Empowering the Fact-checkers! Automatic Identification of Claim Spans on Twitter

Megha Sundriyal[1*]**, Atharva Kulkarni**[1*]**, Vaibhav Pulastya**[1]**,**
Md Shad Akhtar[1]**, Tanmoy Chakraborty**[2]
[1]IIIT Delhi, India, [2]IIT Delhi, India
{meghas, atharvak, vaibhav17271, shad.akhtar}@iiitd.ac.in,
tanchak@ee.iitd.ac.in

Abstract

The widespread diffusion of medical and political claims in the wake of COVID-19 has led to a voluminous rise in misinformation and fake news. The current vogue is to employ manual fact-checkers to efficiently classify and verify such data to combat this avalanche of claim-ridden misinformation. However, the rate of information dissemination is such that it vastly outpaces the fact-checkers' strength. Therefore, to aid manual fact-checkers in eliminating the superfluous content, it becomes imperative to automatically identify and extract the snippets of claim-worthy (*mis*)*information* present in a post. In this work, we introduce the novel task of *Claim Span Identification* (*CSI*). We propose **CURT**, a large-scale Twitter corpus with token-level claim spans on more than $7.5k$ tweets. Furthermore, along with the standard token classification baselines, we benchmark our dataset with **DABERTa**, an adapter-based variation of RoBERTa. The experimental results attest that DABERTa outperforms the baseline systems across several evaluation metrics, improving by about 1.5 points. We also report detailed error analysis to validate the model's performance along with the ablation studies. Lastly, we release our comprehensive span annotation guidelines for public use.

1 Introduction

The swift acceleration of Online Social Media (OSM) platforms has led to tremendous democratized content creation and information exchange. Consequently, these platforms serve as ideal breeding grounds for malicious rumormongers and talebearers, abetting a colossal upsurge of misinformation. Such misinformation manifests in many ways, including bogus claims, fabricated information, and rumors. The massive COVID-19 *'Infodemic'* (Naeem and Bhatti, 2020) is one such malignant byproduct that led to the rampant spread

*Equal contribution

Figure 1: Examples of claim tweets and their ground truth claim spans highlighted in boldface text (blue).

of political and social calumny (Ferrara, 2020; Margolin, 2020; Ziems et al., 2020), accompanied by counterfeit pharmaceutical claims (O'Connor and Murphy, 2020). Therefore, finding such claim-ridden posts on OSM platforms, investigating their plausibility, and differentiating the credible claims from the apocryphal ones has risen to be a pertinent research problem in Argument Mining (AM).

'Claim', as coined by Toulmin (2003), is *'an assertion that deserves our attention'*. It is the key component of any argument (Daxenberger et al., 2017). Consider the second tweet, *'We don't have evidence...'*, as given in Figure 1. For the task of claim identification at the coarse level, the entire tweet will be marked as a claim. However, on closer inspection, we find that the text fragments of *'our wine keeps you from getting #COVID19'* and *'Better alternative to #DisinfectantInjection'* represent the finer argumentative units of claim and form the set of evidence, based on which this tweet is considered a claim. Segregating such argumentative units of misinformed claims from their benign counterparts fosters many benefits. To begin with, it partitions the otherwise independent claims in a single post, enabling us to retrieve a larger number of claims. Secondly, it acts as a precursor to the downstream tasks of claim check-

worthiness and claim verification. Thirdly, it will also bring in the angle of *explainability* in coarse-grained claim identification. Finally, it will serve the manual fact-checkers and hoax-debunkers[1,2] to conveniently strain out the unnecessary shreds of text from further processing. We further elaborate on the necessity of claim span identification and exemplify it in Section 2.

Though the recent literature reflects extensive work done in claim detection (Daxenberger et al., 2017; Chakrabarty et al., 2019; Gupta et al., 2021), limited forays have been made in claim span identification i.e., recognizing the argumentative components of a claim (Wührl and Klinger, 2021). In the recent past, commendable work has been done on span-level argument unit recognition pertaining to other computational counterparts under the umbrella of AM, such as hate speech (Mathew et al., 2021), toxic language (Pavlopoulos et al., 2021) etc. Such study, however, has eluded the realm of claims, owning to the lack of quality annotated datasets. This heralds a specialized corpus creation on claim span identification.

To this end, we propose **CURT** (**C**laim **U**nit **R**ecognition in **T**weets), a large-scale, claim span annotated Twitter corpus. We also present several baseline models for solving claim span identification as a token classification task and evaluate them on CURT. Furthermore, we introduce *claim descriptions*, which are generic prompts aimed to assist the model in focusing on the most significant regions of the input text using explicit instructions on what to designate as a '*claim*'. They are elucidated later in detail. Finally, we benchmark our dataset with **DABERTa** (**D**escription **A**ware **R**oBERTa), a plug-and-play adapter-based variant of RoBERTa (Liu et al., 2019), endeavored to infuse the Pre-trained Language Model (PLM) with the description information. Empirical results attest that DABERTa outperforms the conventional baselines and generic PLMs for our task consistently across various metrics.

Contributions. Through this work, we make the following tangible contributions:

1. **Formulation of a novel problem statement**: We propose the novel task of *Claim Span Identification* that aims to identify argument units of claims in the given text.

2. **Claim span identification dataset and extensive annotation guidelines**: We posit a large-scale Twitter dataset, the first of its kind, with $7.5k$ claim span annotated tweets, to placate the absence of the annotated dataset for claim span identification. Additionally, we develop comprehensive annotation guidelines for the same.

3. **Claim span identification system**: We propose a robust claim span identification framework based on *Compositional De-Attention* (*CoDA*) and *Interactive Gating Mechanism* (*IGM*).

4. **Extensive evaluation and analysis**: We evaluate our model against different baselines to confirm sizable improvements over them. We also report thorough qualitative and quantitative analysis along with the ablation studies.

Reproducibility. We release our dataset and code for DABERTa at https://github.com/LCS2-IIITD/DABERTA-EMNLP-2022. We present detailed dataset annotation guidelines in the Appendix (A.1).

2 Why Claim Span Identification?

As stated in Section 1, we hypothesize that claim span identification would aid fact-checkers to quickly segregate claim-ridden content from the rest of the post. Moreover, we suppose that it will be a propitious precursor for claim verification and fact-checking, facilitating better retrieval of relevant evidences. We back our hypothesis with a small experiment of evidence-based document retrieval. We collect 50 random samples from CURT, along with their corresponding ground-truth claim spans. Further, for both the tweets and the claim spans, we extract top-k relevant articles from a knowledge-base leveraging the traditional retrieval system, BM25 (Robertson et al., 1995). We use the recently released publicly available CORD19 corpus (Wang et al., 2020) to retrieve factual documents. Finally, we present retrieved documents to three evaluators and ask them to mark whether or not the retrieved shreds of evidence are relevant to the given input tweet/span from our dataset. All three annotators label each text-evidence pair independently. Eventually, to obtain the final relevancy score, majority voting is employed. We obtain a high inter-annotator score (*Fleiss Kappa*) of 0.63 and 0.67 for tweets and spans, respectively.

We compare the performance of tweet-based and span-based retrievals in terms of precision (P) and

---

[1]https://www.snopes.com/
[2]https://www.politifact.com/

Input	P@5	P@3	nDCG@5	nDCG@3
Tweets	0.3922	0.2745	0.2733	0.2280
Spans	**0.4407**	**0.3390**	**0.3038**	**0.2521**

Table 1: nDCG@k and P@k scores for tweet and spans using BM25 retrieval system and CORD19 dataset.

normalized Discounted Cumulative Gain (nDCG) scores and report them in Table 1. For comparison, we consider two different top-k settings (k=3 and 5). We begin by examining the retrieval performance using P@k, which measures the fraction of relevant documents extracted in the top-k set. Span-based document retrieval consistently improves precision scores when compared to tweets. For nDCG@5, we discover that span-based retrieval outperforms tweet-based retrieval by more than 3%. When we limit the retrieval depth to 3, we see a similar pattern. This, in turn, demonstrates that entire posts contain much extraneous information, frequently impeding the performance of evidence retrieval systems that are a prerequisite for both automated and manual fact-checking. In summary, we reinforce that our hypothesis positively stands true, as span-based document retrieval results in a better score for precision as well as nDCG. This attests to the task's feasibility and importance in the realm of claims.

3 Related Work

Claims on Social Media. The prevailing research on claims could be cleft into three categories – claim detection (Levy et al., 2014; Chakrabarty et al., 2019; Gupta et al., 2021), claim check-worthiness (Jaradat et al., 2018; Wright and Augenstein, 2020), and claim verification (Zhi et al., 2017; Hanselowski et al., 2018; Soleimani et al., 2020). Bender et al. (2011) pioneered the efforts in claim detection by introducing the *AAWD* corpus. Subsequent studies largely relied on using linguistically motivated features such as sentiment, syntax, context-free grammars, and parse-trees (Rosenthal and McKeown, 2012; Levy et al., 2014; Lippi and Torroni, 2015). Recent works in claim detection have engendered the use of large language models (LMs). Chakrabarty et al. (2019) re-enforced the power of fine-tuning, as their ULMFiT LM, fine-tuned on a large Reddit corpus of about 5M opinionated claims, showed notable improvements in claim detection benchmark. Gupta et al. (2021) proposed a generalized claim detection model for detecting claims independent of its source. They handled structured and unstructured data in con-

junction by training a blend of linguistic encoders (POS and dependency trees) and a contextual encoder (BERT) to exploit the input text's semantics and syntax. As LMs account for significant computational overheads, Sundriyal et al. (2021) addressed this quandary and proposed a lighter framework that attempted to fabricate discernible feature spaces. The *CheckThat! Lab's CLEF*-2020 shared task (Barrón-Cedeno et al., 2020) has garnered the attention of several researchers. Williams et al. (2020) won the task by fine-tuning the RoBERTa (Liu et al., 2019) accentuated by mean pooling and dropout. Nikolov et al. (2020) ranked second with their out-of-the-box RoBERTa vectors supplemented with Twitter meta-data.

Span Identification. Zaidan et al. (2007) introduced the concept of rationales, which highlighted text segments that supported their label's judgment. Trautmann et al. (2020) released *AURC*-8 dataset with token-level span annotations for the argumentative components of stance along with their corresponding label. Mathew et al. (2021) proposed a quality corpus for explainable hate identification with token-level annotations. The *SemEval* community has initiated fine-grained span identification concerning other domains of argument mining such as toxic comments (Pavlopoulos et al., 2021) and propaganda techniques (Da San Martino et al., 2020). These shared tasks amassed many solutions constituting transformers (Chhablani et al., 2021), convolutional neural networks (Coope et al., 2020), data augmentation techniques (Rusert, 2021; Pluciński and Klimczak, 2021), and ensemble frameworks (Zhu et al., 2021a; Nguyen et al., 2021). Wührl and Klinger (2021) resembled the closest study to ours, wherein they compiled a corpus of around $1.2k$ biomedical tweets with claim phrases.

In summary, existing literature on claims concentrates entirely on sentence-level claim identification and does not investigate on eliciting fine-grained claim spans. In this work, we endeavor to move from coarse-grained claim detection to fine-grained claim span identification. We consolidate a large manually annotated Twitter dataset for claim span identification task and benchmark it with various baselines and a dedicated description-based model.

4 Dataset

Over the past few years, several claim detection datasets have been released (Rosenthal and McKeown, 2012; Chakrabarty et al., 2019). However,

Dataset	Train	Test	Validation
Total no. of claims	6044	755	756
Avg. length of tweets	27.40	26.93	27.29
Avg. length of spans	10.90	10.97	10.71
No. of span per tweet	1.25	1.20	1.27
No. of single span tweets	4817	629	593
No. of multiple span tweets	1201	121	161

Table 2: Dataset statistics. All the lengths are in tokens.

none of these corpora come with claim-based rationales that quantify a post as a claim. To bridge this gap, we propose **CURT** (**C**laim **U**nit **R**ecognition in **T**weets), a large scale Twitter corpus with token-level claim span annotations.

Data Selection. We annotate claim detection Twitter dataset released by Gupta et al. (2021) for our task. However, the guidelines they presented have certain reservations, wherein they do not explicitly account for benedictions, proverbs, warnings, advice, predictions, and indirect questions. As a result, tweets such as *'Dear God, Please put an end to the Coronavirus. Amen'* and *'@FLOTUS Melania, do you approve of ingesting bleach and shining a bright light in the rectal area as a quick cure for #COVID19? #BeBest'* have been mislabeled claims. This prompted us to extend the existing guidelines and introduce a more exclusive and nuanced set of definitions based on claim span identification. We present details of the extended annotation guidelines and guideline development procedure in Appendix (A.1). In total, we annotated 7555 tweets from the Twitter corpus by Gupta et al. (2021) which met our guidelines.

Dataset Statistics and Analysis. We segment CURT into three partitions – training set, validation set, and test set, in the split of 80:10:10. Dataset related statistics are given in Table 2. One important point to note here is that while a claim tweet is typically 27 tokens long, a claim span is only around 10 tokens long. This implies that the claim-ridden tweets have a lot of extraneous information. Arguments can also perhaps comprise several claims that may or may not be related to each other. Around 19% of the claim tweets in our dataset contain multiple claim spans. As a result, in total, we obtain 9458 claim spans from 7555 tweets. We observe that the majority of the tweets contain single claims. Out of 7555 tweets, 6039 include a single claim, demonstrating that the majority of tweets contemplate single assertions at a time.

5 Proposed Methodology

In this section, we outline DABERTa and its intricacies. The main aim is to seamlessly coalesce critical domain-specific information into Pre-trained Language Models (PLM). To this end, we introduce *Description Infuser Network* (*DescNet*), a plug-and-play adapter module that conditions the LM representations with respect to the handcrafted descriptions. The underlying principle behind this theoretical formalization is to link a claim span to a claim description to guide the model on what to focus on explicitly. As shown in Figure 2, DescNet houses two sub-components, namely, *Compositional De-Attention block* (*CoDA*) and *Interactive Gating Mechanism* (*IGM*). The particulars of each component are delineated in the following sections.

Claim Descriptions. Before delving into *CoDA* and *IGM*, we first examine Claim Descriptions, which are the cornerstone of the proposed model. Claim Descriptions are handcrafted templates that guide the model where to concentrate its focus. The inclusion of claim description encourages the model to focus on the most essential phrases in the input tweet, which may be thought of as guided attention that leads to increased performance. We judiciously curated our claim descriptions in accordance with the annotation guidelines for claims and non-claims offered by Gupta et al. (2021). In Table 3 we list some of the claim descriptions along with the claims that they most align with. It is noteworthy that a claim can align with more than one claim descriptions as well.

Overview of PLMs for Token Classification. To begin with the details of the proposed framework, DABERTa, we present the working of PLMs for the token classification task. PLMs such as BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2019), and RoBERTa (Liu et al., 2019) are widely used for various downstream NLP tasks owing to their strong contextual language representation capabilities and fine-tuning ease. As the input to these PLMs, each ith input text is first tokenized into a sequence of sub-word embeddings $X_i \in \mathbb{R}^{N \times d}$, where N is the maximum sequence length and d is the feature dimension. Then a positional embedding vector $PE_{pos} \in \mathbb{R}^{N \times d}$ is added to the token embeddings in a pointwise fashion to retain the positional information (Vaswani et al., 2017).

The vector $Z_i \in \mathbb{R}^{N \times d}$, hence obtained, is fed to a stack of transformer encoder blocks. Each
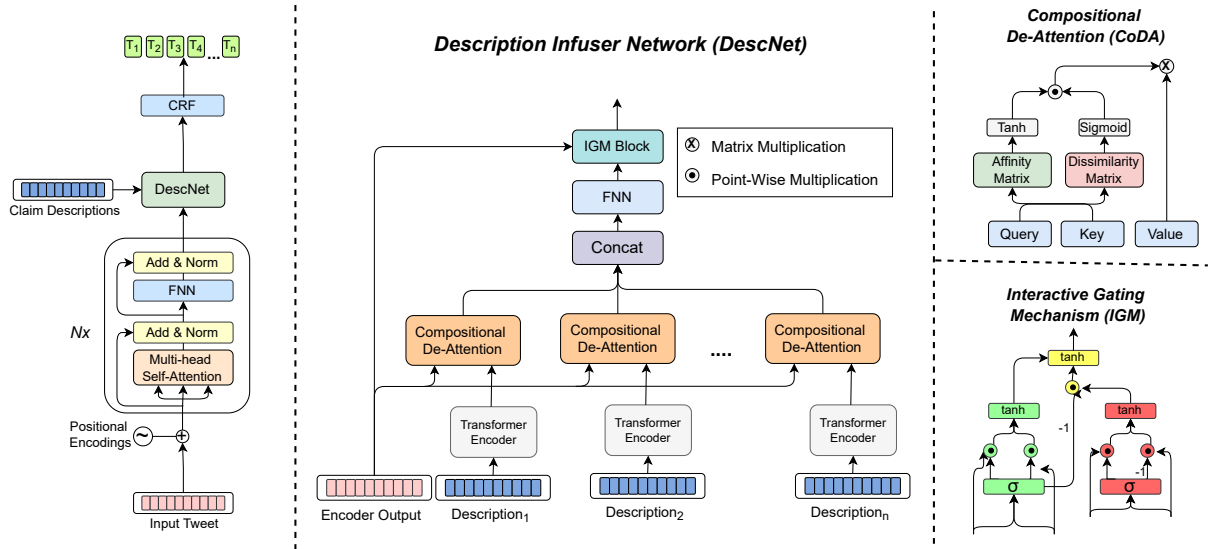
Figure 2: A schematic diagram of `DABERTa` for the claim span identification. $\odot$ represents point-wise multiplication, and $\otimes$ represents matrix multiplication.

Claim Description	Example
Texts in the tweet mentioning statistics, dates or numbers	Another case for more testing for #coronavirus! *Blood tests show 14% of people are now immune to covid-19 in one town in Germany* https://t.co/MVOq3nc4hn
Texts in the tweet that negate a possibly false claim	No! *#Bleach won't cure #COVID19. Disinfectants can't kill the #coronavirus in your body*. In fact, they will hurt you. If you or someone you know has been exposed to bleach, call Poison Control for help (1-800-222-1***). https://t.co/DtIfi77vLz https://t.co/9MxSFoVM0L
Texts in the tweet made in sarcasm or humour	@username I think the *cure to coronavirus is a 6 pack of corona* only. yeah
Texts in the tweet containing opinions that have societal implications	@username @username I think *it's a bio weapon made by China* so I'm not surprised it has a lot of carriers.
Texts in the tweet in the form of conditional statement	*if you smoke weed you are immune to coronavirus*
Texts in the tweet containing a quote from someone	The president said *injecting disinfectant into the body can cure the virus*. What in the holy hell? And @Lysol issued a statement that people should not ingest Lysol. WTF? #Covid_19 #lysol #DontDrinkLysol

Table 3: Examples of handcrafted claim descriptions, along with some aligning examples. Claim spans are highlighted in italics.

encoder block is a modular unit consisting of two sub-layers: (i) *Multi-Headed Self-Attention*, and (ii) *Feed-Forward Network*. Furthermore, each sub-layer contains a residual connection, followed by dropout and layer normalization. For the task of token classification, the output of the last encoder layer is passed to a CRF layer (Lafferty et al., 2001). This modularity of PLMs enables easy integration of adapter modules in their architecture for making these PLMs task-specific and domain-dependent. We choose RoBERTa (Liu et al., 2019) as our backbone network as it is the best-performing baseline (see Table 4).

Description Infuser Network (*DescNet*). Desc-Net is designed to facilitate deep semantic interaction among the input text and claim descriptions, and help underline the key fragments of claims. It consists of precisely engineered components of CoDA and IGM, each devised to augment the pro-

cess of claim span identification.

To put formally, consider $D = \{d_1, d_2, ..., d_m\}$ as the set of m claim descriptions and $T = \{t_1, t_2, ..., t_n\}$ as the corpus of n input texts. The description representations are extracted from pre-trained RoBERTa (Liu et al., 2019) and passed through a transformer encoder layer. To begin with, each ith PLM generated vector $Z_i \in \mathbb{R}^{N \times d}$ of input text t_i interacts with each jth description vector $D_j \in \mathbb{R}^{M \times d}$ via the CoDA block. Here the vector Z_i forms the query, which is processed against the vector D_j acting as the key and value (Equation 1).

$$Z_{ij}^C = CoDA(Z_i, D_j)D_j \qquad (1)$$

All such compositionally manipulated vectors Z_{ij}^C, after interacting with each jth description vectors are concatenated and passed through a dropout layer before going through a non-linear transformation for dimensionality reduction (Equation 2). The resultant vector Z_i' along with the vector Z_i is

passed to the *IGM* module to extract the semantically appropriate features pertinent for fine-grained claim span identification (Equation 3).

$$Z_i' = Concat(Z_{i1}^C, ..., Z_{im}^C) \qquad (2)$$

$$\hat{Z}_i = IGM(Z_i'W, Z_i) \qquad (3)$$

The vector $\hat{Z}_i$ is then passed to a CRF layer.

Compositional De-Attention Block (*CoDA*). The traditional narrative on attention mechanism (Bahdanau et al., 2015; Parikh et al., 2016; Seo et al., 2016; Vaswani et al., 2017) is heavily biased on the use of *Softmax* operator where the attention weights are always bounded between $[0, 1]$. Such a convex weighted addition scheme allows the vectors to only contribute in an additive manner. To counter this bottleneck, Tay et al. (2019) devised a quasi-attention technique that enables learning of additive as well as subtractive attention weights, allowing the input vectors to add to $(+1)$, not contribute to (0), and even subtract from (-1) the output vector. They decomposed the original Softmax-based self-attention as pointwise multiplication between two matrices as shown in Equation 4, where $G(.)$ is the negative pointwise L_1 distance between query Q and key K.

$$A_{quasi} = \left(tanh(\frac{QK^T}{\sqrt{d_k}}) \odot \sigma(\frac{G(Q,K)}{\sqrt{d_k}}) \right) V \qquad (4)$$

We adopt this quasi-attention strategy to promote more meaningful interaction between the input text and claim descriptions and generate more precise claim-relevant representations.

Interactive Gating Mechanism (*IGM*). To further distinguish salient tokens inclusive in claim spans, we posit *Interactive Gating Mechanism*. To begin with, the vectors Z_i and Z_i' are max pooled to obtain $Z_{ip}, Z_{ip}' \in \mathbb{R}^d$. These vectors are passed through a series of gates, the first of them being the *conflict gate* C, aimed at capturing the semantically conflicting features in Z_i and Z_i' (Equation 6).

$$\mu_c = \sigma(Z_{ip}W_{c1} + Z_{ip}'W_{c2} + b_{c1}) \qquad (5)$$

$$C = \tanh(Z_{ip} \odot \mu_c W_{c3} + Z_{ip}' \odot (1 - \mu_c)W_{c4} + b_{c2}) \qquad (6)$$

The *refine gate* R, on the other hand, endeavors to capture the semantically similar features between Z_{ip} and Z_{ip}' (Equation 8).

$$\mu_r = \sigma(Z_{ip}W_{r1} + Z_{ip}'W_{r2} + b_{r1}) \qquad (7)$$

$$R = \tanh(Z_{ip} \odot \mu_r W_{r3} + Z_{ip}' \odot \mu_r W_{r4} + b_{r2}) \qquad (8)$$

To congregate the conflicting and similar semantic representations spawned by the gates C and R, we employ an adaptive gating scheme to retain maximum differential information from each gate. It is given by Equation 10.

$$A = R + (1 - \mu_r) \odot C \qquad (9)$$

$$\hat{Z}_i = tanh(AW_a + b_a) \odot Z_i \qquad (10)$$

Finally, this vector $\hat{Z}_i$ is passed to a CRF layer for token classification.

6 Experiments and Results

Baseline Models. We employ the following baseline systems. ▷ **CNN+CRF**: A Convolutional Neural Network (CNN) trained with GloVe (Pennington et al., 2014) and a CRF head on top. ▷ **BiLSTM+CRF** (Huang et al., 2015): A sequence labeling model comprising Bidirectional Long Short-Term Memory (BiLSTM) and CRF layer. ▷ **BERT** (Devlin et al., 2019): A bidirectional transformer-inspired auto-encoder language model fine-tune for our span identification task. ▷ **DistilBERT** (Sanh et al., 2019): A smaller, faster, and lighter version of BERT fine-tune on our dataset for the task at hand. ▷ **SpanBERT** (Joshi et al., 2020): An enhanced version of the BERT trained on span prediction objective. ▷ **RoBERTa** (Liu et al., 2019): A robustly optimized BERT approach, RoBERTa, is a variant of BERT with improved training methodology. We fine-tune it on our dataset. ▷ **NLRG** (Chhablani et al., 2021): A system proposed at SemEval-2021 Task 5 on toxic span detection (Pavlopoulos et al., 2021). It is a combination of SpanBERT and RoBERTa where the former model is used for predicting the span start and end, while the latter is used for token classification. ▷ **HITSZ-HLT** (Zhu et al., 2021b): The system topped the SemEval-2021 task on toxic span detection. They approached the task as a combination of sequence labeling and span extraction and proposed an ensemble of three BERT-based models.

Evaluation Metrics. In concordance with Pavlopoulos et al. (2021), we evaluate the performance of all the systems, based on token-level precision (P), recall (R), and F1 scores. To further put a lens over how the models fare for different token types, we calculate the micro-level precision, recall, and F1 score for each of the *'B'*, *'I'*, and *'O'* tokens.[3] Lastly, to quantify the number of tokens included in the spans, we also report the Dice Similarity Coefficient (DSC) (Dice, 1945).

---

[3] Each token in the tweet is *BIO (Begin-Inside-Outside)* encoded to mark the claim spans (Ramshaw and Marcus, 1999).

Model Name	F1	P	R	F1			Precision			Recall			DSC
				B	I	O	B	I	O	B	I	0	
CNN+CRF	0.6635	0.6709	0.6947	0.3877	0.6766	0.9263	0.3952	0.6725	**0.9450**	0.3953	0.7718	0.9171	0.6964
BiLSTM+CRF	0.6825	0.6928	0.7048	0.4401	0.6717	**0.9356**	0.4653	0.6703	0.9428	0.4302	0.7459	**0.9382**	0.6884
DistilBERT	0.7645	0.7811	0.8068	0.6677	0.8164	0.7510	0.6560	0.7979	0.8310	0.7227	0.8989	0.7402	0.8277
BERT	0.7807	0.7996	0.8154	0.6900	0.8266	0.7699	0.6863	0.8163	0.8403	0.7302	0.8971	0.7634	0.8356
SpanBERT	0.7914	0.8093	0.8182	0.6971	0.8299	0.7901	0.7047	0.8384	0.8271	0.7203	0.8724	0.8048	0.8377
RoBERTa	0.8020	0.8163	0.8337	0.7221	0.8297	0.7942	0.7165	0.8371	0.8351	0.7624	0.8764	0.8022	0.8399
DistilBERT + CRF	0.8288	0.8581	0.8526	0.8722	0.8148	0.7431	0.8914	0.7852	0.8400	0.8621	0.9181	0.7219	0.8222
BERT + CRF	0.8368	0.8631	0.8556	0.8531	0.8284	0.7666	0.8781	0.8101	0.8375	0.8408	0.9042	0.7597	0.8343
SpanBERT + CRF	0.8390	0.8625	0.8562	0.8507	0.8253	0.7806	0.8742	0.8221	0.8302	0.8394	0.8827	0.7867	0.8316
RoBERTa + CRF	0.8457	0.8706	0.8635	0.8613	0.8340	0.7805	0.8874	0.8301	0.8321	0.8485	0.8972	0.7841	0.8402
NLRG	0.7494	0.7750	0.7832	0.6584	0.7892	0.7398	0.6631	0.7891	0.8119	0.6805	0.8600	0.7486	0.8087
HITSZ-HLT	0.7758	0.7966	0.8037	0.6754	0.8201	0.7780	0.6834	0.8230	0.8291	0.6978	0.8743	0.7850	0.8314
DABERTa	**0.8604**	**0.8814**	**0.8789**	0.9035	**0.8354**	0.7816	0.9205	**0.8242**	0.8379	0.8950	**0.9044**	0.7771	**0.8433**
− {*IGM*}	0.8539	0.8795	0.8768	**0.9121**	0.8310	0.7563	**0.9258**	0.8017	0.8473	**0.9051**	0.9277	0.7358	0.8401
− {*CoDA*} + {*DPA*}	0.8558	0.8788	0.8738	0.8886	0.8319	0.7821	0.9086	0.8184	0.8439	0.8785	0.9032	0.7752	0.8416

Table 4: Experimental results of `DABERTa`, its variants (last two rows), and baselines. DSC, P, and R denote Dice Similarity Coefficient, Precision, and Recall respectively.

Performance Comparison. We summarize our collated results in Table 4. Evidently, `DABERTa` outperforms all the baseline systems against majority of the evaluation metrics.[4] We analyze all the systems based on the following questions.

How accurately do the models predict? To gauge how well each model performs for the token classification task, we monitor precision, recall, and F1 scores. As it can be inferred from Table 4, the traditional word embedding-based deep learning models of CNN and BiLSTM give the poorest token classification performance. An appreciable improvement of about 10-14% across all three metrics is observed when we move from the classical deep learning architectures to the transformer-based models of DistilBERT, BERT, SpanBERT, and RoBERTa. This underlines the importance of using contextual word embeddings and transformer-based architectures for the task at hand. The addition of the CRF layer further amplifies the performance of these models. SpanBERT also fares better than BERT as it is trained using span prediction objective. We also notice that employing the CRF layer results in a somewhat better balance of precision and recall when compared to using a basic linear layer. The ensemble-based models of NLRG and HITSZ-HLT also give admissible results for our task. Our proposed model, `DABERTa`, surpasses all the models in terms of the precision, recall, and F1 scores. An improvement of about 1.5% is observed between RoBERTa and `DABERTa` in terms of these metrics. This justifies the inclusion of *claim descriptions* that amalgamate domain-specific semantic information into RoBERTa architecture via the deftly crafted adapter module. In summary, we see that
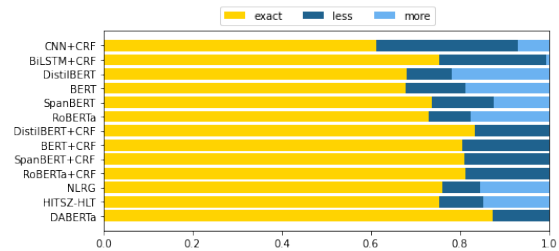

Figure 3: A comparative study among `DABERTa` and baselines. The horizontal bar signifies the ration of number of predicted spans and number of gold spans.

all the models show a good trade-off between precision and recall.

Are the models aggressive or defensive? Observing the precision, recall, and F1 scores for each of the 'B', 'I', and 'O' tags, as shown in Table 4, we get an idea of how aggressive or defensive the models are at predicting claim spans. CNN and BiLSTM show considerable resistance in predicting the claim spans, as evidenced by high precision, recall, and F1 scores for the token 'O' and less for the tokens 'B' and 'I'. The BERT-based models show a sizable improvement of about 22% and 15% for predicting tokens 'B' and 'I', respectively, over the traditional deep learning models. The addition of CRF layers further bolsters the predictive power for the token 'B'. `DABERTa` offers an improvement of about 4-5% over its traditional counterpart for predicting the token 'B'. Upon close inspection, we observe that the ranges of precision, recall, and F1 scores for predicting the tokens 'I' and 'O' vary by not more than 3%. However, the predictive power for the token 'B' varies vastly by about 25%. Hence, we hypothesize that the inclusion of *descriptions* makes our model cognizant of the syntactic and semantic constructs of claims.

How the models behave for multiple spans? Figure 3 illustrates how well the models identify

---

[4]Data preprocessing details are presented in Appendix (A.2).

multiple spans. It is observed that the models of CNN and BiLSTM find it challenging to identify multiple spans. The transformer-based models with a linear head tend to predict more claim spans in the tweet than required. This issue is mitigated when the linear head is replaced with a CRF layer. Still, these models can identify roughly only 80% of the time the occurrence of multiple spans. On the other hand, our model, DABERTa, correctly predicts multiple spans almost more than 85% of the time. Moreover, it does not predict more claim spans than required. Thus, the addition of domain-specific *claim descriptions* appropriately guides DABERTa in identifying the correct occurrence of spans.

Ablation Study. Table 4 also reports the ablation studies. Replacing *CoDA* with a naïve *Dot-Product Attention* (*DPA*), we observe a drop in the performance across almost all the metrics. Amongst all, the performance drop in predicting the token *'B'* is the most prominent ($\sim 1.5\%$ across precision, recall, and F1). Thus, we conjecture that the quasi-attention mechanism is better able to spot the starting of a claim fragment than *DPA*. When *IGM* is removed, the performance for predicting token *'B'* slightly improves. However, it leads to a decrease in the predictive power for *'O'* token ($\sim 2.5\%$ in F1). Therefore, the combination of *CoDA* and *IGM* obtains the most balanced performance.

Hyper-parameter Tuning. We utilize the *base* version of RoBERTa (Liu et al., 2019) to propose DABERTa. The model is trained end-to-end using the Adam optimizer (Kingma and Ba, 2014), learning rate of $4e-5$, and batch size of 32 for 20 epochs with early stopping if the dice score does not improve after 5 epochs. We used the Nvidia Tesla v100 32 GB GPU. The hyper-parameter tuning is done with respect to the validation dataset.
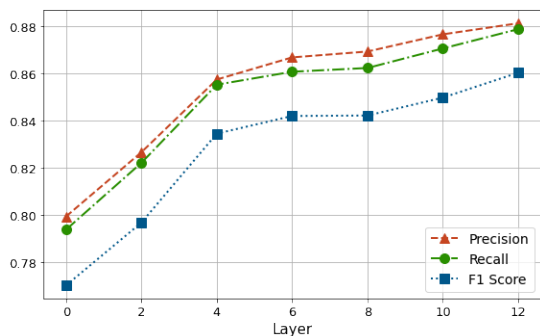

Figure 4: Performance of DABERTa when the adapter module is inserted at different layers of RoBERTa.

Figure 4 reflects the effect of integrating the adapter *DescNet* at different layer of RoBERTa. It is observed that the performance consistently increases as the integration is done at a higher level of RoBERTa layers. This is admissible as studies on probing the PLM layers suggest that different layers encode distinct linguistic properties (Tenney et al., 2019). Furthermore, evidence by Peters et al. (2018) suggests that the lower layers of a language model encode the syntactic information, whereas the higher layers capture the complex semantics. As we strive to employ deep semantic interaction between the PLM representations and the *claim descriptions*, our results are consistent with their findings.

Error Analysis. In this section, we manually analyze the errors the models are prone to make. Table 5 highlights randomly sampled tweets from our dataset, CURT, along with their gold spans and predictions from DABERTa. In addition, we also consider the predictions from the best-performing baseline, RoBERTa, for a fair comparison. We analyze the errors committed by both the systems and divide them into three different categories: (i) tweets with a single-claim span, (ii) tweets with claim-like premises, and (iii) tweets with claims that can be inferred from the underlying undertone of the tweet but no explicit span can be marked to highlight the claim-specific connotation, e.g., figurative sentences, satire, indirect questions etc. (Note: For simplicity, we refer to such claims as *implicit* claims.) In the most straightforward situation where the tweets only contain a single claim DABERTa makes more precise predictions than the baseline system as shown in the first example of Table 5. We observe that both the models identify the claim-span correctly; however, RoBERTa identifies some unnecessary spans, which trespasses our objective of equipping the fact-checkers with only relevant information. The second type of error related to spans is the presence of claim-like premises. Claims and premises[5] are closely related components of argument mining, and differentiating them is strenuous, even for humans. Example 2 in Table 5, exhibits a post containing claim-premise pair. There are two conclusive claims in the tweet — *'#coronavirus was used by the #CCP as a bio weapon'* and *'CCP is kicking out black people from hotels even if they don't have covid'*. Even though *'not only to kill people but to encourage racism*

---

[5](sub)sentences used to support the concluding claim

	Model	Tweet
1	*Gold*	Gold Truly sobering analysis: **US more vulnerable than many countries to #coronavirus** owing to combination of high numbers of uninsured, many w/o paid sick leave, and a leadership that has downplayed the challenge while not preparing the country for it.
	RoBERTa	Gold Truly sobering analysis: **US more vulnerable than many countries to #coronavirus** owing to combination of *high numbers of uninsured*, many w/o *paid sick leave*, and a leadership that has downplayed the challenge while not preparing the country for it.
	DABERTa	Gold Truly sobering analysis: **US more vulnerable than many countries to #coronavirus** owing to combination of high numbers of uninsured, many w/o paid sick leave, and a leadership that has downplayed the challenge while not preparing the country for it.
2	*Gold*	Whether made on purpose or not **#coronavirus was used by the #CCP as a bio weapon**, not only to kill people but to encourage racism among their citizens against foreigners. Especially black people, **CCP is kicking out black people from hotels even if they dont have covid.**
	RoBERTa	Whether made on purpose or not **#coronavirus was used by the #CCP as a bio weapon**, *not only to kill people but to encourage racism among their citizens against foreigners*. Especially black people, CCP is kicking out black people from hotels even if they dont have covid.
	DABERTa	Whether made on purpose or not **#coronavirus was used by the #CCP as a bio weapon**, *not only to kill people but to encourage racism among their citizens against foreigners*. Especially black people, **CCP is kicking out black people from hotels even if they dont have covid.**
3	*Gold*	RT @HealtheNews: Can honey, ginger, garlic or turmeric or any other home remedies cure #Covid19? No, here's why.
	RoBERTa	RT @HealtheNews: Can *honey, ginger, garlic or turmeric or any other home remedies cure #Covid19?* No, here's why.
	DABERTa	RT @HealtheNews: Can *honey, ginger, garlic or turmeric or any other home remedies cure #Covid19?* No, here's why.

Table 5: Error analysis of the outputs. Bold text (green) highlights the correct claim span whereas text in italics (red) represents the mistakes committed by our model, DABERTa, and vanilla RoBERTa as baseline.

among their citizens against foreigners' appears to be a claim at first glance, it serves as the premise to support the conclusive part of the arguments brought forward in the tweet. In most cases, we discern that both the systems identify the claim-spans correctly, but they are easily fooled by the premises, hence leaving room for significant improvement in this regard.

Another prominent class of errors is implicit claims. Extracting the claim-spans in implicit claims is arduous. We observe that both the systems strive to understand the linguistic structure of the implicit claims. For instance, in sample 3, the user intends to assert that honey, ginger, garlic, or turmeric do not cure COVID19; however, DABERTa fails to understand the user's intention and yields the wrong span. We perceive similar behavior from the best-performing baseline, RoBERTa, as well. A plausible reason is the skewed nature of the dataset , which is lopsided with a significant bias toward explicit claims. According to our observations, DABERTa outperforms the best-performing baseline system significantly ($\sim 4\%$; $p < 0.0004$).[6] Hence, furnishing us with empirical shreds of evidence that DABERTa can be efficiently used for claim span identification.

7 Conclusion

Through this systematic research, we introduced the novel task of *Claim Span Identification*, which is valuable on various fronts. We conducted an evidence-based document retrieval experiment, demonstrating that employing claim spans retrieves more relevant evidence than using the entire tweet. Furthermore, as there exists no specialized corpus for claim span identification, we compiled **CURT**, a large-scale Twitter corpus consisting of around $7.5k$ tweets annotated with token-level claim spans. We showed convincing results using various token classification baselines on our dataset. Moreover, we benchmarked CURT with **DABERTa**, an adapter-based variant of RoBERTa, that encapsulates critical domain-specific information into the pre-trained model via *claim descriptions*. Through extensive qualitative, quantitative, and empirical results, we illustrated how DABERTa outperforms the other models on different fronts. Lastly, we also developed an extensive set of annotation guidelines and released them for further research.

8 Limitations

Though DABERTa yields the state-of-the-art performance in claim span identification; there are a few cases where it falls short. Even for humans, recognizing claim spans in figurative or metaphorical sentences is arduous; consequently, our suggested model also struggles with them. As a result, our future study will focus on boosting the claim span identification performance, especially for such sentences. Our analysis also bestowed that the high resemblance between claims and premises confuses the model, making it difficult to distinguish between the two. DABERTa shares the said limitation with other baseline systems as well. As a result, this could be another alluring open challenge to work on.

---

[6]We also perform significance t-test on F1 scores per input tweet comparing DABERTa and RoBERTa, as best-performing baseline.

Acknowledgement

References

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68, Baltimore, Maryland. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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. In *European Conference on Information Retrieval*, pages 499–507. Springer, Nature Publishing Group.

Emily M. Bender, Jonathan T. Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf. 2011. Annotating social acts: Authority claims and alignment moves in Wikipedia talk pages. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 48–57, Portland, Oregon. Association for Computational Linguistics.

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)*, pages 558–563, Minneapolis, Minnesota. Association for Computational Linguistics.

Gunjan Chhablani, Abheesht Sharma, Harshit Pandey, Yash Bhartia, and Shan Suthaharan. 2021. NLRG at SemEval-2021 task 5: Toxic spans detection leveraging BERT-based token classification and span prediction techniques. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 233–242, Online. Association for Computational Linguistics.

Samuel Coope, Tyler Farghly, Daniela Gerz, Ivan Vulić, and Matthew Henderson. 2020. Span-ConveRT: Few-shot span extraction for dialog with pretrained conversational representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 107–121, Online. Association for Computational Linguistics.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lee R Dice. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Emilio Ferrara. 2020. What types of covid-19 conspiracies are populated by twitter bots? *arXiv preprint arXiv:2004.09531*.

Shreya Gupta, Parantak Singh, Megha Sundriyal, Md. Shad Akhtar, and Tanmoy Chakraborty. 2021. LESA: Linguistic encapsulation and semantic amalgamation based generalised claim detection from online content. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3178–3188, Online. Association for Computational Linguistics.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. UKP-athene: Multi-sentence textual entailment for claim verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108, Brussels, Belgium. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. 2018. ClaimRank: Detecting check-worthy claims in Arabic and English. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for*

Computational Linguistics: Demonstrations, pages 26–30, New Orleans, Louisiana. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Span-BERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Klaus Krippendorff, Yann Mathet, Stéphane Bouvry, and Antoine Widlöcher. 2016. On the reliability of unitizing textual continua: Further developments. *Quality & Quantity*, 50(6):2347–2364.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Marco Lippi and Paolo Torroni. 2015. Context-independent claim detection for argument mining. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 185–191. AAAI Press.

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. *arXiv preprint arXiv:1907.11692*.

Josh Margolin. 2020. Fbi warns of potential surge in hate crimes against asian americans amid coronavirus. *ABC News*, 27.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

Salman Bin Naeem and Rubina Bhatti. 2020. The covid-19 'infodemic': a new front for information professionals. *Health information and libraries journal*, 37(3):233—239.

Viet Anh Nguyen, Tam Minh Nguyen, Huy Quang Dao, and Quang Huu Pham. 2021. S-NLP at SemEval-2021 task 5: An analysis of dual networks for sequence tagging. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 888–897, Online. Association for Computational Linguistics.

Alex Nikolov, Giovanni Da San Martino, Ivan Koychev, and Preslav Nakov. 2020. Team alex at clef checkthat! 2020: Identifying check-worthy tweets with transformer models. *arXiv:2009.02931*.

Cathal O'Connor and Michelle Murphy. 2020. Going viral: doctors must tackle fake news in the covid-19 pandemic. *BMJ*, 369.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online. Association for Computational Linguistics.

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)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Kamil Pluciński and Hanna Klimczak. 2021. GHOST at SemEval-2021 task 5: Is explanation all you need? In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 852–859, Online. Association for Computational Linguistics.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Sara Rosenthal and Kathleen McKeown. 2012. Detecting opinionated claims in online discussions. In *Proceedings of the 2012 IEEE Sixth International*

Conference on Semantic Computing, ICSC '12, page 30–37, USA. IEEE Computer Society.

Jonathan Rusert. 2021. NLP_UIOWA at Semeval-2021 task 5: Transferring toxic sets to tag toxic spans. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 881–887, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Amir Soleimani, Christof Monz, and Marcel Worring. 2020. Bert for evidence retrieval and claim verification. *Advances in Information Retrieval*, 12036:359.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Megha Sundriyal, Parantak Singh, Md Shad Akhtar, Shubhashis Sengupta, and Tanmoy Chakraborty. 2021. Desyr: Definition and syntactic representation based claim detection on the web. *arXiv preprint arXiv:2108.08759*.

Yi Tay, Anh Tuan Luu, Aston Zhang, Shuohang Wang, and Siu Cheung Hui. 2019. Compositional de-attention networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Stephen E Toulmin. 2003. *The uses of argument*. Cambridge university press.

Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schütze, and Iryna Gurevych. 2020. Fine-grained argument unit recognition and classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9048–9056.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Evan Williams, Paul Rodrigues, and Valerie Novak. 2020. Accenture at checkthat! 2020: If you say so: Post-hoc fact-checking of claims using transformer-based models. *arXiv: 2009.02431*.

Dustin Wright and Isabelle Augenstein. 2020. Claim check-worthiness detection as positive unlabelled learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 476–488, Online. Association for Computational Linguistics.

Amelie Wührl and Roman Klinger. 2021. Claim detection in biomedical Twitter posts. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 131–142, Online. Association for Computational Linguistics.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267, Rochester, New York. Association for Computational Linguistics.

Shi Zhi, Yicheng Sun, Jiayi Liu, Chao Zhang, and Jiawei Han. 2017. Claimverif: A real-time claim verification system using the web and fact databases. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 2555–2558, New York, NY, USA. Association for Computing Machinery.

Qinglin Zhu, Zijie Lin, Yice Zhang, Jingyi Sun, Xiang Li, Qihui Lin, Yixue Dang, and Ruifeng Xu. 2021a. HITSZ-HLT at SemEval-2021 task 5: Ensemble sequence labeling and span boundary detection for toxic span detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 521–526, Online. Association for Computational Linguistics.

Qinglin Zhu, Zijie Lin, Yice Zhang, Jingyi Sun, Xiang Li, Qihui Lin, Yixue Dang, and Ruifeng Xu. 2021b. HITSZ-HLT at SemEval-2021 task 5: Ensemble sequence labeling and span boundary detection for toxic span detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 521–526, Online. Association for Computational Linguistics.

Caleb Ziems, Bing He, Sandeep Soni, and Srijan Kumar. 2020. Racism is a virus: Anti-asian hate and counter-hate in social media during the covid-19 crisis. *arXiv preprint arXiv:2005.12423*.

A Appendix

A.1 Annotations

A.1.1 Guideline Development

While different frameworks and models of argumentation range in intricacy and claim conceptualization, the claim element is colloquially perceived as a principal component of an argument. Following Stab and Gurevych (2017), we define the claim as '*the argumentative component in which the speaker or writer conveys the central, contentious conclusion of their argument*'. Aharoni et al. (2014) proposed a framework in which an argument is often divided into two parts: claim and premise. The premise, which is another crucial component of an argument, encompasses all shreds of evidence obliged to either corroborate or refute the claim. We confine our corpus to claim components only. However, claims and premises are usually indistinguishable and frequently blend together. As a result, distinguishing them can be challenging, especially when authors use claim-like statements as a premise.

Due to the highly subjective nature of claims, it is imperative to devise structured annotation guidelines to annotate a new dataset for the claim span identification task. Therefore, after rigorous analysis and discussion, we established an initial set of annotation guidelines. To acclimate better with the dataset, we progressed through iterations of improvements. In every iteration, 100 random tweets were annotated by three annotators[7] following the initial set of annotation guidelines. The annotators resolved the ambiguous cases mutually. In successive iterations, we further addressed the unsettled tweets that necessitated clarifications in the annotation guidelines. We reconsidered all prior annotations for every change in the guideline to ensure that the annotations emulated the most advanced version of the annotation guidelines. The final sprint of pilot annotation included annotating another set of randomly chosen 100 tweets with the final guidelines. Following Trautmann et al. (2020), we calculated the inter-annotator agreement using the α_u agreement measure (Krippendorff et al., 2016). We computed the mean pairwise value per post, where each token can be being classified into two classes, claim, and non-claim. We obtained a more than satisfactory agreement score of 0.87.

---

[7]They are linguistic experts and their age ranges between 20-35 years.

Finally, the entire Twitter dataset was annotated by the same annotators that carried out the prefatory pilot annotations.

A.1.2 General Instructions

- A claim is a statement that says you strongly believe that something is true. The action of showing, using or stating something strongly.

- We use tweets that are annotated with a binary label using LESA guidelines (Gupta et al., 2021), which indicates whether a tweet is a claim or not.

- The claim span is that part of a sentence that contains the semantic representation of the claim. Example: @realDonaldTrump A lot of people are saying cocaine cures COVID-19. Claim span: cocaine cures COVID-19.

- Since our primary goal is to tackle misinformation in OSM, we majorly focus on claims that have some social impact.

A.1.3 Guidelines and Examples

- In the case of facts, we annotate the fact/span that may not be known by everyone, for example, scientific facts or legal (law) facts, and doesn't involve any commonsense. However, we do not include universal facts in the claim span.
  Example 1: "Water is colorless" is a universally known fact and hence should not be marked as claim span.
  Example 2: "Virus always mutate" is a scientific fact that may not be known by everyone. Hence we will annotate this fact as a claim.

- An assertion about future eventualities/predictions will not be included in the claim span. Prediction is an extrapolation based on an assertion and is associated with a confidence level that can never be greater than or equal to 100%. Thus, we will not consider predictions as a part of the claim span.
  Example: @realDonaldTrump Uh no actually The virus will never go away Scientists will develop a vaccine for it that should be ready by next June which will allow nearly everyone to be immune to #CoronaVirus This really isn't hard to understand even for a very stable genius.

- A proverb is a simple, concrete, traditional saying that expresses a perceived truth based on common sense or experience which contains wisdom,

7713

truth, morals, and traditional views in a metaphorical, fixed, and memorizable form. The proverbs are not facts. The elements of proverbs should be annotated as claim span.
Example: "Prevention is better than cure" is not a claim.

- If a claim contains statistics or dates, they should be included in the span. But not all numbers are important.
  Example 1: "@FernandoSVZLA @AP So far 50 people outside China have it with no deaths. If China was hiding information and it was more lethal, we would see that fairly quickly". Here the claim span is [50 people outside China have it with no deaths]
  Example 2: "57 round trip to LA thanks coronavirus". The number is not important here.

- In case there are multiple conclusive independent claims in one tweet, we annotate each one of them separately.
  Example: "5 million left Wuhan before the lockdown. If they were really interested in knowing, they'd be testing at least 1 in 100 cases of all viral pneumonia. They're limiting who's being tested so they aren't accused of lying. Oh, and they might be asked to actually do something."
  Claim span would consist of: [1] "5 million left Wuhan before the lockdown" and [2] "They're limiting who's being tested so they aren't accused of lying"

- Tweets that negate a possibly false claim are also considered to be claims.
  Example: "disinfectants are not a cure for coronavirus".

- Tweets 'reporting' something to be true or an instance to have happened or will happen are claims.

- In cases of claims made in the form of a conditional sentence, the premise/context would be included in the span.
  Examples: if you've been in the McDonald's play place you're immune to the coronavirus.

- For claims containing humor/sarcasm, only the humorous phrase will be considered as a claim span if it has some social impact. For satire, the complete sentence will be considered.
  Example: @TheRickWilson Drinking bleach and/or injecting Disinfectant will cure COVID19.

And cancer, heart disease, OCD, schizophrenia and AIDS. And life. #Covid19 #COVID Claim: Drinking bleach and/or injecting Disinfectant will cure COVID19. And cancer, heart disease, OCD, schizophrenia and AIDS. And life.

- Personal experience will only be part of the claim phrase if they are opinions with societal impacts/implications.
  Example: Story about how #HydroxyChloroquine likely help people recover from #Coronavirus. IMO, it was never touted as the cure but as option for treatment doctors should consider and it appears to work in some cases....39 in one place. https://t.co/2hhi6aSVrY
  Claim: [it was never touted as the cure but as option for treatment doctors should consider and it appears to work in some cases....39 in one place.]

- A claim can be a sub-part of a question, only if it is not a direct question.
  Example: @FLOTUS Melania, do you approve of ingesting bleach and shining a bright light in the rectal area as a quick cure for #COVID19 ? #BeBest"
  Claim: [ingesting bleach and shining a bright light in the rectal area as a quick cure for #COVID19]

- Ground/Reasoning to justify a claim will not be a part of the claim phrase.
  Example: Covid-19 vaccine development and deployment in China, when available, will be made a global public good, which will be China's contribution to ensuring vaccine accessibility and affordability in developing countries
  Claim phrase: [Covid-19 vaccine development and deployment in China, when available, will be made a global public good]

- Mocking/attacking a group or individual is not a part of the claim phrase.
  Example: Because #coronavirus has tremendous chances of getting cured but your anti-national agenda is worse than death
  Claim: [coronavirus has tremendous chances of getting cured]

- Claim phrases do not include the predicate part that does not contribute to it being a claim.
  Example: I firmly believe that [if they found a way to bottle the @andersoncooper giggle, it would cure the corona virus]

A.2 Data Preprocessing

We employ *NLTK*[8] to tokenize the tweets. Each token in the tweet is *BIO (Begin-Inside-Outside)* encoded to generate the labels (Ramshaw and Marcus, 1999). Tag *'B'* indicates that the token is at the start of a span, tag *'I'* indicates that the token is within the span, while tag *'O'* denotes that the token is outside the span. As RoBERTa tokenizes each word into subwords (Liu et al., 2019), each subword is given the BIO tag as per their parent word. We eliminate tokens made of non-ASCII and special characters, as well as remove the URLs provided in the tweets. Finally, we split hashtag terms by underscore delimiter and over non-consecutive uppercase character. For instance, *#WuhanLab* splits into *'Wuhan'* and *'Lab'*.

---

[8]https://www.nltk.org/