Detecting COVID-19 Misinformation on Social Media

Tamanna Hossain[*,1] **Robert L. Logan IV**[*,1] **Arjuna Ugarte**[*,2] **Yoshitomo Matsubara**[*,1]
Sean Young[2] **Sameer Singh**[1]
[1] Dept of Computer Science, University of Califonia, Irvine
[2] Dept of Emergency Medicine, University of Califonia, Irvine
{tthossai, rlogan, dugarte, yoshitom, syoung5, sameer}@uci.edu

Abstract

The ongoing pandemic has heightened the need for developing tools to flag COVID-19-related misinformation on the internet, specifically on social media such as Twitter. However, due to novel language and the rapid change of information, existing misinformation detection datasets will not be effective for evaluating systems designed to detect misinformation on this topic. To facilitate research on this task, we release a dataset of 4.8K expert-annotated social media posts to evaluate the performance of misinformation detection systems on 86 different pieces of misinformation relating to COVID-19. We evaluate existing NLP systems on this dataset, identifying key challenges for future models to improve upon.

Post: "Coronavirus CV19 was a top secret biological warfare experiment. That is why it is only affecting the poor." **Misconception:** "Coronavirus is genetically engineered." **Label:** Misinformative
Post: "It looks like we are all going to have to wait much longer for a #COVID19 vaccine." **Misconception:** "We're very close to a vaccine." **Label:** Informative
Post: "CDC: Coronavirus spreads rapidly in dense populations with public transit and regular social gatherings." **Misconception:** "Coronavirus cannot live in warm and tropical temperatures." **Label:** Irrelevant

Table 1: **Dataset Examples:** Given a user *post*, we want to identify whether any of the known *misconceptions* are expressed in the post, in particular, is the post spreads misinformation for a given misconception, is informative by contradicting it, or is irrelevant.

1 Introduction

Detecting spread of misinformation such as, rumors, hoaxes, fake news, propaganda, spear phishing, and conspiracy theories, is an important task for natural language processing (Thorne et al., 2017; Shu et al., 2017; Thorne and Vlachos, 2018). Online social media networks provide particularly fertile ground for the spread of misinformation—they lack gate-keeping and regulations, users publish content without having to go through an editor, peer review, verification of qualification, or providing sources, and social networks tend to create "echo chambers" or closed networks of communication insulated from disagreements.

The COVID-19 pandemic has created a pressing need for the development of tools to combat the spread of misinformation. Since the pandemic affects the global community, there is a wide target audience that is seeking information about the topic, whose safety is threatened by adversarial agents invested in spreading of misinformation for political and economic reasons. Furthermore, due to the complexity of the medical and public health issues involved, it is also difficult to be completely accurate and factual about the information, leading to disagreements that get exacerbated with misinformation. This difficulty is compounded by the rapid evolution of knowledge regarding the disease. As researchers learn more about the virus, statements that previously seemed true may turn out to be false, and vice versa. Detecting this spread of pandemic-related misinformation, thus, has become a critical problem, and received significant attention from government and public health organizations (WHO, 2020), online social media platforms (TechCrunch, 2020), and news agencies ().

To facilitate research in automatic COVID-19 misinformation detection, we have collected a dataset of 86 common misconceptions about the disease along with 4.8K related social media posts, identified and annotated by researchers from the UCI School of Medicine. Given an online user post, our data identifies whether any of the known misconceptions are expressed by the post, and if

[*] First four authors contributed equally.

so, whether the post propagates the misconception (*misinformative*) or is *informative* by contradicting it. Example misconception-post pairs are provided in Table 1 for illustration.

We additionally provide benchmark results measuring the performance of existing NLP models on this task. We evaluate text similarity models on their ability to detect whether the post is relevant to the misconception or not (or identify the most relevant misconception for the post, if any), as textual similarity cannot be used to detect whether the post is informative, or misinformative, about a misconception. Additionally, since the class labels in misinformation detection (*misinformative*, *informative*, *irrelevant*) are somewhat analogous to the *entailment*, *contradiction*, and *neutral* labels in natural language inference (NLI), we also evaluate existing models for this task on how well they are able to detect COVID-19 related misinformation. Our results show that existing NLP models, when used without annotated dataset for the task, fare quite poorly in detecting misinformation, and we thus hope to initiate research in this area.

2 Problem Setup

Given a collection of positively phrased misconceptions $M = \{m_1, \ldots, m_{|M|}\}$ (e.g., "Wearing masks does not prevent spread of COVID-19." is a misconception), and a collection of sentences (e.g., social media posts) $P = \{p_1, \ldots, p_{|P|}\}$, the task is to determine, for each sentence p, whether there exists a misconception $m \in M$ that is being discussed, and if so, whether the discussion is *informative* (e.g., identifies m as false) or *misinformative* (e.g., identifies m as true). This task can be naturally separated into the two following steps:

1. **Misconception Retrieval:** Given p return a subset $M_p \subseteq M$ of relevant misconceptions.
2. **Pairwise Classification:** For each (m, p) pair ($m \in M_p$), predict whether the text is *informative*, *uninformative*, or *irrelevant*.

Due to limited availability of labeled data, we only apply existing NLP models to these sub-tasks. For the misconception retrieval sub-task we rank relevant misconceptions by measuring the semantic similarity between the post and each misconception. For the pairwise classification sub-task, we recast the problem as a natural language inference (NLI) problem, mapping p to the premise, m to the hypothesis, and Misinformative, Informative, and Irrelevant to entailment, contradiction, and neutral labels in NLI, respectively.

These techniques fall within the framework of detecting misinformation using content features (Volkova et al., 2017; Wei and Wan, 2017). Other approaches include using crowd behaviour (Tschiatschek et al., 2018; Mendoza et al., 2010), reliability of the source (Lumezanu et al., 2012; Li et al., 2015), knowledge graphs (Ciampaglia et al., 2015), or a combination of these approaches (Castillo et al., 2011; Kumar et al., 2016).

3 Dataset Collection

Due to novel language used to describe the disease and its associated misconceptions, existing misinformation detection dataset are unlikely to be effective for evaluating systems designed to detect COVID-19-related misinformation on social media. We collect an evaluation dataset for this task, and describe the collection process below.

Misconceptions We extract a set of misconceptions from a Wikipedia article about misinformation related to the COVID-19 pandemic (Wikipedia, 2020). The extracted statements are manually examined, and statements that are not misinformation are removed. Misinformation statements are then manually rephrased to a positive expression of that misinformation, e.g. "Some conspiracy theorists also alleged that the coronavirus outbreak was cover-up for a 5G-related illness" is shortened to "Coronavirus is caused by 5G". Sources of these misconceptions are vetted for reliability and given a reliability score between 0-5 (see Appendix A).

Tweets Our main source of tweets is from COVID-19-related tweets identified by Chen et al. (2020). We only use tweets from March and April 2020, and filter out non-English tweets.

Annotation Process To help identify tweets related to our list of misconceptions, we use BERTScore (Zhang et al., 2019) to compute a similarity metric on tweet-misconception pairs. For each given misconception, the 100 most similar tweets are selected for annotation. Each of these tweet-misconception pairs is manually labeled by researchers in the UCI School of Medicine as either: *misinformative* (tweet is a positive expression of the misconception), *informative* (tweet contradicts/disagrees with the misconception), or *irrelevant* (tweet is not relevant to the misconception).

Class	Count	Percentage
Misinformative	465	9.7 %
Informative	164	3.4 %
Irrelevant	4,161	86.9 %
Total	4,790	100 %

Table 2: Distribution of labels in the annotations

Dataset Statistics The current dataset contains 86 misconceptions, along with 4,790 annotated tweet-misconception pairs. Statistics about the distribution of labels are provided in Table 2. The balance of labels is heavily skewed, containing mostly irrelevant tweets, reflecting the relative infrequency of misinformation spread vs. other COVID-19-related discussion on social media. This dataset, however, is an evolving dataset; we are continually identifying additional misconceptions, as well as collecting more tweet annotations.

4 Performance of Benchmark Models

Supervised classifiers have been used extensively for detecting misinformation, such as LIAR (Wang, 2017; Karimi et al., 2018), BuzzFeedNews (Shu et al., 2017), and FakeNewsNet (Shu et al., 2019), including the framing as NLI in FEVER (Thorne et al., 2018) and the Fake News Classification challenge (Yang et al., 2019). However, these tasks operate with static or slowly evolving domains, on topics that do not require specific expertise to annotate. It is very challenging to gather an annotated dataset large enough to be a useful for detecting misinformation related to COVID-19 - misconceptions and how they are expressed in the language evolve very quickly, and identifying whether something is a misconception requires expertise in public health and medicine. Instead, we evaluate the performance of existing NLP models that have been trained on related tasks and datasets, ported to our setup as described in Section 2.

4.1 Evaluation Metrics

For a given social media post p, we need to identify the misconception that is expressed, and whether p is misinformative or not towards it, with the *true* relevant misconception, m_p^* and its expression given by the labeled data (we omit the tweets for which there is no relevant misconception from this evaluation). We rank all the misconceptions in M for the post p in decreasing order of the score, and observe the rank of m_p^*. For similarity models, we rank the misconceptions based on relevance, whereas for the NLI models, we use the probability of the annotated class, i.e. by entailment for Misinformative, and report these metrics on the all posts that are relevant to *some* misconception, as well as on a subset of posts that are Misinformative. Performance on the misconception retrieval sub-task is evaluated using Hits@k for $k = 1, 5, 10$ and Mean Reciprocal Rank (MRR). For classification, we additionally compute Precision, recall, and F1 by using the class with the highest probability as the prediction, reported separately for *misinformative* and *informative* classes.

4.2 Sentence Similarity Models

Word Representations Non-contextual word embeddings provide static vectorized representations of word tokens. Here we use TF-IDF and GloVe embeddings (Pennington et al., 2014) to obtain vectorized representations, $\vec{p}$ and $\vec{m}$, for posts and misconceptions before computing the cosine similarity score between them. We use 300D GloVe embeddings pretrained on 2014-Wikipedia and Gigaword, and average over token embeddings to compute sentence vectors. NLTK is used for tokenization and vectorization.

Contextual Embeddings Unlike static word embeddings like GloVe, contextualized word embeddings incorporate the context of a word's usage into its vectorized representation. We use an open RoBERTa-base (Liu et al., 2019) implementation[1] to obtain contextual word embeddings for each token in p and m, and use two models of textual similarity: (1) cosine similarity between sentence vectors $\vec{p}$ and $\vec{m}$, computed by averaging over the token vectors, and (2) BERTScore (Zhang et al., 2019) between p and m, which involves adding cosine similarities between RoBERTa token embeddings of p and m to obtain precision and recall values, and using the F1-score as similarity.

Domain Adaptation Since pretrained language models have not been trained on COVID-19 related text or on posts from Twitter, we use domain-adaptive pretraining that has been used to adapt these models to different domains (Gururangan et al., 2020). We finetune RoBERTa-base using a collection of tweets associated with COVID-19 (Chen et al., 2020), and recompute the two similarities that use contextual embeddings.

[1] Available at: https://huggingface.co/roberta-large

Model	Misinformative				*Relevant*			
	H@1	H@5	H@10	MRR	H@1	H@5	H@10	MRR
Cosine Sim., TF-IDF	30.3	61.1	79.6	0.46	31.5	61.5	79.0	0.47
Cosine Sim., Avg. GloVe	12.3	48.8	62.2	0.28	14.5	49.6	63.6	0.30
Cosine Sim., Avg. RoBERTa Embds.	11.4	37.8	52.5	0.24	10.0	34.0	48.8	0.22
BERTScore	32.3	58.5	72.7	0.46	34.2	59.5	73.9	0.48
with Domain Adaptation								
Cosine Sim., Avg. RoBERTa Embds.	15.3	52.7	63.2	0.31	13.0	49.3	63.8	0.29
BERTScore	44.7	78.9	85.4	0.61	48.2	79.3	86.5	0.63

Table 3: **Semantic similarity models**. We present evaluation only for detection of the misinformation class, along with an evaluation on relevance detection, i.e. on tweets that are either Misinformative or Informative.

Model	Misinformative					Informative				
	P	R	F1	H@5	MRR	P	R	F1	H@5	MRR
Trained on SNLI										
Linear, Bag-of-Words	6.9	17.8	9.9	10.3	0.06	6.1	57.3	11.1	36.6	0.19
Linear, Avg. GloVe Embeddings	14.1	30.3	19.2	12.0	0.10	1.9	15.9	3.4	0.0	0.03
Sentence-BERT	8.0	9.0	8.5	9.5	0.10	2.6	26.8	4.7	0.0	0.02
Trained on MNLI										
Linear, Bag-of-Words	8.8	52.7	15.1	9.0	0.08	10.1	48.8	16.7	16.5	0.15
Linear, Avg. GloVe Embeddings	16.7	64.3	26.5	3.9	0.05	2.7	23.8	4.8	17.7	0.09
Sentence-BERT	16.2	30.5	21.2	32.0	0.22	5.1	48.8	9.3	7.9	0.09

Table 4: **NLI models**. Classification and ranking evaluation metrics for Misinformative or Informative classes.

Results Among the similarity models, the domain-adapted BERTScore performs the best at misconception retrieval, achieving the highest Hits@k and MRR across both misinformative and relevant classes. Cosine similarity with TF-IDF and BERTScore without domain adaptation perform similarly. Hits@1 is higher for BERTScore (non-DA) for both classes, however, Hits@10 is higher for TF-IDF cosine similarity. We see from this that domain adaptation is salient for performing misconception retrieval using semantic similarity. However, even though BERTScore (DA) had the highest ranking metrics there is room for improvement since Hits@1 is only 44.7%.

4.3 Textual Entailment Models

Since the classes in misinformation detection correspond to those in natural language inference (NLI), we evaluate classifiers trained on existing datasets for this task. Specifically, we train three NLI models on SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) with a linear classifiers on top of: (1) concatenated unigram and bigram TF-IDF vectors for each input, (2) concatenated average GloVe embeddings for each input, and (3) the Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) representation that uses siamese and triplet networks to obtain semantically meaningful sentence embeddings.

Unfortunately, none of these fared well at the task of pairwise classification (Table 4), with higher recalls than precision across models and for both classes. Models trained on MultiNLI generally performed better, likely benefiting from the varied sources of text in that dataset. The highest precision (16.7 %) and recall (64.3 %) for the misinformative class uses avg. GloVe embeddings trained on MNLI, while the SNLI version obtains the highest Hits@5 of 12 %. This, still, is much worse than the semantic similarity models (Table 3).

5 Conclusions and Future Work

In this paper, we introduced a benchmark for detecting COVID-19 related misinformation on social media, containing known misconceptions and their *misinformative* and *informative* expressions on Twitter, annotated by experts. Off-the-shelf NLP models, however, do not perform well on this data, indicating a need for further research and development on this topic. We plan to continually expand our annotated dataset by including posts from other domains such as news articles, and misconceptions from sources beyond Wikipedia, such as Poynter (2020).

Acknowledgements

We would like to acknowledge Nicole Woodruff, an undergraduate at UCLA, and Aileen Guillen, a medical student at UCI, for volunteering to help annotate data for this project. We would like to acknowledge Lidia Flores, a staff research associate in Dr. Young's Lab, for her contributions in compiling tweets for our future research.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684.

Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, 6(2):e19273.

Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PloS one*, 10(6):e0128193.

Suchin Gururangan, Ana Marasovi, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*.

Hamid Karimi, Proteek Roy, Sari Saba-Sadiya, and Jiliang Tang. 2018. Multi-source multi-class fake news detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1546–1557.

Srijan Kumar, Robert West, and Jure Leskovec. 2016. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*, pages 591–602.

Yaliang Li, Qi Li, Jing Gao, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. 2015. On the discovery of evolving truth. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 675–684.

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*.

Cristian Lumezanu, Nick Feamster, and Hans Klein. 2012. # bias: Measuring the tweeting behavior of propagandists. In *Sixth International AAAI Conference on Weblogs and Social Media*.

Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. 2010. Twitter under crisis: Can we trust what we rt? In *Proceedings of the first workshop on social media analytics*, pages 71–79.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Poynter. 2020. The coronavirusfacts/datoscoronavirus alliance database: https://www.poynter.org/ifcn-covid-19-misinformation/. [Accessed on June 30, 2020].

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 312–320.

TechCrunch. 2020. Facebook, reddit, google, linkedin, microsoft, twitter and youtube issue joint statement on misinformation. [Accessed on June 30, 2020].

James Thorne, Mingjie Chen, Giorgos Myrianthous, Jiashu Pu, Xiaoxuan Wang, and Andreas Vlachos. 2017. Fake news stance detection using stacked ensemble of classifiers. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 80–83, Copenhagen, Denmark. Association for Computational Linguistics.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The fact extraction and verification (fever) shared task. *arXiv preprint arXiv:1811.10971*.

Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. 2018. Fake news detection in social networks via crowd signals. In *Companion Proceedings of the The Web Conference 2018*, pages 517–524.

Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 647–653.

William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Wei Wei and Xiaojun Wan. 2017. Learning to identify ambiguous and misleading news headlines. *arXiv preprint arXiv:1705.06031*.

WHO. 2020. Novel coronavirus(2019-ncov) situation report - 13. [Accessed on June 30, 2020].

Wikipedia. 2020. Misinformation related to the covid-19 pandemic: https://en.wikipedia.org/wiki/Misinformation_related_to_the_COVID-19_pandemic.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Kai-Chou Yang, Timothy Niven, and Hung-Yu Kao. 2019. Fake news detection as natural language inference. *arXiv preprint arXiv:1907.07347*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Reliability Codes

- A score of 5 were sources that clearly debunked the paired misinformation, cited evidence, and were from a reputable source or well-known fact-check website (e.g. CDC, Snopes).
- A score of 4 were sources that debunked the paired misinformation and were from a reliable source (e.g. News sites).
- A score of 3 were sources that refuted the misinformation but were not a well-known source or contained a lot of filler in their article not about the misinformation.
- A score of 2 were sources that labeled the misinformation as false or untrue but did not really provide evidence.
- A score of 1 were sources that did not refute the misinformation or were more descriptive.
- A score of 0 were sources that did not refute the misinformation and may actually support it.
- A score of NA for sources not in English.