Identifying the Adoption or Rejection of Misinformation Targeting COVID-19 Vaccines in Twitter Discourse

Maxwell Weinzierl
Human Language Technology Research Institute,
University of Texas at Dallas
Richardson, Texas, USA
maxwell.weinzierl@utdallas.edu

Sanda Harabagiu
Human Language Technology Research Institute,
University of Texas at Dallas
Richardson, Texas, USA
sanda@utdallas.edu

ABSTRACT

Although billions of COVID-19 vaccines have been administered, too many people remain hesitant. Misinformation about the COVID-19 vaccines, propagating on social media, is believed to drive hesitancy towards vaccination. However, exposure to misinformation does not necessarily indicate misinformation adoption. In this paper we describe a novel framework for identifying the stance towards misinformation, relying on *attitude consistency* and its properties. The interactions between attitude consistency, adoption or rejection of misinformation and the content of microblogs are exploited in a novel neural architecture, where the stance towards misinformation is organized in a knowledge graph. This new neural framework is enabling the identification of stance towards misinformation about COVID-19 vaccines with state-of-the-art results. The experiments are performed on a new dataset of misinformation towards COVID-19 vaccines, called CoVaxLies, collected from recent Twitter discourse. Because CoVaxLies provides a taxonomy of the misinformation about COVID-19 vaccines, we are able to show which type of misinformation is mostly adopted and which is mostly rejected.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**; **Artificial intelligence**.

KEYWORDS

COVID-19, vaccine, misinformation, twitter, social media, stance

ACM Reference Format:
Maxwell Weinzierl and Sanda Harabagiu. 2022. Identifying the Adoption or Rejection of Misinformation Targeting COVID-19 Vaccines in Twitter Discourse. In *Proceedings of the ACM Web Conference 2022 (WWW '22), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3485447.3512039

1 INTRODUCTION

Although billions of inoculations against the SARS-CoV-2 virus, the causative agent of COVID-19, have been administered around the world starting with 2020, too many remain hesitant about this vaccine. It is believed that hesitancy is driven by misinformation about the COVID-19 vaccines that is spread on social media. Recent research by Loomba et al. [17] has shown that exposure to online misinformation around COVID-19 vaccines affects intent to vaccinate in order to protect oneself and others. However, exposure to misinformation about the COVID-19 vaccines does not mean that those exposed adopt the misinformation. This is why knowing if misinformation is adopted or rejected when encountered in social media discourse will enable public health experts to perform interventions at the right time and in the right place on social media, addressing vaccine hesitancy successfully.

Misinformation detection on social media platforms, such as Twitter, is performed in two steps: (1) the recognition whether a social media posting contains any misconception, reference to conspiracy theories or faulty reasoning; and (2) the recognition of the *stance* towards the targeted misinformation. The stance defines the attitude the author of the micro-blog manifests towards the misinformation target, as exemplified in Table 1. When the misinformation is adopted, an *Accept* stance is observed, whereas when it is rejected, the *Reject* stance reflects the attitude towards the targeted misinformation.

Misinformation Target: *The COVID vaccine renders pregnancies risky.*
Stance: **Accept** *Tweet:* <@USER> Chances of a healthy young woman dying of COVID if they even catch it: 0.003% Chances of COVID vaccine causing miscarriage, birth defects, or future infertility: <Data Unavailable> Risk management would say DON'T TAKE THE VACCINE IF YOU'RE PREGNANT.
Stance: **Reject** *Tweet:* Vaccinated women who breastfeed can pass #COVID19 protection to their babies. COVID-19 #vaccines aren't considered a risk to infants during pregnancy or from breastfeeding. During the study, none of the women or infants experienced serious adverse events. <URL>

Table 1: Examples of tweets with different stance towards misinformation targeting COVID-19 vaccines.

Although the identification of misinformation about COVID-19 vaccines in the Twitter discourse is fundamental in understanding its impact on vaccine hesitancy, we consider that efforts focusing on this first step of misinformation detection have made important progress recently, generating high-quality results [16, 18–20]. In this paper we focus on the second step of misinformation detection, namely the identification of the stance towards misinformation, which still needs improvements.

WWW '22, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9096-5/22/04…$15.00
https://doi.org/10.1145/3485447.3512039

A significant barrier in the identification of stance towards misinformation targeting the COVID-19 vaccines stems from the absence of large Twitter datasets which cover misinformation about these vaccines. To address this limitation, we present in this paper a new Twitter dataset, called CoVaxLies, inspired by the recently released COVIDLies dataset [11]. CoVaxLies consists of (1) multiple known Misinformation Targets (MisTs) towards COVID-19 vaccines; (2) a large set of [tweet, MisT] pairs, indicating when the tweet has the stance of: (a) *Accept* towards the MisT; (b) *Reject* towards the MisT; or (c) *No Stance* towards the MisT. In addition, we provide a taxonomy of the misinformation about COVID-19 vaccines, informed by the MisTs available in CoVaxLies, enabling the interpretation of the adopted or rejected misinformation about COVID-19 vaccines.

As it can be noticed from the examples listed in Table 1, identifying the stance of a tweet with respect to a given MisT is not a trivial language processing task. The framework for stance identification presented in this paper makes several contributions that address the Twitter discourse referring to misinformation. First, it takes into account the *attitude consistency* (AC) observed throughout the Twitter discourse between tweet authors that adopt or reject a MisT. AC is informing the equivalence between stance identification and the recognition of *agree* or *disagree* relations between pairs of tweets. Second, this stance identification framework captures the interactions between discourse AC, the stance values of tweets towards a MisT, and the language used in the articulation of the MisT and the content of the tweets. Third, it considers that the Twitter discourse about a MisT encapsulates knowledge that can be represented by learning knowledge embeddings. This knowledge contributes, along with the neural representation of the content language of tweets, to the prediction of agreement or disagreement between pairs of tweets referring to the same MisT. Finally, the system implementing this novel stance identification framework has produced in our experiments very promising results on the CoVaxLies dataset.

The remainder of the paper is organized as follows. Section 2 describes the related work, while Section 3 details the CoVaxLies dataset. Section 4 describes stance identification informed by attitude consistency (AC). Section 5 presents the experimental results, while Section 6 discusses the results. Section 7 summarizes the conclusions.

2 RELATED WORK

In previous work stance identification on Twitter was cast either as (1) a classification problem, learning to predict the stance value of a tweet towards a given target claim; or (2) an inference problem, when a tweet may entail, contradict or does not imply the claim.

Stance identification as a classification problem: Several datasets were used in prior work aiming towards stance classification on Twitter. The PHEME dataset [36] consists of Twitter conversation threads associated with 9 different newsworthy events such as the Ferguson unrest, the shooting at Charlie Hebdo, or Michael Essien contracting Ebola. A conversation thread consists of a tweet making a true and false claim, and a series of replies. There are 6,425 conversation threads in PHEME, 1,067 were annotated as true, 638 were annotated as false and 697 as unverified. A fraction of the PHEME dataset was used in the RumourEval task [8]. The stance labels are 'support', 'deny', 'comment' and 'query'. There are 865 tweets annotated with the 'support' stance label; 325 tweets annotated with the 'deny' stance label; 341 tweets annotated with the 'query' stance label and 2789 tweets annotated with the 'comment' stance label. Several neural classification architectures for stance identification were designed by participants in RumourEval [1, 15, 29]. However, Ghosh et al. [10] showed that the original pre-trained BERT [9] without any further fine-tuning outperforms all the other models on the RumourEval dataset, including the model that utilizes both text and user information [7].

More recently, another dataset containing stance annotations was released, namely the COVIDLies dataset [11]. The starting point was provided by 86 common misconceptions about COVID-19 available from the Wikipedia page dedicated to COVID-19 misinformation, which became **Misinformation Targets (MisTs)**. For each known MisT, a set of tweets were annotated with three possible stance values: (1) agree, when the tweet adopts the MisT; (2) disagree, when the tweet contradicts/rejects the MisT; and (3) no stance when the tweet is either neutral or is irrelevant to the MisT. 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. Recently, using this dataset, Weinzierl et al. [30] used a neural language processing model that exploits the pre-trained domain-specific language model COVID-Twitter-BERT-v2 [22] and refined it by stacking several layers of lexico-syntactic, semantic, and emotion Graph Attention Networks (GATs) [28] to learn and all the possible interactions between these different linguistic phenomena, before classifying the stance of each tweet.

Stance identification as an inference problem: When the COVIDLies dataset of stance annotations was released in [11], stance identification was presented as a natural language inference problem which can benefit from existing textual inference datasets. In fact, Bidirectional LSTM encoders and Sentence-BERT (SBERT) [24] were trained on three common NLI datasets—SNLI [6], MultiNLI [33], and MedNLI [25]. We were intrigued and inspired by the COVIDLies dataset, and believed that we could create a similar dataset containing misinformation about COVID-19 vaccines, which would not only complement the COVIDLies data, but it would also enable the development of novel techniques for identifying the stance towards misinformation targeting COVID-19 vaccines.

3 STANCE ANNOTATIONS IN COVAXLIES

3.1 CoVaxLies: A Twitter Dataset of Misinformation about COVID-19 Vaccines

The CoVaxLies Twitter dataset, which is publicly available - as detailed in Appendix B, contains misinformation about COVID-19 vaccines represented as (1) several *known* Misinformation Targets (MisTs); (2) a collection of tweets paired with the MisTs they evoke, annotated with stance values, indicating whether the tweet agrees, disagrees or has no stance towards the MisT; and (3) a taxonomy of misinformation about the COVID-19 vaccines, revealing the themes and the concerns addressed by the MisTs from CoVaxLies. We used two information sources for identifying Misinformation Targets (MisTs) for COVID-19 vaccines. First, we have considered

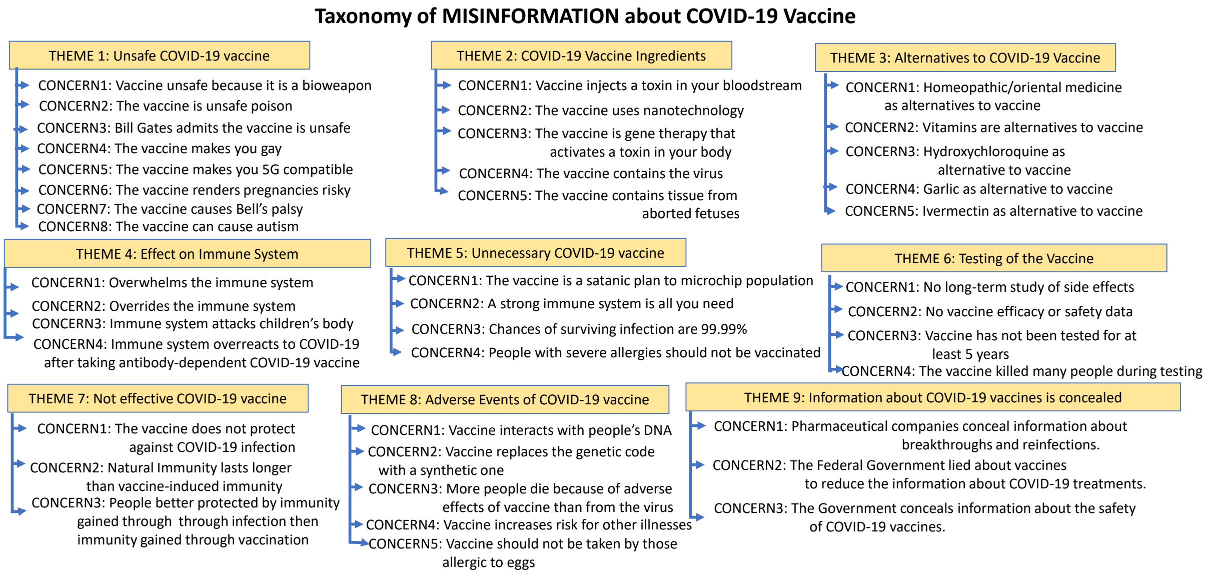

Figure 1: Taxonomy of Misinformation

(a) the Wikipedia page[1] which collects many misconception claims referring to the vaccines developed for immunization against SARS-CoV-2; and (b) MisTs identified by organizations such as the Mayo Clinic, University of Missouri Health Care, University of California (UC) Davis Health, University of Alabama at Birmingham, Science-Based Medicine, PublicHealth.org, Snopes, and the British Broadcast Corporation (BBC), which have been actively collecting misinformation about the COVID-19 vaccines and debunking them on public websites. There are 17 MisTs about COVID-19 vaccines identified in this way in CoVaxLies. Secondly, we have used 19 questions from the Vaccine Confidence Repository [26] to retrieve answers from an index of 5,865,046 unique original tweets obtained from the Twitter streaming API as a result of the query "(covid OR coronavirus) vaccine lang:en". These tweets were authored in the time frame from December 18th, 2019, to January 4th, 2021. Many answers that were retrieved as responding to questions about vaccine confidence contained misinformation, and those answers were considered MisTs as well. In this way we identified an additional set of 37 MisTs, out of which 7 MisTs were already known to us from the first source of information. Therefore, CoVaxLies relies on 47 MisTs about COVID-19 vaccines. Examples of the MisTs are provided in Appendix A. Before using the Twitter streaming API to collect tweets discussing the COVID-19 vaccine, approval from the Institutional Review Board at the University of Texas at Dallas was obtained: IRB-21-515 stipulated that our research met the criteria for exemption.

In order to identify $\mathcal{T}_{\mathcal{R}}$, the collection of tweets which evoke the MisTs from CoVaxLies, we relied on two information retrieval systems: (1) a retrieval system using the BM25 [3] scoring function; and (2) a retrieval system using BERTScore [35] with Domain Adaptation (DA), identical to the one used by Hossain et al. [11]. Both these retrieval systems operated on an index of $C_{\mathcal{T}}$, processing the CoVaxLies MisTs as queries. Researchers from the Human Language Technology Research Institute (HLTRI) at the University of Texas at Dallas judged that 7,346 of the retrieved tweets are relevant to the MisTs from CoVaxLies and organized them in [tweet, MisT] *pairs*. These pairs were also annotated with stance information, creating $\mathcal{T}_{\mathcal{R}}$. In $\mathcal{T}_{\mathcal{R}}$ there are 3,720 tweets which *Accept* their evoked MisT, 2,194 tweets which *Reject* it, and 1,238 tweets that have *No Stance*. We note that CoVaxLies contains an order of magnitude more stance annotations than PHEME [36], the most popular Twitter dataset containing stance annotations, and therefore it presents clear advantages for neural learning methods.

To enable the usage of CoVaxLies in neural learning frameworks, we split $\mathcal{T}_{\mathcal{R}}$ into three distinct collections: (a) a training collection; (b) a development collection; and (c) a test collection. The training collection, which consists of 5,267 [tweet, MisT] pairs, was utilized to train our automatic stance identification systems, described in Section 4. The development collection, which consists of 527 [tweet, MisT] pairs, was used to select system hyperparameters, as discussed in Appendix C. The test collection, which consists of 1,452 [tweet, MisT] pairs, was used to evaluate the stance identification approaches, enabling us to report the results in Section 5.

3.2 The Misinformation Taxonomy from CoVaxLies

Figure 1 illustrates the taxonomy of misinformation available in CoVaxLies. The themes represent the highest level of abstraction, while the concerns differentiate the various MisTs from CoVaxLies. The taxonomy emerged from discussion between public health experts from the University of California, Irvine School of Medicine and computational linguists from HLTRI. Nine misinformation

[1]https://en.wikipedia.org/wiki/COVID-19_misinformation#Vaccines

Figure 2: Distribution of Misinformation Themes and Concerns in the tweets available from CoVaxLies.

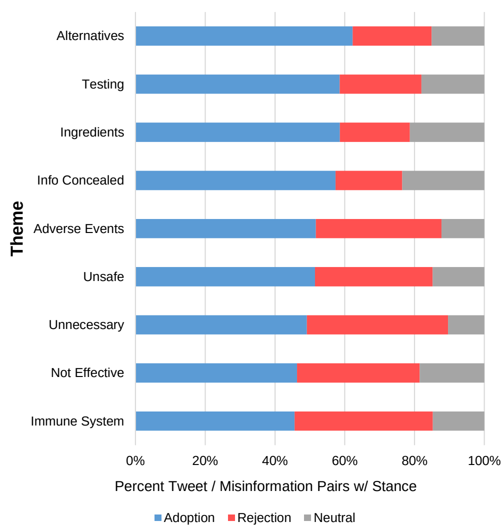

Figure 3: Distribution of Stance Values across Misinformation Themes in CoVaxLies.

themes were revealed, all characterizing aspects that impact confidence in the COVID-19 vaccine. Confidence, along with convenience and complacency, are well known universal factors contributing to vaccine hesitancy, according to the 3C model [21]. For each misinformation theme, as shown in Figure 1, a different number of concerns were revealed: the largest number of concerns pertain to the theme predicating the fact that the COVID-19 vaccines are unsafe (8 concerns) while the smallest number of concerns pertain to the themes claiming that the vaccines are not effective or that information about the vaccines is concealed. Using the information provided by the taxonomy illustrated in Figure 1, we notice in Figure 2 that the misinformation themes that dominate the tweets from CoVaxLies are those about the ingredients of the COVID-19 vaccines, about the adverse events and the fact that the vaccines are unsafe. Moreover, the dominant misinformation regarding the vaccine ingredients claims that the vaccines contain the virus, while the dominant concerns of the lack of safety of the vaccines indicates risky pregnancies or Bell's palsy.

When considering the distribution of tweets that adopt the misinformation, those that reject it and those that are neutral (because of having no stance) for the tweets across all the misinformation themes, we noticed, as illustrated in Figure 3, that the misinformation that is most adopted has the theme of considering alternatives to the COVID-19 vaccines, immediately followed by misinformation regarding the testing of the vaccines and the ingredients used in the vaccines. Interestingly, most of the misinformation that is rejected has to do with the theme indicating that the COVID-19 vaccines are unnecessary, or that they affect the immune system.

4 STANCE IDENTIFICATION THROUGH ATTITUDE CONSISTENCY

4.1 Attitude Consistency and Stance

Central to our stance identification framework is the belief that the stance of any tweet t_j towards a particular MisT m_i should not be considered in isolation. Because t_j participates in the Twitter discourse about m_i, its stance should be consistent with the attitude of the other tweet authors towards m_i. We observed that all the authors of tweets that *Accept* m_i agree among themselves with regard to m_i. Similarly, all the authors of tweets that *Reject* m_i agree among themselves with regard to m_i. But also, any author of a tweet t_j that has an *Accept* stance towards m_i must disagree with the author of any tweet t_k that has a *Reject* stance towards m_i. Therefore, all these tweet authors have Attitude Consistency (AC) towards m_i.

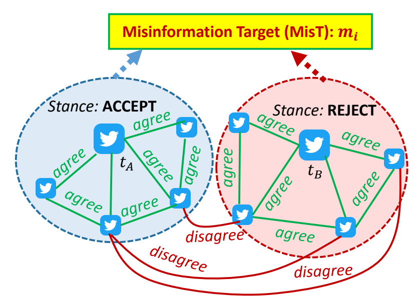

Figure 4: Stance Misinformation Knowledge Graph

AC can be illustrated as in Figure 4, by linking all the tweets that have the *same stance* towards a MisT m_i through implicit *agree* relations, and all tweets that have *opposing stances* towards m_i with implicit *disagree* relations. In this way, all the tweets that have an *Accept* stance towards m_i are organized in a fully connected graph spanned by *agree* relations and similarly, all the tweets having a *Reject* stance towards m_i are organized in a fully connected graph spanned also by *agree* relations. In addition, *disagree* relations are established between all pairs of tweets that have opposing stance towards m_i. Moreover, all tweets that do not have either an *Accept*

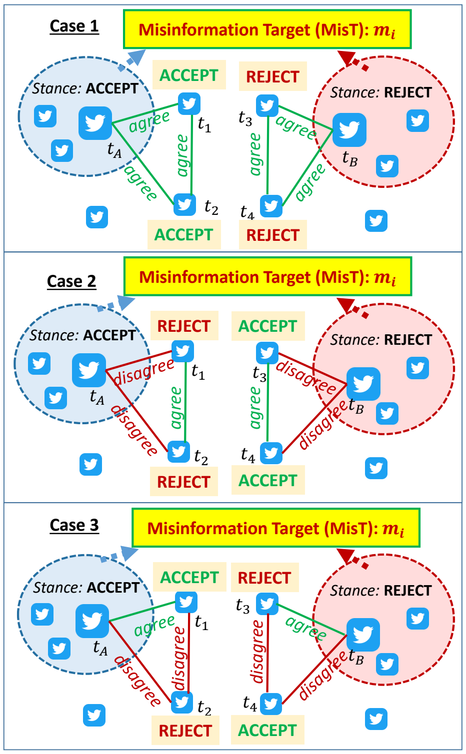

Figure 5: Attitude Consistency Examples

or *Reject* stance towards m_i are considered to automatically have *No Stance* towards m_i. Hence. the stance values $SV = \{Accept, Reject\}$ are the only ones informing AC.

As shown in Figure 4, a Stance Misinformation Knowledge Graph is organized for each m_i, referred to as SMKG(m_i). For clarity, the SMKG(m_i) illustrated in Figure 4 shows only several of the *agree* and *disagree* relations. For each MisT m_i available in the CoVaxLies dataset, we generate an SMKG(m_i) when considering only the tweets annotated with *Accept* or *Reject* stance information, available from the training set of CoVaxLies. However, there are many other tweets in CoVaxLies with no known stance towards any of the MisTs available in the dataset. We refer to the entire set of such tweets as the Tweets with Unknown Stance towards Misinformation (TUSM).

To identify the stance of tweets from TUSM we assume that AC is preserved. This entails three possible cases when considering in addition to the SMKG(m_i), tweets from TUSM, e.g. t_1, t_2, t_3 or t_4, as illustrated in Figure 5. All the three cases of AC show that the unknown stance of any tweet $t_x \in$ TUSM can be identified as *Accept* when knowing if (a) an *agree* relation is predicted between t_x and t_A, a tweet known to have an *Accept* stance towards m_i; or (b) a *disagree* relation is predicted between t_x and t_B, a tweet known to have a *Reject* stance towards m_i. Similarly, the unknown stance of any tweet $t_x \in$ TUSM can be identified as *Reject* when knowing if (a) a *disagree* relation is predicted between t_x and t_A, a tweet known to have an *Accept* stance towards m_i; or (b) an *agree* relation is predicted between t_x and t_B, a tweet known to have a *Reject* stance towards m_i. If none of these relations can be predicted, then the stance of t_x is identified as *No Stance* towards m_i. To formalize the interaction between the implicit relation types and the values of the stance towards a MisT m_i identified for a pair of tweets t_x and t_y we considered a function that selects the Relation Type that preserves AC (RTAC), defined as:

$$RTAC(s_x, s_y) = \begin{cases} agree & \text{if } s_x = s_y \\ disagree & \text{if } s_x \neq s_y \end{cases} \quad (1)$$

where the value of the stance of t_x towards m_i is s_x while the value of the stance of t_y is s_y. Moreover, we believe that AC can be further extended to account for an entire chain of *agree* and *disagree* relations spanning tweets with unknown stance towards m_i.

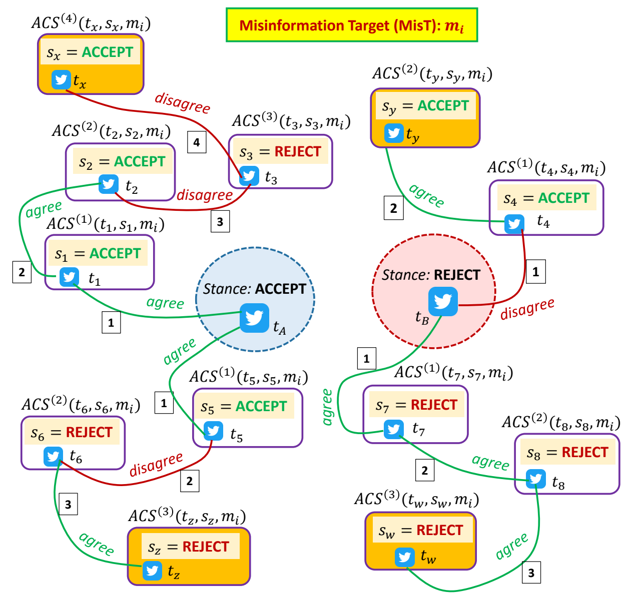

Figure 6: Stance Identification with Transitive Attitude Consistency and Attitude Consistency Scores.

4.2 Transitive Attitude Consistency

Transitive Attitude Consistency extends the interaction between the values of the stance towards a MisT m_i and the binary *agree* and *disagree* relations to an entire chain of such implicit relations that may connect a tweet from TUSM to a tweet from SMKG(m_i), whose stance is known. For example, Figure 6 shows how the identified stance towards m_i of tweets t_x, t_y, t_z and t_w is informed by chains of *agree* or *disagree* relations originating either in t_A or t_B, tweets from SMKG(m_i). It is important to note that this extension has to take into account that every time a new stance s_x towards a MisT m_i is identified for a tweet $t_x \in$ TUSM, the confidence that the AC is preserved is computed by an Attitude Consistency Score (*ACS*). *ACS* depends on l, the number of relations in the chain originating at a tweet with known stance, available from SMKG(m_i) and ending at a tweet $t_x \in$ TUSM, with unknown stance: $ACS^l(t_x, s_x, m_i)$. To compute $ACS^l(t_x, s_x, m_i)$ we first need to consider the way in which we can represent the SMKG(m_i).

The knowledge graph of SMKG(m_i) can be represented in a continuous vector space called the *embedding space* by learning *knowledge embeddings* for its nodes and edges. When formalizing the SMKG(m_i)= $(V; E)$, each node $v_k \in V$ can be seen as $v_k = (t_k, s_k)$, where a tweet t_k is paired with its stance s_k towards m_i; and each edge $e_{ij} \in E$ is either an *agree* or a *disagree* relation. Knowledge embedding models learn an embedding te_k for each tweet t_k as well as an embedding me_i^{agree} for the *agree* relation in SMKG(m_i) and an embedding $me_i^{disagree}$ for the *disagree* relation in SMKG(m_i). But more importantly, knowledge embedding models use a relation scoring function f for assigning a plausibility score to any potential link between two tweets t_x and t_y, given their knowledge embeddings te_x and te_y and the embedding of the relation they share. Because the relation between t_x and t_y must preserve AC between the stance s_x identified for t_x and the stance s_y identified for t_y, the relation between these two tweets is provided by the function $RTAC(s_x, s_y)$. The embedding of the relation indicated by $RTAC(s_x, s_y)$ is computed as:

$$RE(s_x, s_y, m_i) = \begin{cases} me_i^{agree} & \text{if } RTAC(s_x, s_y) = agree \\ me_i^{disagree} & \text{if } RTAC(s_x, s_y) = disagree \end{cases} \tag{2}$$

Hence, the scoring function of the relation between the pair of tweets t_x and t_y is defined as $f(te_x, RE(s_x, s_y, m_i), te_y)$, where f is provided by various knowledge embedding models, such as those that we discuss in Section 4.3,

Given the representation of SMKG(m_i) through knowledge embeddings, we can define $ACS^l(t_x, s_x, m_i)$, starting with the chains of length $l = 1$:

$$ACS^1(t_x, s_x, m_i) = \sum_{(t_y, s_y) \in SMKG(m_i)} \frac{f(te_x, RE(s_x, s_y, m_i), te_y)}{|SMKG(m_i)|} \tag{3}$$

Then, $ACS^l(t_x, s_x, m_i)$ for chains of length $l > 1$ is computed by considering that we have defined already $SV = \{Accept, Reject\}$ and that we shall take into account all tweets from TUSM when generating chains of *agree* and/or *disagree* relations originating in SMKG. We compute $ACS^l(t_x, s_x, m_i)$ as:

$$ACS^l(t_x, s_x, m_i) = \sum_{\substack{t_z \in TUSM \\ t_z \neq t_x}} \sum_{s_z \in SV} \frac{ACS^{l-1}(t_z, s_z, m_i) + f(te_x, RE(s_x, s_z, m_i), te_z)}{|TUSM| - 1} \tag{4}$$

To consider the overall ACS^* of any tweet t_x with stance s_x towards m_i we average the ACS across all possible chains of relations, of varying lengths, up to a maximum length L:

$$ACS^*(t_x, s_x, m_i) = \frac{1}{L} \sum_{l=1}^{L} ACS^l(t_x, s_x, m_i) \tag{5}$$

Finally, stance s_x towards m_i of a tweet $t_x \in$ TUSM is assigned the value corresponding to the maximum ACS^*:

$$s_x = \operatorname*{argmax}_{s_k \in SV} ACS^*(t_x, s_k, m_i) \tag{6}$$

However, Equation 5 shows how we assign stance of value *Accept* or *Reject* to tweets with previously unknown stance towards a MisT m_i. To also assign the stance value *No Stance*, we relied on the development set from CoVaxLies to assign a threshold value $T(m_i)$ for each MisT m_i, such that when $ACS^*(t_x, s_x, m_i) \leq T(m_i)$, for stance values *Accept* and *Reject*, we can finalize the stance s_x of a tweet t_x as having the value *No Stance*. With all stance values finalized for tweets from TUSM towards any MisT m_i from CoVaxLies, we update SMKG(m_i) to contain all the tweets from TUSM that have either an *Accept* or a *Reject* stance towards m_i.

4.3 Learning Knowledge Embeddings for the Stance Misinformation Knowledge Graph

Knowledge embedding models such as TransE [5] and TransD [12] have had significant success in modeling relations in knowledge graphs. More recently, new knowledge embeddings models capture more complex interactions from the knowledge graph, e.g. TransMS [34], TuckER [2], and RotatE [27]. Each knowledge embedding model provides a different method of scoring the likelihood of relations in the knowledge graph SMKG(m_i), as shown in Table 2. The scoring of a relation in each knowledge embedding model relies on me_i^r, the embedding of a relation that maintains AC with the stance towards a MisT m_i of the tweets connected by the relation, and on the embeddings of these tweets, te_x and te_y.

KE Model	Scoring Function $f(te_x, me_i^r, te_y)$
TransE [5]	$-\|\|te_x + me_i^r - te_y\|\|$
TransD [12]	$-\left\Vert (\boldsymbol{I} + me_i^{r,p} \times (te_x^p)^\top) \times te_x + me_i^r - (\boldsymbol{I} + me_i^{r,p} \times (te_y^p)^\top) \times te_y \right\Vert$
TransMS [34]	$-\left\Vert -tanh(te_y \odot me_i^r) \odot te_x + me_i^r + \alpha_i^r \cdot (te_x \odot te_y) - tanh(te_x \odot me_i^r) \odot te_y \right\Vert$
TuckER [2]	$\mathcal{W} \times_1 te_x \times_2 me_i^r \times_3 te_y$
RotatE [27]	$-\|\|te_x \odot me_i^r - te_y\|\|$

Table 2: Knowledge Embedding Scoring Functions.

In Table 2, we denote $||\cdot||$ as the $L1$ norm, $\boldsymbol{I}$ is the identity matrix, $tanh(x)$ is the non-linear hyperbolic tangent function and α_i^r is a real numbered parameter dependent on each MisT. The operator $\odot$ represents the Hadamard product, and $\times_n$ indicates the tensor product along the n-th mode. Any of the scoring functions listed in Table 2 measure the likelihood of an *agree* or *disagree* relation between a pair of tweets which preserves the AC with the stance of the tweets. However, the content of the tweets, communicated in natural language, with the subtleties and deep connections expressed in language, also need to be captured when scoring these relations.

4.4 Interactions between Tweet Language, Stance towards Misinformation and Attitude Consistency

The AC of various tweet authors is expressed through the language they use in their tweets. Therefore, it is imperative to also consider the interaction of the language of tweets with the stance towards misinformation and the attitude consistency of the tweet author's discourse. Because the identification of stance towards misinformation is equivalent to discovering the type of relation, either *agree* or *disagree*, shared by a pair of tweets that preserves AC, we designed a neural language architecture which considers (1) the contextual embeddings of each MisT m_i, as well as each pair of tweets t_x and t_y having a stance towards m_i; and (2) knowledge embeddings learned for the SMKG(m_i) such that we predict the likelihood of a relation between t_x and t_y to be of type *agree* or to be of type *disagree*. This neural architecture for Language-informed Attitude Consistency-preserving Relation scoring (LACRscore) is illustrated in Figure 7.

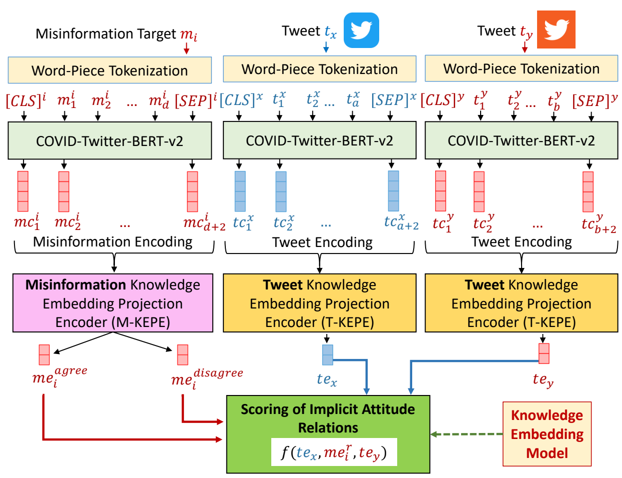

Figure 7: Neural Architecture for Language-informed Attitude Consistency-preserving Relation scoring (LACRscore).

Given a MisT m_i, the LACRscore system first performs Word-Piece Tokenization [9] on (a) the textual description of m_i, producing tokens $m_1^i, m_2^i, ..., m_d^i$, as well as on the text of tweets t_x and t_y, which are then passed through the BERT [9] COVID-19 Language Model COVID-Twitter-BERT-v2 [22] pre-trained on the masked language modeling task [9] for 97 million COVID-19 tweets. This process of further pre-training has been shown to improve performance on downstream tasks in various scientific [4], biomedical [14], and social media [23] domains. COVID-Twitter-BERT-v2 produces contextualized embeddings $mc_1^i, mc_2^i, ..., mc_{d+2}^i$ for the word-piece tokens in the MisT m_i along with the $[CLS]^i$ and $[SEP]^i$ tokens. In this way, we encode the language describing the MisT m_i using a contextualized embedding $mc_1^i \in \mathbb{R}^{1024}$, where 1024 is the contextual embedding size for COVID-Twitter-BERT-V2. Similarly, the language used in the tweets t_x and t_y is represented by contextual embeddings tc_1^x and tc_1^y after being processed through COVID-Twitter-BERT-v2. But, it is important to note, that the scoring function f from any of the knowledge embedding models provided in Table 2, cannot operate directly on the contextual embeddings tc_1^x, mc_1^i or tc_1^y, as they do not have the same dimensions as the knowledge embeddings these models learn. Additionally, we need to produce two knowledge embeddings for the MisT m_i to represent both the *agree* and *disagree* relation embeddings. Therefore, in LACRscore we needed to consider two forms of Knowledge Embedding Projection Encoders (KEPEs) [32], capable of projecting from the contextualized embedding space into the knowledge embedding space. For this purpose, we have relied on the Misinformation Knowledge Embedding Projection Encoder (M-KEPE), using two separate fully-connected layers, to project from mc_1^i into the necessary knowledge embeddings me_i^{agree} and $me_i^{disagree}$ from any of the knowledge embedding models considered. Similarly, the Tweet Knowledge Embedding Projection Encoder (T-KEPE) uses a different fully-connected layer than M-KEPE to project from tc_1^x and tc_1^y to te_x and te_y respectively. As shown in Figure 7, these encoders produce the arguments of the scoring function f, provided by some knowledge embedding model. The likelihood of an *agree* or *disagree* relation between tweets t_x and t_y with respect to the MisT m_i is computed by $f(te_x, me_i^{agree}, te_y)$ and $f(te_x, me_i^{disagree}, te_y)$.

LACRscore was trained on the SMKG(m_i) derived from the training collection of CoVaxLies, described in Section 3.1. Relations from each SMKG(m_i) were used as positive examples, and we performed negative sampling to construct "Attitude Inconsistent" examples. Negative sampling consists of corrupting a relation r between tweets t_x with stance s_x and t_y with stance s_y towards MisT m_j, which preserves AC. This corruption process is performed by randomly sampling either (1): a different tweet $(t_z, s_z) \in SMKG(m_i)$ with the same relation $\hat{r} = r$, to replace t_y such that $RTAC(s_z, s_x) \neq r$, or (2): flipping r from an *agree* relation to $\hat{r}$ =*disagree* relation, or vice versa. The negative sampling will ensure that AC relations will be scored higher than non-AC relations. Moreover, we optimized the following margin loss to train LACRscore when scoring relations:

$$\mathcal{L} = \sum \left[\gamma - f(te_x, me_i^r, te_y) + f(te_x, me_i^{\hat{r}}, te_z) \right]_+ \quad (7)$$

where γ is a training score threshold which represents the differences between the score of AC relations and the non-AC relations. The loss $\mathcal{L}$ is minimized with the ADAM[13] optimizer, a variant of gradient descent. The development collection was used to select all system hyperparameters, which are provided in Appendix C.

5 EXPERIMENTAL RESULTS

To evaluate the quality of stance identification on the test collection from CoVaxLies we use the Precision (P), Recall (R), and F_1 metrics for detecting the *Accept* and *Reject* values of stance. We also compute a Macro averaged Precision, Recall, and F_1 score. The evaluation results are listed in Table 3. The bolded numbers represent the best results obtained. When evaluating the LACRscore system, we have considered (1) five possible knowledge embedding models (TransE; TransD; TuckER; RotatE; and TransMS), which provide different relation scoring functions; and (2) two possible options of stance prediction: (a) using the Attitude Consistency Scoring (ACS) approach described in Section 4.2; and (b) ignoring ACS by and constraining $L = 1$ for any chain of relations, thus ignoring the transitive property of AC.

In addition, we have evaluated several baselines. First, we considered the system introduced by Hossain et al. [11], listed as the Natural Language Inference between Tweet text and MisT text (NLI-Tweet-MisT) system. As a baseline, we have also considered

System	Accept F1	Accept P	Accept R	Reject F1	Reject P	Reject R	Macro F1	Macro P	Macro R
NLI-Tweet-MisT [11]	45.9	72.9	33.5	54.6	38.6	**93.2**	50.2	55.8	63.3
DS-StanceId [30]	86.2	88.3	84.2	79.1	82.7	75.8	82.7	85.5	80.0
LES-GAT-StanceId [30]	86.7	84.6	88.9	80.7	**83.2**	78.3	83.7	83.9	83.6
LACRscore									
+ TransE	69.4	65.6	73.7	47.7	52.3	43.9	58.6	59.0	58.8
+ TransE + ACS	60.1	64.0	56.7	50.5	44.7	58.1	55.3	54.4	57.4
+ TransD	54.9	59.4	51.0	46.6	40.3	55.2	50.7	49.9	53.1
+ TransD + ACS	51.6	56.7	47.4	41.5	35.3	50.5	46.6	46.0	48.9
+ TuckER	87.7	86.7	88.7	82.3	79.3	85.5	85.0	82.0	87.1
+ TuckER + ACS	86.1	85.6	86.6	80.9	73.5	89.8	83.5	79.6	88.2
+ RotatE	86.6	83.6	89.9	80.9	73.5	89.8	83.7	78.5	89.9
+ RotatE + ACS	86.6	85.7	87.5	83.0	80.5	85.8	84.8	83.1	86.6
+ TransMS	85.7	81.8	**90.0**	78.4	69.3	90.3	82.1	75.6	**90.1**
+ TransMS + ACS	**88.7**	**89.8**	87.6	**85.6**	**83.2**	88.2	**87.1**	**86.5**	87.9

Table 3: Results from the stance identification experiments on the CoVaxLies test collection.

the Domain-Specific Stance Identification (DS-StanceId) [30] system, which utilizes the "[CLS]" embedding from COVID-Twitter-BERT-v2 to directly perform stance classification. In addition, we considered the Lexical, Emotion, and Semantic Graph Attention Network for Stance Identification (LES-GAT-StanceId) [30] system which relies on Lexical, Emotion, and Semantic Graph Attention Networks.

The NLI-Tweet-MisT system produced a Macro F_1 score of 50.2, indicating that stance identification as inference over language is not sufficient. Far superior results were obtained by the DS-StanceId system with a Macro F_1 score of 82.7, showcasing the advantage of fine-tuning stance identification systems. The LES-GAT-StanceId system produced a Macro F_1 score of 83.7, which indicates that integrating Lexical, Emotional, and Semantic Graphs further improves stance identification. The LACRscore system with the TuckER configuration produced a Macro F_1 score of 85.0, indicating that identifying the stance towards misinformation through AC presents performance advantages over previous methods. Unsurprisingly. the LACRscore system with the TransMS + ACS configuration performed best, producing a Macro F_1 score of 87.1, which indicates that the transitive nature of AC should not be ignored. The results also show that detecting misinformation rejection tends to be more difficult than the identification of misinformation adoption.

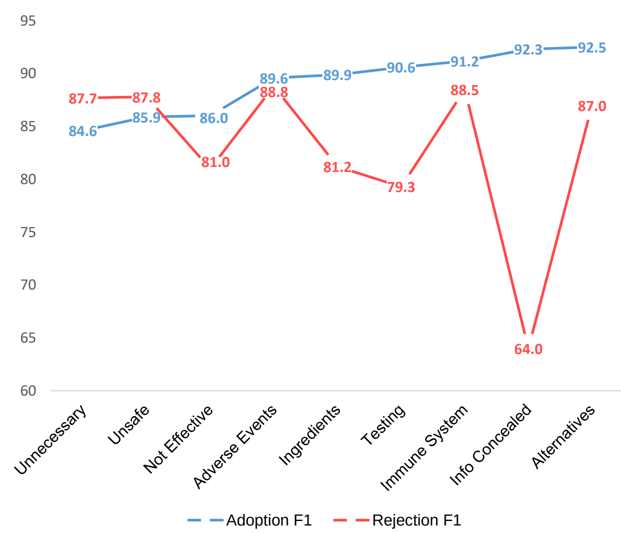

Figure 8: F_1-scores of the misinformation adoption vs. rejection discovered by the LACRscore system with the TransMS and ACS configuration across misinformation Themes from the CoVaxLies dataset.

6 DISCUSSION

Because the LACRscore system produced the best results with the TransMS and ACS configuration, we performed an analysis of the F_1 scores of this system across each of the themes available in the CoVaxLies Misinformation Hierarchy, considering both the *adoption* and *rejection* of misinformation, as illustrated in Figure 8. The identification of adopted misinformation has remarkable performance, across all themes. Moreover, misinformation rejection is identified quite well too, except for the theme of concealing information about vaccines. This is explained by the observation that this theme is addressed by few tweets in CoVaxLies, as illustrated in Figure 2, and moreover, it has the smallest percentage of *rejection* stance values, as illustrated in Figure 3.

7 CONCLUSION

In this paper we present a new method for identifying the stance towards misinformation informed by attitude consistency (AC), which accounts for very promising results on CoVaxLies, a new Twitter dataset of misinformation targeting the COVID-19 vaccines. AC proves to be a stronger signal for stance identification than lexical, emotional and semantic knowledge alone. Moreover, AC informs the knowledge encapsulated in the misinformation discourse on Twitter, which explains the promising results produced by this method, both for the adoption and rejection of misinformation about COVID-19 vaccines.

REFERENCES

[1] Isabelle Augenstein, Andreas Vlachos, and Kalina Bontcheva. 2016. USFD at SemEval-2016 Task 6: Any-Target Stance Detection on Twitter with Autoencoders. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, 389–393.

[2] Ivana Balazevic, Carl Allen, and Timothy Hospedales. 2019. TuckER: Tensor Factorization for Knowledge Graph Completion. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 5185–5194. https://doi.org/10.18653/v1/D19-1522

[3] M. M. Beaulieu, M. Gatford, Xuedong Huang, Stephen Robertson, S. Walker, and P. Williams. 1997. Okapi at TREC-5. In *The Fifth Text REtrieval Conference (TREC-5)* (the fifth text retrieval conference (trec–5) ed.). Gaithersburg, MD: NIST, 143–165. https://www.microsoft.com/en-us/research/publication/okapi-at-trec-5/

[4] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3615–3620.

[5] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems*, Vol. 26. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper.pdf

[6] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 632–642.

[7] Marco Del Tredici, Diego Marcheggiani, Sabine Schulte im Walde, and Raquel Fernández. 2019. You Shall Know a User by the Company It Keeps: Dynamic Representations for Social Media Users in NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 4707–4717.

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

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*. Association for Computational Linguistics, 4171–4186.

[10] Shalmoli Ghosh, Prajwal Singhania, Siddharth Singh, Koustav Rudra, and Saptarshi Ghosh. 2019. Stance Detection in Web and Social Media: A Comparative Study. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Springer International Publishing, Cham, 75–87.

[11] Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 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*. Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2020.nlpcovid19-2.11

[12] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge Graph Embedding via Dynamic Mapping Matrix. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, 687–696. https://doi.org/10.3115/v1/P15-1067

[13] Diederik P. Kingma and Jimmy Ba. 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*. http://arxiv.org/abs/1412.6980

[14] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (09 2019), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682 arXiv:https://academic.oup.com/bioinformatics/article-pdf/36/4/1234/32527770/btz682.pdf

[15] Can Liu, Wen Li, Bradford Demarest, Yue Chen, Sara Couture, Daniel Dakota, Nikita Haduong, Noah Kaufman, Andrew Lamont, Manan Pancholi, Kenneth Steimel, and Sandra Kübler. 2016. IUCL at SemEval-2016 Task 6: An Ensemble Model for Stance Detection in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, 394–400.

[16] Yang Liu and Yi-Fang Wu. 2018. Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (Apr. 2018). https://ojs.aaai.org/index.php/AAAI/article/view/11268

[17] Sahil Loomba, Alexandre de Figueiredo, Simon J. Piatek, Kristen de Graaf, and Heidi J. Larson. 2021. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour* 5, 3 (01 Mar 2021), 337–348.

[18] Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 708–717.

[19] Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor Detection on Twitter with Tree-structured Recursive Neural Networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 1980–1989.

[20] Jing Ma, Wei Gao, and Kam-Fai Wong. 2019. Detect Rumors on Twitter by Promoting Information Campaigns with Generative Adversarial Learning. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 3049–3055.

[21] Noni E. MacDonald. 2015. Vaccine hesitancy: Definition, scope and determinants. *Vaccine* 33, 34 (2015), 4161–4164. https://doi.org/10.1016/j.vaccine.2015.04.036 WHO Recommendations Regarding Vaccine Hesitancy.

[22] Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter. *https://arxiv.org/abs/2005.07503* (2020).

[23] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 9–14.

[24] Nils Reimers and Iryna Gurevych. 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)*. Association for Computational Linguistics, Hong Kong, China, 3982–3992.

[25] Alexey Romanov and Chaitanya Shivade. 2018. Lessons from Natural Language Inference in the Clinical Domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 1586–1596. https://doi.org/10.18653/v1/D18-1187

[26] Isabel Rossen, Mark J. Hurlstone, Patrick D. Dunlop, and Carmen Lawrence. 2019. Accepters, fence sitters, or rejecters: Moral profiles of vaccination attitudes. *Social Science & Medicine* 224 (2019), 23–27. https://doi.org/10.1016/j.socscimed.2019.01.038

[27] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. https://openreview.net/forum?id=HkgEQnRqYQ

[28] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=rJXMpikCZ

[29] Wan Wei, Xiao Zhang, Xuqin Liu, Wei Chen, and Tengjiao Wang. 2016. pkudblab at SemEval-2016 Task 6 : A Specific Convolutional Neural Network System for Effective Stance Detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, 384–388.

[30] Maxwell Weinzierl, Suellen Hopfer, and Sanda M. Harabagiu. 2021. Misinformation Adoption or Rejection in the Era of COVID-19. *Proceedings of the International AAAI Conference on Web and Social Media* 15, 1 (May 2021), 787–795. https://ojs.aaai.org/index.php/ICWSM/article/view/18103

[31] Maxwell A. Weinzierl and Sanda M. Harabagiu. 2021. Automatic detection of COVID-19 vaccine misinformation with graph link prediction. *Journal of Biomedical Informatics* 124 (2021), 103955. https://doi.org/10.1016/j.jbi.2021.103955

[32] Maxwell A Weinzierl, Ramon Maldonado, and Sanda M Harabagiu. 2020. The impact of learning Unified Medical Language System knowledge embeddings in relation extraction from biomedical texts. *Journal of the American Medical Informatics Association* 27, 10 (10 2020), 1556–1567. https://doi.org/10.1093/jamia/ocaa205 arXiv:https://academic.oup.com/jamia/article-pdf/27/10/1556/34153877/ocaa205.pdf

[33] Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 1112–1122. https://doi.org/10.18653/v1/N18-1101

[34] Shihui Yang, Jidong Tian, Honglun Zhang, Junchi Yan, Hao He, and Yaohui Jin. 2019. TransMS: Knowledge Graph Embedding for Complex Relations by Multidirectional Semantics. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 1935–1942.
[35] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.
[36] A. Zubiaga, Geraldine Wong Sak Hoi, Maria Liakata, R. Procter, and P. Tolmie. 2016. Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads. *PLoS ONE* 11 (2016).

A MISINFORMATION TARGETS IN COVAXLIES

Misinformation Targets (MisTs), which represent common misconceptions about the COVID-19 vaccines or refer to conspiracy theories associated with these vaccines, have two different sources. In Table 4, all examples marked with ⋄ correspond to some of the MisTs identified as known misinformation from Wikipedia and other trusted sources, while all examples marked with □ correspond to some of the answers to questions about vaccine confidence, originating from Rossen et al. [26].

⋄ RNA alters a person's DNA when taking the COVID-19 vaccine.
⋄ The COVID-19 vaccine causes infertility or miscarriages in women.
⋄ The COVID-19 vaccine causes Bell's palsy.
⋄ The COVID-19 vaccine contains tissue from aborted fetuses.
⋄ The COVID-19 vaccine can cause autism.
⋄ Hydroxychloroquine protects against COVID-19.
⋄The COVID-19 Vaccine is a satanic plan to microchip people
□ There are severe side effects of the COVID-19 vaccines, worse than having the virus.
□ The COVID-19 vaccine is not safe because it was rapidly developed and tested.
□ The COVID-19 vaccine can increase risk for other illnesses.
□ Vaccines contain unsafe toxins such as formaldehyde, mercury or aluminum.
□ Governments hide COVID-19 vaccine safety information
□ The COVID-19 Vaccine will make you gay.

Table 4: Examples of COVID-19 Misinformation Targets

B CODE AND COVAXLIES DATA AVAILABILITY

The CoVaxLies dataset, comprising the Misinformation Targets (MisTs), the misinformation taxonomy, and the $[tweet_i, MisT_j]$ pairs, which associate a $tweet_i$ with its evoked $MisT_j$ along with stance annotations. The CoVaxLies dataset is publicly available at the following GitHub repository[2].

Code needed to reproduce the experiments described in this paper is also publicly available at the following GitHub repository[3].

We note that an early version of CoVaxLies was presented in Weinzierl and Harabagiu [31], but in that version of CoVaxLies only 17 Misinformation Targets (MisTs) were available, namely the MisTs discovered from Wikipedia and other trusted sources, which are available in this later version as well. Moreover, the previous version of CoVaxLies did not contain any *stance* annotations, and it did not contain the misinformation taxonomy which were made available in the current version.

[2] https://github.com/Supermaxman/vaccine-lies/tree/master/covid19
[3] https://github.com/Supermaxman/covid19-vaccine-nlp

C SYSTEM HYPERPARAMETER SELECTION

System hyperparameters were selected by maximizing the F_1-score of each system on the development set. The LACRscore system was trained with the following hyperparameters: a linearly decayed learning rate of $1e-4$ which was warmed up over the first 10% of the 36 total epochs, an attention drop-out rate of 10%, a batch size of 32, and the tweet and MisT knowledge embedding size was set to 8 for all knowledge embedding models, as we found that to perform best on the development set.

The LACRscore system utilized the training set for learning to score AC-preserving relations by optimizing the margin loss, described in Equation 7. The LACRscore system with the ACS configuration utilized a maximum chain length L of 32, the length value performing best on the development set.

The γ hyperparameter is set to 4.0 for all knowledge graph embedding models, and we sampled 1 negative corrupted relation for each AC relation in the SMKG(m_i).

Threshold values $T(m_i)$ were also automatically selected by maximizing the F_1 score of the LACRscore system on each MisT m_i on the development set.