LOCO: the topic-matched corpus for studying conspiracy theories

Alessandro Miani[a]

[a]University of Bristol, School of Psychological Science, 12a, Priory Road, Bristol BS8 1TU, United Kingdom

Abstract

LOCO is an 88-million-token corpus of conspiracy and mainstream standalone documents (N = 96,743) gathered from 150 websites. LOCO has been built as a freely available resource to help researchers extract features, test hypotheses, and generate predictive and classification models. With its rich set of additional data (freely available at https://osf.io/snpcg), LOCO allows for the study of both the content and spread of CTs, permitting comparisons with a topic-matched corpus of mainstream documents.

1 Introduction

Conspiracy theories (CTs) explain significant social events as being secretly orchestrated by powerful and malicious elites (Douglas et al., 2019). Associated with detrimental societal outcomes, such as vaccine hesitancy, climate skepticism, and non-normative political engagement (Jolley et al., 2022), CTs pose substantial threats to democracy and public health (Ecker et al., 2024). Therefore, understanding the content of CTs is crucial to combating their spread and mitigating their consequences. To achieve this goal, the first step is to build comprehensive corpora of CTs.

Since the advent of Web 2.0, studies devoted to understanding CTs have typically focused on user-generated texts from social media (e.g., Klein et al., 2019; Samory & Mitra, 2018). However, posts and comments found on social media are not CTs per se, as they generally consist of short texts and are often tied to a thread. A better approach to address these limitations is to study CTs as texts from web pages. Compared to posts and tweets on social media, web pages provide space for in-depth and elaborated discourse, constituting standalone and structured texts.

Yet, even when researchers consider web pages, the resource-intensive nature of manual coding prevents the creation of corpora with sufficient statistical power for language analysis, typically limiting studies to a few hundred documents (see e.g., Sak et al., 2015). Thus, a solution is to automate text extraction by first compiling a list of websites delivering CTs and then extracting their content. For this purpose, pre-compiled lists of misinformation websites exist, such as the proprietary NewsGuard[1] and other freely available lists, like MediaBiasFactCheck (MBFC),[2] the Misinformation Domains dataset,[3] and the work of Lin et al. (2023). Works relying on these lists have typically reached large sample sizes, consisting of hundreds of thousands to millions of documents (Carrella et al., 2023; Miani, Carrella, & Lewandowsky, 2024).

However, building a corpus solely of conspiracy documents is not sufficient because it does not enable comparisons beyond descriptive analyses. To systematically study the language of CTs, researchers need a control corpus that allows for between-group comparisons. Ideally, these two corpora should be matched by topics, meaning that for each topic, there is both a conspiracy and a non-conspiracy (mainstream) version. Such a structure facilitates language comparison and helps identify discriminating features of conspiratorial language. We thus created LOCO, a topic-matched corpus of web pages delivering CTs.

2 LOCO

LOCO (the **L**anguage **O**f **CO**nspiracy theories; Miani et al., 2021) is a turnkey resource for understanding the language of CTs that includes 23,937 conspiracy and 72,806 mainstream documents matched by topics that revolve around events that have generated CTs. Each document in LOCO is associated with both fine- and coarse-grained semantic indexing as well as a measure of spread.

2.1 Building LOCO

LOCO was built by retrieving documents via Google searches using seeds associated with events that generated CTs on websites previously classified as either conspiratorial or mainstream. Figure 1 illustrates the workflow to build LOCO. We paired a set of websites (W_i) with a set of seeds (S_j) to generate a series of Google queries ($Q_{ij} = W_i \times S_j$). From each query, the HTML results page was parsed to extract the URLs needed to extract texts.

We started by compiling the two lists of seeds and websites. Seeds are search terms used to retrieve topic-specific text associated with popular and timely CTs (N = 47; e.g., *Princess Diana's death*, *Moon Landing*, *9/11*). The list of conspiracy websites (N = 58; e.g., `infowars.com`) was obtained by selecting websites with the highest conspiratorial rating from MBFC. The list of mainstream websites (N = 92; e.g., `bbc.com`) was generated in a data-driven fashion by extracting websites returned by Google for each seed.

We combined the two lists to generate Google search queries that return web pages containing the seed's terms within a website (e.g., "`site:bbc.com climate change`"). Google was chosen because it allows for automated URL extraction and provides consistent search criteria across all websites, avoiding biases introduced by site-specific search engines. To ensure the results were in English, we added "`hl=en`" to the queries. For each results page, we parsed the HTML to extract URLs pointing to candidate web pages. The R scripts for reproducing LOCO's construction are available at https://osf.io/4vk7f.

The human-readable text was extracted from the HTML page via the Python package *goose*,[4] chosen for its superior performance from our tests compared to other packages. Next, we cleaned the corpus by removing documents where the percentage of the top 1,000 English words was below 40%, and those with word counts outside ± 2.5 standard deviations around the mean. The final word count range for each document was between 100 and 10,000 words.

[1] https://www.newsguardtech.com
[2] https://mediabiasfactcheck.com
[3] https://github.com/JanaLasser/misinformation_domains
[4] https://github.com/goose3/goose3

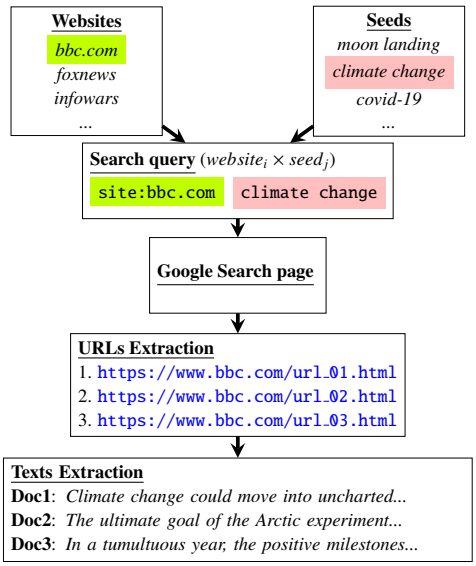

Fig. 1: LOCO's construction workflow

2.2 Extracting Features from LOCO

LOCO includes a comprehensive set of additional data. At the document level, beyond typical text statistics (such as word, sentence, and paragraph counts) and the seeds used to retrieve the web pages, each text is provided with fine-grained semantic fingerprinting (see Box .1), including lexical features and topics. Lexical features (N = 286) were extracted using two popular word-counting tools, LIWC and Empath. Topics (N = 600) were extracted via topic modeling at three different resolutions (100, 200, and 300 topics) and labeled using the top 15 most important terms. We also included a measure of whether a text is a representative instance of the conspiracy subcorpus (N = 4,277). This was achieved by computing the similarity of each document to the overall conspiracy subcorpus. High similarity indicates that the document exhibits prototypical conspiratorial language. Each document also includes metrics of social media engagement, such as Facebook shares, comments, and reactions, obtained via the web tool SharedCount.[5]

At the website level, we gathered information from MBFC about each website's political bias, factual reporting, and level of pseudoscientific claims. From the web tool SimilarWeb,[6] we collected metrics such as monthly global visits and rank, as well as the percentages for each type of incoming traffic (e.g., direct access, Google search, or social media).

2.3 Using LOCO

Given the large number of variables offered in LOCO, there is a risk of *fishing expeditions*, i.e., testing numerous variables simultaneously and then HARKing (Hypothesizing After the Results are Known; Hills & Miani, in press). When testing multiple variables at once, researchers should consider correcting for multiple tests to limit the rate of false positives. Furthermore, the unrestricted availability of LOCO complicates pre-registrations, as researchers might already be familiar with the data before pre-registering (Søgaard et al., 2023). Leveraging LOCO's topic-matching structure, researchers might consider multilevel modeling, cross-nesting documents within topics (or seeds) and/or websites.[7] Note also that the average explained variance (R^2) of individual lexical features is typically low (with an average marginal R^2 of .00055; Miani, van der Plas, & Bangerter 2024).

Semantic fingerprinting

Each document in LOCO is provided with two types of fine-grained semantic fingerprinting: lexical features and topic modeling.

The extraction of lexical features is usually done by counting words using dictionaries, which are pre-compiled lists of words or features categorized by topic (e.g., health), psychological dimension (e.g., affection), or typographical/grammatical category (e.g., punctuation, past-tense verbs). Two widely used and complementary sets of dictionaries are LIWC (a proprietary standalone program; Tausczik & Pennebaker, 2009) and Empath (an open-source Python package; Fast et al., 2016). The main difference between the two is that LIWC was constructed based on human coding, whereas Empath was assembled in a data-driven fashion and can generate ad hoc categories. The two dictionaries highly correlate with each other on similar categories (for a comprehensive discussion, see Hills & Miani, in press).

Topic modeling is based on Latent Dirichlet Allocation (LDA, Blei et al., 2003), which is an unsupervised probabilistic machine learning model capable of identifying co-occurring word patterns and extracting the underlying topic distribution for each text document within a corpus. Extracting LDA topics from a corpus requires researchers to set the desired number of topics (k): a larger number of topics provides a fine-grained resolution, while a smaller number yields more general topics (Colin & Murdock, 2020). Efforts are underway to develop unsupervised algorithms for identifying the optimal k, such as fitting an LDA model for each value of k provided by the researcher. This method has been used to estimate k from tweets, press articles, and web pages (Lasser et al., 2023; Mayor & Miani, 2023; Miani et al., 2022). However, applying such algorithms to large corpora can be resource- and time-intensive. Therefore, for large corpora, researchers have typically chosen to provide different topic resolutions (Carrella et al., 2023; Miani, Carrella, & Lewandowsky, 2024).

3 Works using LOCO

Since its release, LOCO has been a valuable resource for various research endeavors. Comprising texts with additional metadata, LOCO provides a comprehensive tool for analyzing the content and spread of CTs. LOCO's construction method is versatile and can be applied to various types of corpora beyond CTs. For example, it was employed to create DONALD, the topic-matched corpus of 2 million documents focused on politically polarizing topics across four types of ideological biases, including a large set of CTs (Miani, Carrella, & Lewandowsky, 2024).

Using LOCO's texts, researchers have applied various computational techniques to investigate conspiratorial discourse. Unlike the typically short posts on social media, the longer texts from LOCO provide a richer context for understanding conspiratorial incoherence (Miani et al., 2022). Results support the psychological notion that incoherence is a hallmark of conspiracism (Miani & Lewandowsky, 2024). Additionally, other studies have used LOCO to analyze term co-occurrence in gender identity issues (Fleckenstein, 2024) and to examine word-formation patterns in nominal

[5] https://www.sharedcount.com
[6] https://www.similarweb.com/corp/ourdata/
[7] In R, using *lme4*: `lmer(DV ~ subcorpus + (1|topic) + (1|website))`

compounds (Miani, van der Plas, & Bangerter, 2024). Future research could focus on parsing documents to identify narrative elements using syntactic rules, replicating methods applied to social media texts (Samory & Mitra, 2018; Tangherlini et al., 2020).

By simply counting words, researchers can investigate the psychological dimensions of conspiracism. Using Empath, we found that conspiracy (vs. mainstream) documents frequently use language related to deception, dominance, terrorism, and power (Miani et al., 2021). This aligns with language patterns observed among conspiracy believers on social media (Klein et al., 2019). We also discovered that representative (vs. other) conspiratorial documents tend to exhibit exaggerated conspiratorial language associated with social identification, negative emotions, and refutational rhetoric. These documents are also more frequently shared on Facebook. Perhaps unsurprisingly, conspiratorial language appears intensified in both mainstream and conspiracy documents that mention (real or theorized) conspiracies.

Besides texts, LOCO includes a rich set of supplementary data. Topics, seeds, and lexical features can be used as independent or dependent variables, or to further subset LOCO for specific analyses, such as examining health-related content across different ideologies (Reiter-Haas, 2023) or constructing networks based on topic co-occurrence patterns (Miani et al., 2022). On the website level, analysis of incoming traffic to LOCO's websites revealed that users of conspiratorial sites are more likely to gather information via bookmarked websites or through online social networks, rather than impartial search engines (Miani et al., 2021; Carrella et al., 2023). This behavior suggests the presence of confirmation bias.

Annotation schemes for CTs have been developed using LOCO (Fort et al., 2023; Mompelat et al., 2022). Future research should focus on developing tools for the automatic identification of CTs. For instance, leveraging LOCO's topic-matched structure, researchers can create dictionaries of CTs by contrasting mainstream and conspiracy narratives within each seed topic (e.g., *Lady Diana*), so to isolate conspiracy-specific terms, filtered from topic-specific words (e.g., *car crash*, *Paris*).

4 Conclusions

LOCO is a freely available topic-matched corpus from which researchers can extract features and generate predictive and classification models. The strength of LOCO lies in its richness in (meta-)data, which allows for the study of both the content and spread of CTs and facilitates comparisons with a topic-matched set of mainstream documents. The multilevel structure of LOCO allows researchers to consider the natural hierarchical grouping of documents cross-nested within websites and topics. This is useful for extracting conspiracy-specific linguistic markers. Concluding, LOCO is a rich resource that, while primarily providing data for lexical analysis and document spread, can also help to reveal psychological processes.

Acknowledgments

AM (ORCID: 0000-0001-6610-3510) is supported by the Swiss National Science Foundation (SNSF, project number 214293, "In/coherent worldviews").

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003, March). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*(null), 993-1022.

Carrella, F., Miani, A., & Lewandowsky, S. (2023, May). IRMA: the 335-million-word Italian coRpus for studying MisinformAtion. In A. Vlachos & I. Augenstein (Eds.), *Proceedings of the 17th conference of the european chapter of the association for computational linguistics* (pp. 2339–2349). Dubrovnik, Croatia: Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.171.

Colin, A., & Murdock, J. (2020). LDA Topic Modeling: Contexts for the History & Philosophy of Science. In G. Ramsey & A. de Block (Eds.), *Dynamics Of Science: Computational Frontiers in History and Philosophy of Science*. S.l.: Pittsburgh University Press.

Douglas, K. M., Uscinski, J., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., & Deravi, F. (2019, 2). Understanding conspiracy theories. *Political Psychology*, *40*(S1), 3–35. doi: 10.1111/pops.12568.

Ecker, U., Roozenbeek, J., van der Linden, S., Tay, L. Q., Cook, J., Oreskes, N., & Lewandowsky, S. (2024, June). Misinformation poses a bigger threat to democracy than you might think. *Nature*, *630*(8015), 29–32. doi: 10.1038/d41586-024-01587-3.

Fast, E., Chen, B., & Bernstein, M. S. (2016, May). Empath. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM. doi: 10.1145/2858036.2858535.

Fleckenstein, K. (2024). Representations of gender in conspiracy theories: a corpus-assisted critical discourse analysis. *Critical Discourse Studies*,(), 1–17. doi: 10.1080/17405904.2024.2334263.

Fort, M., Tian, Z., Indiana University, USA, Gabel, E., Indiana University, USA, Georgiades, N., . . . Indiana University, USA (2023). Bigfoot in Big Tech: Detecting Out of Domain Conspiracy Theories. In *Proceedings of the Conference Recent Advances in Natural Language Processing - Large Language Models for Natural Language Processings* (pp. 353–363). INCOMA Ltd., Shoumen, BULGARIA. doi: 10.26615/978-954-452-092-2_040.

Hills, T., & Miani, A. (in press). A short primer on historical natural language processing. In T. Hills & G. Pogrebna (Eds.), *Cambridge handbook of behavioral data science*. Cambridge University Press.

Jolley, D., Marques, M. D., & Cookson, D. (2022, October). Shining a spotlight on the dangerous consequences of conspiracy theories. *Current Opinion in Psychology*, *47*(), 101363. doi: 10.1016/j.copsyc.2022.101363.

Klein, C., Clutton, P., & Dunn, A. G. (2019, June). Pathways to conspiracy: The social and linguistic precursors of involvement in Reddit's conspiracy theory forum. *PLOS ONE*, *14*(11), e0225098. doi: 10.1371/journal.pone.0225098.

Lasser, J., Aroyehun, S. T., Carrella, F., Simchon, A., Garcia, D., & Lewandowsky, S. (2023). From alternative conceptions of honesty to alternative facts in communications by us politicians. *Nature human behaviour*, *7*(12), 2140–2151. doi: 10.1038/s41562-023-01691-w.

Lin, H., Lasser, J., Lewandowsky, S., Cole, R., Gully, A., Rand, D. G., & Pennycook, G. (2023, September). High level of correspondence across different news domain quality rating sets. *PNAS Nexus*, *2*(9), . doi: 10.1093/pnasnexus/pgad286.

Mayor, E., & Miani, A. (2023). A topic models analysis of the news coverage of the omicron variant in the united kingdom press. *BMC Public Health*, *23*(1), 1-18. doi: 10.1186/s12889-023-16444-7.

Miani, A., Carrella, F., & Lewandowsky, S. (2024). DONALD: the 2m-document dataset of news articles for studying the language of dubious information.. Retrieved from `https://osf.io/6xpa2`

Miani, A., Hills, T., & Bangerter, A. (2021). LOCO: The 88-million-word language of conspiracy corpus. *Behavior Research Methods*, *54*(4), 1794–1817. doi: 10.3758/s13428-021-01698-z.

Miani, A., Hills, T., & Bangerter, A. (2022). Interconnectedness and (in)coherence as a signature of conspiracy worldviews. *Science Advances*, *8*(43), 1–9. doi: 10.1126/sciadv.abq3668.

Miani, A., & Lewandowsky, S. (2024). Still very much dead and alive: Re-reconsidering belief in contradictory conspiracy theories. doi: 10.31219/osf.io/t6a54.

Miani, A., van der Plas, L., & Bangerter, A. (2024). Loose and tight: Creative formation but rigid use of nominal compounds in conspiracist texts. *The Journal of Creative Behavior*, *58*(1), 114–127. doi:

10.1002/jocb.633.
Mompelat, L., Tian, Z., Kessler, A., Luettgen, M., Rajanala, A., Kübler, S., & Seelig, M. (2022). How “loco” is the loco corpus? annotating the language of conspiracy theories. In *Proceedings of the 16th linguistic annotation workshop (law-xvi) within lrec2022*.
Reiter-Haas, M. (2023). Exploration of Framing Biases in Polarized Online Content Consumption. In *Companion Proceedings of the ACM Web Conference 2023* (pp. 560–564). Austin TX USA: ACM. doi: 10.1145/3543873.3587534.
Sak, G., Diviani, N., Allam, A., & Schulz, P. J. (2015, December). Comparing the quality of pro- and anti-vaccination online information: a content analysis of vaccination-related webpages. *BMC Public Health*, *16*(1), . doi: 10.1186/s12889-016-2722-9.
Samory, M., & Mitra, T. (2018). “the government spies using our webcams”: The language of conspiracy theories in online discussions. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–24. doi: 10.1145/3274421.
Søgaard, A., Hershcovich, D., & de Lhoneux, M. (2023). A two-sided discussion of preregistration of NLP research.
Tangherlini, T. R., Shahsavari, S., Shahbazi, B., Ebrahimzadeh, E., & Roychowdhury, V. (2020, June). An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: Bridgegate, Pizzagate and storytelling on the web. *PLOS ONE*, *15*(6), e0233879. doi: 10.1371/journal.pone.0233879.
Tausczik, Y. R., & Pennebaker, J. W. (2009, December). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. doi: 10.1177/0261927x09351676.