Claim Extraction and Dynamic Stance Detection in COVID-19 Tweets

Noushin Salek Faramarzi[*]
Computer Science
Stony Brook University
Stony Brook, New York, USA
nsalekfarama@cs.stonybrook.edu

Fateme Hashemi Chaleshtori[*]
Computer Science
Colorado State University
Fort Collins, Colorado, USA
fatemeh@colostate.edu

Hossein Shirazi
Management Information Systems
San Diego State University
San Diego, California, USA
hshirazi@sdsu.edu

Indrakshi Ray
Computer Science
Colorado State University
Fort Collins, Colorado, USA
indrakshi.ray@colostate.edu

Ritwik Banerjee
Computer Science
Stony Brook University
Stony Brook, New York, USA
rbanerjee@cs.stonybrook.edu

ABSTRACT

The information ecosystem today is noisy, and rife with messages that contain a mix of objective claims and subjective remarks or reactions. Any automated system that intends to capture the social, cultural, or political zeitgeist, must be able to analyze the claims as well as the remarks. Due to the deluge of such messages on social media, and their tremendous power to shape our perceptions, there has never been a greater need to automate these analyses, which play a pivotal role in fact-checking, opinion mining, understanding opinion trends, and other such downstream tasks of social consequence. In this noisy ecosystem, not all claims are worth checking for veracity. Such a *check-worthy* claim, moreover, must be accurately distilled from subjective remarks surrounding it. Finally, and especially for understanding opinion trends, it is important to understand the stance of the remarks or reactions towards *that* specific claim. To this end, we introduce a COVID-19 Twitter dataset, and present a three-stage process to (i) determine whether a given Tweet is indeed check-worthy, and if so, (ii) which portion of the Tweet ought to be checked for veracity, and finally, (iii) determine the author's stance towards the claim in that Tweet, thus introducing the novel task of topic-agnostic stance detection.

CCS CONCEPTS

• **Information systems → Social networking sites**; **Information systems applications**; **Collaborative and social computing systems and tools**;

KEYWORDS

Stance Detection, Claim Extraction, COVID-19

Table 1: Two Tweets, each containing a check-worthy claim, and additional remarks by the authors expressing their stance vis-à-vis *that* claim: (A) opposition to a political party and its policies; (B) support for the action of a sportsperson.

(A)	{Rahm Emanuel literally said a Biden White House should tell people laid off from retail stores like JC Penny to learn to code.}$_{\text{claim}}$ {He actually said this! Dems dont care!}$_{\text{commentary}}$
(B)	{Italian tennis star Camila Giorgi has been accused of using a fake Covid vaccine certificate}$_{\text{claim}}$ … {Smart girl.}$_{\text{commentary}}$

ACM Reference Format:
Noushin Salek Faramarzi, Fateme Hashemi Chaleshtori, Hossein Shirazi, Indrakshi Ray, and Ritwik Banerjee. 2023. Claim Extraction and Dynamic Stance Detection in COVID-19 Tweets. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3543873.3587643

1 INTRODUCTION

Nearly a month before the World Health Organization declared COVID-19 a pandemic, the agency stated, "we're not just fighting an epidemic; we're fighting an infodemic"[1]. Two major consequences of this have since become apparent: endangering individual life due to misinformed decisions, and a general distrust of media.

The primary approach to thwart misinformation has been fact-checking, where the veracity of a claim is verified against a trustworthy corpus [16, 36, 40]. Media articles, and social media posts in particular, however, often include factual claims as well as commentary in the form of remarks, opinions, and emotive reactions. Accurate identification and extraction of factual claims from such posts should thus be an important component of fact-checking. Furthermore, discriminating facts from commentary within a single post is critically important in understanding the author's stance on a specific issue mentioned in the post, and thus, understanding the

[*] Both authors contributed equally to this work.

[1] www.who.int/director-general/speeches/detail/munich-security-conference

WWW '23 Companion, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9419-2/23/04…$15.00
https://doi.org/10.1145/3543873.3587643

Table 2: From the 80 most frequent content words found in a random sample of 10K tweets in a COVID-19 Twitter corpus [5], we manually select the most relevant terms. Our dataset comprises tweets that contain at least one such term.

booster, china, corona, coronavirus, coronavirusoutbreak, coronaviruspandemic, coronavirusupdates, (covid-19), covid, covid-19, covid__19, covid_19, covid19, covid19https, covid19outbreak, covid19pandemic, covid2019, covidoutbreak, covidpandemic, delta variant, johnson and johnson, lockdown, masks4all, moderna, omicron, omicronvariant, outbreak, pandemic, pfizer, quarantine, sanitizer, selfIsolation, socialdistancing, stayhome, staysafe, vaccine, vaxxed, virus, washyourhands, wfh, workfromhome

zeitgeist surrounding any issue. This paper focuses on three key concerns in this arc, and develops the following pipeline:

Claim existence: Has the author presented an objective (factual) claim in the post? If so, should it be considered *check-worthy*? In the very first step, we identify posts that contain an objective (factual) claim. Further, along the lines of Barrón-Cedeno et al. [6], we determine if the claim is worth fact-checking, and discard from further analyses posts with no check-worthy claims.

Claim extraction: identify which parts of the post correspond to factual claims, and which correspond to the author's commentary (see Table 1).

Dynamic stance detection: identify the author's stance regarding the factual claim. In contrast to prior work on stance detection, we do not have a fixed set of topics. This detection is *dynamic* in the sense that the author's stance is determined with respect to the objective claim presented in that very same post. The topic of this claim is neither fixed nor explicitly labeled, and it may have never been encountered before during training.

We realize this three-part pipeline by developing and releasing a new manually annotated dataset (§2) of 1.3K Tweets related to the COVID-19 infodemic, where our annotations serve to answer the three key research questions of the pipeline (claim existence, claim extraction, and dynamic stance detection). In our experiments for the first two research questions, described in §3 and §4, we demonstrate significant performance gains over popular deep contextualized language models such as BERT [13]. In the third step of our pipeline, we present the novel task of *dynamic* stance detection (§5). We discuss notable related research (§6) before concluding with an outline of possible future developments based on our work presented here.

2 DATASET

Along the lines of research studying the development of pandemic-related discourse [2, 10], we collected tweets based on a list of keywords related to COVID-19 (Table 2) over a long period of 17 months – from May 2020 to September 2021. We then removed duplicates and filtered out tweets with less than five words. In the retained tweets, we replaced URLs and usernames with the "URL" and "USER" tags, respectively. We then added two more preprocessing steps, replacing (i) emojis with their corresponding text representations[2], and (ii) slang and colloquial acronyms with their formal counterparts[3]. To illustrate, this second step converts acronyms like "wml" to "wish me luck" and slangs like "wochit" to "watch it".

Table 3: The three questions answered by annotators, and their distribution over the various labels: Yes/No, or [A]gree, [D]isagree, [N]eutral.

Question	Total	Label		
Does the tweet contain a factual check-worthy claim?	1,348	Yes (*59.4%*)	No (*40.6%*)	
Does the claim span the entire tweet (answer 'no' if the claim spans only part of the tweet)?	801	Yes (*11.9%*)	No (*88.1%*)	
		[A]gree/[D]isagree/[N]eutral		
1-2 *What is the stance of the author with respect to the claim made in the tweet?*	801	[A] (*65.5%*)	[D] (*17.4%*)	[N] (*17.1%*)

2.1 Data Annotation

We annotate this collection of tweets to precisely determine the objective (factual) claim component of each tweet. In the case such a claim exists, the annotators provide one of three labels – *agree*, *disagree*, or *neutral* – for the stance of the tweet with respect to *that very same* claim.

For this work, we developed a web interface, which was designed so that if annotators are unclear about the label for a tweet, they are able to skip that instance. Further, the interface back-end was built to provide those tweets to an annotator which have already been labeled by another. Each tweet is annotated by two individuals working independently. A total of 12 annotators are used in our collection, each proficient in English[4]. Moreover, each annotator is equipped with at least an undergraduate college education. To resolve inter-annotator disagreements, a discussion followed by an external expert adjudicator is used.

The annotation task is formulated by posing three questions, shown in Table 3, along with the distribution over their labels. The first two questions require binary yes/no labels, while the third question requires one of three labels: *agree*, *disagree*, *neutral.*

2.2 Datasets utilized

In this work, we not only use the annotated collection described above, but also an augmented version of it. Additionally, we utilize two datasets of tweets related to COVID-19, released by Alam et al. [2] and Glandt et al. [15]. Here, we describe them briefly and introduce the notations we use to refer to them.

Our **annotated dataset (DS1)** consists of more than 1.3K annotated tweets labeled with the span of the claim in the tweet along with the author's stance towards that claim, gleaned from the remaining portion with the commentary (for example, see Table 1).

Our **augmented dataset (DS1-Aug)** is created by artificially enhancing the size and diversity of the data of our annotated data by augmenting it with back-translation [34]. In this work, we translate

[2] This was done using the `demoji` library, available at pypi.org/project/demoji.

[3] The Text Slang dictinoary, available at www.noslang.com/dictionary

[4] We use the term *proficient user* to reflect the highest level, C2, as described by the Common European Framework of Reference for Languages (CEFR) [12].

Table 4: Dataset statistics about the *check-worthy* (cw) and *not check-worthy* (ncw) classes. Beyond the datasets already described in §2, we use the following to fine-tune and test various models for Task 1: DS2-Eng (only the English-language tweets of DS2), DS1-DS2 (the union of DS1-Aug and DS2), DS1-Test (the test set of our annotated corpus), and DS2-Test (the English-language tweets from the test set of DS-2).

Dataset	cw		ncw		Total
	#	%	#	%	#
DS1	605	60.10	402	40	1007
DS1-Aug	3612	60.10	2398	40	6010
DS2-Eng	387	12.70	2661	87.3	3048
DS2	6876	65.05	3695	35	10571
DS1-DS2	10488	63.25	6093	36.80	16581
DS1-Test	196	57.50	145	42.50	341
DS2-Test	306	36.70	528	63.30	834

an English tweet to a second language, translate that to a third language, and then translate it back to English. We obtain all these translations using the Google Translate API[5]. The intermediate languages are chosen uniformly at random from among the 100-plus languages available from this API. This back-translation method is applied five times to each tweet, thereby generating an augmented corpus five times the size of DS1.

The **COVID-19 infodemic (DS2)** corpus, provided by Alam et al. [2], comprises more than 16K tweets (in Arabic, Bulgarian, Dutch, and English) related to the COVID-19 pandemic. In this dataset, each tweet is labeled to denote whether or not it contains a claim that may cause harm. Their work determines harm by asking seven questions to their annotators, the first being (**Q1**) "*Does the tweet contain a verifiable factual claim?*". Since this question has a yes/no answer, we opt to employ this data to train our models for the claim existence task in our pipeline. We are able to hydrate[6] more than 14K such tweets, and use the Google Translate API to translate the non-English tweets to English.

The **COVID-19-Stance Dataset (DS3)**, provided by Glandt et al. [15], consists of more than 7K tweets. This collection is designed to detect the stance of the tweet's author about one of the four topics: *Stay at Home Orders*, *Keeping Schools Closed*, *Wearing a Face Mask*, and *Anthony S. Fauci, M.D.*. Even though there are no fixed target topics in our work, we utilize this dataset for the dynamic stance detection task in our pipeline.

3 TASK 1: CLAIM EXISTENCE

The first component in our pipeline is a module capable of determining whether a given tweet makes any objective (factual) claim. We consider these to be *check-worthy* (cw). As shown by Zuo et al. [42], nearly half of the tweets (43.4%) are not factual and, thus, are *not check-worthy* (ncw). Table 4 shows the distribution over cw and ncw tweets in the various datasets we use in our experiments on checking for the existence of claims. We describe these next.

Table 5: Experimental results on Task 1: the [P]recision and F1 scores upon fine-tuning three pretrained language models (BERT, RoBERTa, XLNet) using 5 datasets, and testing on DS1-Test and DS2-Test.

Dataset		BERT		RoBERTa		XLNET	
Train	Test	P	F1	P	F1	P	F1
DS2-Eng	DS2-Test	78.3	78.37	80.0	79.91	81.8	81.78
DS2	DS2-Test	79.6	79.52	81.6	81.43	82.0	81.84
DS1	DS1-Test	74.16	69.24	76.43	71.62	72.88	68.35
DS1-Aug	DS1-Test	75.20	73.53	77.37	74.39	74.8	72.77
DS2-Eng	DS1-Test	67.28	66.74	66.05	66.11	69.77	69.78
DS2	DS1-Test	70.17	70.20	70.34	70.18	73.13	73.08
DS1-DS2	DS1-Test	74.63	74.58	75.68	75.24	76.75	76.59

3.1 Experiments

We conduct two sets of experiments. In the first, we exclusively use the data provided by Alam et al. [2], and in the second, we include our own annotated corpus.

3.1.1 Experiments on the COVID-19 Infomedic corpus DS2. We fine-tune various pretrained language models on the training sets of DS2 and DS2-Eng, and test them on a fixed set of English Tweets published by Alam et al. [2]. This test set (**DS2-Test**) consists of 834 tweets, which we were able to re-hydrate. Since we are able to re-hydrate more than 92% of the tweets of the original English-language test set, our experiments using hyperparameters identical to Alam et al. [2] are comparable to their original findings. These results are shown in Table 5. The results reported by Alam et al. [2] show that the best weighted F_1 score for Q1, when training on DS2-Eng, is achieved using the RoBERTa [24] model (78.6%). For the same setup, we obtain an F_1 score of 79.91%. When training on DS2 (*i.e.*, including the translated non-English tweets), we observe an improvement to 81.43% for RoBERTa. Additionally, we report the performance of XLNet [41] here, which is comparable to RoBERTa (F_1 score of 81.84%). A more detailed table of these reproduction experiments is presented in Appendix 8, Table 10.

3.1.2 Experiments incorporating our annotated corpus DS1. Next, we add our annotated corpus to the COVID-19 infodemic collection. The main objective of this series of experiments is to identify models capable of achieving high precision as well as high recall scores on our dataset in Task 1. In other words, we identify models that discard the tweets that do not contain an objective (factual) claim, while retaining the tweets that do. Thus, we test these models only on DS1-Test and focus on the F_1 score for the ncw class.

The results of this second series of experiments are shown in the lower half of Table 5, from which we can glean that training on the augmented collection, DS1-Aug, offers significant improvement across all models. A more detailed analysis shown in Table 6 discloses that while the improvements in cw are modest, training on the augmented corpus provides a significant fillip to the F_1 score for ncw: RoBERTa shows the lowest improvement (5.16%, from

[5] translate.google.com

[6] Due to restrictions on sharing Twitter data, datasets can only share tweet IDs. Given a collection of tweet IDs, *hydration* refers to the process of obtaining their details, including the actual content [39].

Table 6: Class-wise performance measures of training BERT, RoBERTa, and XLNet on various datasets and testing on DS1-Test.

Dataset	Model	BERT			RoBERTa			XLNet		
	Class	Precision	Recall	F_1-Score	Precision	Recall	F_1-Score	Precision	Recall	F_1-Score
DS1	ncw	81.58	42.76	56.11	84.81	46.21	59.82	79.22	42.07	54.95
	cw	68.68	92.86	78.96	70.23	93.88	80.35	68.18	91.84	78.26
	Macro Avg	75.10	67.81	67.53	77.52	70.04	70.09	73.70	66.95	66.61
	Micro Avg	74.16	71.55	69.24	76.43	73.61	71.62	72.88	70.67	68.35
DS1-Aug	ncw	78.43	55.17	64.78	83.70	53.10	64.98	78.57	53.10	63.37
	cw	72.80	88.78	80.00	72.69	92.35	81.35	72.02	89.29	79.73
	Macro Avg	75.60	71.97	72.39	78.19	72.73	73.16	75.3	71.19	71.55
	Micro Avg	75.20	74.49	73.53	77.37	75.66	74.39	74.80	73.90	72.77
DS2-Eng	ncw	59.51	66.90	62.99	61.03	57.24	59.07	64.58	64.14	64.36
	cw	73.03	66.33	69.52	69.76	72.96	71.32	73.60	73.98	73.79
	Macro Avg	66.30	66.61	66.25	65.39	65.10	65.20	69.10	69.06	69.08
	Micro Avg	67.28	66.57	66.74	66.05	66.28	66.11	69.77	69.79	69.78
DS2	ncw	66.42	61.38	63.80	64.05	67.59	65.77	70.77	63.45	66.91
	cw	72.95	77.04	74.94	75.00	71.94	73.44	74.88	80.61	77.64
	Macro Avg	69.70	69.21	69.37	69.53	69.76	69.60	72.83	72.03	72.28
	Micro Avg	70.17	70.38	70.20	70.34	70.09	70.18	73.13	73.31	73.08
DS1-DS2	ncw	72.52	65.52	68.84	75.83	62.76	68.68	75.78	66.90	71.06
	cw	76.19	81.63	78.82	75.57	85.20	80.10	77.46	84.18	80.68
	Macro Avg	74.40	73.57	73.83	75.70	73.98	74.39	76.62	75.54	75.87
	Micro Avg	74.63	74.78	74.58	75.68	75.66	75.24	76.75	76.83	76.59

59.82% to 64.98%) while BERT [13] improves by 8.67% (from 56.11% to 64.78%). This improvement is perhaps not a surprise, since prior work on various natural language understanding tasks has shown improvements upon training with augmented data [23, 33].

We also observe that the performance of the models trained on DS2 and DS2-Eng correlates with the performances of the same models trained on DS1 and DS1-Aug, providing a hearty indication that the COVID-19 infodemic corpus is highly relevant to our task. Training only on DS2 or DS2-Eng shows little or no improvement over the same model trained on DS1-Aug, but we do see the best F_1 scores being attained upon training on DS1-DS2 (the union of DS1-Aug and DS2). The highest improvement can be seen in XLNet, where the F_1 score jumps by nearly 4% (from 72.77% to 76.59%). Thus, as expected, training on additional corpora designed for a similar task yields significant gains.

In Table 7, we share the hyperparameters for the second series of experiments described above.

4 TASK 2: CLAIM EXTRACTION

We treat the extraction of objective (factual) claims in a tweet as a sequence labeling task. The tweet text represents the entire text sequence, and each token is labeled as either being part of the claim or not. We use the IOB2 schema [38] with the following tags:

- *B-Claim* indicates that the token represents the beginning of a claim,
- *I-Claim* indicates that the token is a part of a claim (this tag is only used when the preceding label is B-Claim), and
- *O* indicates that the token is outside the scope of the claim.

If t = $\{w_1, ..., w_n\}$ is a sequence of tokens in a given tweet t, where w_i represents the i^{th} token in t, the task is to assign one of three labels y = {B-Claim, I-Claim, O} to each $w_i \in t$. Figure 1 illustrates the IOB2 tagging schema on our annotated collection, with three categorically distinct examples.

4.1 Experiments

In our baseline model, we fine-tune pretrained BERT embeddings with an added linear layer and the softmax activation function to obtain the class labels for the tokens. During this, the BERT parameters as well as the added layer weights are trained. Our implementation is done using PyTorch[7] and the Transformers library[8].

The BERT architecture has a 0.4 dropout probability for all fully connected layers. The additional linear layer takes BERT's output (*i.e.*, a vector of size 768) as its input, and provides a 3-dimensional output. We use cross-entropy as the loss function here, and train for 5 epochs with a batch size of 32. We use a maximum sequence length of 128, and set the learning rate to $\eta = 5 \times 10^{-5}$.

We compare the baseline results to Flair [1], which is an off-the-shelf framework for training neural networks for natural language processing tasks. It has demonstrated impressive results for many sequence labeling problems such as the named entity recognition task introduced in the CoNLL-2003 [37]. The model is pretrained on a billion words of text, to learn the parameters of a multi-layer long short-term memory (LSTM) network. Given the pretrained network, the tokens are passed as an input sequence, and values from the deepest hidden layer at each index are returned. These values form the contextual embeddings. Beyond the contextual embeddings of Flair, we also use GloVe [30], a classical pretrained

[7] pytorch.org

[8] huggingface.co/transformers

Table 7: Hyperparameters used in Task 1 experiments incorporating our annotated corpus DS1: learning rate (η) and the number of epochs (n_e)

Training dataset	Hyperparameter	BERT	XLNet	RoBERTa
DS1	η	1×10^{-5}	2×10^{-5}	5×10^{-5}
	n_e	10	10	5
DS1-Aug	η	3×10^{-5}	1×10^{-5}	3×10^{-5}
	n_e	3	3	2
DS2-Eng	η	1×10^{-5}	1×10^{-5}	5×10^{-6}
	n_e	10	10	5
DS2	η	3×10^{-5}	1×10^{-5}	1×10^{-5}
	n_e	2	3	3
DS1-DS2	η	1×10^{-5}	2×10^{-5}	2×10^{-5}
	n_e	2	3	2

language model, and the stacked embeddings of GloVe + Flair, GloVe + BERT, and Flair + BERT.

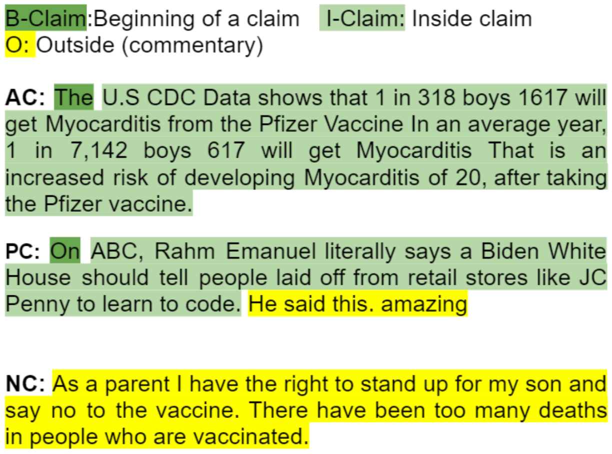
4.1.1 Influential users. Our work pertains to claims made in social media posts, and the users' stance toward such claims. Given that influential users with a large number of followers are able to disseminate claims as well as opinions to a wider audience [9], it is fitting for us to analyze the performance of our models on such users. This is all the more important because the cascading nature of information (or misinformation) propagation tends to center around such users [4].

In this leg of our work, from our entire corpus, we identify influential Twitter accounts that are verified. These include health organizations, political personalities, well-known news sources, and users with more than 10K followers. We then examine the performance of our claim extraction models on the entire test set, as well as a specifically selected small test set of 100 tweets from these influential accounts.

4.1.2 Evaluation measures. Claim extraction being a sequence labeling task, it is worth noting that the standard evaluation techniques are not immediately applicable. If the ground truth annotation of a claim is not an exact token-to-token match of the model's prediction, it is not obvious whether that should be treated as a complete mistake. To illustrate, consider the following mismatch between the ground truth label of a claim and a model's prediction:

> {Coronavirus cases are on the rise nationwide and we have to everything we can to stop the spread.}$_{\text{ground truth}}$ Practice social distancing. Wash your hands. Wear a mask. It saves lives.
>
> {Coronavirus cases are on the rise nationwide and we have to everything we can to stop the spread. Practice social distancing.}$_{\text{predicted}}$ Wash your hands. Wear a mask. It saves lives.

It is our view that such partial extractions should only be partially penalized. Otherwise, not only do the numeric scores appear more severe, but requiring a strict matching between the predicted sequence and the ground truth sequence may lead to overcorrection over different but reasonable extractions. Indeed, this is the view taken by a large body of prior work on sequence labeling and information extraction tasks (see, for example, [20, 26]).



B-Claim:Beginning of a claim I-Claim: Inside claim
O: Outside (commentary)

AC: The U.S CDC Data shows that 1 in 318 boys 1617 will get Myocarditis from the Pfizer Vaccine In an average year, 1 in 7,142 boys 617 will get Myocarditis That is an increased risk of developing Myocarditis of 20, after taking the Pfizer vaccine.

PC: On ABC, Rahm Emanuel literally says a Biden White House should tell people laid off from retail stores like JC Penny to learn to code. He said this. amazing

NC: As a parent I have the right to stand up for my son and say no to the vaccine. There have been too many deaths in people who are vaccinated.

Figure 1: A tweet may entirely or partially consist of an objective claim, or it may not have any objective claim at all. These three types of tweets are shown as all-claim (AC), partial-claim (PC), and no-claim (NC), along with the IOB2 tagging: B-Claim (dark green), I-Claim (light green), and O (yellow).

We thus develop a relaxed evaluation, where the incorrect inclusion of additional tokens or the incorrect exclusion of the claim's tokens are somewhat tolerable. Such a relaxation toward more tolerant measures has seen extensive use in earlier sequence labeling tasks [7, 20, 43], since they offer a more meaningful evaluation of such models. Since our models need to identify token sequences in tweets of varying lengths, we calculate the weighted average of scores based on the tweet lengths. Accordingly, our measure infuses both the impact of the tweet length and the partial penalization in the evaluation. For each tweet, we compute the weighted precision and F_1, and divide by the total number of tokens in the corpus. We call this the *dataset-wise* evaluation of precision and F_1:

$$P_{dw} = \frac{\sum_{i=1}^{T} P_i * |t_i|}{\sum_{i=1}^{T} |t_i|} \qquad F_{1(dw)} = \frac{\sum_{i=1}^{T} F_{1_i} * |t_i|}{\sum_{i=1}^{T} |t_i|},$$

We also compute these measures without considering the length of each tweet, and call it the *tweet-wise* evaluation of precision and F_1

$$P_{tw} = \frac{\sum_{i=1}^{T} P_i}{T} \qquad F_{1(tw)} = \frac{\sum_{i=1}^{T} F_{1_i}}{T}$$

where T and t_i denote the total number of tweets in the test set and the tokens in the i^{th} tweet, respectively, and P_i denotes the precision calculated for the tokens in the i^{th} tweet.

4.2 A discussion of the experimental results

Our experimental results are shown in Table 8. We display them in two sections, with the second showing the performance of various models on the test set of tweets from influential Twitter accounts.

Overall, the stacked embeddings of Flair + BERT show significantly better performance in terms of both precision and F_1 score,

Table 8: Claim extraction results on the test set of tweets from regular Twitter users, and on the test set of tweets from influential Twitter users with verified accounts.

	tweet-wise		*dataset-wise*	
Embedding	**P**	**F1**	**P**	**F1**
		Regular		
BERT (baseline)	0.55	0.55	0.62	0.65
GloVe	0.56	0.55	0.63	0.66
Flair	0.57	0.57	0.65	0.68
Stacked embeddings				
Flair + BERT	**0.72**	**0.67**	**0.74**	**0.70**
		Highly Influential		
BERT (baseline)	0.64	0.74	0.63	0.73
GloVe	0.71	0.76	0.75	0.75
Flair	0.80	0.80	0.79	0.76
Stacked embeddings				
GloVe + Flair	0.82	0.81	0.80	0.76
GloVe + BERT	0.83	**0.82**	0.81	**0.80**
Flair + BERT	**0.84**	0.81	**0.82**	**0.80**

whether or not we account for the length of individual tweets. Based on this success, we run evaluate two other stacked embeddings – GloVe + Flair, and GloVe + BERT – on the test set of influential tweets. All three perform significantly better than the standalone language models, and Flair + BERT continues to have the best overall performance.

It is worth noting, however, that the standalone models perform significantly better on the influential tweets when compared to the regular tweets. A qualitative comparison shows that the influential users tend to use a more formal writing style. This includes, among other differences, proper punctuation, grammatically correct sentences, and correctly spelled words. Regular users often post without paying much attention to these aspects of language, as we can see in the following example:

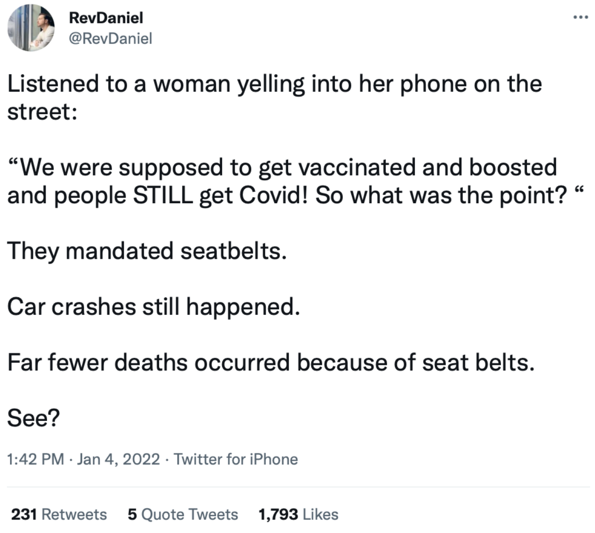
> smh When are stupid media going to get that a pandemic didnt happen . . .

Furthermore, we also notice that influential tweets generally have a grammatically clear distinction between the claim and their additional commentary. For instance, the claim and the commentary may be two separate sentences. Regular users, on the other hand, often intertwine the two (and often, with some ambiguity). The complete tweet in the above example illustrates this phenomenon:

> smh When are stupid media going to get that a pandemic didnt happen, {*it was predicted to sweep across the world killing tens of millions and infecting between 2 and 6 billion pple by now*}$_{\text{claim}}$, well it peaked at .1 of pop. before acute phase ended in May, its now a cold."

5 TASK 3: DYNAMIC STANCE DETECTION

Among many definitions of stance detection (for a detailed discussion, we point the reader to Küçük and Can [21]), we adopt the one proposed by Mohammad et al. [27] in the SemEval task on detecting stance in tweets. They define stance detection as a classification task, where the goal is to determine the position of the author of a given text towards a specific target. The position is represented by one of the following category labels: "Favor", "Against" or "Neither". Many aspects of stance detection have since been explored. For instance, Ng and Carley [29] explored the whether stance classification models can be generalized across datasets. In all such prior work, however, the target topics were always fixed *a prior*. We, on the other hand, explore the possibility of detecting the stance without any fixed set of targets, and indeed, without even the explicit notion of targets. Rather, the target is obtained "on the fly", from each individual instance. In our work, this dynamic target is the objective (factual) claim extracted from that same tweet.



RevDaniel
@RevDaniel

Listened to a woman yelling into her phone on the street:

"We were supposed to get vaccinated and boosted and people STILL get Covid! So what was the point? "

They mandated seatbelts.

Car crashes still happened.

Far fewer deaths occurred because of seat belts.

See?

1:42 PM · Jan 4, 2022 · Twitter for iPhone

231 Retweets **5** Quote Tweets **1,793** Likes

Figure 2: An author's disagreement toward the primary argument presented in the same tweet.

5.1 Experiments & Results

Dynamic stance detection is an important task. Its most direct application is perhaps in understanding the social and cultural zeitgeist surrounding opinion-laden issues, and thus, in understanding the propagation of misinformation. In many cases, we find that users post incorrect statements or arguments, but counter it in an effort to illustrate the fallacy. Figure 2 shows such an example, where the additional commentary clearly shows that the author is not misinformed. Without discriminating between this additional remark and the rest of the tweet, however, automated methods may mistakenly label such posts as misinformation.

Even though stance detection is sometimes considered as a type of sentiment analysis (because the aim is to identify the stance toward the target), it is worth noting that dynamic stance detection is distinct from sentiment analysis, since a tweet may express opposition to a claim using various figurative tools such as sarcasm, humor, or irony. It is also common to find social media posts where the opinion expressed in the post is not literally directed at a target, but the opposition (or support) can be deduced implicitly. As we

Table 9: Results on dynamic stance detection, showing the (P)recision and F_1 scores. Results denoted by * are on the binary classification experiments. The best results on the DS1 test set for the 3-class and binary classification experiments are shown in bold.

Dataset		BERT		RoBERTa		XLNET	
Train	Test	P	F1	P	F1	P	F1
DS3	DS3-Test	0.71	0.71	0.71	0.71	0.73	0.72
DS1	DS3-Test	0.26	0.27	0.27	0.26	0.24	0.24
DS3+DS1	DS3-Test	0.68	0.68	0.70	0.70	0.71	0.71
DS3	DS1-Test	0.39	0.33	0.40	0.32	0.45	0.34
DS1	DS1-Test	**0.63**	0.31	0.59	0.40	0.63	**0.46**
DS1	DS1-Test*	**0.66**	**0.60**	0.62	0.55	0.65	0.51
DS3+DS1	DS1-Test	0.52	0.41	0.59	0.42	0.49	0.36
DS3+DS1	DS1-Test*	0.61	0.57	0.56	0.54	0.56	0.56

can see from Figure 2, this often takes a dialectic and didactic form, and requires a deep understanding of natural language pragmatics. Without the development of such pragmatic analyses, however, dynamic stance detection remains distinct from sentiment analysis.

In a manner similar to the first two tasks, we opt for three different state-of-the-art pretrained language models (BERT, RoBERTa, and XLNet). In particular, our approach is much like the first task on identifying the existence of a claim, since both are essentially classification tasks. Thus, here too, we use the same transformer models to examine the impact of domain-specific datasets. The architecture incorporates a 0.4 dropout probability for all fully connected layers. We add a single linear layer, which converts the model's output to a 3-dimensional output. We use cross-entropy as the loss function and train for 5 epochs with a batch size of 32. A maximum sequence length of 128 is used, and the learning rate is set to $\eta = 5 \times 10^{-5}$.

We utilize not only our annotated corpus (**DS1**) but also the dataset (**DS3**) provided by Glandt et al. [15] for this task. The latter comprises tweets labeled with the stance expressed in them with regard to topics relevant to the pandemic. DS3 differs from DS1 in that it contains four predetermined target topics. Table 9 shows the performance of these three models on two different test sets: DS1-Test, the test partition of our annotated collection, and DS3-Test, the test partition of the collection obtained from DS3 (with four fixed targets).

Our results are not directly comparable to those obtained by Glandt et al. [15], since we combine all the targets together into a single group, in order to partially simulate dynamic stance detection. Thus, even though their work reports a variant of BERT achieving $F_1 > 0.8$ for one of the four targets, our experiments find XLNet to be the best performer, with $F_1 = 0.72$, when testing on DS3-Test.

Next, we combine the two collections to see if the use of additional datasets designed for traditional stance detection lead to improvements. However, we actually observe a decline in the results. This is probably due to the extreme target-imbalance created due to the inclusion of DS3 in the training of these models. In essence, training on a small number of fixed targets makes the models worse for detecting stance where the target appears dynamically, often never seen during training.

Finally, we also conduct experiments on binary stance detection. Here, we remove the neutral stance and keep the positive and negative stances as the two classes. BERT achieves the best results in terms of both precision and F_1 measure. As expected, the binary classification results are a moderate improvement. The key bottleneck, however, remains the *zero-shot* style of stance detection in this third task.

6 RELATED WORK

Since 2019, several Twitter datasets pertaining to the COVID-19 pandemic have been released by researchers all over the world. These collections have seen manifold use in natural language understanding and social network analysis tasks. A corpus of over 8 million tweets was provided by Dimitrov et al. [14] for a range of knowledge discovery tasks. They use an established RDF schema to provide data-modeling vocabularies, where the data includes metadata about the tweets, as well as about the extracted entities, hashtags, user mentions, sentiments, and URLs. Further, they describe protoypical use cases of this corpus, such as trend anlysis, citation discovery models, prediction of a tweet's virality, and most relevant to our work – stance detection. Larger multilingual datasets have also been developed, such as the collection of over 123 million tweets to track misinformation and rumor (Chen et al. [11]). For a more fine-grained analysis of fake news (political as well as medical) surrounding the pandemic, Alam et al. [2] developed a manually annotated dataset of 16K tweets. Their dataset has been incorporated in the development of the claim detection models in this work (§3). While our task is clearly different, we underscore that our annotated dataset, too, offers the distinction of having explicitly labeled claim spans within each individual tweet.

Beyond the provision of large datasets, a sizeable body of work exists on the identification of claims. This may be due, at least in part, to how claims are woven into argumentation (fallacious or otherwise). Indeed, the central component of an argument is its claims [8]. In spite of this central importance of claims in natural language processing tasks, there is no universally accepted procedure or substantial agreement – not even among human readers – defining what constitutes a claim [22]. Nevertheless, ClaimBuster – a claim detection system proposed by Hassan et al. [18], based on a large corpus of annotated debates – has gained widespread attention. Their study offers an end-to-end pipeline for fact-checking, including a claim spotter that scores sentences for check-worthy factual claims based on a scoring model trained on the token and part-of-speech features, a claim matcher that uses token-based and semantic similarity to retrieve fact-checked claims from a curated database, and a claim checker that collects evidence using external resources and APIs. Their work, however, scores entire sentences instead of precisely extracting claims from within a sentence.

A precise extraction of claims from within a sentence calls for sequence labeling, an important NLP task recently discussed by He et al. [19], among others. Their survey reviews state-of-the-art deep learning techniques applied to three fundamental sequence labeling tasks: named entity recognition (NER), part-of-speech (POS) tagging, and text chunking. Many deep learning systems use bidirectional long short-term memory networks (BiLSTM) to encode the input and a CRF layer for the final prediction (Ma and Hovy

[25], among others). More recent work extended this approach to use context-dependent embeddings such as ELMo, BERT, and Flair [31, 35]. In this work, our sequence labeling task for claim extraction was designed along these lines (see §4.1).

Even though our work is distinct from traditional stance detection, we find it prudent to offer a brief discussion of prior work in this area. As a research topic, detecting an author's stance predates the COVID-19 pandemic. Mohammad et al. [28], for instance, present a dataset of tweet-target pairs annotated for stance as well as sentiment. This was one of the first studies to demonstrate that while the knowledge of sentiment benefits stance detection, sentiment alone is not sufficient. Their use of support vector machines to classify the stance toward five fixed targets was further improved with distant supervision and the incorporation of word embeddings.

Since 2019, a majority of improvements in stance detection has used targets related to COVID-19. In this work, we have utilized the dataset presented by one such work, by Glandt et al. [15], who provide annotations on over 7K tweets about the author's stance toward one of four fixed targets. Their labels were obtained by posing three questions to the annotators, including one question explicitly about the sentiment expressed in the tweet. In their work, it was demonstrated that significant improvements can be achieved by the use of (i) domain-specific pretrained language models, and (ii) domain adaptation incorporate previous stance detection datasets.

While prior research has not formulated dynamic stance detection as in this work, there are two specific stance detection studies in the spirit of *low-shot* learning, that look into a similar task. One study on *cross-topic* stance detection learns to identify the stance towards one target (say, the Pfizer vaccine) from texts that only mention other targets (say, Anthony Fauci) [3]. Thus, while the targets are not seen "on the fly", as in our work, this work demonstrates that the stance toward a target can sometimes be learned without explicit mention of that same target. Along a different line, Hardalov et al. [17] survey stance detection and its relation to fact-checking and fake news detection. They define the target in terms of the textual context toward which the stance is expressed, and show that stance can be used as a component of fact-checking. Perhaps the most pertinent work in this direction is the very first fake news challenge, organized by Pomerleau and Rao [32]. In this challenge, instead of true/fake labels, the task was to determine the stance of a news article's headline vis-à-vis its body. This challenge can be viewed as a one-shot stance learning task, and in that spirit, is similar to our third task. The latter, it is worth noting, may be regarded as zero-shot stance learning because there are no fixed targets, and the model encounters entirely targets it never saw during training or development.

7 CONCLUSION AND FUTURE WORK

We present a new Twitter dataset related to COVID-19, annotated for claim extraction from within tweets and also for dynamic stance detection with no targets fixed *a priori*. Given that the current information ecosystem is filled with messages that mix objective claims with subjective commentary and remarks, it is both difficult and imperative that we strive toward the development of models that can accurately distinguish between the two. This distinction is critically important for fact-checking, as argued by Hardalov et al. [17], we hope that our work has brought the two fields of research on misinformation detection and stance detection closer.

This work has demonstrated that accurate claim extraction is possible by modeling it as a sequence labeling problem. We have also provided the first step toward dynamic stance detection. Our experiments on the incorporation of traditional stance detection datasets demonstrates that there is significant room for improvement in this direction, and we hope that this first study on dynamic stance detection paves the way for future research. We also note that the limited success in our third task is due at least in part to the size of the dataset. We intend to develop much larger corpora for the tasks explored in this work, and hope to see others developing similar datasets as well.

ACKNOWLEDGMENTS

This work was supported in part by the U.S. National Science Foundation (NSF) under the awards IIS 2027750, CNS 1822118, and SES 1834597, and by NIST, ARL, Statnett, AMI, Cyber Risk Research, NewPush, and State of Colorado Cybersecurity Center.

REFERENCES

[1] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Association for Computational Linguistics, Minneapolis, Minnesota, 54–59. https://doi.org/10.18653/v1/N19-4010

[2] Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, Abdulaziz Al-Homaid, Wajdi Zaghouani, Tommaso Caselli, Gijs Danoe, Friso Stolk, Britt Bruntink, and Preslav Nakov. 2021. Fighting the COVID-19 Infodemic: Modeling the Perspective of Journalists, Fact-Checkers, Social Media Platforms, Policy Makers, and the Society. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 611–649. https://doi.org/10.18653/v1/2021.findings-emnlp.56

[3] Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance Detection with Bidirectional Conditional Encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 876–885. https://doi.org/10.18653/v1/D16-1084

[4] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. Everyone's an Influencer: Quantifying Influence on Twitter. In *WSDM '11* (Hong Kong, China). Association for Computing Machinery, New York, NY, USA, 65–74. https://doi.org/10.1145/1935826.1935845

[5] Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, and Gerardo Chowell. 2020. A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration. arXiv:2004.03688 https://arxiv.org/abs/2004.03688

[6] Alberto Barrón-Cedeno, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, and Fatima Haouari. 2020. Checkthat! at clef 2020: Enabling the automatic identification and verification of claims in social media. *Advances in Information Retrieval* 12036 (2020), 499.

[7] Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2019. Sub-event detection from twitter streams as a sequence labeling problem. 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, Minneapolis, Minnesota, 745–750. https://doi.org/10.18653/v1/N19-1081

[8] Tuhin Chakrabarty, Christopher Hidey, and Kathy McKeown. 2019. IMHO Fine-Tuning Improves Claim Detection. 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, Minneapolis, Minnesota, 558–563. https://doi.org/10.18653/v1/N19-1054

[9] Arpan Chaudhury, Partha Basuchowdhuri, and Subhashis Majumder. 2012. Spread of Information in a Social Network Using Influential Nodes. In *Advances in Knowledge Discovery and Data Mining*, Pang-Ning Tan, Sanjay Chawla, Chin Kuan

Ho, and James Bailey (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 121–132.

[10] Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health Surveill* 6, 2 (29 May 2020), e19273. https://doi.org/10.2196/19273

[11] Emily Chen, Kristina Lerman, Emilio Ferrara, et al. 2020. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health and Surveillance* 6, 2 (2020), e19273.

[12] Council of Europe 2023. *Global scale - Table 1 (CEFR 3.3): Common Reference levels.* Council of Europe. Retrieved March 13, 2023 from https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale

[13] 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, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[14] Dimitar Dimitrov, Erdal Baran, Pavlos Fafalios, Ran Yu, Xiaofei Zhu, Matthäus Zloch, and Stefan Dietze. 2020. TweetsCOV19 - A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM '20).* Association for Computing Machinery, New York, NY, USA, 2991–2998. https://doi.org/10.1145/3340531.3412765

[15] Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance Detection in COVID-19 Tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).* Association for Computational Linguistics, Online, 1596–1611. https://doi.org/10.18653/v1/2021.acl-long.127

[16] Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A Richly Annotated Corpus for Different Tasks in Automated Fact-Checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL).* Association for Computational Linguistics, Hong Kong, China, 493–503. https://doi.org/10.18653/v1/K19-1046

[17] Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A Survey on Stance Detection for Mis- and Disinformation Identification. In *Findings of the Association for Computational Linguistics: NAACL 2022.* Association for Computational Linguistics, Seattle, United States, 1259–1277. https://doi.org/10.18653/v1/2022.findings-naacl.94

[18] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, et al. 2017. Claimbuster: The first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment* 10, 12 (2017), 1945–1948.

[19] Zhiyong He, Zanbo Wang, Wei Wei, Shanshan Feng, Xianling Mao, and Sheng Jiang. 2020. A Survey on Recent Advances in Sequence Labeling from Deep Learning Models. , 16 pages. https://doi.org/10.48550/ARXIV.2011.06727 arXiv:arXiv:2011.06727

[20] Abhyuday Jagannatha and Hong Yu. 2016. Structured prediction models for RNN based sequence labeling in clinical text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, Austin, Texas, 856–865. https://doi.org/10.18653/v1/D16-1082

[21] Dilek Küçük and Fazli Can. 2020. Stance Detection: A Survey. *ACM Comput. Surv.* 53, 1, Article 12 (feb 2020), 37 pages. https://doi.org/10.1145/3369026

[22] Namhee Kwon, Liang Zhou, Eduard Hovy, and Stuart W. Shulman. 2007. Identifying and Classifying Subjective Claims. In *Proceedings of the 8th Annual International Conference on Digital Government Research: Bridging Disciplines & Domains (dg.o '07).* Digital Government Society of North America, Philadelphia, Pennsylvania, USA, 76—-81.

[23] Hankyol Lee, Youngjae Yu, and Gunhee Kim. 2020. Augmenting Data for Sarcasm Detection with Unlabeled Conversation Context. In *Proceedings of the Second Workshop on Figurative Language Processing.* Association for Computational Linguistics, Online, 12–17. https://doi.org/10.18653/v1/2020.figlang-1.2

[24] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. , 13 pages. https://doi.org/10.48550/arXiv.1907.11692

[25] Xuezhe Ma and Eduard Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Association for Computational Linguistics, Berlin, Germany, 1064–1074. https://doi.org/10.18653/v1/P16-1101

[26] Zulfat Miftahutdinov, Ilseyar Alimova, and Elena Tutubalina. 2019. KFU NLP Team at SMM4H 2019 Tasks: Want to Extract Adverse Drugs Reactions from Tweets? BERT to The Rescue. In *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task.* Association for Computational Linguistics, Florence, Italy, 52–57. https://doi.org/10.18653/v1/W19-3207

[27] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016).* Association for Computational Linguistics, San Diego, California, 31–41. https://doi.org/10.18653/v1/S16-1003

[28] Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)* 17, 3 (2017), 1–23.

[29] Lynnette Hui Xian Ng and Kathleen M Carley. 2022. Is my stance the same as your stance? A cross validation study of stance detection datasets. *Information Processing & Management* 59, 6 (2022), 103070.

[30] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[31] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. 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, 2227–2237. https://doi.org/10.18653/v1/N18-1202

[32] Dean Pomerleau and Delip Rao. 2017. Fake news challenge stage 1 (FNC-I): Stance detection. Retrieved March 15, 2023 from https://www.fakenewschallenge.org/

[33] Sina Mahdipour Saravani, Ritwik Banerjee, and Indrakshi Ray. 2021. An Investigation into the Contribution of Locally Aggregated Descriptors to Figurative Language Identification. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP.* Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 103–109. https://doi.org/10.18653/v1/2021.insights-1.15

[34] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Association for Computational Linguistics, Berlin, Germany, 86–96. https://doi.org/10.18653/v1/P16-1009

[35] Jana Straková, Milan Straka, and Jan Hajic. 2019. Neural Architectures for Nested NER through Linearization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, Florence, Italy, 5326–5331. https://doi.org/10.18653/v1/P19-1527

[36] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. 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, 809–819. https://doi.org/10.18653/v1/N18-1074

[37] Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003.* Association for Computational Linguistics, Edmonton, Canada, 142–147. https://aclanthology.org/W03-0419

[38] Erik F. Tjong Kim Sang and Jorn Veenstra. 1999. Representing Text Chunks. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics.* Association for Computational Linguistics, Bergen, Norway, 173–179. https://aclanthology.org/E99-1023

[39] Twitter, Inc. 2023. *GET statuses/lookup.* Twitter, Inc. Retrieved March 13, 2023 from https://developer.twitter.com/en/docs/twitter-api/v1/tweets/post-and-engage/api-reference/get-statuses-lookup

[40] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Association for Computational Linguistics, Online, 7534–7550. https://doi.org/10.18653/v1/2020.emnlp-main.609

[41] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc., Vancouver, Canada, 5753–5763. https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf

[42] Chaoyuan Zuo, Ritwik Banerjee, Fateme Hashemi Chaleshtori, Hossein Shirazi, and Indrakshi Ray. 2022. Seeing Should Probably Not Be Believing: The Role of Deceptive Support in COVID-19 Misinformation on Twitter. *J. Data and Information Quality* 15, 1, Article 9 (dec 2022), 26 pages. https://doi.org/10.1145/3546914

[43] Chaoyuan Zuo, Kritik Mathur, Dhruv Kela, Noushin Salek Faramarzi, and Ritwik Banerjee. 2022. Beyond belief: a cross-genre study on perception and validation of health information online. *International Journal of Data Science and Analytics* 13, 4 (2022), 299–314.

8 APPENDIX

Table 10: Reproducing the results of Alam et al. [2], using binary data for whether a tweet contains a verifiable factual claim, and testing on DS2-Test.

Dataset	Class	Precision	Recall	F_1
		BERT		
DS2-Eng	NCW	72.28	67.32	69.71
	CW	81.79	85.04	83.38
	Macro Avg	77.03	76.18	76.55
	Micro Avg	78.30	78.54	78.37
DS2	NCW	75.94	66.01	70.63
	CW	81.69	87.88	84.67
	Macro Avg	78.81	76.95	77.65
	Micro Avg	79.60	79.86	79.52
		RoBERTa		
DS2-Eng	NCW	71.97	73.86	72.90
	CW	84.62	83.33	83.97
	Macro Avg	78.29	78.59	78.44
	Micro Avg	80.00	79.86	79.91
DS2	NCW	79.39	67.97	73.24
	CW	82.27	89.77	86.18
	Macro Avg	81.13	78.87	79.71
	Micro Avg	81.60	81.77	81.43
		XLNet		
DS2-Eng	NCW	78.47	70.26	74.14
	CW	83.75	88.83	86.21
	Macro Avg	81.11	79.54	80.18
	Micro Avg	81.80	82.01	**81.78**
DS2	NCW	79.40	69.28	74.00
	CW	83.42	89.58	86.39
	Macro Avg	81.41	79.43	80.19
	Micro Avg	82.00	82.13	81.84