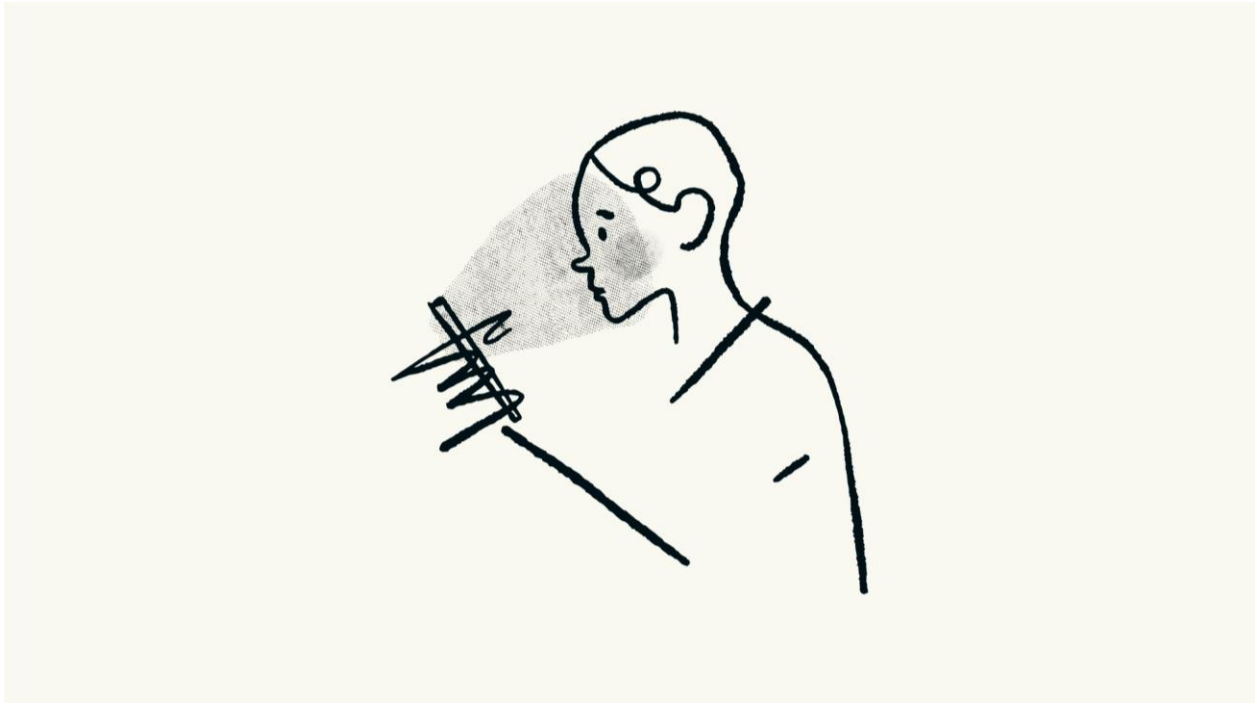
Claim Extraction and Dynamic Stance Detection in COVID-19 Tweets

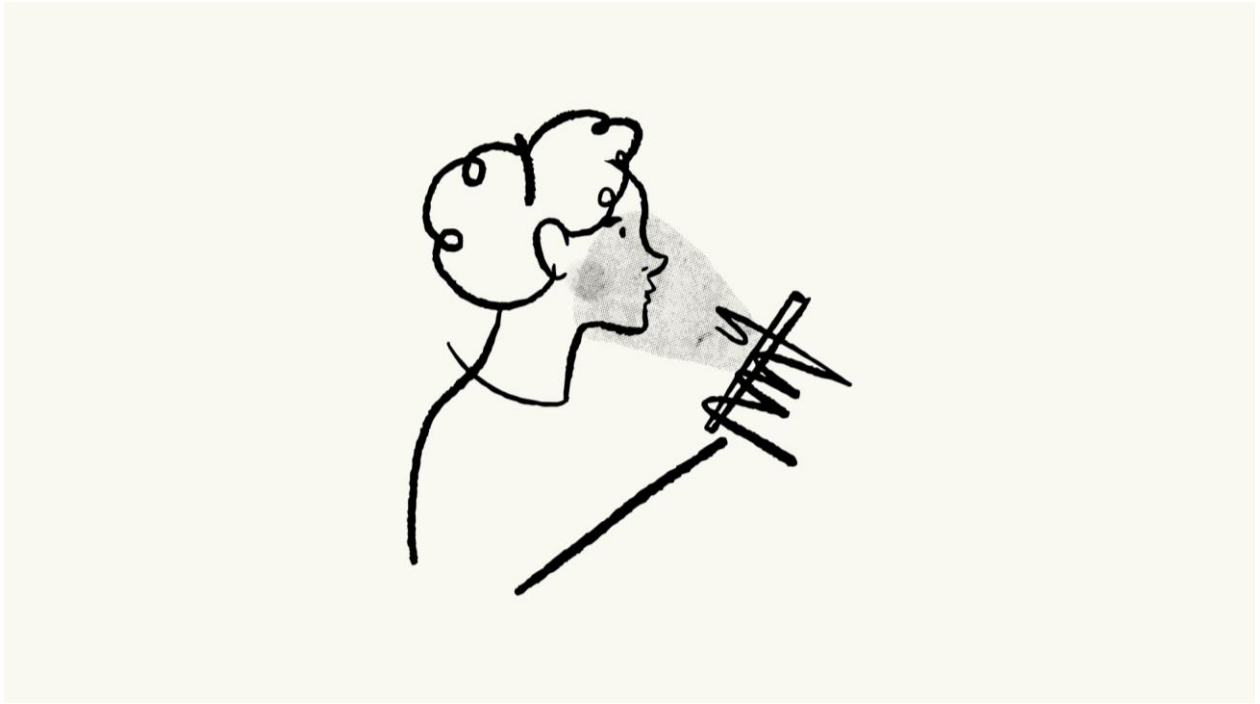
Noushin Salek Faramarzi*[1] , Fateme Hashemi Chaleshtori*[2] , Hossein Shirazi[3]
Indrakshi Ray[2], Ritwik Banerjee[1]

[1]Stony Brook University, [2]Colorado State University, [3]San Diego State University

*Equal Contribution

We are not just fighting an **epidemic**;
we're fighting an **infodemic**





Introduction

- COVID-19 pandemic led to an "infodemic" with two major consequences:
  - endangering lives due to misinformed decisions
  - general distrust of media

- **Fact-checking** has been the primary approach to combat misinformation
- Accurate **identification** and **extraction** of factual claims from media articles and social media posts is crucial for effective fact-checking

(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.}$_{\text{commentary}}$

(A) opposition to a political party and its policies
(B) support for the action of a sportsperson

PAPER FOCUS

Identifying posts that contain objective (factual) claims and determining if they are worth fact-checking

Determining the author's stance towards the factual claim presented in the post, without a fixed set of topics.

Differentiating factual claims from the author's commentary in the post

DATASET

Data Collection & Preprocessing

- Tweets were collected over a 17-month period (May 2020 to September 2021) using keywords related to COVID-19.
- **Duplicates** and tweets with less than five words were removed
- **URLs** and usernames were replaced with "URL" and "USER" tags, respectively
- **Emojis** were replaced with their corresponding text representations
- **Slang** and colloquial acronyms were replaced with their formal counterparts. Example: "wml" is converted to "wish me luck" and "wochit" is converted to "watch it."

- The most relevant terms manually selected from the 80 most frequent content words found in a random sample of 10,000 tweets in a COVID-19 Twitter corpus.

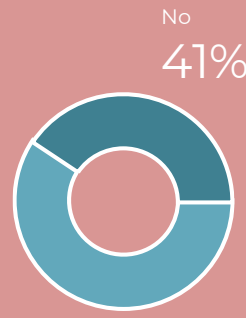
booster, china, corona, coronavirus, coronavirusoutbreak, coronaviruspandemic, coronavirusupdates, (covid-19), covid, covid-19, covid__19, covid_19, covid19, covid19https, covid19outbreak, covid19pandemic, covid2019, covidoutbreak, covidpandemic, delta variant, johnson and johnson, lockdown, masks4all,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

DATASET

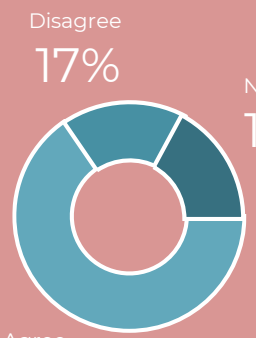
Data Annotation

- The objective (factual) claim component of each tweet is annotated, if present. Annotators provide one of three labels **(agree, disagree, or neutral**) to indicate the stance of the tweet towards the factual claim.
- A **web interface** was developed for the annotation task, allowing annotators to skip instances if they were unsure about the label for a tweet.
- Each tweet was independently annotated by **two** individuals. A total of 12 annotators, proficient in **English** and with at least an undergraduate college education, were involved in the annotation process.

Does the tweet contain a factual check-worthy claim?



What is the stance of the author with respect to the claim made in the tweet?



CLAIM EXISTENCE

Objective: Identify models that discard the tweets that do not contain an objective (factual) claim, while retaining the tweets that do.

Datasets	• Our annotated corpus **(DS1)** • Augmented version of our annotated corpus using Back-translation technique **(DS1-aug)** • COVID-19 Infomedic corpus by Alam et al. **(DS2)**

Models & setup	• fine-tuning on diverse pretrained language models like BERT, RoBERTa, and XLNET, employing distinct dataset setups for the Train and Test sets.
	  • Setup 1: Train on DS2 and DS2-Eng, and test them on a fixed set of English Tweets published by Alam et al.
	  • Setup 2: Train on any of the experimented datasets, test on only our annotated set (**DS1-test**)

CLAIM EXISTENCE RESULTS

Results on Setup 1

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

- The best weighted $F1$ score for "**Does the tweet contain a factual check-worthy claim?**" question , when training on DS2-Eng, is achieved using the RoBERTa model (78.6%) by Alam et al. For the same setup, we obtain an $F1$ score of **79.91**%.
- Training on DS2 (i.e., including the translated non-English tweets), we observe an improvement to 81.43% for RoBERTa.

Results on Setup 2

Dataset		BERT		RoBERTa		XLNET	
Train	Test	P	F1	P	F1	P	F1
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

- 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 this task.
- Training on the augmented collection, **DS1-Aug**, offers significant improvement across all models
- 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.
- The best $F1$ 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 $F1$ score jumps by nearly 4% (from 72.77% to 76.59%).

CLAIM EXTRACTION

Objective: Identify models that differentiate factual claims from the author's commentary in the post

- 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
  - **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)
  - **O** indicates that the token is outside the scope of the claim.
- For 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
- We compare the baseline results to Flair which is an off-the shelf framework for training neural networks for natural language processing tasks.
- Experimented with different versions of stacked embedding such as **Flair + BERT** and more

CLAIM EXTRACTION RESULTS

Influential users:

- Analyze model performance on influential users with large followings.
- specifically selected small test set of 100 tweets from these influential accounts

Evaluation Metric

- The standard evaluation techniques are **not** immediately applicable.
- We develop a relaxed evaluation, where the incorrect inclusion of additional tokens or the incorrect exclusion of the claim's tokens are somewhat tolerable.
- 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**.

- When utilizing the standalone model BERT, the regular tweet set achieves an F1 score of **0.55**, while the influential tweets yield an F1 score of **0.74**.
- The performance of stacked embeddings Flair + BERT exhibits notable improvement in terms of F1 score.
- Specifically, the regular tweet set achieves an F1 score of **0.67** and a precision of **0.72**, while the influential users' tweet sets attain a precision of **0.81** and **0.84** respectively.
- 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.

DYNAMIC STANCE DETECTION

Objective: Determining the author's stance towards the factual claim presented in the post, without a fixed set of topics.

.

An author’s disagreement toward the primary argument presented in the same tweet

- The position is represented by one of the following category labels: “Favor”, “Against” or “Neither”.
- We explore the possibility of detecting the stance without any fixed set of targets, and indeed, without even the explicit notion of targets.
- 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.

DYNAMIC STANCE EXPERIMENTS

Datasets

- Our annotated corpus **(DS1)**
- COVID-19-Stance Dataset by Glandt et al. (**DS3**), 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.

Models & Results

- fine-tuning on diverse pretrained language models like BERT, RoBERTa, and XLNET, employing distinct dataset setups for the Train and Test sets.
- In particular, our approach is much like the first task of identifying the existence of a claim, since both are essentially classification tasks.
- **XLNET** yielded the best result with a precision of **0.63** and an F1 score of **0.46**. while trained and tested on our dataset.
- Our results are not directly comparable to those obtained by Glandt et al., 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 $F1 > 0.8$ for one of the four targets, our experiments find XLNet to be the best performer, with $F1 = 0.72$, when testing on DS3-Test.

Understanding the **Results**

Key Findings

•Presenting a new Twitter dataset focused on COVID-19, annotated for claim extraction and dynamic stance detection without predefined targets, aiming to accurately differentiate objective claims from subjective commentary.

•Demonstrating the feasibility of accurate claim extraction through a sequence labeling approach and providing an initial exploration of dynamic stance detection.

•Incorporating traditional stance detection datasets revealed potential for improvement and sets the stage for future research in this area.

•Acknowledging the limited success in the third task, which can be attributed partly to the dataset's size. Plans to develop larger corpora for the explored tasks and encouraging the development of similar datasets by other researchers.