Drink bleach or do what now? Covid-HeRA: A dataset for risk-informed health decision making in the presence of COVID19 misinformation

Anonymous ACL submission

Abstract

Given the wide spread of inaccurate medical advice related to the 2019 coronavirus pandemic (COVID-19), such as fake remedies, treatments and prevention suggestions, misinformation detection has emerged as an open problem of high importance and interest for the NLP community. To combat potential harm of COVID19-related misinformation, we release Covid-HeRA, a dataset for health risk assessment of COVID-19-related social media posts. More specifically, we study the severity of each misinformation story, i.e., how harmful a message believed by the audience can be and what type of signals can be used to discover high malicious fake news and detect refuted claims. We present a detailed analysis, evaluate several simple and advanced classification models, and conclude with our experimental analysis that presents open challenges and future directions.

1 Introduction

While an increasing percentage of the population relies on social media platforms for news consumption, the reliability of the information shared remains an open problem. Fake news and other types of misinformation have been widely prevalent in social media, putting audiences at great risks globally. Detecting and mitigating the impact of misinformation is therefore a crucial task that has attracted research interest, with a variety of approaches proposed, from linguistic indicators to deep learning models (Bal et al., 2020). Several research endeavors tackle key issues, such as mitigating label scarcity with additional weak social supervision signals, improving intractability with attention mechanisms, leverage network, group and/or user information, etc. (Jin et al., 2016; Ruchansky et al., 2017; Shu et al., 2019; Wang et al., 2020; Lu and Li, 2020).

Fake news frequently emerge for certain phenomena and topics, e.g., public health issues, politics etc. (Allcott et al., 2019; Shin et al., 2018; Bode and Vraga, 2018). Unsurprisingly, the same applies for the current global pandemic, where inaccurate stories are surfacing daily. It is often difficult for users, that decide to take action based on health advice found online, to understand the consequences and potential risks from following unreliable guidance, especially when all information spread by influential users is perceived as equally credible (Morales et al., 2020). This adds to the worry and anxiety felt by many, already in a difficult situation (Kleinberg et al., 2020).

While much work has been focused on identifying health-related misinformation, there has been little attention on making a distinction between the seriousness of misinformation (Fernandez and Alani, 2018). Severity varies greatly across each message: some might be jokes, some might be discussing the impact of fake news or refute the claim, others might be highly malicious, or others might be simply inaccurate information with little impact that results in no harm. The severity of each message can vary depending on its content, e.g., urging users to eat garlic is less severe than urging users to drink bleach. Several news articles are posted daily related to COVID19, which capture weak signals of misinformation severity. However, identifying the severity level for each misinformation story is a challenging task that has not been previously studied.

To help reduce the impact of COVID19 misinformation on potential health-related decisions of users, we study the severity of misinformation detection. In contrast with previous works that treat misinformation as a binary classification task, we build a novel health risk assessment misinformation benchmark dataset, Covid-HeRA, that contains social media posts annotated on a finer scale, based on whether the message content is: a) real news, b) inaccurate or misinformation or c) refutes/rebuts

a specific claim or news article. In addition, posts labeled as misinformation are judged based on their potential to impact user's health, assuming the individual might be making decisions upon the advice and suggestions read in the post. In other words, our goal is to produce a label that reflects the level of risk factors in the presence of inaccurate claims and news, conditioned on the worst-case assumption that the user will follow the advice.

We present our data analysis that reveals several key insights about the most prominent unreliable news and evaluate several baselines, as well as state-of-the-art models and variations. We hope to guide research on developing risk-aware misinformation deterrence algorithms. To facilitate future research, the Covid-HeRA data, the data analysis and the baseline models are open sourced for public usage[1].

2 Related Work

Health-related misinformation research spans a broad range of disciplines including computer science, social science, journalism, psychology, and so on (Dhoju et al., 2019; Castelo et al., 2019; Fard and Lingeswaran, 2020). While health-related misinformation is only a facet of misinformation research, there has been much work analysing misinformation in different medical domains, such as cancer (Bal et al., 2020; Loeb et al., 2019), orthodonics (Kılınç and Sayar, 2019), sexually transmitted disease and infections (Zimet et al., 2013; Joshi et al., 2018), autism (Baumer and McGee, 2019), influenza (Culotta, 2010; Signorini et al., 2011), and more recently COVID-19 (Garrett, 2020; Brennen et al., 2020; Cinelli et al., 2020; Cui and Lee, 2020).

Health Misinformation on Social Media The web and social media data have been used to monitor influenza prevalence and awareness online (Smith et al., 2016; Ji et al., 2013; Huang et al., 2017). Systems such as Google Flu Trends use real-time signals, such as search queries, to detect influenza epidemics (Ginsberg et al., 2009; Preis and Moat, 2014; Santillana et al., 2014; Kandula and Shaman, 2019). However, relying solely on the search queries leads to an overestimation of influenza, namely because there is no distinction between general awareness about the flu and searches for treatment methods (Smith et al., 2016; Klembczyk et al., 2016). Our work focuses on social media, in particular health misinformation on micro-blogging sites, such as Twitter. Tomeny et al. (2017) examined geographical and demographical trends in anti-vaccine tweets. They analyzed anti-vaccine tweets with respect to autism spectrum disorder, and trained a classifier to predict a binary label for anti-vaccine using features such as unigrams, bigrams, word occurrence counts, punctuation, and location. Our work goes beyond such a binary classification, as our model is able to further categorize misinformation on a set of fine-grained severity scale.

Baumer and McGee (2019) apply topic modeling to an autism spectrum disorder (ASD) blogging community dataset with the goal of understanding the representation, delegation and authority of such a method. In a recent workshop on automatic classification of influenza (flu) vaccine behavior mentions in tweets (Weissenbacher et al., 2018), the top performing system compared deep learning models with pre-trained language models with an LSTM classifier and an ensemble of statistical classifiers with task-specific features which resulted in comparable performances. An error analysis showed that vaccine hesitancy was conflicting with vaccination behavior (Joshi et al., 2018). Huang et al. (2017) examine the geographic and demographic patterns of the flu vaccine in social media. Recent work, has also focused on identifying users disseminating misinformation, in the case of cancer treatments (Ghenai and Mejova, 2018), as well as hybrid approaches combining user-related features with content features (Ruchansky et al., 2017).

Early works on COVID-19 misinformation With the threat of COVID-19 misinformation to public health organizations, there has been several call-to-actions (Chung et al., 2020; Mian and Khan, 2020; Calisher et al., 2020) to underscore the gravity and impact of COVID-19 misinformation Garrett (2020). Tasnim et al. (2020) outlines several potential strategies to ensure effective communication on COVID-19. Among the recommendations are ensuring up-to-date reliable information, via identifying fake news, or misinformation.

Brennen et al. (2020) analyse the different types, sources, and claims of COVID-19 misinformation, and show that the majority appear on social media outlets. As the dialog on the pandemic evolves, so does the need of reliable and trustworthy information online (Cuan-Baltazar et al., 2020). Pennycook et al. (2020) show that, people tend to believe false

[1]Code available at https://github.com/TIMAN-group/covid19_misinformation

claims about COVID-19 and share false information when they do not think as critically about the accuracy and veracity of the information.

Kouzy et al. (2020) annotated about 600 messages containing hashtags about COVID-19, they discovered that about 25% of messages contain some form of misinformation and about 17% contain some unverifiable information. Singh et al. (2020) provide a large-scale exploratory analysis of how myths and COVID-19 themes are propagated on Twitter, by analysing how users share URL links. Cinelli et al. (2020) cluster word embeddings to identify topics and measure engagement of users on several social media platforms. They provide a comparative study of information reproduction and provide rumor amplification parameters for COVID-19 on these platforms.

The coronavirus pandemic has lead to several measures enforced across the globe, from social distancing and shelter-in-place orders to budget cuts and travel bans (Nicola et al., 2020). In addition, news circulate daily advice for the public, with suggestions that help prevent the spread and precautions to keep the infection and mortality rates low. Some articles, however, contain fake remedies that reportedly cure or prevent COVID19, promote false diagnostic procedures, report incorrect news about the virus properties or urge the audience to avoid specific food or treatments that might make symptoms worse or the reader more likely to contract the virus[3]. With such information overload, any decision making procedures based on misinformation have high likelihood of severly impacting one's health (Ingraham and Tignanelli, 2020). Therefore, we aim to predict the severity of incorrect information released on social media, as well as detect any posts that refute or rebut unreliable claims and suggestions on coronavirus misinformation.

In the next sections, we first describe our data collection and annotation methodology. We present statistics and examples of our dataset (Section 3). Subsequently, in Section 4 we present experimental results with several baseline and state-of-the-art machine learning models. Finally, we conclude with a discussion and possible future extensions (Section 5).

[3] https://www.cmu.edu/ideas-social-cybersecurity/research/coronavirus.html

3 Dataset

We introduce our data source and annotation strategy. Moreover, we present detailed statistics and data analysis that shows key insights on the most prevalent harmful misinformation online.

3.1 Data Construction

The goal of creating a new misinformation benchmark dataset is two-fold. First, we want to highlight the importance of understanding the impact of COVID-19 misinformation in health-related decision making and which behavioral aspects are affected by the digital spread of inaccurate harmful advice. More importantly, we aim to flag unreliable posts based on the potential risk and severity of the statements, so that users stay informed on the consequences of incorrect health advice when making decisions.

Thus, we seek to target misinformation on a finer annotation scale, based on whether it has the potential to guide the audience towards health-related decisions or behavioral changes with high risk factor, i.e., high likelihood of severely impacting one's health. To this end, we frame the task as a multi-class classification problem, where each social media post is categorized as: a) **Real News/Claims**, i.e., reliable correct information, b) **Refutes/Rebuts**, i.e., refutation or rebuttal of an incorrect statement, c) **Not severe**, i.e., misinformation but unlikely to result in risky behavioral changes or harmful decisions, d) **Possibly severe** misinformation, with possible severe health-related impact and e) **Highly severe** misinformation with increased potential risks for any individual following the advice & suggestions expressed in the social media post content.

These categories enable researchers to study the impact of coronavirus health misinformation at a finer granular level, to develop algorithms that caution the audience on the potential risks and to design systems that present unbiased information, i.e., both the original - potentially unreliable - post, along with any possible rebutting claims expressed online. In Table 1, we present example posts for each category and further describe our annotation process in the following paragraphs.

We make use of CoAID, a large scale healthcare misinformation data collection related to COVID19 with *binary* ground truth labeled news articles and claims, accompanied with associated tweets and user replies (Cui and Lee, 2020). This dataset pro-

COVID-19 tweets by category

Category	Tweet	Reasoning
Real News/Claims	"*What is a coronavirus? Large family of viruses. Some can cause illness in people or animals. In humans, it's known to cause respiratory infections*"	This category includes correct facts and accurate news.
Not severe	"*Vatican confirms Pope Francis and two aides test positive for Coronavirus*"	This is misinformation, but behavioral changes of are less likely to occur.
Possibly severe	"*Vitamin C Protects Against Coronavirus*"	Although an individual may decide to take daily doses of vitamin C, it is unlikely to be harmful and potential risks are less significant than for other actionable items.
Highly severe	"*Good News: Coronavirus Destroyed By Chlorine Dioxide _ Kerri Rivera*" "*Flu Vaccine Increases Coronavirus Risk 36% Says Military Study*" "*People of color may be immune to the Coronavirus because of Melanin blackmentravels*"	These tweets either promote specific behavioral changes and fake remedies [2] with increased health risks, or may result in increased exposure for certain socioeconomic groups.
Refutes/Rebuts Misinf/tion	"*For those who think COVID-19 is just like the flu: In the 2018-19 flu season, there were 34,000 deaths over a period of months and months. In contrast, there have been over 40,000 COVID-19 deaths in a little more than a month and half even with social distancing.*"	These type of posts are useful in identifying fake claims, as well as presenting opposing views.

Table 1: COVID-19 example tweets labeled on a severity scale, alongside with a brief explanation of each category.

vides us with a large amount of reliable twitter data and alleviates the need for labeling tweets as real or fake. Furthermore, it has the potential to be updated automatically with additional instances, enabling Covid-HeRA semi-supervised models as future work.

To obtain annotations based on our defined severity categorization, all tweets labeled in CoAID as misinformation are shuffled and distributed to two different annotators. Each annotator is asked to judge whether any decisions or other actionable items can be taken based on the expressed content, and whether those could result in harmful choices, risky behavioral changes or other severe health impacts.

Additionally, we asked annotators to flag any post that expresses an opinion or argument against the unreliable claims, i.e., refutes or rebuts misinformation (see Figure 1 for a screenshot of our annotation interface). To assess agreement levels, an external validator was asked to annotate a random sample of the labeled tweets. The kappa score between the annotators and the validator was 0.7037, which shows good agreement on the task (Hunt, 1986). A final round was introduced as an additional step, in order to resolve conflicts on ambiguous instances. The total number of tweets labeled

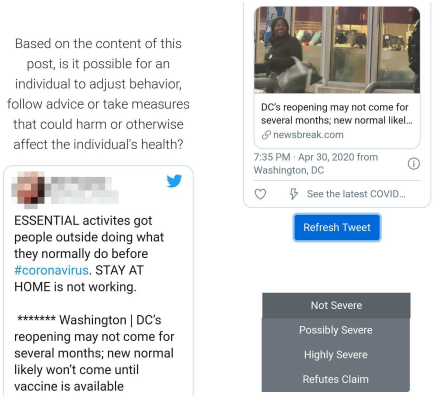


Figure 1: A screenshot of a tweet example and possible annotation options.

Category	#Tweets	#Tokens (Vocab)
Not severe	1,851	3,324
Possibly severe	439	11,171
Highly severe	568	16,328
Refutes/Rebuts	447	2,244
Real News/Claims	57,981	51,478
Total	61,286	84,545

Table 2: Covid-HeRA Dataset Statistics

per category, alongside with the number of unique words, are presented in Table 2.

3.2 Data Analysis

We first identify the most frequent discriminative terms per category, i.e., terms that appear very often in a category but are infrequent in the remaining categories. We use a document frequency threshold of 0.5% to discard terms that are very commons across the whole data collection, for example "COVID19" or "virus" appearing in more than half of the tweets. In Figure 2, we visualize the top-30 terms per category, with each term weighted by its representativeness. When comparing the "Not severe" category with the other severe categories, we see that many of the terms here refer to conspiracies about COVID-19, such as "artificially", "labmade", "bioweapon" which pertains to the conspiracy that COVID-19 is a man-made virus. The top terms for the "Highly severe" category seem to be about treatments and are more risky words such as "risk", "mask", "cure", "vaccines", and "hydroxychloroquine". The top terms for the "Refutes/Rebuts" contains terms such as "myths", "weaponized", "lying", and "antibiotics" as the messages in this category addressed and debunked conspiracies and misinformation, while the top terms for the "True News/Claims" are "resources", "symptoms", "testing", and "guidance", as these messages are generally informative and provide advice about COVID-19.

We visualize the compactness of each category in Figure 3. We use pre-trained BERT embeddings (Devlin et al., 2018) to measure how close each tweet is by the centroid of its corresponding category, based on their document vectors. We hypothesize that compact categories are more likely to be well-formed and thus easier to classify. We compute the skewness and kurtosis for each distribution. Each distribution shows a positive skewness, and as we can see, they are right-tailed distributions. The "Possibly severe" category is the most skewed and the "Highly severe" category is the least. The negative kurtosis of both the "Highly severe" and "Refutes/Rebuts" categories shows that these categories have less of a peak and appear more uniform, which we is also evident by the flatness of these curves. This may be due to the broad range to topics covered in both of these categories compared to the rest.

In Figure 4, we analyze the top-10 frequent hashtags per category. We remove common hashtags such as "#covid_19", "#coronavirus", etc. The length of each bar indicates how frequently the hashtag appears. We find that the "Not severe" category follows a similar pattern to Figure 2, in that the top hashtags are pertaining more to rumors and conspiracies, such as the "#pope" tested positive for COVID-19, or that COVID-19 is a "#bioweapon". This maybe attributed to the fact that those susceptible to misinformation, are less likely to think critically about news sources and thus tend to believe more false claims (Pennycook et al., 2020). Both severe categories focus on remedies, e.g. "#vitaminc" and vaccination. Interestingly, the Refutes/Rebuts top hashtags had terms associated to computation such as "#dataviz", "#tableau", as well as hashtag to promote social distancing "#stayhome". These hashtags may be evidence of several infographics and data visualizations shared in social media, often used as arguments against misinformation.

Finally, we present the most common claims and news per category, and their corresponding frequency (Table 3).

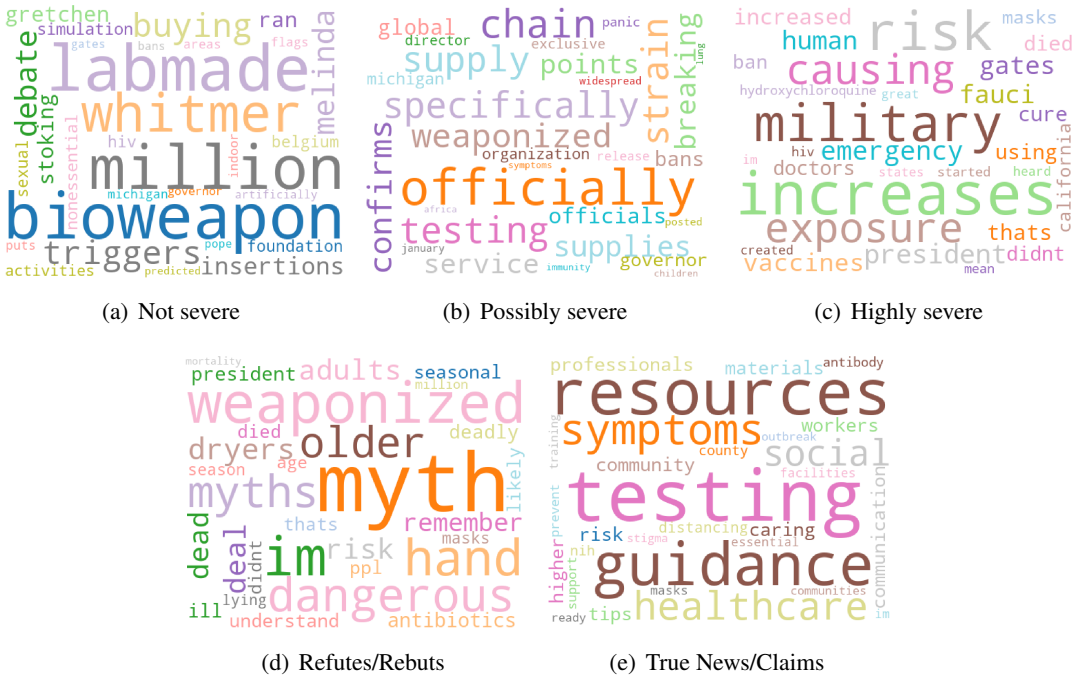
4 Experiments

To showcase the open challenges of risk-based labeling of misinformation, we perform experiments with several baselines and state-of-the-art multi-class classification models. We pre-process tweets to filter out reserved tokens, such as *RT* or *retweet*, urls and mentions. We then split the data into 80% training and 20% testing, keeping the same splits across all models for fair comparison.

The algorithms we experiment with are the following:

Random Forests with bag-of-words (RF-TFIDF) or 100-dimensional pre-trained Glove embeddings (RF-Glove) as text representation

Support Vector Machine with bag-of-words



(a) Not severe (b) Possibly severe (c) Highly severe

(d) Refutes/Rebuts (e) True News/Claims

Figure 2: Most common terms per category

Common COVID-19 Twitter misinformation

Claims/News	Category	Frequency
COVID-19 testing (viral test procedures and information)	Real News/Claims	481
COVID-19 is more contagious than the flu	Real News/Claims	466
Coronavirus Hoax	Highly Severe	338
COVID-19 is just like the flu	Refutes/Rebuts	287
Lab-Made Coronavirus Triggers Debate	Not severe	152
Michigan Governor Gretchen Whitmer Bans Buying US Flags During Lockdown	Not severe	148
Shanghai Government Officially Recommends Vitamin C	Possibly severe	102
Flu Vaccine Increases Coronavirus Risk 36% Says Military Study	Highly severe	99
Vitamin C Protects Against Coronavirus	Possibly severe	79
Coronavirus [is not a] Hoax	Refutes/Rebuts	73

Table 3: Most common claims with their corresponding frequency and gold-truth labels.

(SVM-TFIDF) or 100-dimensional pre-trained Glove embeddings (SVM-Glove)

Logistic Regression with bag-of-words (LR-TFIDF) or 100-dimensional pre-trained Glove embeddings (LR-Glove), same as for SVM and RF.

Bi-directional LSTM model (Schuster and Paliwal, 1997) with 100-dimensional pre-trained Glove embeddings as initial representation (LSTM).

Multichannel CNN convolutional neural network with multiple kernel sizes and 100-dimensional pre-trained Glove embeddings as initial representation, similar to (Kim, 2014) (CNN).

Task-specific BERT fine-tuned on our downstream text classification task, initialized with general-purpose BERT embeddings (Devlin et al., 2018).

We additionally test whether incorporating additional sources of information, such as user replies or news articles with related content can improve predictive performance. To this end, we train **dEFEND** (Shu et al., 2019), a state-of-the-art fake news detection model that builds upon a Hierarchical Attention Network (Yang et al., 2016), by adding co-attention between two textual sources; in our case either tweet replies or corresponding news content.

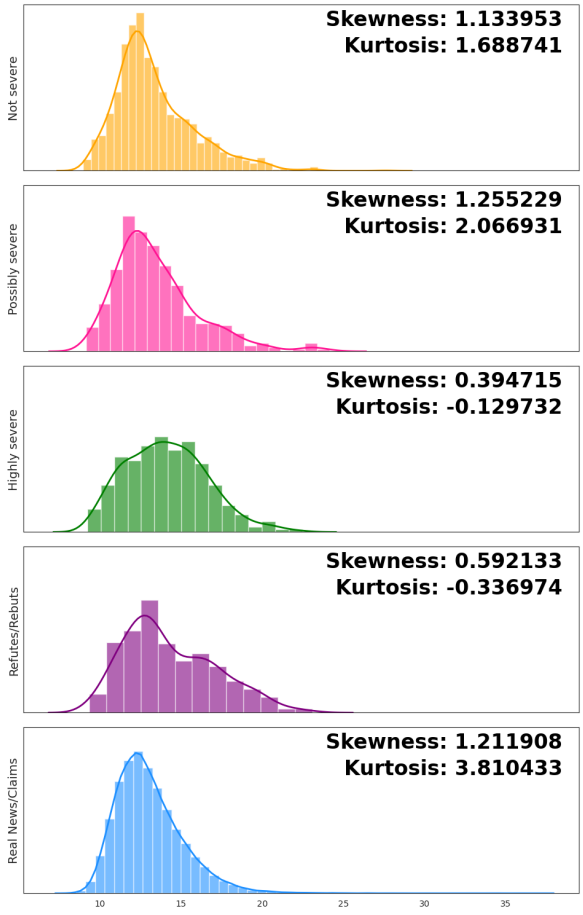


Figure 3: Category compactness, measured by the distance distribution of twitter posts, relative to its category centroids

Our evaluation metrics are accuracy, precision, recall, and F1 score. We train with cross-entropy, Adam and 50 minibatch size for all models. In Table 4, we report the average score of 3 independent trials, i.e., run each model three times with different seeds.

In terms of F1 score, CNN and LSTM models perform slightly better than simpler baselines. Surprisingly, incorporating user engagement features, news content or contextualized pretrained embeddings did not help [4]. We note however that BERT and dEFEND (co-attending on news content) have higher recall, suggesting that ensemble models could further improve performance.

One of the main challenges of health-related misinformation with finer granularity is that some categories might be substantially underrepresented, i.e., risk assessment in misinformation creates a high imbalanced problem, especially for some topics, that is even more difficult to tackle than misinformation detection. To test this hypothesis, we perform the same experiments in a common binary classification setting. More specifically, we discard all refutation and rebuttal tweets and collapse all tweets labeled as misinformation in a common label, irrespective of the severity label, essentially backtracking to a real vs. fake traditional framework. We evaluate the same set of algorithms and present results in Table 5. Compared to the fine-grained labels of Covid-HeRA, the binary classification task produces higher performance across all evaluation metrics, highlighting the limitations of our finer categorization setting. Despite being an important task, i.e., including key goals such as distinguishing between harmful social media medical advice and refuted claims, health misinformation risk-assessment presents many challenges.

	Accuracy	Precision	Recall	F1
RF-TFIDF	0.956	0.286	0.236	0.246
RF-Glove	0.963	0.256	0.200	0.200
SVM-TFIDF	0.946	0.243	0.203	0.203
SVM-Glove	0.960	0.330	0.233	0.246
LR-TFIDF	0.946	0.276	0.220	0.230
LR-Glove	0.963	0.313	0.233	0.250
BiLSTM	0.960	**0.396**	0.273	0.286
CNN	0.946	0.340	0.280	**0.290**
BERT	0.956	0.370	0.305	0.275
dEFEND w.replies	0.950	0.250	0.220	0.234
dEFEND w.news	0.954	0.350	**0.310**	0.250

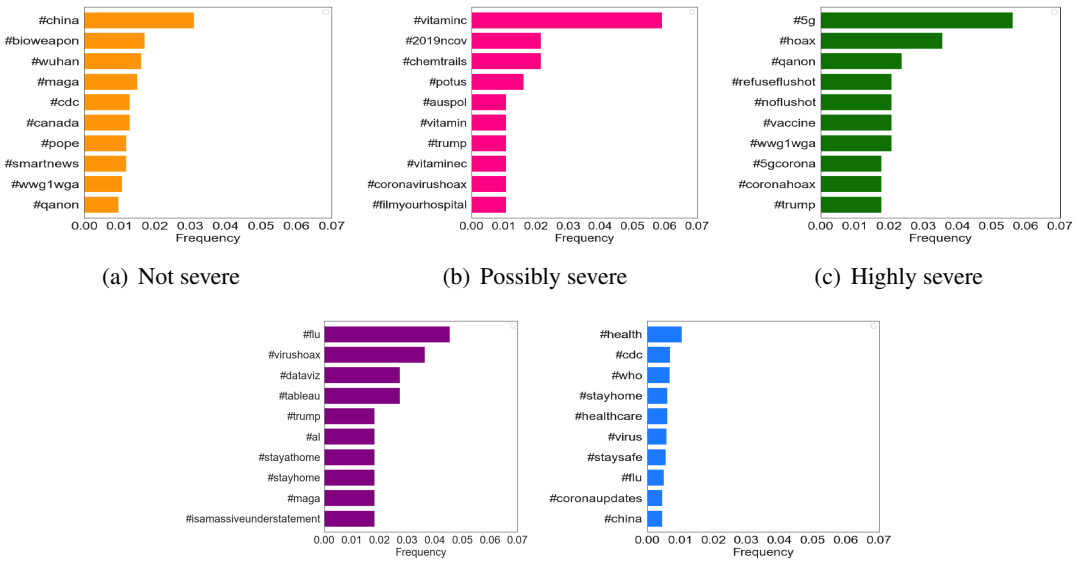
Table 4: Predictive performance of evaluated models on the Covid-HeRA dataset

	Accuracy	Precision	Recall	F1
RF-TFIDF	0.956	0.690	0.580	0.606
RF-Glove	0.960	0.646	0.500	0.493
SVM-TFIDF	0.963	0.713	0.580	0.613
SVM-Glove	0.956	0.690	0.580	0.606
LR-TFIDF	0.950	0.680	0.580	0.633
LR-Glove	0.956	0.712	0.573	0.603
BiLSTM	0.966	0.860	0.580	0.626
CNN	0.966	0.850	0.623	0.670
BERT	0.960	0.850	0.615	0.660
dEFEND w.replies	0.980	0.800	0.540	0.645
dEFEND w.news	**0.980**	**0.920**	**0.680**	**0.750**

Table 5: Predictive performance of evaluated models on the Covid-HeRA dataset - binary classification task, i.e., with posts labeled as misinformation or not.

Finally, based on the per-class evaluation, we

[4] We also performed experiments with COVID-Twitter-BERT, a Transformer model pre-trained on 22.5M COVID19-related Twitter messages (Müller et al., 2020), unfortunately with much lower performance than general BERT. We leave further analysis on the reasons why fine-tuned embeddings on COVID-19 posts were not helpful as future work.



(a) Not severe (b) Possibly severe (c) Highly severe

(d) Refutes/Rebuts (e) True News/Claims

Figure 4: Most frequent hashtags and emoticons/emojis used per category

note another challenging difference, apart from the imbalance discussed above. In Figure 5, we present the confusion matrix for the best performing model on the severity multi-class classification task (CNN). Tweets labeled as "Not severe" and "Possibly severe" are more likely to be predicted as "Real News/Claims", probably due to the true latent semantics being similar for these categories. Further research on handling imbalance as well as integrating auxiliary signals is required. We conclude with future directions in the next section.

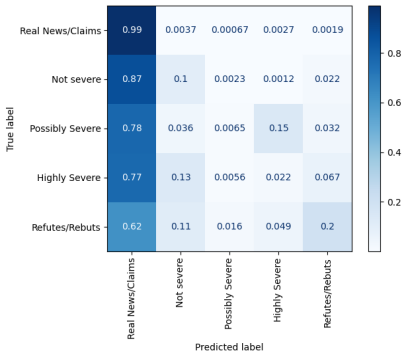


Figure 5: Confusion matrix for the CNN model trained on Covid-HeRA dataset.

5 Conclusion

In this work, we release Covid-HeRA, a new benchmark dataset for risk-aware health misinformation on COVID-19 related social media posts. We describe our data collection and conduct thorough data analysis and extensive experiments with baseline methods and state-of-the-art text classification and misinformation models. Our experimental results demonstrate the usefulness and challenges of finer-grained multi-class classification in healthcare misinformation detection. We hope Covid-HeRA will enable researchers to design advanced models for risk scoring of misinformation spread and to develop systems that inform the social media audience on the respective dangers of following unreliable advice from inaccurate sources.

There are several possible future directions. First and foremost, we hope to take into account the substantial imbalance of misinformation, compared to the overall number of tweets online. By leveraging advancements in relevant research, we can build custom loss functions or data sampling methods to mitigate the challenges of sparsity in underrepresented categories. Additionally, to alleviate the need of large training sets, future research could focus on the exploration of weak supervision signals, semi-supervised and self-supervised algorithms. Few-shot models can also handle distribution shift and novel classes with fewer examples,e.g., adding a scale or category in our annotation set. Finally, the task of identifying rebuttal and refutation posts, which present arguments against misinformation spread, is something we aim to tackle on future research, exploring additional linguistic signals and auxiliary tasks, e.g., applying controversy detection algorithms (Lourentzou et al.).

References

Hunt Allcott, Matthew Gentzkow, and Chuan Yu. 2019. Trends in the diffusion of misinformation on social media. *Research & Politics*, 6(2):2053168019848554.

Rakesh Bal, Sayan Sinha, Swastika Dutta, Risabh Joshi, Sayan Ghosh, and Ritam Dutt. 2020. Analysing the extent of misinformation in cancer related tweets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 924–928.

Eric PS Baumer and Micki McGee. 2019. Speaking on behalf of: Representation, delegation, and authority in computational text analysis. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 163–169. ACM.

Leticia Bode and Emily K Vraga. 2018. See something, say something: Correction of global health misinformation on social media. *Health communication*, 33(9):1131–1140.

J Scott Brennen, Felix M Simon, Philip N Howard, and Rasmus Kleis Nielsen. 2020. Types, sources, and claims of covid-19 misinformation. *Reuters Institute*.

Charles Calisher, Dennis Carroll, Rita Colwell, Ronald B Corley, Peter Daszak, Christian Drosten, Luis Enjuanes, Jeremy Farrar, Hume Field, Josie Golding, et al. 2020. Statement in support of the scientists, public health professionals, and medical professionals of china combatting covid-19. *The Lancet*, 395(10226):e42–e43.

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.

Michael Chung, Adam Bernheim, Xueyan Mei, Ning Zhang, Mingqian Huang, Xianjun Zeng, Jiufa Cui, Wenjian Xu, Yang Yang, Zahi A Fayad, et al. 2020. Ct imaging features of 2019 novel coronavirus (2019-ncov). *Radiology*, 295(1):202–207.

Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. The covid-19 social media infodemic. *arXiv preprint arXiv:2003.05004*.

Jose Yunam Cuan-Baltazar, Maria José Muñoz-Perez, Carolina Robledo-Vega, Maria Fernanda Pérez-Zepeda, and Elena Soto-Vega. 2020. Misinformation of covid-19 on the internet: Infodemiology study. *JMIR Public Health and Surveillance*, 6(2):e18444.

Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset.

Aron Culotta. 2010. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the first workshop on social media analytics*, pages 115–122. acm.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Sameer Dhoju, Md Main Uddin Rony, Muhammad Ashad Kabir, and Naeemul Hassan. 2019. Differences in health news from reliable and unreliable media. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 981–987.

Amir Ebrahimi Fard and Shajeeshan Lingeswaran. 2020. Misinformation battle revisited: Counter strategies from clinics to artificial intelligence. In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 510–519, New York, NY, USA. Association for Computing Machinery.

Miriam Fernandez and Harith Alani. 2018. Online misinformation: Challenges and future directions. In *Companion Proceedings of the The Web Conference 2018*, pages 595–602.

Laurie Garrett. 2020. Covid-19: the medium is the message. *The Lancet*, 395(10228):942–943.

Amira Ghenai and Yelena Mejova. 2018. Fake cures: user-centric modeling of health misinformation in social media. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–20.

Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012.

Xiaolei Huang, Michael C Smith, Michael J Paul, Dmytro Ryzhkov, Sandra C Quinn, David A Broniatowski, and Mark Dredze. 2017. Examining patterns of influenza vaccination in social media. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*.

Ronald J Hunt. 1986. Percent agreement, pearson's correlation, and kappa as measures of inter-examiner reliability. *Journal of Dental Research*, 65(2):128–130.

Nicholas E Ingraham and Christopher J Tignanelli. 2020. Fact versus science fiction: fighting coronavirus disease 2019 requires the wisdom to know the difference. *Critical Care Explorations*, 2(4).

Xiang Ji, Soon Ae Chun, and James Geller. 2013. Monitoring public health concerns using twitter sentiment classifications. In *2013 IEEE International Conference on Healthcare Informatics*, pages 335–344. IEEE.

Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. 2016. News verification by exploiting conflicting social viewpoints in microblogs. In *Thirtieth AAAI conference on artificial intelligence*.

Aditya Joshi, Xiang Dai, Sarvnaz Karimi, Ross Sparks, Cecile Paris, and C Raina MacIntyre. 2018. Shot or not: Comparison of nlp approaches for vaccination behaviour detection. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 43–47.

Sasikiran Kandula and Jeffrey Shaman. 2019. Reappraising the utility of google flu trends. *PLOS Computational Biology*, 15(8):e1007258.

Delal Dara Kılınç and Gülşilay Sayar. 2019. Assessment of reliability of youtube videos on orthodontics. *Turkish journal of orthodontics*, 32(3):145.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Bennett Kleinberg, Isabelle van der Vegt, and Maximilian Mozes. 2020. Measuring emotions in the covid-19 real world worry dataset. *arXiv preprint arXiv:2004.04225*.

Joseph Jeffrey Klembczyk, Mehdi Jalalpour, Scott Levin, Raynard E Washington, Jesse M Pines, Richard E Rothman, and Andrea Freyer Dugas. 2016. Google flu trends spatial variability validated against emergency department influenza-related visits. *Journal of medical Internet research*, 18(6):e175.

Ramez Kouzy, Joseph Abi Jaoude, Afif Kraitem, Molly B El Alam, Basil Karam, Elio Adib, Jabra Zarka, Cindy Traboulsi, Elie W Akl, and Khalil Baddour. 2020. Coronavirus goes viral: Quantifying the covid-19 misinformation epidemic on twitter. *Cureus*, 12(3).

Stacy Loeb, Shomik Sengupta, Mohit Butaney, Joseph N Macaluso Jr, Stefan W Czarniecki, Rebecca Robbins, R Scott Braithwaite, Lingshan Gao, Nataliya Byrne, Dawn Walter, et al. 2019. Dissemination of misinformative and biased information about prostate cancer on youtube. *European urology*, 75(4):564–567.

Ismini Lourentzou, Graham Dyer, Abhishek Sharma, and ChengXiang Zhai. Hotspots of news articles: Joint mining of news text & social media to discover controversial points in news. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 2948–2950. IEEE.

Yi-Ju Lu and Cheng-Te Li. 2020. GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 505–514, Online. Association for Computational Linguistics.

Areeb Mian and Shujhat Khan. 2020. Coronavirus: the spread of misinformation. *BMC medicine*, 18(1):1–2.

Alex Morales, Kanika Narang, Hari Sundaram, and Chengxiang Zhai. 2020. Crowdqm: Learning aspect-level user reliability and comment trustworthiness in discussion forums. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 592–605. Springer.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503*.

Maria Nicola, Zaid Alsafi, Catrin Sohrabi, Ahmed Kerwan, Ahmed Al-Jabir, Christos Iosifidis, Maliha Agha, and Riaz Agha. 2020. The socio-economic implications of the coronavirus and covid-19 pandemic: a review. *International Journal of Surgery*.

Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, and David Rand. 2020. Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy nudge intervention.

Tobias Preis and Helen Susannah Moat. 2014. Adaptive nowcasting of influenza outbreaks using google searches. *Royal Society open science*, 1(2):140095.

Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806.

Mauricio Santillana, D Wendong Zhang, Benjamin M Althouse, and John W Ayers. 2014. What can digital disease detection learn from (an external revision to) google flu trends? *American journal of preventive medicine*, 47(3):341–347.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Jieun Shin, Lian Jian, Kevin Driscoll, and François Bar. 2018. The diffusion of misinformation on social media: Temporal pattern, message, and source. *Computers in Human Behavior*, 83:278–287.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 395–405.

Alessio Signorini, Alberto Maria Segre, and Philip M Polgreen. 2011. The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS one*, 6(5):e19467.

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*.

Michael Smith, David A Broniatowski, Michael J Paul, and Mark Dredze. 2016. Towards real-time measurement of public epidemic awareness: Monitoring influenza awareness through twitter. In *AAAI Spring Symposium on Observational Studies through Social Media and Other Human-Generated Content*.

Samia Tasnim, Md Mahbub Hossain, and Hoimonty Mazumder. 2020. Impact of rumors or misinformation on coronavirus disease (covid-19) in social media.

Theodore S Tomeny, Christopher J Vargo, and Sherine El-Toukhy. 2017. Geographic and demographic correlates of autism-related anti-vaccine beliefs on twitter, 2009-15. *Social Science & Medicine*, 191:168–175.

Yaqing Wang, Weifeng Yang, Fenglong Ma, Jin Xu, Bin Zhong, Qiang Deng, and Jing Gao. 2020. Weak supervision for fake news detection via reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 516–523.

Davy Weissenbacher, Abeed Sarker, Michael J Paul, and Graciela Gonzalez-Hernandez. 2018. Overview of the third social media mining for health (smm4h) shared tasks at emnlp 2018. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 13–16.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Gregory D Zimet, Zeev Rosberger, William A Fisher, Samara Perez, and Nathan W Stupiansky. 2013. Beliefs, behaviors and hpv vaccine: correcting the myths and the misinformation. *Preventive medicine*, 57(5):414–418.