Proceedings of the Fifteenth International AAAI Conference on Web and Social Media (ICWSM 2021)

Misinformation Adoption or Rejection in the Era of COVID-19

Maxwell Weinzierl, [1] **Suellen Hopfer,** [2] **Sanda M. Harabagiu** [1]
[1]Human Language Technology Research Institute, University of Texas at Dallas
[2] Public Health, School of Medicine, University of California, Irvine
maxwell.weinzierl@utdallas.edu, shopfer@hs.uci.edu, sanda@utdallas.edu

Abstract

The COVID-19 pandemic has led to a misinformation avalanche on social media, which produced confusion and insecurity in netizens. Learning how to automatically recognize adoption or rejection of misinformation about COVID-19 enables the understanding of the effects of exposure to misinformation and the threats it presents. By casting the problem of recognizing misinformation adoption or rejection as *stance* classification, we have designed a neural language processing system operating on micro-blogs which takes advantage of Graph Attention Networks relying on lexical, emotion, and semantic knowledge to discern the stance of each micro-blog with respect to COVID-19 misinformation. This enabled us not only to obtain promising results, but also allowed us to use a taxonomy of COVID-19 misinformation themes and concerns to characterize the misinformation adoption or rejection that can be best recognized automatically.

Introduction

With the outbreak of the COVID-19 pandemic, people turned to social media platforms, such as Twitter, to find information about this infectious virus, in search to understand a range of issues, including its origin, efficient preventive measures, and eventual treatments. However, as information needs were surging, so was misinformation diffusion. People were exposed to false claims, rumors, and conspiracy theories. Therefore, as noted in (Tagliabue, Galassi, and Mariani 2020), misinformation influenced the public perception of COVID-19 risks. Hence, not only is it important to know how misinformation spreads but it is essential to understand when it is adopted or rejected.

Recently, epidemiological models originally designed for the study of the spread of biological viruses (Serrano, Iglesias, and Garijo 2015) have been used for in-depth analysis of the propagation of information about COVID-19 (Cinelli et al. 2020) on several social media platforms, such a Twitter, Instagram, Gab, Reddit, and YouTube, finding that unreliable information was amplified at almost the same rate as reliable facts. An exploratory study into the propagation, content, and authorship of misinformation on Twitter around the topic of COVID-19 was reported in (Shahi, Dirkson, and Majchrzak 2020), revealing that misinformation authors are driven by affiliations to others, promoting false claims under the guise of protecting others. The study also finds that COVID-19 misinformation tweets use relatively less informal language. e.g. less so-called netspeak (e.g. lol and thx), swearing and assent (e.g. OK) than tweets referring to factual information. Such stylistic features along with other content-based characteristics are known to inform the recognition of misinformation in social media platforms, together with signals of the social context of the misinformation, represented, as suggested in (Shu et al. 2017), by the features of the misinformation propagation and the observed *stance*.

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Misinformation Target: *Shaking hands cannot infect anyone since it is Sunnah.*
STANCE: **Agree** *Tweet:* NEW CASE: Illinois Health Officials say despite this new case, people do not need to alter their daily lives, but should work to stop the spread of germs.
STANCE: **Disagree** *Tweet:* Still many Indians follow this dharma intentionally and also unintentionally as well. Now WHO says, to maintain 3 feet distance from one individual to another to prevent from attack of Corona virus. Never touch anyone. Wash ur hands everytime.
Misinformation Target: *The coronavirus outbreak is a cover-up for a 5G-related illness.*
STANCE: **Agree** *Tweet:* @DineshDSouza Who ever said the Coronavirus is a hoax is correct. It's 5G radiation disease and its only going to get worse!!
STANCE: **Disagree** *Tweet:* I just read a tweet where someone claimed Coronavirus was actually a result of 5g exposure. These idiots walk among us.

Table 1: Examples of COVID-19 misinformation and tweets adopting or rejecting it from the COVIDLIES dataset.

The stance defines the attitude the author of a micro-blog manifests towards the misinformation target, as exemplified in Table 1. When the misinformation is adopted, an *agreement* stance is observed, whereas when it is rejected, the stance reflects the *disagreement* with the targeted misinformation. Automatic stance recognition is an affect identification task aiming to capture the attitudes expressed in texts with

respect to some target proposition. The examples listed in Table 1 originate from COVIDLIES (Hossain et al. 2020), the first benchmark of COVID-19 misinformation annotated with stance information, recently released. It is not the first dataset containing stance annotations. In 2017 and 2019 the data available in RumourEval (Gorrell et al. 2019) enabled multiple teams to compete in the task of rumor stance prediction on Twitter and Reddit.

As it can be seen from the examples listed in Table 1, identifying the stance of a tweet with respect to a target misinformation is a not trivial language processing task. In (Hossain et al. 2020), this problem was cast as an inference problem, a methodology also used for stance identification in(Augenstein et al. 2016) as well as for recognizing the stance in the Fake News Challenge (*www.fakenewschallenge.org*), as reported in (Mohtarami et al. 2018). However, much recent work, e.g. (Sun et al. 2018), (Siddiqua, Chy, and Aono 2019), has cast the problem of stance identification as a classification problem that benefits from lexical, sentiment, syntactic and contextual knowledge.

To find which methodology best serves the recognition of stance on the COVIDLIES dataset, we have designed a novel neural stance classification architecture which produced superior results on the same dataset compared with the results generated when casting stance identification as a textual inference problem, as reported in (Hossain et al. 2020). But more importantly, we discerned a taxonomy of themes and concerns from the misinformation targets available in COVIDLIES to find what kind of misinformation is more adopted and what kind is more rejected. Interestingly, we found that misinformation themes for which adoption is easier to detect automatically are the same for which misinformation rejection is harder to identify automatically, and vice versa. This important finding can guide future efforts on inoculation against misinformation in the era of COVID-19. Moreover, recent research (Loomba et al. 2021) has shown that exposure to online misinformation around COVID-19 vaccines affects intent to vaccinate in order to protect oneself or others. Therefore, knowledge about misinformation adoption or rejection can further explain the impact of misinformation on COVID-19 vaccine hesitancy.

Stance Twitter Annotations for COVID-19 Misinformation

The ongoing pandemic has led several researchers to develop datasets containing social media narratives about COVID-19, including the multilingual COVID-19 Twitter dataset presented in (Chen, Lerman, and Ferrara 2020), which is available to the research community via a COVID-19-TweetIDs GitHub. In addition, the COAID (Covid-19 heAlthcare mIsinformation Dataset) is another public benchmark dataset which includes (a) confirmed fake and true news articles about COVID-19 from fact-checked or reliable websites; and (b) postings from social platforms (e.g. Facebook, Twitter, Instagram, Youtube, and TikTok) that contain links to the articles. This enables the social media postings to be categorized as containing misinformation or not based on their links to

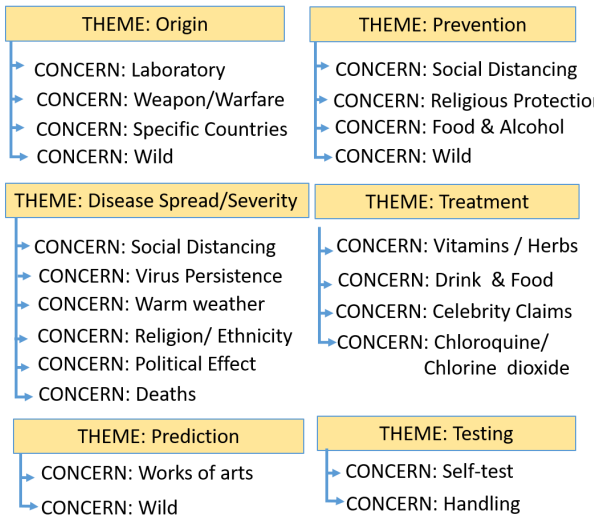

Figure 1: Taxonomy of Themes and Concerns discerned from the COVIDLIES Misinformation Targets.

fake or true news articles. However, none of these datasets have any stance annotations available.

A very different and interesting perspective on COVID-19 misinformation detection on social media was taken in (Hossain et al. 2020), which introduced the COVIDLIES dataset. The starting point was 86 common misconceptions about COVID-19 available from the Wikipedia page dedicated to COVID-19 misinformation, which became *misinformation targets*. For each known misinformation target, a set of relevant tweets were automatically retrieved, using BERTSCORE (Zhang et al. 2020b). The 100 most relevant tweets for each misinformation target were selected for stance annotation, performed by researchers in the UCI School of Medicine. There are three possible values for stance: (1) *agree*, when the tweet adopts the target misinformation; (2) *disagree*, when the tweet contradicts/rejects the target misinformation; and (3) *no stance* when the tweet is either neutral or is irrelevant to the targeted misinformation. It is to be noted that the annotation decisions for *no stance* conflate the stance decision with the relevance decision, combining two distinct linguistic phenomena: attitude and relevance. This may pose significant issues to the problem of stance identification, as some training and testing data could have nothing to do with the stance of the tweet, but only with the relevance of the tweet against a misinformation target. Of the 6761 annotated tweets, 5,748 (85.02%) received a label of *no stance*; 670 (9.91%) received a label of *agree* and 343 (5.07%) received a label of *disagree*.

When analyzing the misinformation targets available from the COVIDLIES dataset, we were able to find that six different *themes* of misinformation emerged. They regarded (1) the origin of the COVID-19 virus; (2) prevention against infection with COVID-19; (3) treatment; (4) the spread and severity of the disease caused by COVID-19; (5) predictions related to COVID-19 and (6) testing. As it can be seen in Figure 1, the ORIGIN theme of misinformation covers four concerns: (1) the virus originated as a weapon or warfare; (2) it origi-

Figure 2: Count of tweet / misinformation target pairs from COVIDLIES, grouped by Theme and then by Concern.

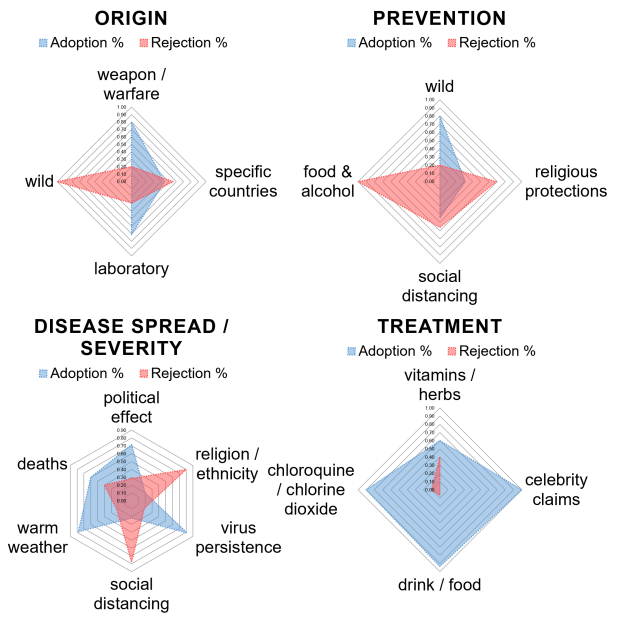

Figure 3: Distribution of Adoption vs Rejection of Misinformation across Themes and Concerns Discerned in the COVIDLIES dataset.

nated in a laboratory; (3) it originated in specific countries, e.g. U.S.A, Canada, China or Israel; and (4) there are some "wild" concerns, e.g. its origin is motivated by population control, or it is the result of a spy operation. Similarly, the PREVENTION and the TREATMENT themes are characterised by four concerns. The DISEASE SPREAD/SEVERITY theme is characterized by five concerns whereas the PREDICTION and TESTING themes cover only two concerns each.

As shown in Figure 2, the largest number of tweets from COVIDLIES address misinformation regarding the ORIGIN of the virus, whereas misinformation about TESTING and PREDICTIONS related to COVID-19 are encountered in the smallest number of tweets. The number of tweets relevant to misinformation with TREATMENT, DISEASE SPREAD/SEVERITY, and PREVENTION is of similar magnitude. However, more interesting is the distribution of tweets for each concern that either adopt or reject the misinformation pertaining to each theme. Figure 3 shows that our taxonomy of misinformation operating on the COVIDLIES annotations of stance indicates that most adopted misinformation about the origin of COVID-19 pertains to its claimed origin in certain countries, whereas, not surprisingly, the most rejected misinformation pertains to "wild" concerns. In contrast, for the misinformation covering the PREVENTION theme, it is not the "wild" concerns that are most rejected, but those about the role of food and alcohol on prevention, while the "wild" concerns are the most adopted forms of misinformation. Misinformation falling under the DISEASE SPREAD/SEVERITY seems to be adopted when addressing concerns expressing the role of warmer weather, political effects (promoted by the Trump administration), the number of deaths recorded, or the virus persistence (on various surfaces). The misinformation is mostly rejected when addressing concerns of social distancing and religious or ethnic issues. Surprisingly, almost all misinformation about treatments is adopted, with few exceptions when it concerns the role of vitamins or herbs (including cannabis) when it is rejected. Because, as illustrated in Figure 1, the misinformation for the PREDICTION and TESTING themes expressed only two concerns, the distribution of adoption vs. rejection of misinformation pertaining to these themes is not displayed in Figure 3. For the PREDICTION themes, most adopted misinformation dealt with the WORKS OF ARTS concern, encompassing claims that the coronavirus was predicted in various works of art, ranging from novels (e.g. "The Eyes of Darkness") to shows (e.g. "The Simpsons"), while the rejected misinformation covered "wild" concerns. The misinformation for the TESTING theme was in general adopted, rather than rejected.

Given the variation of stance annotations across themes and concerns, we also contemplated if any data enhancement should be considered. Because currently there are no additional datasets of COVID-19 misinformation annotated with stance, we experimented with another recently released dataset that annotated the *severity* of COVID-19 misinformation, rather than the stance. The HeRA (Dharawat et al. 2020) dataset provides several additional annotations on a subset of the CoAID (Cui and Lee 2020) dataset, which is annotated with real and misinformation news articles and claims along with tweets which discuss these articles or claims. HeRA refines the CoAID annotations on the tweet-level by further annotating misinformation-discussing tweets with either a severity level or as a rebuttal to misinformation. We hypothesized that the severity and rebuttal annotation scheme implies a likely stance s: if the tweet t is annotated as a rebuttal then the tweet likely rejects the misinformation in the news article

or claim. If the tweet t is annotated with a certain level of severity, we assumed that it is likely that the tweet adopts the misinformation of the news article or claim. In addition, we extract the misinformation target m from the news article headline or claim the tweet is discussing.

Given the COVIDLIES dataset and the taxonomy of misinformation we have created, as well as the possibility of data enhancement based on the alignment of annotations of the HeRA dataset with the COVIDLIES annotations, we were interested to find how well a neural language processing system could find which misinformation is adopted or rejected, with or without any data enhancement.

Automatic Identification of Stance toward COVID-19 Misinformation

While the creators of the COVIDLIES dataset have cast the problem of stance identification as an inference problem, which can benefit from existing textual inference datasets, we believe that the language uses of agreement or disagreement with a certain misinformation target, as expressed in the tweets of the dataset, are a function of the deep pragmatic cues of the context of the tweet, and may be further signaled by lexico-syntactic, semantic, and emotion features. Consequently, we have designed a neural language processing system that exploits the pre-trained domain-specific language model COVID-Twitter-BERT-v2 (Müller, Salathé, and Kummervold 2020) and refines it by stacking several layers of lexico-syntactic, semantic, and emotion Graph Attention Networks (GATs) (Veličković et al. 2018) to learn and refine all the possible interactions between these different linguistic phenomena, before classifying a tweet as (a) agreeing; (b) disagreeing or (c) having no stance towards a misinformation target. For this purpose, we have designed the architecture of the Lexical, Emotion, and Semantic Graph Attention Network system for Stance Identification (LES-GAT-StanceId), illustrated in Figure 4.

Figure 4 shows how, given a misinformation target m composed of a word-piece tokens (Devlin et al. 2019) $m_1, m_2, ..., m_a$ and a tweet t composed of b word-piece tokens $t_1, t_2, ..., t_b$, the pre-trained domain-specific language model COVID-Twitter-BERT-v2 (Müller, Salathé, and Kummervold 2020) produces contextualized embeddings for each word-piece token in the joint misinformation target / tweet sequence $[CLS], m_1, m_2, ..., m_a, [SEP], t_1, t_2, ..., t_b, [SEP]$, where $[CLS]$ represents a special classifier token which marks the beginning of the first sequence of tokens and $[SEP]$ represents a special separator token used to mark the end of a sequence of tokens. COVID-Twitter-BERT-v2 was pre-trained on 97M COVID-19 tweets, providing domain-specific language modeling for tasks concerning COVID-19. The contextualized embeddings $C = \{c_1, c_2, ..., c_L\}$ of size $L = a + b + 3$ generated by COVID-Twitter-BERT-v2 are further enriched with lexico-syntactic, emotion and semantic information through three corresponding Graph Attention Networks (GATs). Each GAT operates on a different graph.

The Lexical Graph, on which the Lexical GAT illustrated in Figure 4 operates, consists of each word from the misinformation target m or the tweet t which are spanned by some

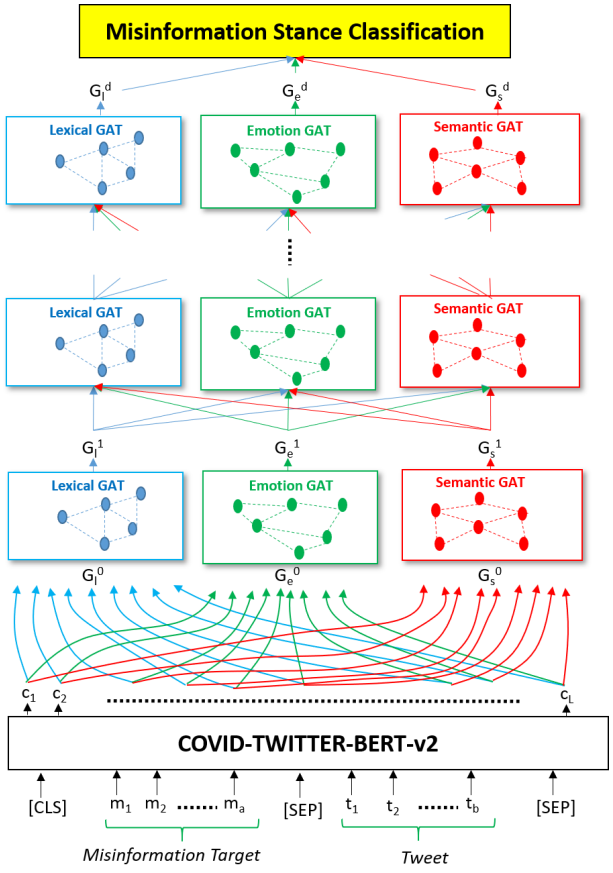

Figure 4: Neural architecture of the Lexical, Emotion, and Semantic Graph Attention Network for Stance Identification (LES-GAT-StanceId) system.

syntactic dependency relation generated with SpaCy (Honnibal et al. 2020), whereas the edges are the product of the dependency parse. For example, the word "shaking" from the first misinformation target illustrated in Table 1 is connected to the words "hands" and "infect". To build the Emotion Graph on which the Emotion GAT operates, words are provided with emotion tags available from SenticNet 5 (Cambria et al. 2018) which follow the Hourglass of Emotions model (Cambria, Livingstone, and Hussain 2011) for emotion categorization of words. For example, the word "virus" has both the "fear" and the "disgust" emotion tags. Edges between words which share one or more emotion tags complete the Emotion Graph. The generation of the Semantic Graph, on which the Semantic GAT operates, is using semantic similarity information. SenticNet 5 (Cambria et al. 2018) provides information about the semantic similarity between pairs of words. For example, the word "contagious" is considered semantically similar to the words "infectious", "communicable", "epidemic", "pandemic", and "epizootic". An edge spans all pairs of semantically similar words. To expand the Semantic Graph, edges are generated between pairs of words which are deemed semantically similar within h edge hops, as in (Zhang et al. 2020a). Furthermore, self-loops are also

added for every word in the Lexical, Emotion, and Semantic Graphs to allow for contextual information to inform the graph representations and to produce stable training when words have no edges. Each Graph Attention Network (GAT) operates on one of these graphs with the purpose to refine the representations of each word from each graph through self-attention, taking into account the contributions of the adjacent connections available in each graph.

For simplicity reasons, Figure 4 does not illustrate the fact that LES-GAT-StanceId first projects the large 1024-dimensional contextualized embeddings C from COVID-Twitter-BERT-v2 down to $F << 1024$ with a linear layer for each of the Lexical, Emotion, and Semantic Graphs. The respective linear layers are implemented as:

$$G_l^0 = CW_l + b_l \quad (1)$$

$$G_e^0 = CW_e + b_e \quad (2)$$

$$G_s^0 = CW_s + b_s \quad (3)$$

where W_l, W_e, W_s are weight matrices of size $1024 \times F$ and b_l, b_e, b_s are weight vectors of size F (where all these matrices and vectors are model parameters that are learned during training). These layers reduce the complexity and increase the efficiency of the following Graph Attention Networks by reducing the number of trainable parameters and improving the speed.

As illustrated in Figure 4, the LES-GAT-StanceId system stacks n layers of GATs. A GAT at layer $n \in \{1, ..., d\}$ computes a hidden representation for every word-piece node embedding $g_i^{n-1} \in G^{n-1}$ as $h_i^n = W^n g_i^{n-1}$, where W^n is a weight matrix of varying size. At $n = 1$ the size of W^1 is of size $F \times F$ from the linear projection layer, while at $n > 1$ we set W^n to size $F \times 3F$, since the input at layers $n > 1$ consider inputs from the Lexical, Emotion, and Semantic GATs at the previous layer $n-1$: $G^{n-1} = [G_l^{n-1}, G_e^{n-1}, G_s^{n-1}]$. This hidden representation is utilized to compute self-attention weights with the following equation:

$$\alpha_{i,j}^n = \frac{exp(LeakyReLU((a^n)^T[h_i^n, h_j^n]))}{\sum_{k \in adj(i)} exp(LeakyReLU((a^n)^T[h_i^n, h_k^n]))} \quad (4)$$

where a^n is a weight vector of size $2F$, $[..,..]$ represents concatenation, $LeakyReLU(x) = max(0.2x, x)$, and $adj(...)$ produces the list of adjacent nodes for a given node in either the Lexical, Emotion, or Semantic graphs. These attention weights are utilized along with the hidden representations of word-piece node embeddings to produce the final GAT layer representation as:

$$g_i^n = \sigma(\sum_{j \in adj(i)} \alpha_{i,j}^n h_j^n) \quad (5)$$

where σ is an exponential linear unit (ELU) nonlinearity (Clevert, Unterthiner, and Hochreiter 2016).

The final output for each GAT on layer n is therefore $G^n = \{g_1^n, g_2^n, ..., g_L^n\}$. Each of the GATs at each layer produces a graph representation G_l^n, G_e^n, and G_s^n respectively, which each utilize Lexical, Emotion, and Semantic edges respectively for their adjacency function. These Lexical, Emotion, and Semantic Graph representations are concatenated together to form $G^n = [G_l^n, G_e^n, G_s^n]$. G^n is then provided as input to all three Lexical, Emotion, and Semantic GATs for the next layer, producing G^{n+1}. The LES-GAT-StanceId system has d layers of separate Lexical, Emotion, and Semantic Graph Attention Networks. This allows each Lexical, Emotion, and Semantic GAT to consider previous Lexical, Emotion, and Semantic Graph representations jointly, learning graph node embeddings which consider interactions between different graphs. This process results in a final LES-GAT-StanceId graph node representation $G^d = [G_l^d, G_e^d, G_s^d]$.

The misinformation stance classification layer of the LES-GAT-StanceId system is provided as a fixed-length representation by taking the average embedding of the final LES-GAT-StanceId layer d as $z = \frac{1}{L}\sum_{i=1}^{L} g_i^d$. This embedding z is provided to the stance classification layer, which employs a fully connected layer with a softmax activation function to produce final probabilities $P(Agree|m,t)$, $P(Disagree|m,t)$, and $P(No\_Stance|m,t)$.

The LES-GAT-StanceId system is trained end-to-end on the cross-entropy loss function:

$$\mathcal{L} = -\sum_{(s,m,t) \in D} logP(s|m,t;\theta) \quad (6)$$

where $s \in \{Agree, Disagree, No\_Stance\}$, D is a set of all training examples of labeled tweet / misinformation pairs, and θ is a set of all trainable parameters from LES-GAT-StanceId. These parameters are optimized with ADAM (Kingma and Ba 2015), a variant of gradient descent, to minimize $\mathcal{L}$.

The HeRA (Dharawat et al. 2020) dataset is also utilized as additional pre-training data. Therefore, a dataset D' is constructed of stance s, misinformation target m, and tweet t triples from the HeRA dataset. This dataset contains likely examples of agreement and disagreement between tweets and misinformation targets, which provides a good source of pre-training for the task of stance identification in COVIDLIES. First, the trainable parameters θ of LES-GAT-StanceId are initialized randomly or to pre-trained values for COVID-Twitter-BERT-v2. The parameters are then optimized on the loss $\mathcal{L}'$ using D', and then fine-tuned on the loss $\mathcal{L}$ using D, the training data from COVIDLIES. This system is referred to as LES-GAT-HeRA-StanceId.

A thresholding technique is also implemented due to the imbalance of data in the COVIDLIES dataset. A tweet / misinformation pair is classified as the maximum probability stance between *Agree* or *Disagree* if $P(Agree|m,t) > T$ or $P(Disagree|m,t) > T$. This thresholding method is similar to prior work (Hossain et al. 2020), but the same joint LES-GAT-StanceId stance system is used as opposed to a separate relevancy model.

Evaluation Results

Stance identification performance on the COVIDLIES dataset was evaluated on three systems: (1) the Domain-Specific Stance Identification (DS-StanceId) system; (2) the Lexical, Emotion, and Semantic Graph Attention Network for Stance Identification (LES-GAT-StanceId) system; and (3) the Lexical, Emotion, and Semantic Graph Attention Network

Stance Recognition System	*Agree*			*Disagree*			*No_Stance*			Macro		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
(Hossain et al. 2020)	63.3	30.6	41.2	14.4	34.1	20.3	90.0	88.0	89.0	55.9	50.9	50.2
DS-StanceId	**73.5**	73.5	73.5	52.3	44.2	46.2	95.3	96.2	95.8	73.7	71.3	71.8
LES-GAT-StanceId	72.2	72.7	72.5	**57.6**	48.7	52.8	95.6	**96.4**	**96.0**	**75.1**	72.6	73.7
LES-GAT-HeRA-StanceId	69.4	**78.8**	**73.8**	54.4	**52.3**	**53.3**	**96.3**	95.2	95.7	73.4	**75.4**	**74.3**

Table 2: Results from the 5-fold cross-validation stance identification experiments on the COVIDLIES dataset.

enhanced with HeRA for Stance Identification (LES-GAT-HeRA-StanceId) system. The DS-StanceId system utilizes the "[CLS]" embedding from COVID-Twitter-BERT-v2 as the misinformation stance classification input embedding z. The LES-GAT-StanceId system utilizes Lexical, Emotion, and Semantic Graph Attention Networks architecture illustrated in Figure 4. The LES-GAT-HeRA-StanceId system utilizes HeRA as pre-training data, but otherwise is the same as the LES-GAT-StanceId system.

We performed 5-fold cross-validation on the unique tweets within the COVIDLIES dataset. Hyper-parameters were selected based on initial experiments on an independent 80/20 split. The number of semantic hops h was set to 3, the size of the GATs F was set to 64, the depth of GAT layers d was set to 6, and the threshold T was set to 0.2. The LES-GAT-HeRA-StanceId system shares the same hyper-parameters as the LES-GAT-StanceId system except for the learning rate of the HeRA pre-training step, which is set to $5e-5$. All systems follow the same training schedule on each of the five splits: 10 epochs, a linearly decayed learning rate of $5e-4$ with a warm-up for 10% of training steps, an attention drop-out rate of 10%, and gradient norm clipping of 1.0. We compare against the best results from Hossain et al. (2020), which utilize a pre-trained domain-adapted (DA) natural language inference (NLI) Sentence-BERT (Reimers and Gurevych 2019) system along with a DA misinformation retrieval model BERTScore (Zhang et al. 2020b). Results are provided in Table 2.

Performance was determined based on Precision (P), Recall (R), and F_1[1] score for detecting the three values of stance, namely *Agree*, *Disagree*, and *No_Stance*. We also compute a Macro averaged Precision, Recall, and F_1 score. Evaluation metrics were computed over detected stances for all 5 cross-validation splits. The DS-StanceId system produced a Macro F_1 score of 71.8, which demonstrates the advantage of fine-tuning stance identification systems. The LES-GAT-StanceId system produced a Macro F_1 score of 73.7, which indicates that integrating Lexical, Emotional, and Semantic Graphs improves stance identification. The LES-GAT-HeRA-StanceId system produced a Macro F_1 score of 74.3, supporting our hypothesis that misinformation deemed as severe or rebutted in the HeRA dataset contributes to the improvement of stance identification. The results also show that detecting disagreement, thus misinformation rejection, is more difficult than detecting agreement, thus misinformation adoption.

Stance identification results for the DS-StanceId system in Table 2 demonstrate that significant performance gains can be attributed to fine-tuning stance identification systems. Major performance gains, from an F_1 score of 41.2 to 73.5 for the *Agree* stance and 20.3 to 46.2 for the *Disagree* stance represent the value of fine-tuning on tweet / misinformation pairs as opposed to using only general Natural Language Inference (NLI) datasets, as it was done in (Hossain et al. 2020).

Improvements in stance identification for the LES-GAT-StanceId system are driven largely by improvements in the *Disagree* stance. The *Disagree* stance has the fewest number of examples, with only 343 instances in a dataset of 6,761 examples. The LES-GAT-StanceId system overcomes this resource constraint by integrating additional Lexical, Emotion, and Semantic information. The LES-GAT-StanceId system gains 6.6 points of F_1 score over the DS-StanceId system for the *Disagree* stance, which can be attributed to the utilization of Lexical, Emotion, and Semantic Graph edges and their learned interactions in the LES-GAT-StanceId system. Finding aligned emotions and shared semantics between words in misinformation targets and words in tweets, along with a lexical consideration of the role of those words, allows the LES-GAT-StanceId system to improve *Disagree* stance identification.

While the *Disagree* stance is the stance with the fewest number of examples, the *Agree* stance also only has 670 examples in the COVIDLIES dataset. The LES-GAT-HeRA-StanceId system adds an additional 7,632 *Agree* and 443 *Disagree* inferred stance tweet / misinformation pairs from the HeRA dataset. The LES-GAT-HeRA-StanceId system is pre-trained on these inferred stance examples and fine-tuned on the examples in COVIDLIES, which produces improvements over the LES-GAT-StanceId system in both the *Agree* and *Disagree* stance F_1 scores, gaining 1.3 points and 0.5 points respectively.

Discussion

Because in our experiments we have obtained the best stance detection when using the LES-GAT-HeRA-StanceId system, as illustrated in Table 2, we performed an analysis of the performance of the system for each of the themes of misinformation that we have discerned on the COVIDLIES dataset, both for adoption and rejection of misinformation, as illustrated in Figure 5. Detecting the *Agree* stance, and thus the adoption of misinformation, is shown to be more difficult for the TESTING, PREVENTION, and ORIGIN themes, while it is easier for misinformation considering PREDICTION, TREATMENT, and DISEASE SPREAD / SEVERITY themes. Detecting the *Disagree* stance, and thus the rejection of misinfor-

[1] F_1 is defined as $F_1 = 2 \times P \times R/(P+R)$

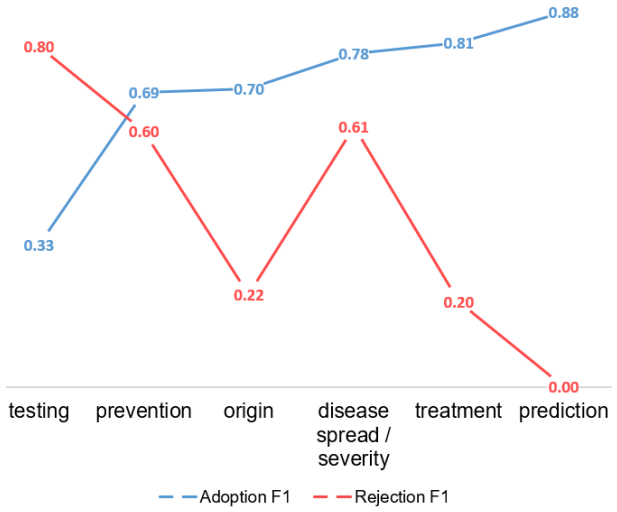

Figure 5: F_1-scores of the misinformation adoption vs. rejection discovered by the LES-GAT-HeRA-StanceId system across misinformation themes from the COVIDLIES dataset.

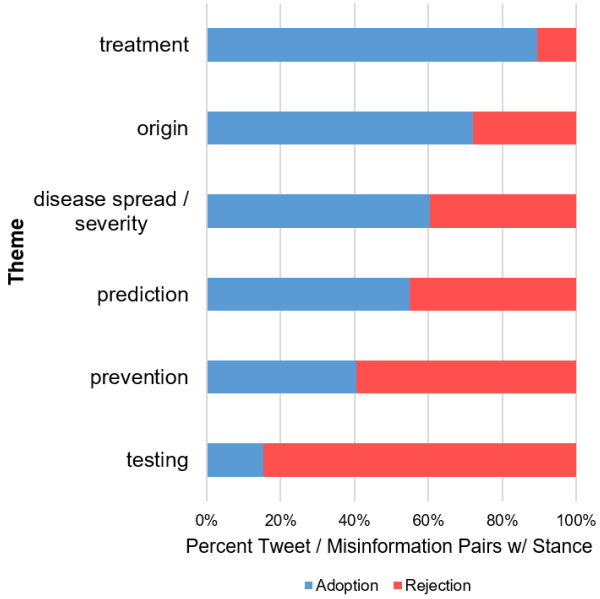

Figure 6: Adoption vs Rejection of misinformation for each [tweet/misinformation target] pair across misinformation themes from the COVIDLIES dataset.

mation is shown to be more difficult for the PREDICTION, TREATMENT, and ORIGIN themes, while it is easier for the TESTING, DISEASE SPREAD / SEVERITY, and PREVENTION themes. Interestingly, misinformation themes for which adoption is easier to detect are the same for which misinformation rejection is harder to identify, and vice versa.

To find the reason for the relative ease of identifying misinformation adoption for some themes, which also present difficulty in identifying misinformation rejection (and vice versa), we investigated the percentage of [tweet / misinformation target] pairs in the COVIDLIES dataset which exhibit misinformation adoption or rejection, illustrating the findings in Figure 6. A high adoption percentage indicates that a given misinformation theme tends to be easily adopted on Twitter, with most users having stances of agreement. For example, misinformation surrounding fake, ineffective, or harmful TREATMENT of COVID-19 appears to be largely accepted on Twitter along with misinformation surrounding the ORIGIN of COVID-19, such as the conspiracy of COVID-19 being a bio-engineered virus. A high rejection percentage indicates a misinformation theme is largely rejected, meaning most Tweets disagree with misinformation falling within that theme. For example, misinformation surrounding TESTING or PREVENTION, such as the availability of tests in the United States or the United States' early response to the pandemic, is largely rejected on Twitter. A balanced acceptance and rejection percentage indicates a highly contentious theme which has balanced levels of adoption and rejection on Twitter. For example, themes of DISEASE SPREAD/ SERIOUSNESS and PREDICTION appear to be relatively balanced on this Twitter dataset.

Figure 5 also helps explain why misinformation rejection is easier to identify in some themes, which also provides a more difficult case for identifying misinformation adoption. Themes such as TREATMENT and ORIGIN show a higher percentage of adopted misinformation than rejected misinformation, while misinformation about themes such as TESTING are predominantly rejected. The theme of PREDICTION stands as an outlier, in that it is the most balanced theme on Twitter with respect to adoption and rejection, but the performance of the LES-GAT-HeRA-StanceId system is extremely poor for rejection. This discrepancy can be partially explained by considering the concerns within the PREDICTION theme, which are provided in Figure 2. The WORKS OF ART concern is comprised entirely of examples of *Agreement* stance or *No_Stance*, while the DEATHS concern only has examples of *Disagreement* stance or *No_Stance*. The LES-GAT-HeRA-StanceId system therefore performs well on the WORKS OF ART concern, with an F_1 score of 0.88, while the DEATHS concern is entirely incorrect, with an F_1 score of 0. An inspection of the 9 examples of the DEATHS concern indicate that these mistakes are largely due to a lack of common sense knowledge available to the system. For example, given the misinformation target: **Misinformation Target 212:** *"There will be 500 deaths at the end of the pandemic"* and the following tweet:
"AureliaCotta @ungubunugu1274 @McBlondeLand @trvrb @seattleflustudy Should people be afraid if we are on the brink of a pandemic which will likely kill more people than died in WW1 and WW2 combined? That's what Spanish flu did, and coronavirus is looking as bad or worse.", in order to recognize that the content of the tweet is expressing a stance of *Disagreement* with the misinformation target, common sense knowledge in the form that much more than 500 people died in World War 1 (WW1) and World War 2 (WW2), is necessary. Commonsense Knowledge (or pragmatics) like this dominates the DEATHS concern's 9 examples of *Disagreement* stance which the LES-GAT-HeRA-StanceId system

Misinformation Target	Tweet	RoB-RT-Sentiment	LES-GAT-HeRA-StanceId	Annotation
216: We're very close to a vaccine.	Everything you said is true plus 1 more thing. @StacyLStiles @drdrew 3....TRUMP: "We are rapidly developing a vaccine. ... The vaccine is coming along well, and in speaking to the doctors, we think this is something that we can develop very rapidly." — news conference Wednesday THE FACTS: No vaccine is imminent for the coronavirus	*Positive*	*Disagree*	*Disagree*
157: Hand sanitiser sold commercially does not destroy the coronavirus.	@MykeCole @KarenFlemingPhD @amayor Hand sanitizer works on enveloped viruses as long as they have at least 60% alcohol content. However, soap is better because you're actually washing away the virus, not just killing it. So when you can't wash, hand sanitizer is a good second option. <link>	*Positive*	*Disagree*	*Disagree*
225: Anybody in the U.S. who wants a COVID-19 test can get a test.	@denmeow The care act I think it's called is to protect people against payments of federally required tests. That means that if you have to take a test for COVID-19 then your insurance company HAS TO PAY In full idk about no insured tho	*Neutral*	*Disagree*	*Disagree*

Table 3: Misinformation Targets and Tweets which demonstrate the advantage of stance identification as compared to sentiment analysis.

incorrectly identifies as *No_Stance*.

Furthermore, to highlight the importance of designing stance detection systems for identifying the adoption or rejection of misinformation in the COVID-19 era, we also considered a sentiment detection system. Often, it is incorrectly believed that sentiment detection is sufficient for identifying a multitude of linguistic affect phenomena. Comparison of the LES-GAT-HeRA-StanceId system with the RoB-RT-Sentiment system (Barbieri et al. 2020), a state-of-the-art Twitter sentiment detection system, is presented in Table 3. This comparison demonstrates that sentiment detection alone is not sufficient for stance identification, as sentiment and stance values can often be misaligned. The sentiment of a tweet entirely ignores the misinformation target, and therefore can miss the user's stance towards that target. The first and second examples from Table 3 demonstrate how a tweet can contain a *Positive* sentiment, in both cases as tweet responses to other users, but it expressed *Disagreement* with respect to certain misinformation targets. The third example demonstrates how the user's uncertainty surrounding various facts leads the sentiment analysis system to decide on a *Neutral* value, while the tweet actually expresses a *Disagreement* stance with respect to the misinformation target.

Conclusion

In this paper we described a neural stance classification system using COVID-Twitter-BERT-v2 and stacked Graph Attention Networks operating on lexico-syntactic, emotion, and semantic knowledge. The system obtains significant improvements over stance detection methods relying on natural language inference. In addition, informed by a misinformation taxonomy that we discerned from the COVIDLIES dataset, we present an analysis of misinformation themes for which misinformation adoption or rejection is easier to automatically identify. This work is a stepping stone in the direction of developing misinformation inoculation interventions on social media platforms in the era of COVID-19.

References

Augenstein, I.; Rocktäschel, T.; Vlachos, A.; and Bontcheva, K. 2016. Stance Detection with Bidirectional Conditional Encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 876–885. Austin, Texas: Association for Computational Linguistics.

Barbieri, F.; Camacho-Collados, J.; Espinosa Anke, L.; and Neves, L. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1644–1650. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.findings-emnlp.148. URL https://www.aclweb.org/anthology/2020.findings-emnlp.148.

Cambria, E.; Livingstone, A.; and Hussain, A. 2011. The Hourglass of Emotions. COST'11, 144–157. Berlin, Heidelberg: Springer-Verlag. ISBN 9783642345838. doi:10.1007/978-3-642-34584-5_11. URL https://doi.org/10.1007/978-3-642-34584-5_11.

Cambria, E.; Poria, S.; Hazarika, D.; and Kwok, K. 2018. SenticNet 5: Discovering Conceptual Primitives for Sentiment Analysis by Means of Context Embeddings. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16839.

Chen, E.; Lerman, K.; and Ferrara, E. 2020. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health Surveillance* 6.

Cinelli, M.; Quattrociocchi, W.; Galeazzi, A.; Valensise, C.; Brugnoli, E.; Schmidt, A.; Zola, P.; Zollo, F.; and Scala, A. 2020. The COVID-19 social media infodemic. *Nature Scientific reports* 10.

Clevert, D.-A.; Unterthiner, T.; and Hochreiter, S. 2016. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs).

Cui, L.; and Lee, D. 2020. CoAID: COVID-19 Healthcare Misinformation Dataset. *https://arxiv.org/abs/2006.00885* .

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.

Dharawat, A.; Lourentzou, I.; Morales, A.; and Zhai, C. 2020. Drink bleach or do what now? Covid-HeRA: A dataset for risk-informed health decision making in the presence of COVID19 misinformation. *https://arxiv.org/abs/2010.08743* .

Gorrell, G.; Kochkina, E.; Liakata, M.; Aker, A.; Zubiaga, A.; Bontcheva, K.; and Derczynski, L. 2019. SemEval-2019 Task 7: RumourEval, Determining Rumour Veracity and Support for Rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, 845–854. Minneapolis, Minnesota, USA: Association for Computational Linguistics.

Honnibal, M.; Montani, I.; Van Landeghem, S.; and Boyd, A. 2020. spaCy: Industrial-strength Natural Language Processing in Python. doi:10.5281/zenodo.1212303. URL https://doi.org/10.5281/zenodo.1212303.

Hossain, T.; Logan IV, R. L.; Ugarte, A.; Matsubara, Y.; Young, S.; and Singh, S. 2020. COVIDLies: Detecting COVID-19 Misinformation on Social Media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Online: Association for Computational Linguistics.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. URL http://arxiv.org/abs/1412.6980.

Loomba, S.; de Figueiredo, A.; Piatek, S. J.; de Graaf, K.; and Larson, H. J. 2021. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behavior* .

Mohtarami, M.; Baly, R.; Glass, J.; Nakov, P.; Màrquez, L.; and Moschitti, A. 2018. Automatic Stance Detection Using End-to-End Memory Networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 767–776. New Orleans, Louisiana: Association for Computational Linguistics.

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

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. Hong Kong, China: Association for Computational Linguistics. doi:10.18653/v1/D19-1410. URL https://www.aclweb.org/anthology/D19-1410.

Serrano, E.; Iglesias, C.; and Garijo, M. 2015. *A Survey of Twitter Rumor Spreading Simulations*, 113–122. ISBN 978-3-319-24068-8.

Shahi, G.; Dirkson, A.; and Majchrzak, T. A. 2020. An Exploratory Study of COVID-19 Misinformation on Twitter. *https://arxiv.org/abs/2005.05710* .

Shu, K.; Sliva, A.; Wang, S.; and Liu, H. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter* 9.

Siddiqua, U. A.; Chy, A. N.; and Aono, M. 2019. Tweet Stance Detection Using an Attention based Neural Ensemble Model. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 1868–1873. Minneapolis, Minnesota: Association for Computational Linguistics. doi:10.18653/v1/N19-1185. URL https://www.aclweb.org/anthology/N19-1185.

Sun, Q.; Wang, Z.; Zhu, Q.; and Zhou, G. 2018. Stance Detection with Hierarchical Attention Network. In *Proceedings of the 27th International Conference on Computational Linguistics*, 2399–2409. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Tagliabue, F.; Galassi, L.; and Mariani, P. 2020. The "Pandemic" of Disinformation in COVID-19. *SN Comprehensive Clinical Medicine* 2. doi:10.1007/s42399-020-00439-1.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph Attention Networks. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=rJXMpikCZ.

Zhang, B.; Yang, M.; Li, X.; Ye, Y.; Xu, X.; and Dai, K. 2020a. Enhancing Cross-target Stance Detection with Transferable Semantic-Emotion Knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3188–3197. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.291. URL https://www.aclweb.org/anthology/2020.acl-main.291.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020b. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=SkeHuCVFDr.