Suspicious Sentence Detection and Claim Verification in the COVID-19 Domain

ROMCIR 2022: 2nd Workshop on Reducing Online Misinformation through Credible Information Retrieval

Elitsa Pankovska[1] Konstantin Schulz[2] Georg Rehm[2]

[1] Technical University Berlin

[2] German Research Center for Artificial Intelligence

April 10, 2022

PanQura

Qurator
Curation Technologies

Overview

1. Introduction
2. Methodology
3. Experiments
4. Prototype
5. Evaluation

Goals

- decrease the **speed** and **spread** of fake news
- high-performance software **component** for fact checking of small- to medium-sized documents
- build upon **existing work** on COVID-19 fake news detection

Vosoughi et al. 2018, Barrón-Cedeño et al. 2020, Das et al. 2021

Approaches

- analytical **target**: content, source, propagation
- **fact checking**: journalists vs. crowd vs. APIs
- **language models** vs. SVM, Random Forests etc.
- claim **verification**: Wikipedia, knowledge graphs, specific markup
- **ClaimBuster**: outdated, multiple separate tools, out of domain

Srivastava et al. 2017, Rehm 2018, Bourgonje et al. 2017, Rehm et al. 2018, Vosoughi et al. 2018, Bhatt et al. 2018, Collins et al. 2020, Antoun et al. 2020, Nguyen et al. 2020, Wise et al.2020, Domingo-Fernández et al. 2020, Vaswani et al. 2017, Li et al. 2021, Gundapu et al.2021, Chernyavskiy et al. 2021, Schulz et al. 2022; https://idir.uta.edu/claimbuster/

Definitions

- **Fake News**: false stories that appear to be news, spread on the internet or using other media, usually created to influence political views or as a joke
- various **types** of misinformation: satire, parody, ..., manipulated or fabricated content
- **veracity** and **intention** (to deceive)
- **trustworthiness**
- **suspiciousness**: claim or statement that contains possibly false/misleading information, or is proved not to be entirely true

Wardle2017, Viviani et al. 2017, Guess et al. 2020;
https://dictionary.cambridge.org/dictionary/english/fake-news

NLP Tasks

- Sentence Classification
- Claim Extraction
- Claim Verification using external Knowledge Bases

Pipeline

1. **segmentation** and claim extraction: spaCy[1]
2. binary **classification** using BERT, DistilBERT, SciBERT, RoBERTa
3. **non-suspicious** claims are discarded
4. removal of punctuation and **stop words**
5. **GET request** to Google Fact Check Tools API

https://spacy.io

Google Fact Check Tools API

- ClaimReview markup
- GET https://factchecktools.googleapis.com/v1alpha1/claims:search?languageCode=en&maxAgeDays=200&query=ginger%20cures%20corona&key=[YOUR_API_KEY]
- mapping of results
  - 1: "**false**", "four pinocchios", "inaccurate", "miscaptioned", "misattributed", "scam"
  - 2: "**mostly false**", "three pinocchios", "misleading"
  - 3: "**mixture**", "two pinocchios", "biased", "cherry-picking", "not the whole story", "exaggerates"
  - 4: "**mostly true**", "half true", "one pinocchio"
  - 5: "**true**", "accurate", "unbiased", "correct"

https://developers.google.com/search/docs/advanced/structured-data/factcheck

Datasets

1. **CORD-19** (COVID-19 Open Research Dataset): >500,000 scholarly articles about COVID-19 and related coronaviruses ⟶ non-suspicious sentences (science)
2. **FakeCovid**: 40 languages (titles mostly English), cross-domain, news, fact-checked, COVID-19 ⟶ suspicious sentences
3. **CoAID** (Covid-19 heAlthcare mIsinformation Dataset): fake news on websites and social media, incl. users' social engagement; large overlap with FakeCovid
4. **COVID19 Fake News Detection in English**: real and fake news on COVID-19 ⟶ non-suspicious sentences (news)

Shahi et al. 2020, Cui et al.2020, Wang et al. 2020, Das et al. 2021

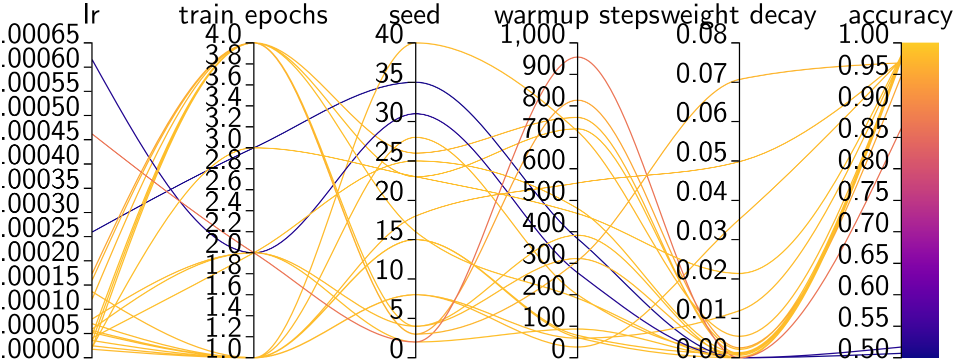
Data Samples

Sentence	Suspicious
A rapid antigen changes and recombination of the viral RNA genome contribute to the reduced effectiveness of vaccination and anti-influenza drugs.	0
Weed (cannabis) cures coronavirus.	1
We conducted a meta-analysis to assess the prevalence of depression, anxiety, distress, and insomnia during the COVID-19 pandemic.	0
To add to the knowledge base, we initiated a regional COVID-19 in pregnancy collaborative observational study with a coordinating center, standardized data collection and a shared database.	0
U.S. House Speaker Nancy Pelosi was in Wuhan, China, six days after the impeachment proceedings against President Trump ended.	1

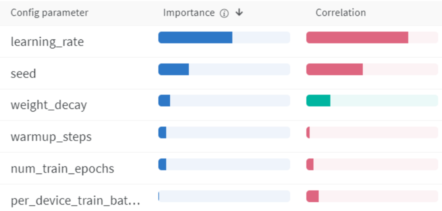
Hyperparameters: Search Space

- **Learning** Rate: [1e-5, ..., 1e-3]
- Number of **Epochs**: [1, ..., 4]
- **Seed**: [1, ..., 42]
- **Batch** Size: [8, 16]
- **Warmup** Steps: [0, ..., 1000]
- Weight **Decay**: [1e-6, ..., 0.1]

Hyperparameters: Results

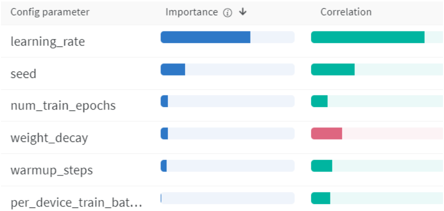


Hyperparameters: Importance



Config parameter	Importance ⓘ ↓	Correlation
learning_rate	🟦	🟥
seed	🟦	🟥
weight_decay	🟦	🟩
warmup_steps	🟦	🟥
num_train_epochs	🟦	🟥
per_device_train_bat...	🟦	🟥

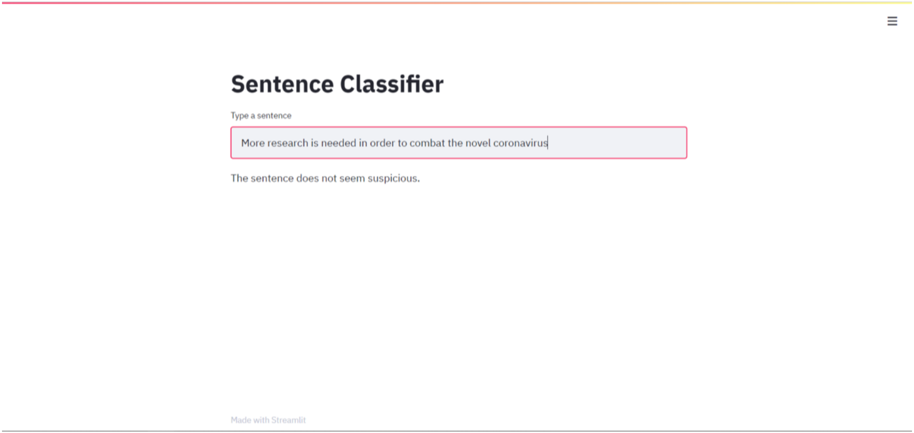
(a) Accuracy



Config parameter	Importance ⓘ ↓	Correlation
learning_rate	🟦	🟩
seed	🟦	🟩
num_train_epochs	🟦	🟩
weight_decay	🟦	🟥
warmup_steps	🟦	🟩
per_device_train_bat...	🟦	🟩

(b) Cross-entropy loss

Sentence Classifier



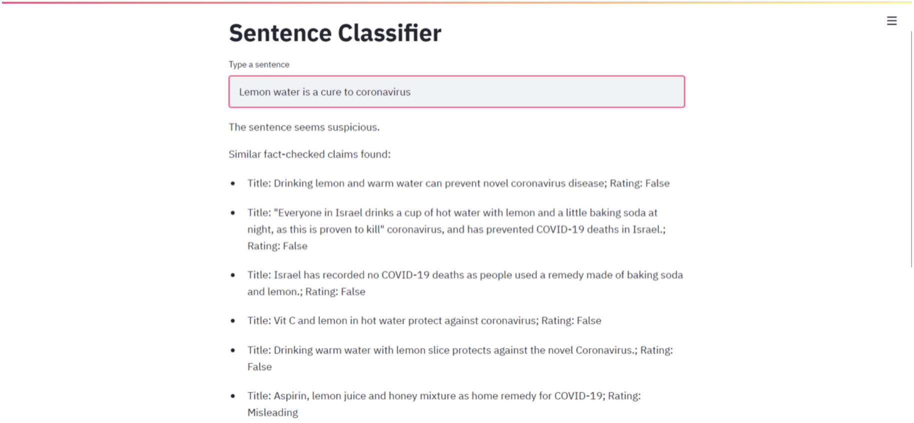
Sentence Classifier

Type a sentence

More research is needed in order to combat the novel coronavirus

The sentence does not seem suspicious.

Fact Checking



Sentence Classifier

Type a sentence

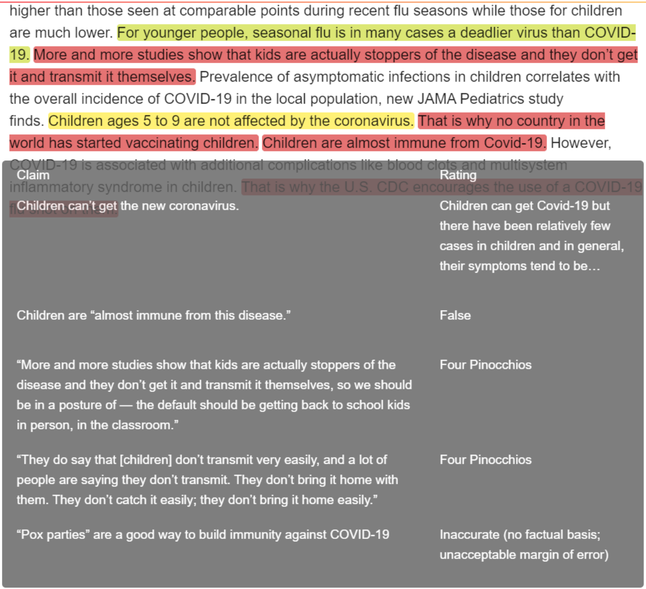
Lemon water is a cure to coronavirus

The sentence seems suspicious.

Similar fact-checked claims found:

- Title: Drinking lemon and warm water can prevent novel coronavirus disease; Rating: False
- Title: "Everyone in Israel drinks a cup of hot water with lemon and a little baking soda at night, as this is proven to kill" coronavirus, and has prevented COVID-19 deaths in Israel.; Rating: False
- Title: Israel has recorded no COVID-19 deaths as people used a remedy made of baking soda and lemon.; Rating: False
- Title: Vit C and lemon in hot water protect against coronavirus; Rating: False
- Title: Drinking warm water with lemon slice protects against the novel Coronavirus.; Rating: False
- Title: Aspirin, lemon juice and honey mixture as home remedy for COVID-19; Rating: Misleading

Visualization of Ratings



higher than those seen at comparable points during recent flu seasons while those for children are much lower. For younger people, seasonal flu is in many cases a deadlier virus than COVID-19. More and more studies show that kids are actually stoppers of the disease and they don't get it and transmit it themselves. Prevalence of asymptomatic infections in children correlates with the overall incidence of COVID-19 in the local population, new JAMA Pediatrics study finds. Children ages 5 to 9 are not affected by the coronavirus. That is why no country in the world has started vaccinating children. Children are almost immune from Covid-19. However, COVID-19 is associated with additional complications like blood clots and multisystem inflammatory syndrome in children. That is why the U.S. CDC encourages the use of a COVID-19 flu shot on them.

Claim	Rating
Children can't get the new coronavirus.	Children can get Covid-19 but there have been relatively few cases in children and in general, their symptoms tend to be…
Children are "almost immune from this disease."	False
"More and more studies show that kids are actually stoppers of the disease and they don't get it and transmit it themselves, so we should be in a posture of — the default should be getting back to school kids in person, in the classroom."	Four Pinocchios
"They do say that [children] don't transmit very easily, and a lot of people are saying they don't transmit. They don't bring it home with them. They don't catch it easily; they don't bring it home easily."	Four Pinocchios
"Pox parties" are a good way to build immunity against COVID-19	Inaccurate (no factual basis; unacceptable margin of error)

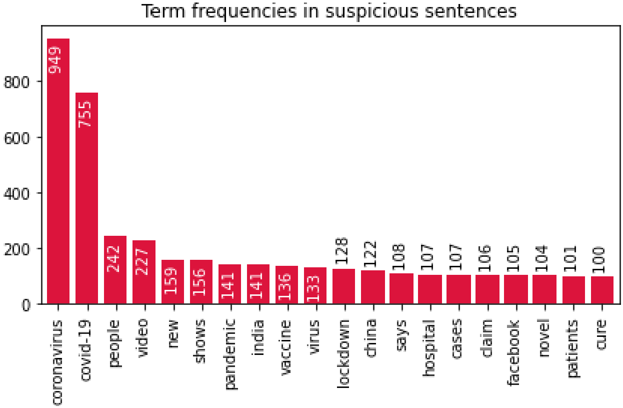
Model Comparison

Model	Accuracy	CE loss	F_1	Precision	Recall
BERT	**98.11%**	**0.0952**	**0.9805**	0.9816	**0.9793**
DistilBERT	97.89%	0.09849	0.9781	0.9796	0.9773
SciBERT	97.64%	0.1197	0.9755	0.9799	0.9711
RoBERTa	97.61%	0.1006	0.975	**0.9818**	0.9684

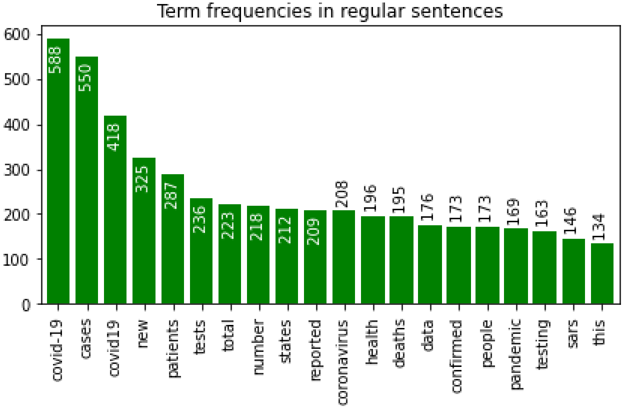
10-Fold Cross-Validation

Model	**Accuracy**	**CE loss**	$\mathbf{F_1}$	**Precision**	**Recall**
BERT	**97.7185%**	0.1216	**0.9769**	0.9762	**0.9777**
DistilBERT	97.692%	**0.0966**	0.9766	**0.9773**	0.976

Most Common Words: Suspicious



Most Common Words: Non-Suspicious



Misclassified Samples

Sentence	**True Label**	**Predicted Label**
There is no one in New Zealand receiving hospital-level care for COVID-19.	regular	suspicious
Even discharged patients could be a long-term asymptomatic carriers.	suspicious	regular

Summary

- integration of **suspiciousness** detection & claim **verification**
- 5-point **scale** of suspiciousness
- multiple datasets with different **registers**
- partial **standardization** of review data
- using **language models** and fact check **API** ⟶ fully automated, fast, cheap

References I

- Bourgonje, P., J. M. Schneider, and G. Rehm (2017). "From Clickbait to Fake News Detection: An Approach based on Detecting the Stance of Headlines to Articles". In: *Proceedings of the Second Workshop on Natural Language Processing meets Journalism – EMNLP 2017 Workshop (NLPMJ 2017)*. Ed. by O. Popescu and C. Strapparava. 7 September. Copenhagen, Denmark, pp. 84–89.
- Srivastava, A., G. Rehm, and J. M. Schneider (2017). "DFKI-DKT at SemEval-2017 Task 8: Rumour Detection and Classification Using Cascading Heuristics". In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, pp. 486–490.
- Vaswani, A. et al. (2017). "Attention is all you need". In: *Advances in neural information processing systems*. Ed. by I. Guyon et al. Vol. 30. Long Beach, California, USA: Curran Associates, Inc., pp. 5998–6008.

References II

Viviani, M. and G. Pasi (2017). “Credibility in Social Media: Opinions, News, and Health Information—a Survey”. In: *Wiley interdisciplinary reviews: Data mining and knowledge discovery* 7.5, e1209.

Wardle, C. (2017). *Fake news. It’s complicated.* https://firstdraftnews.com/fake-news-complicated/. First Draft News.

Bhatt, G. et al. (2018). “Combining Neural, Statistical and External Features for Fake News Stance Identification”. In: *Companion Proceedings of the The Web Conference 2018*. WWW ’18. Lyon, France: International World Wide Web Conferences Steering Committee, pp. 1353–1357. DOI: 10.1145/3184558.3191577.

References III

Rehm, G. (2018). "An Infrastructure for Empowering Internet Users to handle Fake News and other Online Media Phenomena". In: *Language Technologies for the Challenges of the Digital Age: 27th International Conference, GSCL 2017, Berlin, Germany, September 13-14, 2017, Proceedings*. Ed. by G. Rehm and T. Declerck. Lecture Notes in Artificial Intelligence (LNAI) 10713. 13/14 September 2017. Gesellschaft für Sprachtechnologie und Computerlinguistik e.V. Cham, Switzerland: Springer, pp. 216–231.

Rehm, G., J. M. Schneider, and P. Bourgonje (2018). "Automatic and Manual Web Annotations in an Infrastructure to handle Fake News and other Online Media Phenomena". In: *Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018)*. Ed. by N. Calzolari et al. Miyazaki, Japan: European Language Resources Association (ELRA), pp. 2416–2422.

References IV

- Vosoughi, S., D. Roy, and S. Aral (2018). "The spread of true and false news online". In: *Science* 359.6380, pp. 1146–1151. DOI: 10.1126/science.aap9559.
- Antoun, W. et al. (2020). "State of the Art Models for Fake News Detection Tasks". In: *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*. Doha,Qatar: IEEE, pp. 519–524. DOI: 10.1109/ICIoT48696.2020.9089487.
- Barrón-Cedeño, A. et al. (2020). "Overview of CheckThat! 2020: Automatic Identification and Verification of Claims in Social Media". In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Ed. by A. Arampatzis et al. Cham: Springer International Publishing, pp. 215–236.

References V

- Collins, B., D. T. Hoang, N. T. Nguyen, and D. Hwang (2020). "Fake News Types and Detection Models on Social Media A State-of-the-Art Survey". In: *Intelligent Information and Database Systems*. Ed. by P. Sitek, M. Pietranik, M. Krótkiewicz, and C. Srinilta. Singapore: Springer Singapore, pp. 562–573.
- Cui, L. and D. Lee (2020). *CoAID: COVID-19 Healthcare Misinformation Dataset*. arXiv: 2006.00885 [cs.SI].
- Domingo-Fernández, D. et al. (Dec. 2020). "COVID-19 Knowledge Graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology". In: *Bioinformatics* 37.9, pp. 1332–1334. DOI: 10.1093/bioinformatics/btaa834.
- Guess, A. M. and B. A. Lyons (2020). "Misinformation, Disinformation, and Online Propaganda". In: *Social Media and Democracy: The State of the Field, Prospects for Reform*. Ed. by N. Persily and J. A. Tucker. Cambridge: Cambridge University Press, pp. 10–33.

References VI

- Nguyen, V.-H., K. Sugiyama, P. Nakov, and M.-Y. Kan (Oct. 2020). “FANG”. In: *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*. DOI: 10.1145/3340531.3412046.
- Shahi, G. K. and D. Nandini (2020). “FakeCovid – A Multilingual Cross-domain Fact Check News Dataset for COVID-19”. In: *Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media*. Atlanta, Georgia, USA: The AAAI Press.
- Wang, L. L. et al. (July 2020). “CORD-19: The COVID-19 Open Research Dataset”. In: *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Online: Association for Computational Linguistics.
- Wise, C. et al. (2020). *COVID-19 Knowledge Graph: Accelerating Information Retrieval and Discovery for Scientific Literature*. arXiv: 2007.12731 [cs.IR].

References VII

Chernyavskiy, A., D. Ilvovsky, and P. Nakov (2021). "WhatTheWikiFact: Fact-Checking Claims Against Wikipedia". In: *CIKM '21: Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. Virtual Event, Australia: Association for Computing Machinery.

Das, S. D., A. Basak, and S. Dutta (2021). "A Heuristic-Driven Ensemble Framework for COVID-19 Fake News Detection". In: *Combating Online Hostile Posts in Regional Languages during Emergency Situation*. Ed. by T. Chakraborty et al. Cham: Springer International Publishing, pp. 164–176.

Gundapu, S. and R. Mamidi (2021). *Transformer based Automatic COVID-19 Fake News Detection System*. arXiv: 2101.00180 [cs.CL].

References VIII

- Li, X. et al. (2021). "Exploring Text-Transformers in AAAI 2021 Shared Task: COVID-19 Fake News Detection in English". In: *Combating Online Hostile Posts in Regional Languages during Emergency Situation*. Ed. by T. Chakraborty et al. Cham: Springer International Publishing, pp. 106–115.
- Schulz, K. et al. (2022). "User Experience Design for Automatic Credibility Assessment of News Content About COVID-19". en. In: *HCI International 2022 – Late Breaking Papers*. Forthcoming. Virtual: Springer.