Challenges in Automated Detection of COVID-19 Misinformation

DRAHOMIRA HERRMANNOVA[*†], Oak Ridge National Laboratory, USA
GAUTAM THAKUR, Oak Ridge National Laboratory, USA
JOSHUA N. GRANT, Oak Ridge National Laboratory, USA
VARISARA TANSAKUL, Oak Ridge National Laboratory, USA
BRYAN EATON, Oak Ridge National Laboratory, USA
OLIVERA KOTEVSKA, Oak Ridge National Laboratory, USA
JORDAN BURDETTE, Oak Ridge National Laboratory, USA
MARTIN SMYTH, Institute for Advanced Computational Science, Stony Brook University, USA
MONICA SMITH, National Geospatial-Intelligence Agency, USA

The COVID-19 pandemic has made the dangers of spread of misinformation obvious but despite much global effort to curbing its spread, fake information about the pandemic keeps proliferating. In this paper we address the development of automated methods for verification of claims about COVID-19 and discuss the challenges associated with this task. We focus on labeled data collection, limitations of existing models, and difficulties of applying misinformation detection models in practical applications. Our initial analysis indicates label imbalance may be a particular challenge for developing claim verification models and we discuss options for alleviating this issue.

CCS Concepts: • **Computing methodologies** → **Discourse, dialogue and pragmatics**; • **Human-centered computing** → **Web-based interaction**; • **Networks** → *Online social networks.*

Keywords: Misinformation, Disinformation, Fake News, Claim Verification, Misinformation Detection, COVID-19

1 INTRODUCTION

In recent years, the online spread of false and misleading claims has influenced many narratives and affected the world in various ways from moving the markets [12] and helping to incite violence against minorities [23], to shaping political opinions [5]. The COVID-19 pandemic has made the dangers of spread of misinformation even more obvious. To address this, a significant amount of effort has been devoted to both manual fact checking of claims as well as to developing automated claim verification methods. In this paper we address the development of automated methods for predicting the truthfulness of claims about COVID-19 and discuss the challenges associated with this task. We focus on labeled data collection, limitations of existing models, and difficulties of applying claim verification models in practical applications. Our long term goal is to study the relation between misinformation, its spread, and COVID-19 infection prevalence around the world. Consequently, our current focus is on developing accurate claim verification models that can be applied on large amounts of online claims.

[*] This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

[†] Corresponding author: herrmannovad@ornl.gov

2 DATASETS

A significant challenge for automated misinformation detection is the availability of labeled data. As [3] point out, past datasets constructed for claim verification were either small or based on artificially created claims. Recently, a large number of datasets have been released that are built using data collected from fact checking websites, e.g., [3, 6, 9, 19] among many others. A significant advantage of leveraging data from fact checkers is in the quality of labels, the size and scope of data, and the fact these claims represent real-world data. However, leveraging data from fact checkers poses a number of challenges for developing claim verification models.

One downside of leveraging this data is that it can suffer from significant label imbalance with the majority of claims being false or partially false. For example, we collected over 16 thousand claims about COVID-19 from seven fact checkers and mapped their truth labels to the following scale:

- **True:** Correct, verified statement, or a statement that is almost completely true (e.g. claims labeled "mostly true").
- **False/misleading:** Parts or all of statement are false, or a statement that is presented in a way that could be misleading.
- **Other:** Sarcasm, satire, outdated, not enough evidence, etc. as well as claims without a label.

Across our 16,080 claims, only 224 claims (1.4%) can be categorized as true and 721 (4.5%) as other, while the remaining claims are labeled false/misleading (15,135, or 94%). Additional details about our data collection and harmonization approach are presented in [10]. This label imbalance is unsurprising given the nature of fact checkers, but poses an issue for training classification models. To overcome the issue of label imbalance, several works have explored collecting additional COVID-19-related statements from reliable public health and research organizations such as the CDC, UK NHS, and WHO [6, 9]. However, collecting additional true claims in this manner may be much more difficult for other topics of misinformation where such clear sources of reliable information may not exist.

In contrast, Shahi and Nandini [19] did not incorporate additional true claims. They harmonized all labels by assigning them to either "false" or "other" categories (true claims would be included in "other") and focused on predicting these two categories. While the authors did not specify how many claims in their dataset were labeled "true", based on the above analysis of fact checker data we expect that besides true claims, many claims that would fall into the "other" category contain sarcasm, satire, and other types of claims that are not false, but may also not be a good representation of true claims. A somewhat similar approach was previously employed by [3] who did not map all labels to the same scale but instead developed models to predict labels as collected from fact checkers – they overcome the issue of disparate label sets from different fact checkers by employing multi-task learning. However, it is unclear whether a model trained on such imbalanced data could be used on real-world data collected from online sources where the vast majority of claims are likely to be true [2]. Collecting additional true claims may help alleviate this issue, but identifying good sources of true claims is not a trivial task, particularly if collecting claims about multiple topics.

Another difficulty of utilizing data collected from fact checkers is that truth labels and other data such as claim source information, as well as the way claims can be collected automatically are not harmonized across different fact checkers. We also observed that in many cases, links to the original claim no longer work. To overcome the first issue, one possibility is to develop a protocol for data interoperability. For example, the Open Archives Initiative Protocol for Metadata Harvesting[1] has been developed to enable harvesting of metadata descriptions records and is widely implemented by institutional repositories to provide access to scholarly outputs – a similar approach could be leveraged by fact checkers. To overcome the latter issue existing services and protocols for web archiving could be leveraged [11].

[1] https://www.openarchives.org/pmh/

Addressing these challenges has the potential to significantly simplify access to important data to researchers studying misinformation or developing detection systems and consequently to help advance research in this area.

3 CLAIM VERIFICATION APPROACHES

The natural language processing community has in the past invested significant efforts in the development of systems to deal with disinformation and misinformation, and several detailed reviews of literature on misinformation detection and related tasks exist (see for example [13, 17, 21]). Therefore, in this section we focus mainly on the verification of claims surrounding COVID-19.

A significant focus within COVID-19 misinformation detection has been put on Twitter and several existing studies have investigated the spread of COVID-19 related misinformation on Twitter [14, 16, 20], however, other sources of information including YouTube [15] and Reddit [1] have been explored as well. Of the existing studies on detection of COVID-19 misinformation, many have framed the problem as a classification task that focuses on predicting the veracity of claims based on their content [4, 8, 9, 14, 15, 19]. These previous works have experimented with different text classification models including both traditional [8, 9] and deep learning models [9, 14].

Previous studies have shown incorporating both additional metadata [22] such as the speaker (the claim author) and the context (where/on what occation was the claim made), as well as claim evidence [3, 18] can significantly improve the performance of claim verification models. While much additional metadata is available directly from fact checkers, evidence for claims has to be collected separately. Previously, this has been done by leveraging search engines to look for articles most closely related to a given claim [3, 4, 18], or by incorporating Tweets and Wikipedia pages on a given topic [7]. The advantage of models trained on claims, metadata, and evidence is that these models no longer focus solely on surface-level linguistic features, but can also leverage related information from other sources, leading to higher accuracy [3, 18, 22]. The limitation of this approach is the need to obtain data which may be used as evidence for claims. This step can be particularly difficult when collecting evidence for large amounts of real-world online claims and when prediction speed is an issue. Consequently, claim verification using claim content may represent a quick and simple option for getting an initial veracity prediction for a claim, with evidence-based models enabling more accurate prediction in downstream applications.

4 CONCLUSION

As more misinformation and disinformation spreads online, quickly detecting the veracity of a claim is becoming a critical task that can help the public in making more evidence based decisions. While recent years have seen an explosion of research on claim verification and other relevant tasks, many challenges and open questions remain. In this paper we focused on some challenges related to data collection and model development for automated verification of claims; in particular on label imbalance associated with collecting data from fact checkers, and on challenges associated with developing claim verification models. Our future goal is to study the relation between misinformation and COVID-19 infection prevalence. To this end our focus will be on applying automated claim verification models to a large dataset of Tweets, Facebook posts and other claims about COVID-19. We hope our present discussion will prove useful to other researchers wanting to apply automated claim verification in other studies and applications.

REFERENCES

[1] Jai Aggarwal, Ella Rabinovich, and Suzanne Stevenson. 2020. Exploration of Gender Differences in COVID-19 Discourse on Reddit. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Association for Computational Linguistics, Online. https://www.aclweb.org/anthology/2020.

nlpcovid19-acl.13

[2] Jennifer Allen, Baird Howland, Markus Mobius, David Rothschild, and Duncan J. Watts. 2020. Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances* 6, 14 (2020). https://doi.org/10.1126/sciadv.aay3539 arXiv:https://advances.sciencemag.org/content/6/14/eaay3539.full.pdf

[3] Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. 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)*. Association for Computational Linguistics, Hong Kong, China, 4685–4697. https://doi.org/10.18653/v1/D19-1475

[4] Alberto Barron-Cedeno, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, Shaden Shaar, and Zien Sheikh Ali. 2020. Overview of CheckThat! 2020: Automatic Identification and Verification of Claims in Social Media. In *Lecture Notes in Computer Science*. Springer International Publishing, 215–236. https://doi.org/10.1007/978-3-030-58219-7_17

[5] Alexandre Bovet and HernÃąn A. Makse. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications* 10, 1 (Jan 2019). https://doi.org/10.1038/s41467-018-07761-2

[6] Limeng Cui and Dongwon Lee. 2020. CoAID: COVID-19 Healthcare Misinformation Dataset. *ArXiv preprint* (May 2020). arXiv:2006.00885 [cs.SI]

[7] Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, 69–76. https://doi.org/10.18653/v1/S17-2006

[8] Mohamed K. Elhadad, Kin Fun Li, and Fayez Gebali. 2020. Detecting Misleading Information on COVID-19. *IEEE Access* 8 (2020), 165201–165215. https://doi.org/10.1109/access.2020.3022867

[9] Mohamed K. Elhadad, Kin Fun Li, and Fayez Gebali. 2021. COVID-19-FAKES: A Twitter (Arabic/English) dataset for detecting misleading information on COVID-19. *Advances in Intelligent Systems and Computing* 1263 AISC (1 Jan. 2021). https://doi.org/10.1007/978-3-030-57796-4_25

[10] Joshua N. Grant, Varisara Tansakul, Bryan M. Eaton, Gautam Thakur, Martin Smyth, and Monica Smith. 2021. Harmonization Challenges in Data Collection of COVID-19 Misinformation. *Proceedings of 2021 CHI workshop on Opinions, Intentions, Freedom of Expression, ..., and Other Human Aspects of Misinformation Online* (2021).

[11] Shawn Jones, Martin Klein, and Herbert Van de Sompel. 2021. Robustifying Links To Combat Reference Rot. *Code4Lib Journal* 50 (2021).

[12] Tero Karppi and Kate Crawford. 2015. Social Media, Financial Algorithms and the Hack Crash. *Theory, Culture & Society* 33, 1 (May 2015), 73âĂŞ92. https://doi.org/10.1177/0263276415583139

[13] Dilek Kucuk and Fazli Can. 2020. Stance Detection: A Survey. *Comput. Surveys* 53, 1, Article 12 (Feb. 2020), 37 pages. https://doi.org/10.1145/3369026

[14] Sumit Kumar, Raj Ratn Pranesh, and Kathleen M. Carley. 2020. A Fine-Grained Analysis of Misinformation in COVID-19 Tweets. Online. https://www.cmu.edu/ideas-social-cybersecurity/events/2020papers/raj-and-sumet_paper.pdf

[15] Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. 2020. NLP-based Feature Extraction for the Detection of COVID-19 Misinformation Videos on YouTube. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Association for Computational Linguistics, Online. https://www.aclweb.org/anthology/2020.nlpcovid19-acl.17

[16] Shahan Ali Memon and Kathleen M. Carley. 2020. Characterizing COVID-19 Misinformation Communities Using a Novel Twitter Dataset.. In *CIKM (Workshops)*. arXiv:https://arxiv.org/pdf/2008.00791.pdf

[17] Ray Oshikawa, Jing Qian, and William Yang Wang. 2020. A Survey on Natural Language Processing for Fake News Detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 6086–6093. https://www.aclweb.org/anthology/2020.lrec-1.747

[18] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 22–32. https://doi.org/10.18653/v1/D18-1003

[19] Gautam Kishore Shahi and Durgesh Nandini. 2020. FakeCovid – A Multilingual Cross-domain Fact Check News Dataset for COVID-19. *CySoc 2020 International Workshop on Cyber Social Threats, ICWSM 2020* (June 2020). https://doi.org/10.36190/2020.14 arXiv:2006.11343

[20] Lisa Singh, Shweta Bansal, Leticia Bode, Ceren Budak, Guangqing Chi, Kornraphop Kawintiranon, Colton Padden, Rebecca Vanarsdall, Emily Vraga, and Yanchen Wang. 2020. A first look at COVID-19 information and misinformation sharing on Twitter. *arXiv preprint arXiv:2003.13907* (2020). https://arxiv.org/abs/2003.13907

[21] Qi Su, Mingyu Wan, Xiaoqian Liu, and Chu-Ren Huang. 2020. Motivations, Methods and Metrics of Misinformation Detection: An NLP Perspective. *Natural Language Processing Research* 1, 1-2 (June 2020). https://doi.org/10.2991/nlpr.d.200522.001

[22] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Vancouver, Canada, 422–426. https://doi.org/10.18653/v1/P17-2067

[23] Jenifer Whitten-Woodring, Mona S. Kleinberg, Ardeth Thawnghmung, and Myat The Thitsar. 2020. Poison If You DonâĂŹt Know How to Use It: Facebook, Democracy, and Human Rights in Myanmar. *The International Journal of Press/Politics* 25, 3 (May 2020), 407âĂŞ425. https://doi.org/10.1177/1940161220919666