10 DATASETS FOUND

1. P

CoAID Dataset

   - paperswithcode.com
   - opendatalab.com

Updated Jun 30, 2022

□Share

×

Share

Facebook

Twitter

Email

Click to copy link

Link copied

Cite

×

Cite

Limeng Cui; Dongwon Lee (2022). CoAID Dataset [Dataset]. https://paperswithcode.com/dataset/coaid

Copy

Copied to clipboard

Explore at:

- □Papers with Code | paperswi...
- □OpenDataLab | opendatalab.com

Dataset updated

Jun 30, 2022

Authors

Limeng Cui; Dongwon Lee

Description

CoAID include diverse COVID-19 healthcare misinformation, including fake news on websites and social platforms, along with users' social engagement about such news. CoAID includes 4,251 news, 296,000 related user engagements, 926 social platform posts about COVID-19, and ground truth labels.

2. Z

CoAID dataset with multiple extracted features (both sparse and dense)

- data.niaid.nih.gov
- zenodo.org

Updated Jun 10, 2022

□Share

✕

Share

Facebook

Twitter

Email

Click to copy link

Link copied

❞ Cite

✕

Cite

Guillaume Bernard (2022). CoAID dataset with multiple extracted features (both sparse and dense) [Dataset]. https://data.niaid.nih.gov/resources?id=zenodo_6630404

⧉

Copy

⧉

Copied to clipboard

Explore at:

- □Zenodo | data.niaid.nih.gov
- □zenodo.org

Dataset updated

Jun 10, 2022

Dataset authored and provided by

Guillaume Bernard

License

Description

This is a publication of the CoAID dataset originaly dedicated to fake news detection. We changed here the purpose of this dataset in order to use it in the context of event tracking in press documents.

Cui, Limeng, et Dongwon Lee. 2020. « CoAID: COVID-19 Healthcare Misinformation Dataset ». ArXiv:2006.00885 [Cs], novembre. http://arxiv.org/abs/2006.00885.

In this dataset, we provide multiple features extracted from the text itself. Please note the text is missing from the dataset published in the CSV format for copyright reasons. You can download the original datasets and manually add the missing texts from the original publications.

Features are extracted using:

- A corpus of reference articles in multiple languages languages for TF-IDF weighting. (features_news) [1]
- A corpus of tweets reporting news for TF-IDF weighting. (features_tweets) [1]
- A S-BERT model [2] that uses distiluse-base-multilingual-cased-v1 (called features_use) 3
- A S-BERT model [2] that uses paraphrase-multilingual-mpnet-base-v2 (called features_mpnet) 4

References:

[1]: Guillaume Bernard. (2022). Resources to compute TF-IDF weightings on press articles and tweets (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.6610406

[2]: Reimers, Nils, et Iryna Gurevych. 2019. « Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks ». In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3982-92. Hong Kong, China: Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410.

3. t

CoAID - Dataset - LDM

- service.tib.eu

Updated Dec 16, 2024

□Share

×

Share

Facebook

Twitter

Email

Click to copy link

Link copied

❞ Cite

✕

Cite

(2024). CoAID - Dataset - LDM [Dataset]. https://service.tib.eu/ldmservice/dataset/coaid

⧉

Copy

⧉

Copied to clipboard

Explore at:

- □service.tib.eu

Dataset updated

Dec 16, 2024

Description

A dataset of COVID-19 misinformation detection, focusing on healthcare misinformation.

4. Z

CoAID dataset texts with OCR degradations

- data.niaid.nih.gov
- zenodo.org

Updated Jun 10, 2022

□Share

✕

Share

Facebook

Twitter

Email

Click to copy link

Link copied

Cite

✕

Cite

Guillaume Bernard (2022). CoAID dataset texts with OCR degradations [Dataset]. https://data.niaid.nih.gov/resources?id=zenodo_6630709

Copy

Copied to clipboard

Explore at:

- □Zenodo | data.niaid.nih.gov
- □zenodo.org

Dataset updated

Jun 10, 2022

Dataset authored and provided by

Guillaume Bernard

Description

This is the text of the CoAID dataset dedicated to fake news detection that has been updated to be used in event detection.

Cui, Limeng, et Dongwon Lee. 2020. « CoAID: COVID-19 Healthcare Misinformation Dataset ». ArXiv:2006.00885 [Cs], novembre. http://arxiv.org/abs/2006.00885.

Guillaume Bernard. (2022). CoAID dataset with multiple extracted features (both sparse and dense) (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.6630405

Some degradations are applied using the DocCreator [1] tool in order to degrade the text of the tweets and to reproduce some common errors found in OCRised documents [2].

[1]: Journet, Nicholas, Muriel Visani, Boris Mansencal, Kieu Van-Cuong, et Antoine Billy. 2017. « DocCreator: A New Software for Creating Synthetic Ground-Truthed Document Images ». Journal of Imaging 3 (4): 62. https://doi.org/10.3390/jimaging3040062.

[2]: Linhares Pontes, Elvys, Ahmed Hamdi, Nicolas Sidere, et Antoine Doucet. 2019. « Impact of OCR Quality on Named Entity Linking ». In Digital Libraries at the Crossroads of Digital Information for the Future, 11853:102-15. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-34058-2_11.

The results of the OCR degradations are as follow:

CoAID CER/WER

Without
Character degradation
Phantom degradation
Bleed
Blur
All

CoAID
CER
2.105
6.358
2.105
2.122
2.616
7.898

CoAID
WER
2.494
20.230
2.496

```
2.580
3.726
20.230
```

5.

Fake News Spreader Classification - CoAID Extended dataset

- figshare.com
- explore.openaire.eu

txt

Updated Jun 7, 2023

□Share

×

Share

Facebook

Twitter

Email

Click to copy link

Link copied

❞ Cite

×

Cite

Simone Leonardi; Giuseppe Rizzo; Maurizio Morisio (2023). Fake News Spreader Classification - CoAID Extended dataset [Dataset]. http://doi.org/10.6084/m9.figshare.14392859.v1

Copy

Copied to clipboard

Explore at:

- □figshare.com
- □explore.openaire.eu

txtAvailable download formats

Unique identifier

https://doi.org/10.6084/m9.figshare.14392859.v1

Dataset updated

Jun 7, 2023

Dataset provided by

figshare

Authors

Simone Leonardi; Giuseppe Rizzo; Maurizio Morisio

License

Description

This dataset contains a gold standard for the classification of users sharing the misinformation about COVID-19. It presents a list of mapped used id for privacy concerns, the list of real tweet id as retrieved from Twitter and the label classifying the tweet author as spreader or checker. Spreader are users supporting fake news, while checkers are users supporting real news. The list of fake and real news came from the CoAID dataset by Limeng and Dongwon.Data were retrieved from December 1, 2019 to April 1, 2021.For further details look at the paper "Fake News

Spreader Automated Classification for Breaking the Misinformation Chain" in the MDPI Information Journal Special Issue "News Research in Social Networks and Social Media", or open an issue in the GitHub repository.

6.

CoAID dataset with multiple extracted features (both sparse and dense) and...

- zenodo.org
- data.niaid.nih.gov

csv

Updated Jun 10, 2022

□Share

✕

Share

Facebook

Twitter

Email

Click to copy link

Link copied

❞ Cite

✕

Cite

Guillaume Bernard; Guillaume Bernard (2022). CoAID dataset with multiple extracted features (both sparse and dense) and degraded by OCR [Dataset]. http://doi.org/10.5281/zenodo.6630966

Copy

Copied to clipboard

Explore at:

- □zenodo.org
- □Zenodo | data.niaid.nih.gov

csvAvailable download formats

Unique identifier

https://doi.org/10.5281/zenodo.6630966

Dataset updated

Jun 10, 2022

Dataset provided by

Zenodohttp://zenodo.org/

Authors

Guillaume Bernard; Guillaume Bernard

License

Description

This is the same datasets as:

Guillaume Bernard. (2022). CoAID dataset with multiple extracted features (both sparse and dense) (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.6630405

But with texts degraded by OCR as described in:

Guillaume Bernard. (2022). CoAID dataset texts with OCR degradations (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.6630710

7.

CoAID-0.4

- kaggle.com

Updated Nov 25, 2024

□Share

×

Share

Facebook

Twitter

Email

Click to copy link

Link copied

Cite

×

Cite

Ananya Kunisetty (2024). CoAID-0.4 [Dataset]. https://www.kaggle.com/datasets/ananyakunisetty/coaid-0-4

Copy

Copied to clipboard

Explore at:

- □Kaggle | kaggle.com

*bakery_dining*CroissantCroissant is a format for machine-learning datasets. Learn more about this at mlcommons.org/croissant.

Dataset updated

Nov 25, 2024

Dataset provided by

[Kaggle](http://kaggle.com/)http://kaggle.com/

Authors

Ananya Kunisetty

Description

Dataset

This dataset was created by Ananya Kunisetty

Contents

8.

ckpt-CoAID

- kaggle.com

zip

Updated Nov 26, 2024

□Share

×

Share

Facebook

Twitter

✉Email

Click to copy link

Link copied

” Cite

×

Cite

DHEERAJ KURUKUNDA (2024). ckpt-CoAID [Dataset]. https://www.kaggle.com/datasets/dheerajkurukunda/ckpt-coaid/suggestions?status=pending&yourSuggestions=true

⧉

Copy

⧉

Copied to clipboard

Explore at:

- □Kaggle | kaggle.com

zip(124059187 bytes)Available download formats

Dataset updated

Nov 26, 2024

Authors

DHEERAJ KURUKUNDA

License

Description

Dataset

This dataset was created by DHEERAJ KURUKUNDA

Released under Apache 2.0

Contents

9. Z

Data from: PANACEA dataset - Heterogeneous COVID-19 Claims

- data.niaid.nih.gov
- zenodo.org

Updated Jul 15, 2022

□Share

×

Share

Facebook

Twitter

Email

Click to copy link

Link copied

❞ Cite

✕

Cite

Procter, Rob (2022). PANACEA dataset - Heterogeneous COVID-19 Claims [Dataset]. https://data.niaid.nih.gov/resources?id=zenodo_6493846

⧉

Copy

⧉

Copied to clipboard

Explore at:

- □Zenodo | data.niaid.nih.gov
- □zenodo.org

Dataset updated

Jul 15, 2022

Dataset provided by

Procter, Rob
Liakata, Maria
Zubiaga, Arkaitz
Arana-Catania, Miguel
He, Yulan
Kochkina, Elena

License

Description

The peer-reviewed publication for this dataset has been presented in the 2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), and can be accessed here: https://arxiv.org/abs/2205.02596. Please cite this when using the dataset.

This dataset contains a heterogeneous set of True and False COVID claims and online sources of information for each claim.

The claims have been obtained from online fact-checking sources, existing datasets and research challenges. It combines different data sources with different foci, thus enabling a comprehensive approach that combines different media (Twitter, Facebook, general websites, academia), information domains (health, scholar, media), information types (news, claims) and applications (information retrieval, veracity evaluation).

The processing of the claims included an extensive de-duplication process eliminating repeated or very similar claims. The dataset is presented in a LARGE and a SMALL version, accounting for different degrees of similarity between the remaining claims (excluding respectively claims with a 90% and 99% probability of being similar, as obtained through the MonoT5 model). The similarity of claims was analysed using BM25 (Robertson et al., 1995; Crestani et al., 1998; Robertson and Zaragoza, 2009) with MonoT5 re-ranking (Nogueira et al., 2020), and BERTScore (Zhang et al., 2019).

The processing of the content also involved removing claims making only a direct reference to existing content in other media (audio, video, photos); automatically obtained content not representing claims; and entries with claims or fact-checking sources in languages other than English.

The claims were analysed to identify types of claims that may be of particular interest, either for inclusion or exclusion depending on the type of analysis. The following types were identified: (1) Multimodal; (2) Social media references; (3) Claims including questions; (4) Claims including numerical content; (5) Named entities, including: PERSON − People, including fictional; ORGANIZATION − Companies, agencies, institutions, etc.; GPE − Countries, cities, states; FACILITY − Buildings, highways, etc. These entities have been detected using a RoBERTa base English model (Liu et al., 2019) trained on the OntoNotes Release 5.0 dataset (Weischedel et al., 2013) using Spacy.

The original labels for the claims have been reviewed and homogenised from the different criteria used by each original fact-checker into the final True and False labels.

The data sources used are:

- The CoronaVirusFacts/DatosCoronaVirus Alliance Database. https://www.poynter.org/ifcn-covid-19-misinformation/
- CoAID dataset (Cui and Lee, 2020) https://github.com/cuilimeng/CoAID
- MM-COVID (Li et al., 2020) https://github.com/bigheiniu/MM-COVID
- CovidLies (Hossain et al., 2020) https://github.com/ucinlp/covid19-data
- TREC Health Misinformation track https://trec-health-misinfo.github.io/
- TREC COVID challenge (Voorhees et al., 2021; Roberts et al., 2020) https://ir.nist.gov/covidSubmit/data.html

The LARGE dataset contains 5,143 claims (1,810 False and 3,333 True), and the SMALL version 1,709 claims (477 False and 1,232 True).

The entries in the dataset contain the following information:

- Claim. Text of the claim.
- Claim label. The labels are: False, and True.
- Claim source. The sources include mostly fact-checking websites, health information websites, health clinics, public institutions sites, and peer-reviewed scientific journals.
- Original information source. Information about which general information source was used to obtain the claim.
- Claim type. The different types, previously explained, are: Multimodal, Social Media, Questions, Numerical, and Named Entities.

Funding. This work was supported by the UK Engineering and Physical Sciences Research Council (grant no. EP/V048597/1, EP/T017112/1). ML and YH are supported by Turing AI Fellowships funded by the UK Research and Innovation (grant no. EP/V030302/1, EP/V020579/1).

References

- Arana-Catania M., Kochkina E., Zubiaga A., Liakata M., Procter R., He Y.. Natural Language Inference with Self-Attention for Veracity Assessment of Pandemic Claims. NAACL 2022 https://arxiv.org/abs/2205.02596
- Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. Nist Special Publication Sp,109:109.
- Fabio Crestani, Mounia Lalmas, Cornelis J Van Rijsbergen, and Iain Campbell. 1998. “is this document relevant?. . . probably” a survey of probabilistic models in information retrieval. ACM Computing Surveys (CSUR), 30(4):528–552.
- Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. Now Publishers Inc.
- Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pre-trained sequence-to-sequence model. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, pages 708–718.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In International Conference on Learning Representations.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.
- Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. Linguistic Data Consortium, Philadelphia, PA, 23.

- Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset. arXiv preprint arXiv:2006.00885.
- Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. 2020. Mm-covid: A multilingual and multimodal data repository for combating covid-19 disinformation.
- Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. COVIDLies: Detecting COVID-19 misinformation on social media. In Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020, Online. Association for Computational Linguistics.
- Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. Trec-covid: constructing a pandemic information retrieval test collection. In ACM SIGIR Forum, volume 54, pages 1–12. ACM New York, NY, USA.

10.

Co-aid

- kaggle.com

zip

Updated Nov 25, 2024

□Share

✕

Share

Facebook

Twitter

Email

Click to copy link

Link copied

❞ Cite

✕

Cite

Nandanmanjunath I (2024). Co-aid [Dataset].
https://www.kaggle.com/datasets/nandanmanjunathi/co-aid

Copy

Copied to clipboard

Explore at:

- □Kaggle | kaggle.com

zip(5640713 bytes)Available download formats

Dataset updated

Nov 25, 2024

Authors

Nandanmanjunath I

Description

Dataset

This dataset was created by Nandanmanjunath I

Contents

11. 🔍

Not seeing a result you expected?
Learn how you can add new datasets to our index.

□Share

✕
Share
Facebook
Twitter
Email
Click to copy link
Link copied
Cite
✕
Cite
Limeng Cui; Dongwon Lee (2022). CoAID Dataset [Dataset]. https://paperswithcode.com/dataset/coaid
Copy
Copied to clipboard
□
Save

CoAID Dataset

Explore at:

- □Papers with Code | paperswi...
- □OpenDataLab | opendatalab.com

Dataset updated
Jun 30, 2022
Authors
Limeng Cui; Dongwon Lee
Description

CoAID include diverse COVID-19 healthcare misinformation, including fake news on websites and social platforms, along with users' social engagement about such news. CoAID includes 4,251 news, 296,000 related user engagements, 926 social platform posts about COVID-19, and ground truth labels.