IRMA: the 335-million-word Italian coRpus for studying MisinformAtion

Fabio Carrella
School of Psychological Science
University of Bristol
fabio.carrella@bristol.ac.uk

Alessandro Miani
Institute of Work
and Organizational Psychology
University of Neuchâtel
alessandro.miani@unine.ch

Stephan Lewandowsky
School of Psychological Science
University of Bristol
stephan.lewandowsky@bristol.ac.uk

Abstract

The dissemination of false information on the internet has received considerable attention over the last decade. Misinformation often spreads faster than mainstream news, thus making manual fact checking inefficient or, at best, labor-intensive. Therefore, there is an increasing need to develop methods for automatic detection of misinformation. Although resources for creating such methods are available in English, other languages are often underrepresented in this effort. With this contribution, we present IRMA, a corpus containing over 600,000 Italian news articles (335+ million tokens) collected from 56 websites classified as 'untrustworthy' by professional fact-checkers. The corpus is freely available and comprises a rich set of text- and website-level data, representing a turnkey resource to test hypotheses and develop automatic detection algorithms. It contains texts, titles, and dates (from 2004 to 2022), along with three types of semantic measures (i.e., keywords, topics at three different resolutions, and LIWC lexical features). IRMA also includes domain-specific information such as source type (e.g., political, health, conspiracy, etc.), quality, and higher-level metadata, including several metrics of website incoming traffic that allow to investigate user online behavior. IRMA constitutes the largest corpus of misinformation available today in Italian, making it a valid tool for advancing quantitative research on untrustworthy news detection and ultimately helping limit the spread of misinformation.[1]

1 Introduction

Over the last decade, there has been an increase in worry about misinformation, which has led to numerous studies (e.g., Lazer et al. 2018; Pennycook and Rand 2019; Roozenbeek et al. 2020). This level of focus is justified by the threat that misinformation poses to individuals, institutions, and society in an increasingly digitalized world. Helped by social media capillarity and a lack of gatekeeping, misinformation is eroding long-standing institutional barriers, compromising democratic processes, as happened during the last US presidential elections (Allcott and Gentzkow, 2017; Dave et al., 2021), and producing serious sociopolitical uncertainty, as in the examples of global warming and COVID-19 vaccines (van der Linden et al. 2017; Loomba et al. 2021).

Currently, there are two main approaches used to detect misinformation online: manual and automatic. The first relies on human effort, mostly represented by fact-checking services that employ experts to manually verify the accuracy of claims, articles, and entire websites. The second is based on the identification of particular textual content features, usually performed through natural language processing (NLP) tools, e.g., deep learning models. Because misinformation spreads alarmingly faster than reliable news (Vosoughi et al. 2018; Gravino et al. 2022), automatic tools allow to detect and limit the spread of false news quickly and without involving costly human effort. These tools are usually trained on large sets of textual data, which are for the most part in English language (see e.g., Zubiaga et al., 2016; Potthast et al., 2017; Castelo et al., 2019; Miani et al., 2022b).

In a worldwide effort to fight misinformation, resources have been made available for Arabic, Spanish, Portuguese, and German (Alkhair et al., 2019; Posadas-Durán et al., 2019; Monteiro et al., 2018; Vogel and Jiang, 2019). However, to our knowledge, the Italian language has been overlooked. Several attempts have been undertaken to under-

[1] IRMA is freely available at https://osf.io/rywp4/.

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2339–2349

stand misinformation in the Italian context (see e.g., Bessi et al., 2015; Del Vicario et al., 2017), but these works focus on social media, and the data is not made publicly available.

The availability of an open-access dataset would substantially encourage research into the role of misinformation in the Italian context. A recent study conducted in Italy showed how the inability to recognize false information can obstruct public health campaigns (Moro et al., 2021). Misinformation in Italy has also been linked to political parties that have governed in recent years (Monti, 2020), as well as their voters (Cantarella et al., 2020). For example, Caldarelli et al. (2021) showed that right-wing parties were responsible for 96% of COVID-19-related untrustworthy news retweeted by political communities in Italy.

Considering the urgent need to address the societal problems caused by the spread of misinformation in Italy, we created IRMA (the Italian coRpus of MisinformAtion), a corpus containing over 600,000 Italian news articles scraped from websites classified as untrustworthy sources by professional fact-checkers.

2 Method

We decided to use source trustworthiness assessment as a proxy to identify the material of interest (cf. Grinberg et al. 2019; Pennycook et al. 2021). Therefore, we opted for two different misinformation databases, namely NewsGuard (NG, NewsGuard, 2020) and the Misinformation Domains (MD) dataset[2]. NG is a professional fact-checking database that provides indexes of trustworthiness for thousands of news domains. It rates domains in several categories related to news transparency and journalism ethics. The MD dataset is an open-source collection of domains referenced by Gallotti et al. (2020) and extended with other lists curated by fact-checking collectives, individual scholars, and journalists. We decided to use two different databases in order not to be too dependent on one individual source. We chose these two datasets since, differently from other misinformation databases, they comprise a considerable number of Italian sources. We also opted for domain-based rather than article-based fact checking to present a greater variety of data. We recognise that an article-based fact-checking service could have improved the "precision" of the material provided; however, we also think that domain-level fact-checkers represent an optimal balance between quantity and quality of (mis)information. We agree that not all sources deliver only unreliable news. However, having varying degrees of misinformation is an advantage. Future studies could manually annotate documents in IRMA to offer a fine-grained indicator of misinformation, helping the development of classifiers (Mompelat et al., 2022). Finally, the growing number of scholars who have used these two fact-checking databases attests to their reliability (e.g., Edelson et al. 2021; Bhadani et al. 2022; Lasser et al. 2022).

2.1 Corpus construction

We queried both databases (NG and MD) on June 8, 2022. We decided to collect data from a random limited sample of 80 untrustworthy domains in order to keep the database at a manageable size. Once we obtained the list of websites, we started collecting their content using *BeautifulSoup4* (Richardson, 2007), a Python package for parsing HTML documents. Since some of the domains were video-only news sources, paywall-protected websites, or extinct websites, the final number of scraped domains amounted to 56 websites.

Once we obtained the text documents from the websites, we started cleaning the corpus following the pipeline implemented in other works on misinformation (Miani et al., 2022b). In this order, we (1) removed duplicates, (2) selected texts within a word count range between 100 and 10,000 words (counted via white-space tokenization), and (3) removed non-Italian documents by selecting texts in which the percentage of Italian stop words (obtained from Benoit et al., 2021) was above 20% of the whole text (a threshold we chose after visual inspection).

The final corpus, IRMA, is composed of 634,932 documents ($N = 335,021,926$ tokens, $N = 1,137,168$ types) obtained from 56 websites, spanning a date range between 2004 and 2022, with an average document word count of 555 words ($SD = 554$, range: $101 - 9,993$).

2.2 Variables

Although it mostly consists of texts, IRMA also contains metadata such as documents' titles and urls[3], and dates (from 2004 to 2022). Envisioning

[2] https://github.com/JanaLasser/misinformation_domains

[3] Only valid for domains in MD (N = 22).

the possibility of analyzing IRMA without specific training in NLP, we provide a series of measures related to documents' semantic content such as *keywords*, *topics*, and *lexical features*, so that researchers (e.g., social scientists, psychologists) can download the datasets and start testing their hypotheses.

Documents' dates were obtained automatically via the package *BeautifulSoup4* (accounting for 79.74% of documents' dates). When the script was not able to retrieve webpage's date, we extracted the date from the URL of the document via regular expression. This allowed to obtain dates for 92.8% of IRMA's documents (see distribution in Figure 5 in the Appendix).

2.2.1 Pre-processing

Before extracting keywords and topics from documents, texts were pre-processed. Pre-processing was mostly done by removing stop words and infrequent (e.g., misspellings or extremely rare) words. The text cleaning pipeline was done using the *quanteda* R package (Benoit et al., 2018). The pipeline was as follow: (1) lower casing texts; (2) removing URLs, punctuation, numbers, separators, symbols, and split hyphens; (3) separating contractions; (4) removing stop words (obtained from Benoit et al., 2021); (5) lemmatization.[4] We then built the document-term matrix (DTM) and selected the top 10,000 features, reducing sparsity, i.e., removing rare words, from 99.98 to 98.24%. The DTM was finally composed of 634,932 documents and 10,000 terms, for a total of 167,049,425 types (without trimming, the DTM was composed of 1,137,168 terms accounting for 335,021,926 types).

Note that text pre-processing was done only for extracting keywords and topics from documents. IRMA's domains included in the MD dataset (N = 22) come with raw non-pre-processed texts, so researchers can apply any type of pre-processing depending on the task needed and based on specific theoretical grounding (see e.g., Hills and Miani, Forthcoming). Documents from domains classified by NG (N = 34), on the other hand, do not include raw texts, titles or links, due to policy restrictions. The articles for such domains are attached as DTM, and still retain all other features (see Section A.1).

[4] https://raw.githubusercontent.com/michmech/lemmatization-lists/master/lemmatization-it.txt

2.2.2 Keywords

Keywords were extracted from each document by computing the term frequency-inverse document frequency (TF-IDF), a technique that assesses the relevance of a word to a document in a corpus. For each word in a document, TF-IDF is computed by counting how many times a word appears in a document divided by the inverse document frequency of the word in the corpus. TF-IDF was computed using the function `dfm_tfidf` from the R package *quanteda*. Keywords were defined as words with the highest TF-IDF score per document. For all documents in IRMA, we obtained a total of 9,801 unique keywords (see Table 1). In addition, we attach to IRMA the top 10 TF-IDF scores for each document (see top-20 in Table 1).

2.2.3 Topics

Topics were extracted via Latent Dirichlet Allocation, or 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. Different from keywords, topics offer a fine-grained indexing of semantic content. Extracting LDA topics from a corpus requires researchers to set a number of topics (k) desired: if a fine-grained resolution is required, then a large number of topics is better; if the number of topics is small, these topics become more general (Colin and Murdock, 2020). Using the *topicmodels* R package (Grün and Hornik, 2011), we extracted three different topic resolutions, setting k at 20, 100, and 200 topics, hence obtaining a total of 320 different topics. Within a set of k topics, for each document in IRMA, topics are expressed as probabilities, hence summing to 1 (note that if all 320 topics are taken, then the sum is 3). The topic for a document with the highest (γ) value is the topic with the highest probability of being represented in such a document, followed by the probabilities of other topics. Note also that we did not provide labels for topics. Instead, we provide the top 10 words per each topic which, taken together, summarize the topic's content (Nguyen et al., 2020).

2.2.4 Lexical features

Lexical features were extracted from the raw texts with the Linguistic Inquiry and Word Count (LIWC, version 2022, Boyd et al., 2022), relying on the most recent Italian translation (Agosti and Rellini, 2007). LIWC is a widely-used standalone

application that extracts psychologically meaningful features from texts (Tausczik and Pennebaker, 2010), also in Italian (see e.g., Trevisan et al., 2021). LIWC analyzes texts and checks whether words are included in predefined categories (e.g., negative and positive emotions, social ties, etc.); if so, values associated with the matched categories increases. Different from topic modelling (in which topics' probabilities sum for each document), categories in LIWC are expressed as percentages of words in a document associated with a category and hence, if a word appears in two categories, they overlap. For example, the category `anxiety` (composed of words such as *anxious*, *avoid*, *insecure*) is also a subgroup of the category `negative_emotions`.

2.2.5 Websites metadata

Note that due to proprietary data, all NG's websites are anonymized via a unique website ID (e.g., *website1*, *website2*). Nevertheless, for all websites, we provide a measure of website's quality of information (an aggregated measure of bias, factuality, credibility, and transparency where higher scores correspond to higher quality domains, see Lin et al., 2022). For each website, we also extracted (in October 2022, from SimilarWeb[5]) a set of metadata about websites' incoming traffic such as monthly visits, visit duration, bounce rate (the percentage of visitors who leave after visiting only one page), and pages visited. Incoming traffic is further partitioned into direct traffic (reaching the website by typing the URL on the web browser or recalling it from bookmarks), from a search engines (e.g., using Google), from referrals (when a website is reached through another website), and from social media (e.g., a post on Facebook or Twitter). Traffic from social media was further partitioned across the most popular social media platforms (e.g., Facebook, Twitter, YouTube, etc).

3 Exploring IRMA's features

In this section, we explore some of IRMA's features and provide examples replicating previous works.

We checked whether the type of incoming traffic (i.e., direct and search) was related to websites' credibility, as previous works show (Miani et al., 2022b). We found that credibility of websites was related positively with search traffic ($r = .44, p = .0016$) and negatively with direct traffic ($r = -.30, p = .0341$), suggesting that confirmation bias drives traffic towards towards misinformation websites also in this Italian sample.

We also tested the degree to which credibility was linked to interconnectedness, that is how multiple ideas form a dense and highly interconnected network, a property of conspiracy narratives (Miani et al., 2022a). To this purpose, we created networks for each website from the co-occurrence of the top-fifty most frequent words extracted from the TF-IDF (see variable `tfidf10` in Table 3). We fitted a multilevel regression predicting the degree of connectedness by credibility (nesting observations within keywords). Credibility was negatively related to connectedness ($\beta = -.063, t = -3.193, p = .0014$), meaning that low credible sources are more interconnected. Despite using only misinformation websites, these results replicate previous works on conspiracy theories. In Figure 1, we show two networks built from documents with the highest and lowest credibility scores ($N = 100,000$ in each group): the network in the low (vs high) credibility group is visually more interconnected.

Finally, we explored to what extent lexical features were linked to websites' credibility. To this goal, we fitted a series of linear models predicting credibility by the lexical features extracted with LIWC. In Figure 2, we show the 20 highest and 20 lowest beta coefficients from regression (all $ps < .001$, Bonferroni corrected). Results parallel previous works (Miani et al., 2022b; Fong et al., 2021; Klein et al., 2019; Oswald, 2016) showing that low quality sources tend to endorse a language characterized by anger (category `Rabbia`), negative emotions (`Emo_Neg`), causality (`Causa`), and negations (e.g., "do not", `Negazio`) along with use of longer words (indexing sophisticated lexicon, `BigWords`), swear words (`parolac`), and longer texts overall (i.e., word count `WC`).

4 Conclusions

We introduced IRMA, our publicly available corpus of 'untrustworthy' news in Italian. This is, as far as we know, the first Italian corpus of its kind. It consists of over 600,000 texts (335+ million words) and a number of variables to help scholars find the material that meets their needs. It can be used to develop deep learning classifiers as well as conduct different types of qualitative/quantitative research.

IRMA allows for a vast range of textual analyses thanks to the variety and quantity of data and

[5] https://www.similarweb.com/corp/ourdata/

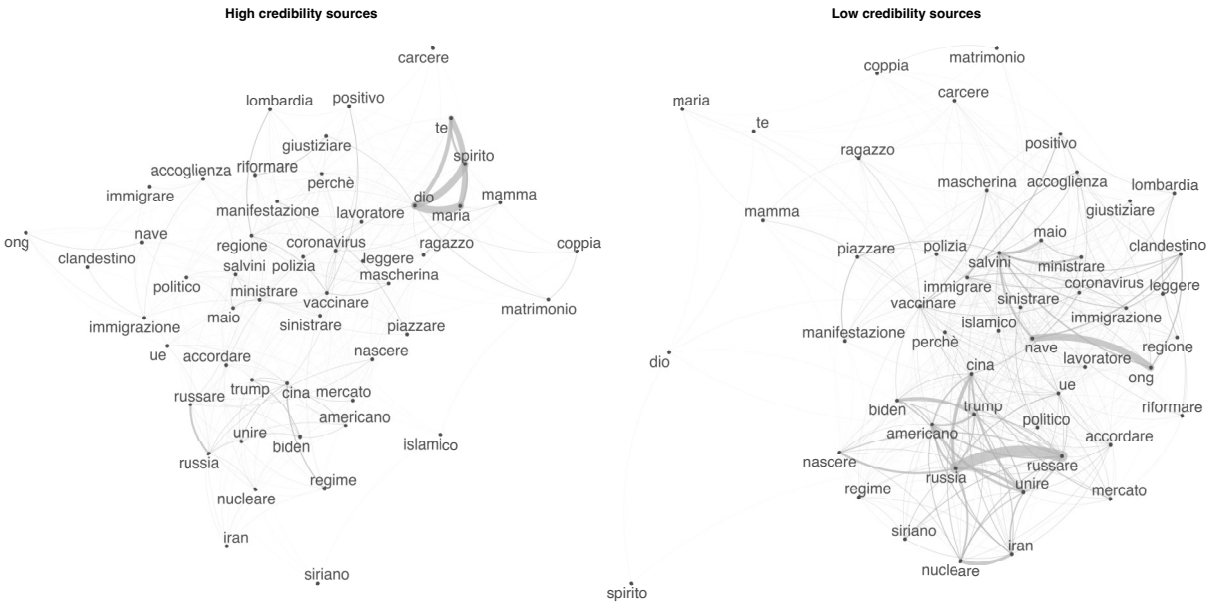

Figure 1: Co-occurrence of the top-fifty most frequent words extracted from the TF-IDF.

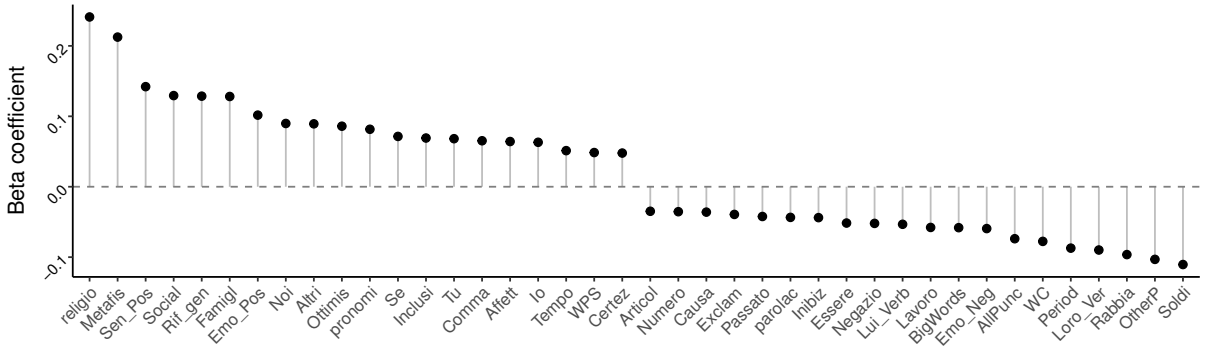

Figure 2: Coefficients (β, Y axis) from regressions predicting LIWC lexical features (on the X axis) by websites credibility scores. Positive values indicate the feature being positively correlated with credibility.

metadata included. For example, time-based data associated with textual data allows for the identification of specific periods for historical analysis (e.g., Hills and Miani, Forthcoming). A set of different semantic indexes in the form of keywords and topics help researchers find data relating to specific topics. Lexical features (specifically, LIWC scores) allow a variety of sociological and psychological studies (e.g., Fong et al., 2021). Topics and lexical features can be traced along a time series to explore their evolution through time (see e.g., Figure 3 in Appendix for topics) exploring cultural and societal trends (e.g., Lansdall-Welfare et al., 2017). IRMA also contains domain-specific features such as the type(s) of news typically shared by a specific source, as well as data on the incoming traffic for a domain, which can be used to study digital community behaviour (e.g., conspiracy websites' incoming traffic in Miani et al., 2022b).

Concluding, IRMA represents a fresh resource in an underrepresented context, such as the Italian one. This corpus was created under PRODEMINFO, an ERC-funded project that also involves other languages (e.g., German, Spanish, Hungarian). This means that the same pipeline employed to generate IRMA can be applied to other languages in the future. As a result, we hope our effort will encourage the creation of new similar corpora and stimulate future research into misinformation.

5 Limitations

Our dataset contains material classified as 'untrustworthy' by two different datasets, which relied on different classification criteria. NG ranks websites based on nine weighted criteria. Each site is assigned a trust score ranging from 0 (very poor) to

100 (exemplary). Domains with less than 60 points are labelled as "not trustworthy". On the other hand, the MD dataset is a curated collection of domain-level fact-checking databases, where the different proprietary rates are mapped onto two unifying labels, namely "accuracy" and "transparency". Although potentially leading to a different alignment of source reliability, depending on the dataset that classified the source, Lasser et al. (2022) found a high degree of agreement between the MD dataset and the NG database scores (Krippendorff's α = 0.84), as well as other collections (Lin et al., 2022).

Despite the fact that the two datasets label the websites in IRMA as "untrustworthy", this does not necessarily imply that they are all actively spreading fake news. This is due to the fact that domains could be rated not just on news quality and reliability, but also on other complimentary factors such as company policies (e.g., whether and how websites disclose information about ownership and financing). However, we are unable to provide the classification standards for the domains in our database, as well as the domains themselves due to restrictions of NG proprietary data policies. Therefore, we suggest prospective users to judge the quality of news for themselves perhaps via data-driven approaches.

Finally, it is important to note that both the NG and the MD datasets can vary over time, thus websites previously deemed untrustworthy may no longer be so in the future (and vice-versa).

6 Acknowledgements

This project has received funding from the European Research Council (ERC) under Advanced Grant PRODEMINFO (101020961). SL was also supported by funding from the Humboldt Foundation in Germany. FC was supported by the Templeton Foundation through a grant awarded to Wake Forest University.

References

Alberto Agosti and Alessandra Rellini. 2007. The Italian LIWC dictionary. *Austin, TX: LIWC. Net*.

Maysoon Alkhair, Karima Meftouh, Kamel Smaïli, and Nouha Othman. 2019. An Arabic corpus of fake news: Collection, analysis and classification. In *International Conference on Arabic Language Processing*, pages 292–302. Springer.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *CSN: Politics (Topic)*.

Kenneth Benoit, David Muhr, and Kohei Watanabe. 2021. *stopwords: Multilingual Stopword Lists*. R package version 2.3.

Kenneth Benoit, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30):774.

Alessandro Bessi, Fabiana Zollo, Michela Del Vicario, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2015. Trend of narratives in the age of misinformation. *PloS one*, 10(8):e0134641.

Saumya Bhadani, Shun Yamaya, Alessandro Flammini, Filippo Menczer, Giovanni Luca Ciampaglia, and Brendan Nyhan. 2022. Political audience diversity and news reliability in algorithmic ranking. *Nature Human Behaviour*, 6(4):495–505.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*.

Guido Caldarelli, Rocco De Nicola, Marinella Petrocchi, Manuel Pratelli, and Fabio Saracco. 2021. Flow of online misinformation during the peak of the COVID-19 pandemic in Italy. *Epj Data Science*, 10.

Michele Cantarella, Nicolò Fraccaroli, and Roberto Volpe. 2020. Does fake news affect voting behaviour? *PSN: Political Behavior (Topic)*.

Sonia Castelo, Thais Almeida, Anas Elghafari, Aécio Santos, Kien Pham, Eduardo Nakamura, and Juliana Freire. 2019. A topic-agnostic approach for identifying fake news pages. In *Companion proceedings of the 2019 World Wide Web conference*, pages 975–980.

Allen Colin and Jamie Murdock. 2020. LDA topic modeling: Contexts for the history & philosophy of science. In Grant Ramsey and Andreas de Block, editors, *Dynamics Of Science: Computational Frontiers in History and Philosophy of Science*. Pittsburgh University Press, S.l.

Dhaval M Dave, Drew McNichols, and Joseph J Sabia. 2021. Political violence, risk aversion, and non-localized disease spread: Evidence from the U.S. capitol riot. Working Paper 28410, National Bureau of Economic Research.

Michela Del Vicario, Sabrina Gaito, Walter Quattrociocchi, Matteo Zignani, and Fabiana Zollo. 2017. News consumption during the Italian referendum: A cross-platform analysis on Facebook and Twitter. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 648–657. IEEE.

Laura Edelson, Minh-Kha Nguyen, Ian Goldstein, Oana Goga, Damon McCoy, and Tobias Lauinger. 2021. Understanding engagement with U.S. (mis)information news sources on Facebook. In *Proceedings of the 21st ACM Internet Measurement Conference*, IMC '21, page 444–463, New York, NY, USA. Association for Computing Machinery.

Amos Fong, Jon Roozenbeek, Danielle Goldwert, Steven Rathje, and Sander van der Linden. 2021. The language of conspiracy: A psychological analysis of speech used by conspiracy theorists and their followers on Twitter. *Group Processes & Intergroup Relations*, 24(4):606–623.

Riccardo Gallotti, Francesco Valle, Nicola Castaldo, Pier Luigi Sacco, and Manlio De Domenico. 2020. Assessing the risks of "infodemics" in response to COVID-19 epidemics. *Nature human behaviour*.

Pietro Gravino, Giulio Prevedello, Martina Galletti, and Vittorio Loreto. 2022. The supply and demand of news during COVID-19 and assessment of questionable sources production. *Nature human behaviour*.

Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on Twitter during the 2016 US presidential election. *Science*, 363(6425):374–378.

Bettina Grün and Kurt Hornik. 2011. topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, 40(13).

Thomas Hills and Alessandro Miani. Forthcoming. A short primer on historical natural language processing. In Thomas Hills and Ganna Pogrebna, editors, *Cambridge Handbook of Behavioral Data Science*. Cambridge University Press.

Colin Klein, Peter Clutton, and Adam G Dunn. 2019. Pathways to conspiracy: The social and linguistic precursors of involvement in reddit's conspiracy theory forum. *PloS one*, 14(11):e0225098.

Thomas Lansdall-Welfare, Saatviga Sudhahar, James Thompson, Justin Lewis, FindMyPast Newspaper Team, and Nello Cristianini. 2017. Content analysis of 150 years of British periodicals. *Proceedings of the National Academy of Sciences*, 114(4):E457–E465.

Jana Lasser, Segun T. Aroyehun, Almog Simchon, Fabio Carrella, David Garcia, and Stephan Lewandowsky. 2022. Social media sharing of low-quality news sources by political elites. *PNAS Nexus*, 1(4). Pgac186.

David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

Hause Lin, Jana Lasser, Stephan Lewandowsky, Rocky Cole, Andrew Gully, David Gertler Rand, and Gordon Pennycook. 2022. High level of agreement across different news domain quality ratings.

Sahil Loomba, Alexandre de Figueiredo, Simon J Piatek, Kristen de Graaf, and Heidi Jane Larson. 2021. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature human behaviour*.

Alessandro Miani, Thomas Hills, and Adrian Bangerter. 2022a. Interconnectedness and (in)coherence as a signature of conspiracy worldviews. *Science Advances*, 8(43):eabq3668.

Alessandro Miani, Thomas Hills, and Adrian Bangerter. 2022b. Loco: The 88-million-word language of conspiracy corpus. *Behavior Research Methods*, 54(4):1794–1817.

Ludovic Mompelat, Zuoyu Tian, Amanda Kessler, Matthew Luettgen, Aaryana Rajanala, Sandra Kübler, and Michelle Seelig. 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*, pages 111–119, Marseille, France. European Language Resources Association.

Rafael A Monteiro, Roney LS Santos, Thiago AS Pardo, Tiago A de Almeida, Evandro ES Ruiz, and Oto A Vale. 2018. Contributions to the study of fake news in Portuguese: New corpus and automatic detection results. In *International Conference on Computational Processing of the Portuguese Language*, pages 324–334. Springer.

Matteo Monti. 2020. Italian populism and fake news on the internet: A new political weapon in the public discourse. In Giacomo Delledonne, Giuseppe Martinico, Matteo Monti, and Fabio Pacini, editors, *Italian Populism and Constitutional Law: Strategies, Conflicts and Dilemmas*, pages 177–197. Springer International Publishing, Cham.

Giuseppina Lo Moro, Fabrizio Bert, Ettore Minutiello, Andrea L. Zacchero, Tiziana Sinigaglia, Gianluca Colli, Rossella Tatti, Giacomo Scaioli, and Roberta Siliquini. 2021. COVID-19 fake news, conspiracy beliefs and the role of eHealth literacy: an Italian nationwide survey. *The European Journal of Public Health*, 31.

Inc. NewsGuard. 2020. Rating process and criteria. Internet Archive, `https://web.archive.org/web/20200630151704/https://www.newsguardtech.com/ratings/rating-process-criteria/`. Accessed: 2022-04-20.

Dong Nguyen, Maria Liakata, Simon DeDeo, Jacob Eisenstein, David Mimno, Rebekah Tromble, and Jane Winters. 2020. How we do things with words: Analyzing text as social and cultural data. *Frontiers in Artificial Intelligence*, 3:62.

Steve Oswald. 2016. Conspiracy and bias: argumentative features and persuasiveness of conspiracy theories. *OSSA Conference Archive*, 168:1–16.

Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855):590–595.

Gordon Pennycook and David G. Rand. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences of the United States of America*, 116:2521 – 2526.

Juan-Pablo Posadas-Durán, Helena Gómez-Adorno, Grigori Sidorov, and Jesús Jaime Moreno Escobar. 2019. Detection of fake news in a new corpus for the Spanish language. *Journal of Intelligent & Fuzzy Systems*, 36(5):4869–4876.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.

Leonard Richardson. 2007. Beautiful soup documentation. *April*.

Jon Roozenbeek, Claudia R. Schneider, Sarah Dryhurst, John R. Kerr, Alexandra L. J. Freeman, Gabriel Recchia, Anne Marthe van der Bles, and Sander van der Linden. 2020. Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science*, 7.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Martino Trevisan, Luca Vassio, and Danilo Giordano. 2021. Debate on online social networks at the time of COVID-19: An Italian case study. *Online Social Networks and Media*, 23:100136.

Sander van der Linden, Anthony Leiserowitz, Seth A. Rosenthal, and Edward W. Maibach. 2017. Inoculating the public against misinformation about climate change. *Global Challenges*, 1.

Inna Vogel and Peter Jiang. 2019. Fake news detection with the new German dataset "GermanFakeNC". In *TPDL*.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989.

A Appendix

A.1 Data availability

IRMA is freely available[6] and include the files:

1. **corpus.csv.zip**. A compressed file (.zip) containing the csv file of the corpus itself: 634,932 rows (documents) x 7 columns. Documents are identified by a hexadecimal unique ID stored in the variable `doc_id`. See dataset's variables in Table 3.
2. **website_description.csv**. A csv file containing detailed descriptions of websites. 56 rows (websites) x 25 columns. Note that due to proprietary data, NG's websites are anonymized via a unique website ID (e.g., *website1*, *website2*). See dataset's variables in Table 4.
3. **IRMA.dfm.rdata**. The IRMA's DTM. No preprocessing has been applied prior to conversion: the file contains punctuation and cased words (634,932 documents; 1,531,576 features). Note that the file was too large to be converted into a matrix, therefore we exported it as a quanteda DFM object, hence it requires the *quanteda* R package (Benoit et al., 2018).
4. **LDA_over_time.pdf**. A PDF file containing each topic's gamma values plotted over time. It contains 320 pages (i.e., the number of topics: k20 + k100 + k200). See Figure 3 for an example (at page 182 of the pdf file). Terms can be searched within the pdf.
5. **corpus_LF.csv**. The Lexical features obtained from LIWC. A csv file of 634,932 rows (documents) x 95 columns (94 LIWC lexical features and 1 documents' ID [`doc_id`]). See dataset's variables in Table 2.
6. **LDA_topic_gamma.csv.zip**. A compressed file (.zip) containing the csv file of 634,932 rows (documents) x 320 columns (LDA gamma values for k20, k100, and k200 topics). Each cell contains gamma value, that is the probability a topic is part of a document.
7. **topic_description.csv**. A file containing detailed descriptions of topics. 320 rows (topics) x 6 columns. See dataset's variables in Table 5.

[6] https://osf.io/rywp4/

Figure 3: LDA topic gamma values (Y axis) over time (X axis). Topic `k200_062` related to Covid-19 restrictions. The 10 top-most important words for topics (in decreasing order) are displayed above the plot (ENG translation: mask, close, activity, contaging, closing, observe, zone, reopening, open).

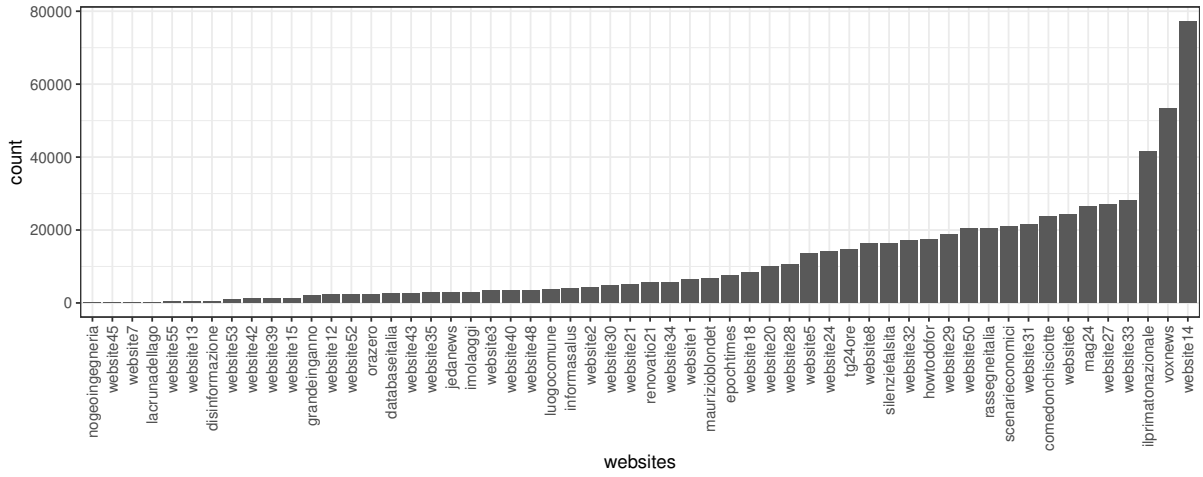

Figure 4: Document count by website.

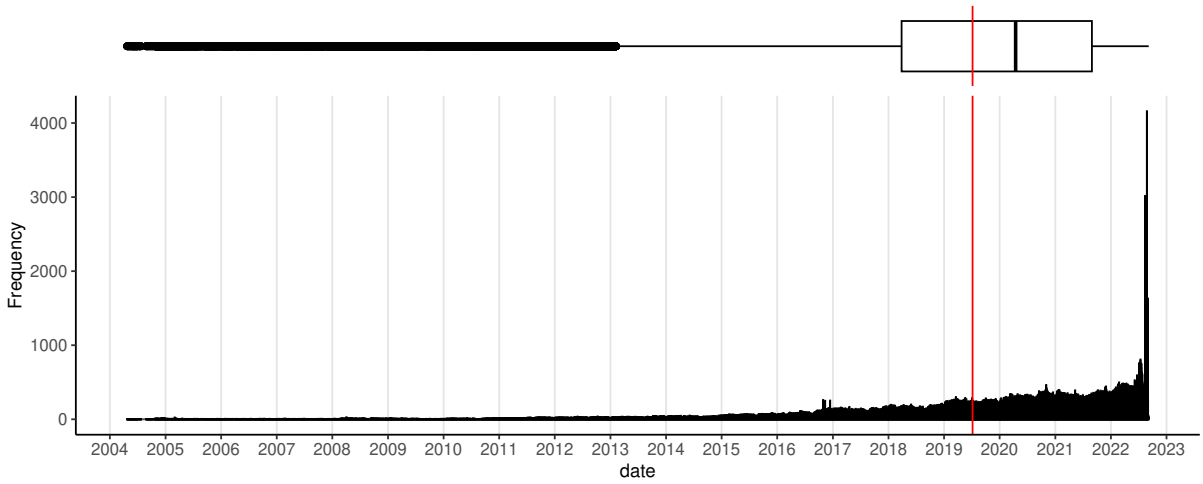

Figure 5: Distribution of documents by date (from "2004-04-22" to "2022-09-06"). The red vertical line represents the mean, the boxplot on top displays the median and the interquartile ranges.

Keyword	(ENG translation)	**Frequency**
vaccinare	(vaccinate)	6,677
ucraino	(Ukrainian)	5,013
trump	(Trump)	2,150
dragare	(Draghi)	2,075
pass	(pass)	1,980
conta	(Conti)	1,767
renzi	(Renzi)	1,669
salvini	(Salvini)	1,628
russia	(Russia)	1,326
mascherina	(mask)	1,314
banca	(bank)	1,295
putin	(Putin)	1,275
berlusconi	(Berlusconi)	1,246
siriano	(Syrian)	1,228
cina	(China)	1,227
cinese	(Chinese)	1,209
scuola	(School)	1,200
gesù	(Jesus)	1,198
papa	(Pope)	1,179
maio	(di Maio)	1,172

Table 1: Top 20 most frequent keywords (expressed as number of documents). Note that due to lemmatization, the words *dragare* and *conta* often refer to Mario Draghi and Giuseppe Conte.

Variable
(1) doc_id, (2) WC, (3) WPS, (4) BigWords, (5) Dic, (6) pronomi, (7) Io, (8) Noi, (9) Se, (10) Tu, (11) Altri, (12) Negazio, (13) Consen, (14) Articol, (15) Prepos, (16) Numero, (17) Affett, (18) Sen_Pos, (19) Emo_Pos, (20) Ottimis, (21) Emo_Neg, (22) Ansia, (23) Rabbia, (24) Tristez, (25) Mec_Cog, (26) Causa, (27) Intros, (28) Discrep, (29) Inibiz, (30) possib, (31) Certez, (32) Proc_Sen, (33) Vista, (34) Udito, (35) Sentim, (36) Social, (37) Comm, (38) Rif_gen, (39) amici, (40) Famigl, (41) Umano, (42) Tempo, (43) Passato, (44) Present, (45) Futuro, (46) Spazio, (47) Sopra, (48) Sotto, (49) Inclusi, (50) Esclusi, (51) Movimen, (52) Occupaz, (53) Scuola, (54) Lavoro, (55) Raggiun, (56) Svago, (57) Casa, (58) Sport, (59) TV_it, (60) Musica, (61) Soldi, (62) Metafis, (63) religio, (64) Morte, (65) Fisico, (66) Corpo, (67) Sesso, (68) Mangiare, (69) Dormire, (70) Cura_cor, (71) parolac, (72) Non_flu, (73) riempiti, (74) Voi, (75) Lui_lei, (76) Loro, (77) Condizio, (78) Transiti, (79) P_pass, (80) gerundio, (81) Essere, (82) Avere, (83) Io_Ver, (84) Tu_Verbo, (85) Lui_Verb, (86) Noi_Verb, (87) Voi_Verb, (88) Loro_Ver, (89) AllPunc, (90) Period, (91) Comma, (92) QMark, (93) Exclam, (94) Apostro, (95) OtherP

Table 2: List of columns for the dataset **corpus_LF.rdata**

Variable	Description
doc_id	Hexadecimal sequence of document unique identification number (e.g., D1d049)
date	The date the webpage was uploaded (format: YYYY-MM-DD, $N_{empty} = 45,185$)
website	The identification number for websites from which the document was extracted (e.g., *website15*, *ilprimatonazionale*; see also Table 4, below)
title	Title of the document ($N_{empty} = 359,526$)
txt	Document text ($N_{empty} = 358,851$)
URL	URL associated with the document ($N_{empty} = 360,137$)
WC	Word count
KW	Keyword associated with the document (see Table 1)
tfidf10	Top-10 words ordered by TF-IDF scores

Table 3: Names and variable descriptions for the dataset **corpus.csv.zip**

Variable	Description
website	Website's identification (e.g., *grandeinganno*, *website21*)
Ndoc	Number of documents for each website
WC_{type}	Word count statistics. Type includes: mean, SD, min, and max
DATE_{type}	Date range. Type includes: min and max ($N_{empty} = 8$)
type_of_news	Website type of content (e.g., conspiracy, political, health-related, religious, general, and/or viral)
Monthly_Visits	Count of visits in the past month (i.e., September 2022). Note that for websites with less than 5,000 monthly visits, SimilarWeb does not collect further traffic data. To those websites, ($N = 7$), we assigned the value 5,000
Visit_Duration	Average of visit duration (in seconds)
Bounce_Rate	The percentage of visitors who enter a site and leave after visiting only one page
Pages_per_Visit	Average of pages visited in each visit
Traffic_{type}	Proportion of incoming traffic. Type includes: Direct, Referrals, Search, and Social
Social_{type}	Proportion of incoming traffic from social media. Type includes: Linkedin, Vkontakte, Others, Telegram_Webapp, Youtube, Twitter, and Facebook
credibility	Websites' credibility scores (obtained from Lin et al., 2022).

Table 4: Names and variable descriptions for the dataset **website_description.csv**

Variable	Description
topic_name	Topic unique ID. It is composed by the topic resolution plus a three-character serial number (e.g., k100_032 is the 32th topic at 100k resolution)
top_words	Top-ten words ordered by importance (for the topic)
topic	Name of the topic with the highest correlation (within the same topic resolution)
topic_cor	Pearson r correlation estimate for the highest correlated topic
LF	Name of the LIWC's lexical feature with the highest correlation
LF_cor	Pearson r correlation estimate

Table 5: Names and variable descriptions for the dataset **topic_description.csv**. Note that correlation are computed on the document level ($N = 634,932$)