopwords. No additional words were added to the list of stopwords. For performing lemmatization on the data, the spaCy package was used with the "es_core_news_sm" pipeline. Using parameter settings of alpha = 0.01, beta = 0.9 topic number = 4, resulted in the final model used which has a coherence score of 0.327, which was chosen to remove significant overlap between topics on the topic visualization map provided by pyLDAvis.

3 Results

Each of our topic models produced human-interpretable topics covering a range of expected vaccine opinions and information, summarized in Table 1.

Table 1. Topic summaries for Twitter samples.

Dataset	Topics
English (pre-rollout)	1) Routine vaccinations, public health, mandates 2) COVID vaccine development cycle 3) Biotechnology methods, including application to COVID vaccines 4) Anger and skepticism, pro- and anti-vaccine. General mistrust, safety concerns, conspiracy theories 5) Minimal coherent themes: International relationships, spying, stealing technology, animal vaccines 6) Global competition for vaccine success, economic impact and financial benefits, supply chains
English (post-rollout)	1) Vaccine publicity and promotion, vaccine distrust 2) Vaccine memes and government vaccine acquisition 3) Vaccine Q&A, information, and research 4) "First shots" – individual reports and prioritization 5) Logistics for vaccine rollout, vaccine supply chain 6) Problems with rollout, critique of government
Spanish (post-rollout)	1) Vaccine distribution 2) Conspiracy theories and complaints 3) Vaccine effectiveness, virus variants, refuting misinformation 4) No coherent theme

3.1 English (pre-rollout, March – September 2020)

After preprocessing the data, the English set included a total of 767,623 tweets. Topics 2 and 4 were the largest topics. Topic 2 focuses primarily on the process of developing and testing vaccines, and consists largely of links to news articles about trial progress and early results. Topic 4 includes messages both for and against vaccination (as well as non-pharmaceutical interventions like mask-wearing), appearing to originate from individuals. This topic includes considerable misleading information and conspiracy theories (e.g. Bill Gates is tracking people through microchips in the vaccine) as well as misleading frameworks on actual scientific reporting (e.g., the virus is never going to go away entirely, so protective measures are pointless).

Topic 1 includes primarily messaging in support of routine vaccinations from institutions and authorities. There is some discussion of at-the-time-hypothetical COVID vaccines, as well as mandates for both these and routine vaccines, including minimal generic anti-vaccination posts. Topic 3 focuses on biotechnology methods including their application to COVID vaccines, with occasional skeptical commentary. Topic 5 has less coherence than other topics, but does include discussion of international espionage in the vaccine development process. Topic 6 focuses largely on the finances of the development process, including allegations over malfeasance related to the head of Operation Warp Speed's industry investments, and concerns that companies might prioritize profit over distribution of effective vaccines.

3.2 English (post-rollout, December 2020 – February 2021)

The topic labels for the second English dataset reveal a marked shift in the discussion relating to vaccines. Generally, the conversation shifted from vaccine development to distribution, with increasing concern about access and the fairness or transparency of the distribution process. COVID vaccines are more clearly being discussed separately from other vaccines, with the latter appearing rarely if at all. Conspiracy theories from the earlier dataset grow less prominent, replaced by misinformation around vaccine impacts (e.g., exaggerating side effects severity, claiming vaccine-related deaths are being hidden, suggesting that politicians are getting fake vaccines for photo ops).

Topic 1 contains extensive discussion of pro- and anti-vaccination arguments, including the majority of misinformation and counterarguments. This topic appears to originate largely with individuals, with minimal repetition of phrasing. Topic 2 contains more automated and repetitive messaging, primarily variations on "If you've ever X, don't worry about what's in the vaccine." It also includes some relatively repetitive discussion of vaccine investments and who profits from development and distribution. Topic 3 contains links to articles, town halls, and Q&As aimed at helping the public understand vaccines in greater detail, with a focus on research findings. This topic includes some mild skepticism, mostly around specific medical conditions (e.g., questions about vaccine interaction with immune disorders) with little reference to more misinformation-laden objections elsewhere. Topic 4 focuses on "first shots," both personal reports of gaining access for oneself or relatives, and arguments for prioritization of access. This topic is extremely pro-vaccine. Topic 5 messages provide a big-picture view of the distribution effort, consisting largely of announcements and articles regarding site openings and needs, shifts between access phases, and COVID-related aid. Finally, Topic 6 focuses on problems with the rollout, including critiques of politicians (largely Trump and the Trump administration), and concerns about vaccines being withheld from specific locations and populations.

3.3 Spanish (post-rollout)

After pre-processing the national U.S. dataset, we analyzed a subset of 44,606 Spanish tweets, detecting 4 topics. Topic 1 was themed around vaccine distribution, with considerable discussion about clinic availability, eligibility for vaccines, and appointments, as well as involvement by government agencies. Topic 2 centered around complaints and conspiracy theories, including the bulk of the Spanish-language misinformation. Topic 3, by contrast, was more positive about vaccines, discussing effectiveness and concerns about virus variants, as well as direct refutation of conspiracy theories and misinformation. Topic 4 did not appear to address a coherent theme.

Spanish discussion covers South American countries and politicians more often than English, while English discussion of international aspects of the pandemic focuses on global competition among major powers. Spanish discussion of vaccine development continued during the post-rollout period, where the English discussion shifted toward vaccine receipt and refusal – potentially reflecting differences in access.

4 Conclusions

Between the two time periods, conversation shifts beyond simply the change from vaccine anticipation to availability and pre- to post-election. Notably, there is discontinuity in the misinformation spreading across time periods, suggesting that anti-vaccination narratives may have been initiated in an ad hoc or opportunistic manner. Rather than substantial obvious input from QAnon and other malinformation sources with a strong “storyline,” we found simpler claims taking advantage of existing hesitancy and/or group identity. Conspiracy theories around Bill Gates and microchips appeared to peter out between time periods, with “Gates” appearing in a full 4.8% in the first English dataset, but only .5% of the second set. That discussion was even sparser in the post-rollout Spanish data, comprising only .2% of those tweets.

Post-rollout, both English and Spanish discussions contained extensive misinformation. There were many similarities in conspiracy theories and misinformation spreading among both communities, but also differences. Both languages, for example, include repeated claims about harm from the vaccine. The English set includes claims of unreported deaths and severe side effects, along with suggestions that the vaccine causes infertility. Similar Spanish claims also include vaccine-related infertility, as well as cancer or changes DNA (though the latter was addressed in counter-arguments more often than it appeared directly). Suggestions that the vaccine doesn’t work at all were more prevalent in Spanish.

In both languages, misinformation often reflected distrust of elites. In English this played out via claims that celebrities and politicians knowingly faked vaccine photo-ops, getting placebos in order to encourage the public to take a risky treatment. Spanish claims suggested that vaccines were part of an experiment by pharmaceutical companies, rich people, or the illuminati; or that they were being released primarily to make money for these elites. Distrust was also reflected in non-misleading discussions (e.g., complaints about politicians getting early access to vaccines).

References

1. Gallotti, R., Valle, F., Castaldo, N., Sacco, P., De Domenico, M.: Assessing the risks of ‘infodemics’ in response to COVID019 epidemics. Nature Human Behavior 4, 1285-1293 (2020)
2. Gottlieb, M., Dyer, S.: Information and disinformation: Social media in the COVID-19 crisis. Academic Emergency Medicine 27(7), 640-641 (2020)
3. Imhoff, R., Lamberty, P.: A Bioweapon or a Hoax? The Link Between Distinct Conspiracy Beliefs About the Coronavirus Disease (COVID-19) Outbreak and Pandemic Behavior. Social Psychology and Personality Science 11(8), 1110-1118 (2020)
4. Tagliabue, F., Galassi, L., Pierpaolo, M.: The “pandemic” of disinformation in COVID-19. SN Comprehensive Clinical Medicine 2, 1287-1289 (2020)
5. Bright, J., Au, H., Bailey, H., Elswah, M., Schliebs, M., Marchal, N., Schweiter, C., Rebello, K., Howard, P.N.: Coronavirus coverage by state-baked English-language news sources: Understanding Chinese, Iranian, Russian and Turkish government media (COMPROB data memo). Oxford Internet Institute, Oxford UK (2020).
6. Digital Forensics Lab: Weaponized: How rumors About COVID-19’s origins led to a narrative arms race. The Atlantic Council, Washington D.C. (2020)
7. Evanega, S., Lynas, M., Adams, J., Smolenyak, K.: Coronavirus misinformation: Quantifying sources and themes in the COVID-19 'infodemic.' The Cornell Alliance for Science, Ithaca NY (2020)
8. Lucas, E., Morris, J., Rebegea, C.: Information bedlam: Russian and Chinese information operations during COVID-19. Center for European Policy Analysis, Washington, D.C. (2020)
9. Nguyen, A., Catalan-Matamoros, D.: Digital mis/disinformation and public engagement with health and science controversies: Fresh perspectives from Covid-19. Media and Communication 8 (2), 323 - 328 (2020)
10. Romer, D., Jamieson, K.H.: Conspiracy theories as barriers to controlling the spread of COVID-19 in the U.S. Social Science and Medicine 263, 113356 (2020)
11. Top Languages Other than English Spoken in 1980 and Changes in Relative Rank, 1990-2010, https://www.census.gov/programs-surveys/sis/resources/visualizations/spoken-languages.html, last accessed 2021/5/26.
12. Roozenbeek, C.R., Schneider, S.D., Kerr, J., Freeman, A.L.J., Recchia, G., van der Bles, A.M., Van der Linden, S.: Susceptibility to misinformation about COVID-19 around the world. Royal Society Open Science 7(10) (2020)
13. Kapadia, S. Evaluate Topic Models: Latent Dirichlet Allocation (LDA). https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0, last accessed 2021/6/14 (2020)