Proceedings of the Sixteenth International AAAI Conference on Web and Social Media (ICWSM 2022)

Drink Bleach or Do What Now? Covid-HeRA: A Study of Risk-Informed Health Decision Making in the Presence of COVID-19 Misinformation

Arkin Dharawat[1], Ismini Lourentzou[2], Alex Morales[3], ChengXiang Zhai[1]

[1] Department of Computer Science, University of Illinois at Urbana - Champaign
[2] Department of Computer Science, Virginia Tech
[3] Twitter
arkin.dharawat@gmail.com, ilourentzou@vt.edu, czhai@illinois.edu, alexmorales@twitter.com

Abstract

Given the widespread dissemination 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 research community. Several works study health misinformation detection, yet little attention has been given to the perceived severity of misinformation posts. In this work, we frame health misinformation as a risk assessment task. More specifically, we study the severity of each misinformation story and how readers perceive this severity, *i.e.*, how harmful a message believed by the audience can be and what type of signals can be used to recognize potentially malicious fake news and detect refuted claims. To address our research questions, we introduce a new benchmark dataset, accompanied by detailed data analysis. We evaluate several traditional and state-of-the-art models and show there is a significant gap in performance when applying traditional misinformation classification models to this task. We conclude with open challenges and future directions.

Introduction

While an increasing percentage of the population relies on social media platforms for news consumption, fake news and other types of misinformation have been also widely prevalent, putting audiences at great risk 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). Fake news frequently emerges for certain phenomena and topics, *e.g.*, public health issues, politics *etc.* (Allcott, Gentzkow, and Yu 2019; Shin et al. 2018; Bode and Vraga 2018). Unsurprisingly, the same applies to the current global pandemic, where inaccurate stories are amplifying hate speech, increasing stigmatization, undermining public health response and polarizing public debate on topics related to COVID-19. It is often difficult for users, who may 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, van der Vegt, and Mozes 2020).

Moreover, reducing the impact of misinformation on potential health-related decisions requires deeper understanding of the target audience, which goes beyond fact-checking. To this end, we aim to capture how social media users view the severity of misinformation circulating online. Although much work has been focused on identifying health-related misinformation, there has been little attention to further making a distinction between the perceived severity 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, while others might be simply inaccurate information with limited effects. Overall, 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 which has led to hundreds of deaths (Islam et al. 2020).

Some news articles related to COVID-19 misinformation reporting hospitalization, injuries, and deaths may be interpreted as weak signals of misinformation severity to some extent. However, how readers view the severity level for each misinformation story remains a challenging task that has not been previously studied, even though building such models would be beneficial, *e.g.*, in prioritizing policy-making towards mitigating severely harmful misinformation, such as potentially deadly fake remedies; a goal that we hope to achieve in this paper.

Furthermore, to mitigate misinformation spread, it is critical to build systems that inform the target audience about the negative impacts and provide resources for properly assessing information early in the process. Creating an early warning system for misinformation requires detection mechanisms directly based on content. Propagation-based methods that rely on social context information by incorporating additional signals such as like, retweets and network information (Zhou et al. 2020) are often inadequate for early detection. In addition, adversarial attacks of misinformation spreaders due to manipulation strategies for both networks and content, *e.g.*, forming groups and increasing influence by connecting with each other, undermine the performance of existing approaches (Wu et al. 2019).

Based on these challenges, our work aims to help re-

duce the impact of health misinformation and enable better decision-support health information systems, by proposing a novel perceived health risk misinformation framework. Specifically, we aim at understanding the impact of unreliable advice from inaccurate sources based on the perceived severity of misinformation. To this end, we also introduce a new benchmark dataset that contains social media posts annotated on a Likert scale, based on whether the content of the post is perceived to be: a) Real News/Claims, b) Refutes/Rebuts, c) Misinformation (Other), d) Misinformation (Possibly severe), and e) Misinformation (Highly severe). The distinction of different reader-perceived severity levels is judged based on their potential to impact a user's health, assuming the individual might be making decisions upon reading the advice and/or suggestions in the post. In other words, our goal is to produce a label that reflects the level of recognized risk factors in the presence of inaccurate claims and news, on a challenging cold-start content-based scenario where additional metadata information such as likes, retweets and user information is not available, and conditioned on the worst-case assumption that the user will follow the advice. We conduct the first empirical study of the effectiveness of representative machine learning methods for predicting the perceived severity of misinformation so as to establish baseline benchmark results for facilitating further research.

Our contributions can be summarized as follows:

- In contrast with prior work that treats misinformation as a fact-checking and/or veracity task, we introduce a new task of modeling the perceived risk of health-related misinformation.
- To facilitate research on this new task and understand how users view harmful unreliable information shared online, we release a novel social media **Covid He**alth **R**isk **A**ssessment (**`Covid-HeRA`**) misinformation dataset that consists of $61,286$ tweets labeled on a severity scale.
- We benchmark several content-based classification models, including state-of-the-art fake news detection models and contextualized embeddings. Experimental results show a significant performance gap when compared to traditional misinformation detection settings.
- We present a detailed data analysis that reveals several key insights about the most prominent unreliable news. Moreover, we present technical and modeling challenges that this new task poses.

Our dataset captures how social media users perceive severe misinformation and can potentially be used to determine the user's understanding and recognition of the severity of misinformation. We hope that **`Covid-HeRA`** will facilitate future research on developing detection models that take into account user views and misinformation severity[1].

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), orthodontics (Kılınç and Sayar 2019), sexually transmitted diseases and infections (Zimet et al. 2013; Lohmann et al. 2018; Joshi et al. 2018; Tomaszewski et al. 2021), autism (Baumer and McGee 2019), influenza (Culotta 2010; Signorini, Segre, and Polgreen 2011), and more recently COVID-19 (Garrett 2020; Brennen et al. 2020; Cinelli et al. 2020; Cui and Lee 2020). Social media data have also been used to monitor influenza prevalence and awareness (Smith et al. 2016; Ji, Chun, and Geller 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 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 microblogging sites, such as Twitter.

Tomeny, Vargo, and El-Toukhy (2017) examined geographic and demographic trends in anti-vaccine tweets. Huang et al. (2017) examine the geographic and demographic patterns of the flu vaccine in social media. Recent research 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, Seo, and Liu 2017). Joshi et al. (2018) make a distinction between *vaccine hesitancy* identification, and *vaccination behavior detection*, in that the former deals with the attitudes, or stance, while the latter is concerned with detecting the action of getting vaccinated. Our work models how users perceive different severity levels of misinformation.

With the threat of COVID-19 misinformation to public health organizations, there have 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, Hossain, and Mazumder (2020) outlines several potential strategies to ensure effective communication on COVID-19, *e.g.*, ensuring up-to-date reliable information via identifying fake news and misinformation. Brennen et al. (2020) analyze 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 for reliable and trustworthy information online (Cuan-Baltazar et al. 2020).

Pennycook et al. (2020) show that people tend to believe false claims about COVID-19 and share false information when they do not think critically about the accuracy and veracity of the information. Recent work aims at understanding reader perception of news reliability (Gabriel et al. 2022). Kouzy et al. (2020) show that $\sim 25\%$ of COVID-19 messages contain some form of misinformation and $\sim 17\%$ contain some unverifiable information. Singh et al. (2020) provide a large-scale exploratory analysis of how myths and COVID-

[1]Covid-HeRA is open-sourced at https://github.com/TIMAN-group/covid19_misinformation

19 themes are propagated on Twitter, by analyzing how users share URL links. Cinelli et al. (2020) cluster word embeddings to identify topics and measure the 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 led 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). News circulates daily advice for the public. Some articles, however, contain fake remedies that reportedly cure or prevent COVID-19, promote false diagnostic procedures, report incorrect news about the virus properties or urge the audience to avoid specific food or treatments that might supposedly either make symptoms worse or the reader more likely to contract the virus[2]. With such information overload, any decision-making procedures based on misinformation have a high likelihood of severely impacting one's health (Ingraham and Tignanelli 2020). Therefore, we aim to capture how users perceive the severity of incorrect information released on social media, as well as to detect any posts that refute or rebut coronavirus-related false claims. To this end, we release a novel dataset, present experimental results with several state-of-the-art machine learning models and conclude with possible future extensions.

Methodology

In this section, we introduce our task formulation, data collection & annotation strategy. Subsequently, we present detailed statistics and analysis with key insights on the most prevalent harmful misinformation online. We frame the task as multi-class classification and design an annotation scale, based on whether a social media post has the potential to guide the audience towards health-related decisions or behavioral changes with high-risk factors, *i.e.*, high likelihood of severely impacting one's health. In other words, we aim to model the *perceived severity* by ordinary users that would potentially follow medical advice found online. Since, to the best of our knowledge, there is no public dataset appropriate for this task, we release `Covid-HeRA` to facilitate future research.

Upon extensive team discussions, and in conjunction with data inspection, we design a five-class framework that is inspired by Likert-type scales (Likert 1932), typically used in risk stratification and ordinal risk rating. To this end, each social media post is categorized as follows: a) **Real News/-Claims**, *i.e.*, messages with reliable correct information, b) **Refutes/Rebuts**, *i.e.*, messages which contain a refutation or rebuttal of an incorrect statement, c) **Other**, *i.e.*, other types of misinformation or misinformation that is unlikely to result in risky behavioral changes, 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 and 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.

We make use of CoAID (Cui and Lee 2020), a large-scale healthcare misinformation data collection related to COVID-19, with binary ground-truth labels for news articles and claims, accompanied with associated tweets and user replies. This dataset provides 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 semi-supervised models as future work. To obtain annotations based on our defined severity categorization, all CoAID tweets labeled as misinformation are shuffled and distributed to two different annotators. Our annotators are ordinary social media users, since we want to capture how regular users perceive potentially harmful misinformation shared online. This formulation is useful in studying target audience behavioral aspects influenced by misinformation messages and in rethinking platform policies for managing misinformation based on perceived severity levels.

Each annotator is asked to judge whether any decisions can be taken or other actionable items can be performed 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). Each tweet is annotated jointly by the two annotators. To resolve conflicts on ambiguous instances, a final round of annotator discussions was introduced as an additional step, where an external validator provided an additional label after the discussion. To assess agreement levels, the external validator was asked to also annotate a random sample of the labeled tweets. Cohen's kappa coefficient between the annotators and the validator was 0.7037, which shows good agreement on the task (Hunt 1986). Both annotators and the validator are proficient in English, and diversified w.r.t. nationality to mitigate systemic bias as much as possible.

The total number of tweets labeled per category, alongside the number of unique words, are presented in Table 2. In addition, the dataset contains a *claim index*, with each tweet mapped to a specific category claim that captures the main topic[3]. The average time between a tweet and a reply is approximately 11hrs for "Real News/Claims" ($40,333$ seconds), 12hrs for "Other" ($44,708$ seconds), 8hrs for "Possibly severe", 9hrs for "Highly severe" ($32,369$ seconds) and 7hrs for "Refutes/Rebuts" ($25,238$ seconds).

Data Analysis

We first identify the most frequent discriminative terms per category, *i.e.*, terms that appear very often in a specific category, but may be infrequent in the remaining categories. We use a 0.5% document frequency threshold to discard terms

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

[3]Such mapping is already provided by the CoAID dataset.

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.
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 with increased health risks, or may result in increased exposure for certain socioeconomic groups.
Other	"*Vatican confirms Pope Francis and two aides test positive for Coronavirus*"	This involves other types of misinformation not related to health, *e.g.*, conspiracy theories and public figure rumors; health-related behavioral changes might be less likely to occur.
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.

Category	#Tweets	#Tokens (Vocab)
Possibly severe	439	11,171
Highly severe	568	16,328
Refutes/Rebuts	447	2,244
Other	1,851	3,324
Real News/Claims	57,981	51,478
Total	61,286	84,545

Table 2: `Covid-HeRA` Dataset Statistics.

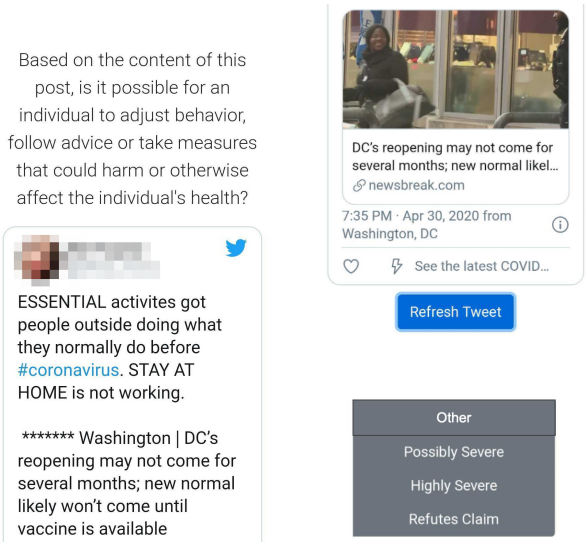


Figure 1: A screenshot of a tweet example and possible annotation options.

that are very common across the whole data collection, *e.g.*, "*COVID-19*" 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 "Other" category with the rest of the categories, we see that many of the terms refer to conspiracy theories 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" contain keywords such as "*myths*", "*weaponized*", "*lying*", and "*antibiotics*" as the messages in this category address and debunk 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.

In Figure 3, we visualize the compactness of each category. We use pre-trained BERT embeddings (Devlin et al. 2018) to measure how close each tweet is by the centroid of its corresponding category. More specifically, we compute the centroid as the arithmetic mean of all tweet embeddings from a specific category. 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 is also evident by the flatness of these curves. This may be due to the broad range of topics covered in both of these categories compared to the rest.

(a) Possibly severe (b) Highly severe (c) Refutes/Rebuts (d) Other (e) Real News/Claims

Figure 2: Most common terms per category

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 "Other" 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 may be 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 contain terms associated with computation such as "#dataviz", "#tableau", as well as hashtags 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, as per the aforementioned claim index (category claim for each tweet), we present the most common claims and news per category, and their corresponding frequency (Table 3).

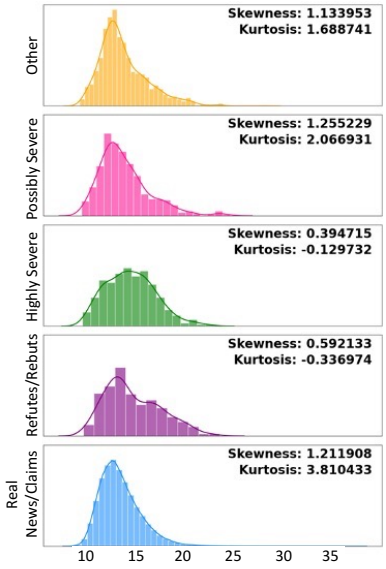


Figure 3: Category compactness, measured by the distance distribution of Twitter posts, relative to its category centroids.

Experiments

We perform experiments with several baselines and state-of-the-art models. We pre-process tweets to filter out reserved tokens, such as *RT* or *retweet*, urls and mentions. The same pre-processing is performed for tweet replies. Additionally, we split the data into 80% training and 20% testing, keeping the same splits across all models for a fair comparison. To avoid any data leakage, we remove duplicate tweets and retweets and stratify based on *claim index*, *i.e.*, tweets discussing the same claim under the same label are kept either in training or test, but not in both. Our evaluation metrics are Precision (P), Recall (R) and F1 score (macro-averaged).

The algorithms we experiment with are:

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

Support Vectors with bag-of-words (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 (BiLSTM).

Multichannel CNN 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).

Hierarchical Attention Networks (HAN), with word-level and sentence-level attention mechanisms (Yang et al. 2016)

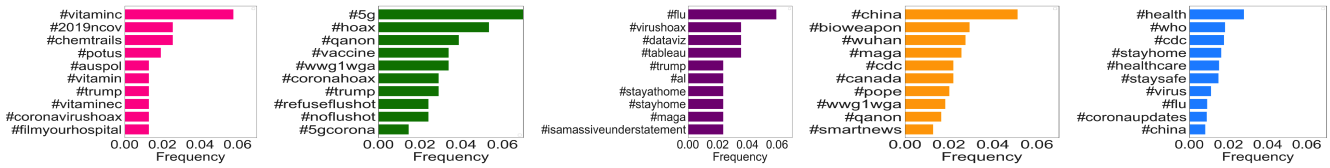
dEFEND, a state-of-the-art fake news detection model that builds upon HAN by adding co-attention between two textual sources (Shu et al. 2019); in our case either tweet replies or corresponding news content. dEFEND allows us to evaluate whether incorporating additional sources of information with related content can improve predictive performance.

Hyper-parameter Details

Hyper-parameter search is performed with Tune (Liaw et al. 2018), Async Successive Halving Scheduler (Li et al. 2018), Adam (Kingma and Ba 2015) and 50 mini-batch size for all models. We trained all the models, except BERT, for a total of 100 epochs. We fine-tuned BERT for a total of 3 epochs. The best performing hyper-parameters were: (1) For the bidirec-

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	Other	152
Michigan Governor Gretchen Whitmer Bans Buying US Flags During Lockdown	Other	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 ground-truth labels.



(a) Possibly severe (b) Highly severe (c) Refutes/Rebuts (d) Other (e) Real News/Claims

Figure 4: Most frequent hashtags per category.

tional LSTM (BiLSTM), two hidden layers with 256 neurons each and tanh activations, followed by the final classification layer. Dropout is set to 0.5. (2) The convolutional neural network (CNN) has 128 filters of kernel size $\{3, 4, 5\}$, stride 1 and ReLU activations, followed by max-pooling, a dense layer of 256 units, and the final classification layer. Dropout is set to 0.5. (3) The hierarchical attention model (HAN) consists of a world-level GRU layer and a sentence-level GRU layer, each with 256 neurons, tanh activations, followed by 0.7 dropout and a final classification layer. dEFEND extends HAN with an additional GRU layer for the replies/news, with the same configuration as for the word and sentence GRU layers. (4) BERT is fine-tuned by adding a 128 linear layer and a final classification layer with ReLU activations and 0.7 dropout. We train all deep learning models with two GeForce GTX 1080 Nvidia GPUs on an 64-bit Ubuntu 16.04.6 machine with Intel(R) Xeon(R) CPU E5-2650@2.20GHz and $\sim$ 260 GB RAM.

Experimental Results

In terms of F1 score, BERT performs better than all baselines (Table 4). Surprisingly, incorporating user engagement features or news content did not help[4]. One of the main challenges of the finer granularity in `Covid-HeRA` is that some categories might be substantially underrepresented. In other words, this task is a highly imbalanced problem. We evaluate most commonly used methods for mitigating class imbalance, including data augmentation for minority classes and custom loss functions[5]. We find that replacing Cross-Entropy with Tversky loss (Salehi, Erdogmus, and Gholipour 2017) results in 6% relative performance gains (Table 4 - **BERT + Tversky**). In addition, we experiment with word-level augmentations (oversampling), such as word insertions and substitutions based on pre-trained embeddings, and synonym replacement based on lexicons and thesauri (*e.g.*, WordNet) (Zhang, Zhao, and LeCun 2015; Wang and Yang 2015; Fadaee, Bisazza, and Monz 2017; Kobayashi 2018). We find that such augmentations produced the best performance (Table 4 - **BERT + DA**) with 8% relative performance gains.

Different from existing misinformation problem formulations, `Covid-HeRA` is more challenging. We perform the same experiments in a traditional binary misinformation detection setting, *i.e.*, predict the veracity of information. More specifically, we discard all refutation and rebuttal tweets and collapse all tweets labeled as misinformation in a common label irrespective of severity, *i.e.*, essentially backtracking to a real versus fake news traditional framework. We evaluate the same set of algorithms. Compared to the perceived severity classification task of `Covid-HeRA`, the traditional misinformation detection task produces much higher performance across all evaluation metrics (Table 4, right). These results leave much room for improvement, showcasing that this task is not yet fully addressed by current state-of-the-art models and highlighting the challenges of accurately distinguishing between general misinformation, harmful social media posts and refuted claims.

Error Analysis and Discussion

In order to identify potential avenues of improvement and limitations of the classifiers explored, we take a closer look

[4]We additionally experiment with COVID-Twitter-BERT, a Transformer model pre-trained on 22.5M COVID-19 Twitter messages (Müller, Salathé, and Kummervold 2020), but we have observed lower performance than with general BERT. We leave further analysis to future work.

[5]Presented in more detail in the appendix.

Method	MULTI-CLASS			BINARY		
	P	R	F1	P	R	F1
RF-TFIDF	0.29	0.24	0.25	0.69	0.58	0.61
RF-Glove	0.26	0.20	0.20	0.65	0.50	0.49
SVM-TFIDF	0.24	0.20	0.20	0.71	0.58	0.61
SVM-Glove	0.33	0.23	0.25	0.69	0.58	0.61
LR-TFIDF	0.28	0.22	0.23	0.68	0.58	0.63
LR-Glove	0.31	0.23	0.25	0.71	0.57	0.60
BiLSTM	0.40	0.27	0.27	0.86	0.58	0.63
CNN	0.34	0.28	0.29	0.85	0.62	0.67
BERT	0.34	0.32	**0.33**	0.85	0.62	0.66
HAN	0.20	0.20	0.20	0.65	0.67	0.67
dEFEND w. replies	0.25	0.22	0.23	0.80	0.54	0.65
dEFEND w. news	0.35	0.31	0.25	**0.92**	**0.68**	**0.75**
BERT + DA	**0.55**	**0.51**	**0.41**	0.93	0.64	0.70
BERT + Tversky	0.40	0.48	0.39	0.68	0.83	0.72

Table 4: Performance of evaluated models on Covid-HeRA: proposed severity categorization (Left) and traditional misinformation (binary classification) task, *i.e.*, with posts labeled as misinformation or not (Right)

at some instances (tweet messages) that are consistently misclassified across the set of evaluated deep learning classifiers, *i.e.*, instances that all neural models are not able to predict correctly. We also perform error analysis on the false negatives from the best performing classifier. We group false negatives based on the misinformation topic referenced in each message. For example, for the category "Other", which makes up the majority of the false negatives, test instances can be further categorized by the message topics, as shown in Table 5. In particular, in the *Misclassified* column, we count the total number of messages that are incorrectly classified by all deep-learning models explored. Our error analysis shows that considering the different topics could potentially improve misinformation detection models on this task.

We also qualitatively examine the false positives of the best performing model (*i.e.*, BERT) and find only 11 false positives classified as misinformation, showing the classifier is conservative but precise in the classification. Future directions would include improving the recall, as the examples missed are usually controversial topics or messages which appear as misinformation but are actually opinions, such as the mean incubation period of the virus and the seriousness of the virus compared to the flu. There is a general trend of poor recall across the levels of severity, among all of the explored classifiers, which highlights the difficulty of predicting the severity of misinformation.

Conclusion and Future Work

In this work, we present **Covid He**alth **R**isk **A**ssessment (**Covid-HeRA**), a novel task that aims to capture the perceived severity of health-related misinformation. We describe our data collection and conduct thorough data analysis and extensive experiments with baseline methods and state-of-the-art models. Our experimental results demonstrate the challenges that computational models face in capturing the true latent semantics in our finer-grained health misinformation dataset. We hope Covid-HeRA will significantly facilitate the research in modeling the perception of misinformation severity from the target audience and will spur the development of more advanced systems that can inform social media users on the respective dangers of following unreliable advice from inaccurate sources.

Topic	Misclassified	Total
COVID-19 was artificially engineered	56	72
Media COVID-19 scare	25	33
False Death Toll Reports	5	23
Trump Reelection	8	9
Surgeon General Projection Reports	6	8

Table 5: Topic themes in test data for the "Other" class. *Misclassified* column: number of instances incorrectly classified by all neural classifiers. *Total* column: total number of instances in test data. Both categorized by topic theme.

There are several new directions and opportunities. Covid-HeRA aims at a deeper understanding of reader views on misinformation severity and can have immediate practical impact, as the potential data user group includes computational linguists, sociologists, psychologists, misinformation researchers, *etc.* Our dataset captures how social media users, the main consumers of misinformation, view and recognize (harmful) fake news and claims. Such new directions that target whether readers understand the impact of misinformation can help reduce misinformation spread.

In particular, the data collected includes a special class "refutes/rebuts claim" that can be used in conjunction with the replies and the news titles, to form a targeted tailor-made refutation of a severe false claim. Since we release the tweet identifiers and code for downloading the raw tweets, all associated metadata can be acquired, *e.g.*, timestamp, likes, retweets, reply time, geolocation, *etc.* As such, Covid-HeRA can be useful for advancing propagation-based methods that incorporate metadata. While our work evaluates content-based methods for early detection, we hope future work will expand predictive modeling methods by incorporating metadata.

The temporal evolution of misinformation is another interesting application to our dataset, which can be accomplished through chronological ordering of tweets based on timestamps and trend forecasting, or by utilizing transfer learning and additional Twitter misinformation datasets on related health domains. Combined with additional resources, Covid-HeRA can be used for addressing research questions on how misinformation and perceived severity affects consumer trends during the pandemic, for example in regards to time-sensitive products such as masks, hand sanitizers, health supplements and even fake remedies.

To alleviate the need for large training sets, future research could focus on weakly-supervised, semi-supervised, and self-supervised algorithms. Few-shot models can also handle distribution shift and classes with fewer examples. Anomaly and outlier detection methods for text data are interesting directions worthy of future exploration that could potentially improve the detection accuracy. As the presence of severe misinformation tweets is rare, our dataset is ideal for bench-

marking text-based anomaly/outlier detection model architectures in realistic scenarios (Ruff et al. 2019; Manolache, Brad, and Burceanu 2021). Finally, the subtask of identifying rebuttal and refutation posts that present useful arguments against misinformation spread, is something we aim to further improve in the future, either by incorporating additional linguistic signals and auxiliary tasks, or by applying controversy detection algorithms (Lourentzou et al. 2015; Timmermans et al. 2017).

Broader Impact

This work aims at understanding how average social media users perceive the severity of misinformation. Our labels are produced by readers rather than medical experts, and hence are not suitable for traditional health analytics and misinformation fact-checking algorithms. Moreover, the detection of misinformation spreaders, as well as additional aspects in which a post might be dangerous, including but not limited to political implications, bias/racism towards specific groups, distrust in science, technological impact, *etc.*, go beyond our task but are certainly noteworthy for discussion and future research. All data crawled through the Twitter API fall under Twitter's privacy policy. Following the Twitter terms of service, we only distribute Twitter-provided identifiers to download the data, and refrain from publicly sharing any tweets or user information. We note that this paper describes a work of research, thus, any definitions and methods described here do not reflect how Twitter classifies misleading information or enforces its misinformation policies. We also note that users can choose to make their posts private, delete their accounts or remove tweets. Hence, as for all Twitter-related research, exact reconstruction of the dataset in later years may not be possible. Misinformation spread on social media has been a long-standing problem with potential public health risks (Ghenai and Mejova 2018; Geeng, Yee, and Roesner 2020). Considering the substantial increase in misinformation during the pandemic, and the lack of methods that aim at understanding the perceived severity of false claims and fake news, this work serves as a timely dataset on COVID-19 misinformation circulating on social media.

Appendix

The dataset statistics show that there exists substantial imbalance in misinformation classes, compared to the real news/-claims class (Table 2). This is a recurring challenge for misinformation tasks, as most tweets online are either irrelevant, opinions and discussions about health topics or news shared across the platform. Generally, there is an abundance of proposed methods in the literature that can handle class imbalance. The most common approaches involve either assigning weights to individual training examples, custom loss functions, re-sampling the data or generating synthetic examples to improve robustness (Chawla et al. 2002; Xie et al. 2019). For a thorough review on class imbalance in machine learning, we refer the reader to literature reviews (Chawla 2009; He and Garcia 2009). Here, we explore some of the aforementioned methods. Nevertheless, our goal is not to exhaustively experiment with all possible options.

	P	R	F1
BERT + vanilla CE	0.34	0.32	0.33
BERT + weighted CE	0.38	0.34	0.35
BERT + Tversky	0.40	0.48	0.39
BERT + Generalized Dice	0.40	0.46	0.34
BERT + Self-Adjusting Dice	0.44	0.45	0.36
BERT + Soft Dice	0.38	0.40	0.37
BERT + DA	**0.55**	**0.51**	**0.41**
BERT + DA + Tversky	0.38	0.45	0.36

Table 6: Comparison of methods that handle class imbalance on the `Covid-HeRA` data, with the best performing model, *i.e.*, BERT. Evaluation with Precision (P), Recall (R) and F1.

In terms of data augmentation, we restrict our exploration to word-level augmentations, since, these methods produce good results for text classification tasks (Wei and Zou 2019; Xie et al. 2019). In terms of custom loss functions, we experiment with a set of the most representative ones, including vanilla Cross-Entropy and weighted Cross-Entropy, in which examples are weighted disproportionately to the number of instances per class (Cui et al. 2019). We also experiment with the **Dice** and **Tversky** loss functions (Milletari, Navab, and Ahmadi 2016; Salehi, Erdogmus, and Gholipour 2017) and their variants such as the **Soft Dice Loss** (Milletari, Navab, and Ahmadi 2016), the **Generalized Dice Loss** (Crum, Camara, and Hill 2016; Sudre et al. 2017) and the **Self-adjusting Dice Loss** (Li et al. 2020).

Regarding data-augmentation, the percentage of words in each sentence to be altered is set to 7%. The number of augmented samples per instance is 3, *i.e.*, for each training instance, we create three additional synthetic examples by randomly sampling variations of word substitutions, insertions, deletions, *etc.* For the Tversky loss we set $\alpha = 0.6$ and $\beta = 0.4$. For the Self-adjusting Dice loss, we set $\alpha = 0.75$. In practice, a small smoothness factor is added to each loss. Here, we set this smoothness to 10^{-5}.

In Table 6, we present results for all aforementioned methods, producing approximately $1 - 8\%$ relative performance gainS. The vanilla Cross-Entropy cannot handle minority classes properly, mostly predicts all tweets as "Real News/-Claims" and confuses "Refutes/Rebuts" with "Highly Severe". Both data augmentation (**BERT + DA**) and Tversky loss (**BERT + Tversky**) improve the prediction for some classes, *e.g.*, "Other" or "Refutes/Rebuts", however, some examples from the "Possibly Severe" and "Highly Severe" classes are misclassified as "Real News/Claims". This is possibly due to the broad range of similar topics, as described in our data analysis. Further research on handling such cases is required. We note that Tversky approximates our evaluation metric (F-measure) when $\alpha + \beta = 1$ (as in our case). The combination of data augmentation and Tversky did not produce better results. In the future, we hope to introduce models that synergistically combine data augmentation and training losses.

References

Allcott, H.; Gentzkow, M.; and Yu, C. 2019. Trends in the diffusion of misinformation on social media. *Research & Politics*.

Bal, R.; Sinha, S.; Dutta, S.; Joshi, R.; Ghosh, S.; and Dutt, R. 2020. Analysing the Extent of Misinformation in Cancer Related Tweets. In *AAAI ICWSM*.

Baumer, E. P.; and McGee, M. 2019. Speaking on Behalf of: Representation, Delegation, and Authority in Computational Text Analysis. In *AAAI/ACM Conference on AI, Ethics, and Society*.

Bode, L.; and Vraga, E. K. 2018. See something, say something: Correction of global health misinformation on social media. *Health communication*.

Brennen, J. S.; Simon, F. M.; Howard, P. N.; and Nielsen, R. K. 2020. Types, sources, and claims of COVID-19 misinformation. *Reuters Institute*.

Calisher, C.; Carroll, D.; Colwell, R.; Corley, R. B.; Daszak, P.; Drosten, C.; Enjuanes, L.; Farrar, J.; Field, H.; Golding, J.; et al. 2020. Statement in support of the scientists, public health professionals, and medical professionals of China combatting COVID-19. *The Lancet*.

Castelo, S.; Almeida, T.; Elghafari, A.; Santos, A.; Pham, K.; Nakamura, E.; and Freire, J. 2019. A topic-agnostic approach for identifying fake news pages. In *Companion of The Web Conf.*

Chawla, N. V. 2009. Data mining for imbalanced datasets: An overview. In *Data Mining and Knowledge Discovery Handbook*. Springer.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *JAIR*.

Chung, M.; Bernheim, A.; Mei, X.; Zhang, N.; Huang, M.; Zeng, X.; Cui, J.; Xu, W.; Yang, Y.; Fayad, Z. A.; et al. 2020. CT imaging features of 2019 novel coronavirus (2019-nCoV). *Radiology*.

Cinelli, M.; Quattrociocchi, W.; Galeazzi, A.; Valensise, C. M.; Brugnoli, E.; Schmidt, A. L.; Zola, P.; Zollo, F.; and Scala, A. 2020. The covid-19 social media infodemic. *arXiv:2003.05004*.

Crum, W. R.; Camara, O.; and Hill, D. L. 2016. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Transactions on Medical Imaging*.

Cuan-Baltazar, J. Y.; Muñoz-Perez, M. J.; Robledo-Vega, C.; Pérez-Zepeda, M. F.; and Soto-Vega, E. 2020. Misinformation of COVID-19 on the Internet: Infodemiology study. *JMIR Public Health and Surveillance*.

Cui, L.; and Lee, D. 2020. CoAID: COVID-19 Healthcare Misinformation Dataset. *arXiv preprint arXiv:2006.00885*.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-balanced loss based on effective number of samples. In *CVPR*.

Culotta, A. 2010. Towards detecting influenza epidemics by analyzing Twitter messages. In *1st workshop on social media analytics*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*.

Dhoju, S.; Main Uddin Rony, M.; Ashad Kabir, M.; and Hassan, N. 2019. Differences in health news from reliable and unreliable media. In *Companion of The Web Conf.*

Fadaee, M.; Bisazza, A.; and Monz, C. 2017. Data augmentation for low-resource neural machine translation. *ACL*.

Fard, A. E.; and Lingeswaran, S. 2020. Misinformation Battle Revisited: Counter Strategies from Clinics to Artificial Intelligence. In *Companion of The Web Conf.*

Fernandez, M.; and Alani, H. 2018. Online misinformation: Challenges and future directions. In *Companion of The Web Conf.*

Gabriel, S.; Hallinan, S.; Sap, M.; Nguyen, P.; Roesner, F.; Choi, E.; and Choi, Y. 2022. Misinfo Reaction Frames: Reasoning about Readers' Reactions to News Headlines. In *ACL*.

Garrett, L. 2020. COVID-19: the medium is the message. *The Lancet*.

Geeng, C.; Yee, S.; and Roesner, F. 2020. Fake News on Facebook and Twitter: Investigating How People (Don't) Investigate. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–14.

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

Ginsberg, J.; Mohebbi, M. H.; Patel, R. S.; Brammer, L.; Smolinski, M. S.; and Brilliant, L. 2009. Detecting influenza epidemics using search engine query data. *Nature*.

He, H.; and Garcia, E. A. 2009. Learning from imbalanced data. *IEEE TKDE*.

Huang, X.; Smith, M. C.; Paul, M. J.; Ryzhkov, D.; Quinn, S. C.; Broniatowski, D. A.; and Dredze, M. 2017. Examining patterns of influenza vaccination in social media. In *AAAI Workshops*.

Hunt, R. J. 1986. Percent agreement, Pearson's correlation, and kappa as measures of inter-examiner reliability. *Journal of Dental Research*.

Ingraham, N. E.; and Tignanelli, C. J. 2020. Fact versus science fiction: fighting coronavirus disease 2019 requires the wisdom to know the difference. *Critical Care Explorations*.

Islam, M. S.; Sarkar, T.; Khan, S. H.; Kamal, A.-H. M.; Hasan, S. M.; Kabir, A.; Yeasmin, D.; Islam, M. A.; Chowdhury, K. I. A.; Anwar, K. S.; et al. 2020. COVID-19–related infodemic and its impact on public health: A global social media analysis. *The American journal of tropical medicine and hygiene*, 103(4): 1621.

Ji, X.; Chun, S. A.; and Geller, J. 2013. Monitoring public health concerns using twitter sentiment classifications. In *IEEE International Conference on Healthcare Informatics*.

Joshi, A.; Dai, X.; Karimi, S.; Sparks, R.; Paris, C.; and MacIntyre, C. R. 2018. Shot or not: Comparison of NLP approaches for vaccination behaviour detection. In *EMNLP SMM4H Workshop*.

Kandula, S.; and Shaman, J. 2019. Reappraising the utility of Google Flu Trends. *PLOS Computational Biology*.

Kılınç, D. D.; and Sayar, G. 2019. Assessment of reliability of youtube videos on orthodontics. *Turkish Journal of Orthodontics*.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. In *EMNLP*.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. *ICLR*.

Kleinberg, B.; van der Vegt, I.; and Mozes, M. 2020. Measuring emotions in covid-19 real world worry dataset. *arXiv:2004.04225*.

Klembczyk, J. J.; Jalalpour, M.; Levin, S.; Washington, R. E.; Pines, J. M.; Rothman, R. E.; and Dugas, A. F. 2016. Google flu trends spatial variability validated against emergency department influenza-related visits. *Journal of medical Internet research*.

Kobayashi, S. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. *NAACL*.

Kouzy, R.; Abi Jaoude, J.; Kraitem, A.; El Alam, M. B.; Karam, B.; Adib, E.; Zarka, J.; Traboulsi, C.; Akl, E. W.; and Baddour, K. 2020. Coronavirus goes viral: Quantifying the covid-19 misinformation epidemic on twitter. *Cureus*.

Li, L.; Jamieson, K.; Rostamizadeh, A.; Gonina, E.; Hardt, M.; Recht, B.; and Talwalkar, A. 2018. Massively parallel hyperparameter tuning. *arXiv preprint arXiv:1810.05934*.

Li, X.; Sun, X.; Meng, Y.; Liang, J.; Wu, F.; and Li, J. 2020. Dice Loss for Data-imbalanced NLP Tasks. In *ACL*.

Liaw, R.; Liang, E.; Nishihara, R.; Moritz, P.; Gonzalez, J. E.; and Stoica, I. 2018. Tune: A Research Platform for Distributed Model Selection and Training. *arXiv preprint arXiv:1807.05118*.

Likert, R. 1932. A technique for measurement of attitude scales. *Archives of psychology*.

Loeb, S.; Sengupta, S.; Butaney, M.; Macaluso Jr, J. N.; Czarniecki, S. W.; Robbins, R.; Braithwaite, R. S.; Gao, L.; Byrne, N.; Walter, D.; et al. 2019. Dissemination of misinformative and biased information about prostate cancer on YouTube. *European urology*.

Lohmann, S.; Lourentzou, I.; Zhai, C.; and Albarracín, D. 2018. Who is saying what on Twitter: an analysis of messages with references to HIV and HIV risk behavior. *Acta de investigación psicológica*, 8(1): 95–100.

Lourentzou, I.; Dyer, G.; Sharma, A.; and Zhai, C. 2015. Hotspots of news articles: Joint mining of news text & social media to discover controversial points in news. In *IEEE BigData*.

Manolache, A.; Brad, F.; and Burceanu, E. 2021. DATE: Detecting Anomalies in Text via Self-Supervision of Transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 267–277.

Mian, A.; and Khan, S. 2020. Coronavirus: the spread of misinformation. *BMC medicine*.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *IEEE International Conference on 3D Vision*.

Morales, A.; Narang, K.; Sundaram, H.; and Zhai, C. 2020. CrowdQM: Learning aspect-level user reliability and comment trustworthiness in discussion forums. In *PAKDD*.

Müller, M.; Salathé, M.; and Kummervold, P. E. 2020. COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter. *arXiv:2005.07503*.

Nicola, M.; Alsafi, Z.; Sohrabi, C.; Kerwan, A.; Al-Jabir, A.; Iosifidis, C.; Agha, M.; and Agha, R. 2020. The socio-economic implications of the coronavirus and COVID-19 pandemic: a review. *International Journal of Surgery*.

Pennycook, G.; McPhetres, J.; Zhang, Y.; and Rand, D. 2020. Fighting COVID19 misinformation on social media: Experimental evidence for a scalable accuracy nudge intervention. *PsyArXiv*.

Preis, T.; and Moat, H. S. 2014. Adaptive nowcasting of influenza outbreaks using Google searches. *Royal Society open science*.

Ruchansky, N.; Seo, S.; and Liu, Y. 2017. Csi: A hybrid deep model for fake news detection. In *CIKM*.

Ruff, L.; Zemlyanskiy, Y.; Vandermeulen, R.; Schnake, T.; and Kloft, M. 2019. Self-attentive, multi-context one-class classification for unsupervised anomaly detection on text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4061–4071.

Salehi, S. S. M.; Erdogmus, D.; and Gholipour, A. 2017. Tversky loss function for image segmentation using 3D fully convolutional deep networks. In *International Workshop on Machine Learning in Medical Imaging*.

Santillana, M.; Zhang, D. W.; Althouse, B. M.; and Ayers, J. W. 2014. What can digital disease detection learn from (an external revision to) Google Flu Trends? *American journal of preventive medicine*.

Schuster, M.; and Paliwal, K. K. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*.

Shin, J.; Jian, L.; Driscoll, K.; and Bar, F. 2018. The diffusion of misinformation on social media: Temporal pattern, message, and source. *Computers in Human Behavior*.

Shu, K.; Cui, L.; Wang, S.; Lee, D.; and Liu, H. 2019. defend: Explainable fake news detection. In *SIGKDD*.

Signorini, A.; Segre, A. M.; and Polgreen, P. M. 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*.

Singh, L.; Bansal, S.; Bode, L.; Budak, C.; Chi, G.; Kawintiranon, K.; Padden, C.; Vanarsdall, R.; Vraga, E.; and Wang, Y. 2020. A first look at COVID-19 information and misinformation sharing on Twitter. *arXiv:2003.13907*.

Smith, M.; Broniatowski, D. A.; Paul, M. J.; and Dredze, M. 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*.

Sudre, C. H.; Li, W.; Vercauteren, T.; Ourselin, S.; and Cardoso, M. J. 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer.

Tasnim, S.; Hossain, M. M.; and Mazumder, H. 2020. Impact of rumors and misinformation on COVID-19 in social media. *Journal of preventive medicine and public health*, 53(3): 171–174.

Timmermans, B.; Kuhn, T.; Beelen, K.; and Aroyo, L. 2017. Computational controversy. In *International Conference on Social Informatics*.

Tomaszewski, T.; Morales, A.; Lourentzou, I.; Caskey, R.; Liu, B.; Schwartz, A.; Chin, J.; et al. 2021. Identifying False Human Papillomavirus (HPV) Vaccine Information and Corresponding Risk Perceptions From Twitter: Advanced Predictive Models. *Journal of medical Internet research*, 23(9): e30451.

Tomeny, T. S.; Vargo, C. J.; and El-Toukhy, S. 2017. Geographic and demographic correlates of autism-related anti-vaccine beliefs on Twitter, 2009-15. *Social Science & Medicine*.

Wang, W. Y.; and Yang, D. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets. In *EMNLP*.

Wei, J.; and Zou, K. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In *EMNLP-IJCNLP*.

Wu, L.; Morstatter, F.; Carley, K. M.; and Liu, H. 2019. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter*, 21(2): 80–90.

Xie, Q.; Dai, Z.; Hovy, E.; Luong, M.-T.; and Le, Q. V. 2019. Unsupervised data augmentation for consistency training. *arXiv:1904.12848*.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; and Hovy, E. 2016. Hierarchical attention networks for document classification. In *NAACL-HLT*.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level convolutional networks for text classification. In *NeurIPS*.

Zhou, X.; Jain, A.; Phoha, V. V.; and Zafarani, R. 2020. Fake news early detection: A theory-driven model. *Digital Threats: Research and Practice*, 1(2): 1–25.

Zimet, G. D.; Rosberger, Z.; Fisher, W. A.; Perez, S.; and Stupiansky, N. W. 2013. Beliefs, behaviors and HPV vaccine: correcting the myths and the misinformation. *Preventive medicine*.