iDiR Department of Computer Science and Engineering
The University of Texas at Arlington

UNIVERSITY OF TEXAS ARLINGTON

Detecting Stance of Tweets Toward Truthfulness of Factual Claims

Zhengyuan Zhu, Zeyu Zhang, Foram Patel, Chengkai Li

The Innovative Data Intelligence Research Laboratory(IDIR Lab)
Department of Computer Science and Engineering

2022 Computation + Journalism Conference
June 10th, 2022

Overview

- A factual claim can be **false information** or **a true fact**.

Image source: https://indepest.com/2020/05/08/fake-news/

Overview

- A factual claim can be **false information** or **a true fact**.
- False information can be **easily internalized** and true fact can be **unacknowledged** by social media users.

Image source: https://indepest.com/2020/05/08/fake-news/

Overview

- A factual claim can be **false information** or **a true fact**.
- False information can be **easily internalized** and true fact can be **unacknowledged** by social media users.
- The spread and influence of factual claims are reflected by public opinion toward their veracity.

Image source: https://indepest.com/2020/05/08/fake-news/

Overview

- The quest to **discern the public opinions** toward the veracity of factual claim.
  - Our society is damaged by the dispute of social media users who believe
    - false information is true
    - a true fact is false

Overview

- The quest to **discern the public opinions** toward the veracity of factual claim.
  - Our society is damaged by the dispute of social media users who believe
    - false information is true
    - a true fact is false
  - Journalists are working hard on
    - comprehending the spreading of misinformation
    - mitigating the vicious impact of misinformation

Overview

- What is stance detection in general?
  - Determining whether a user **supports or refutes** a **target** based on their social media post.

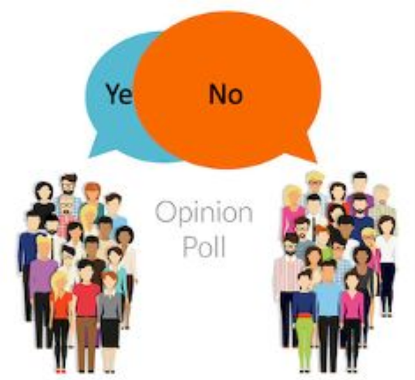


Image source: https://www.jdsupra.com/legalnews/don-t-fully-trust-public-opinion-polls-25406/

Overview

- What is stance detection in general?
  - Determining whether a user **supports or refutes** a **target** based on their social media post.
- General stance detection example:
  - Target: **Immigration**

Overview

- What is stance detection in general?
  - Determining whether a user **supports or refutes** a **target** based on their social media post.
- General stance detection example:
  - Target: **Immigration**
  - Tweet: “Unregulated immigration that hurts working Americans and globalism which only further weakens America!”

Overview

- What is stance detection in general?
    - Determining whether a user **supports or refutes** a **target** based on their social media post.
- General stance detection example:
    - Target: **Immigration**
    - Tweet: “Unregulated immigration that hurts working Americans and globalism which only further weakens America!”
    - Stance: **Refute**

Overview

- The emphasis of our study:
  - **The target of stance is a factual claim**.
  - The stance refers to a Twitter user’s belief or assertion regarding **the truthfulness of a factual claim**.

Overview

- The emphasis of our study:
    - **The target of stance is a factual claim**.
    - The stance refers to a Twitter user's belief or assertion regarding **the truthfulness of a factual claim**.
- An example of truthfulness stance:

Claim: Rep. Paul Gosar asks Capitol Police to arrest illegal immigrants attending State of the Union.

Overview

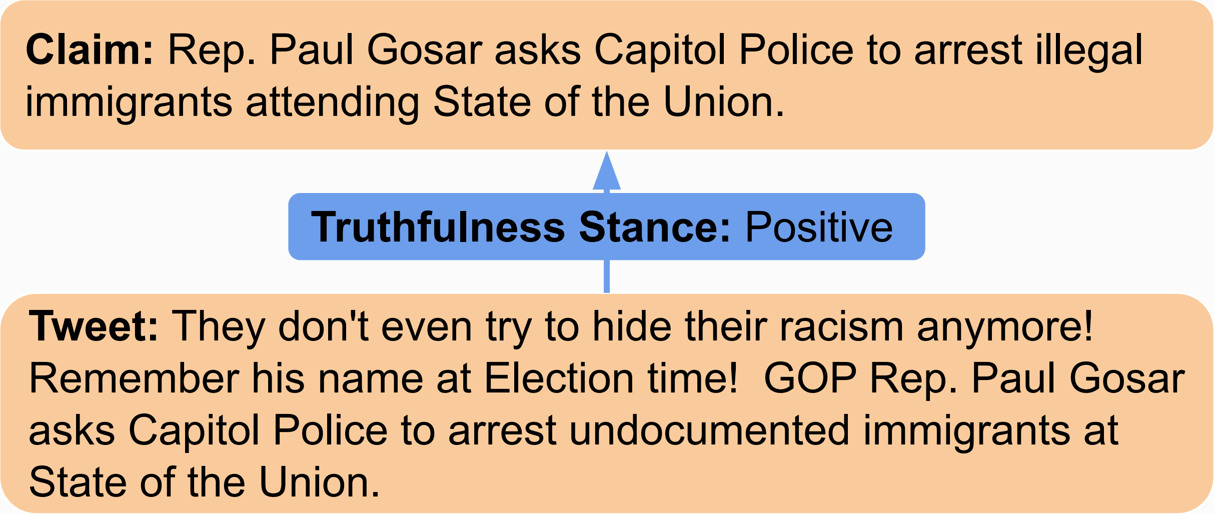
- The emphasis of our study:
  - **The target of stance is a factual claim**.
  - The stance refers to a Twitter user's belief or assertion regarding **the truthfulness of a factual claim**.
- An example of truthfulness stance:

Claim: Rep. Paul Gosar asks Capitol Police to arrest illegal immigrants attending State of the Union.

Tweet: They don't even try to hide their racism anymore! Remember his name at Election time! GOP Rep. Paul Gosar asks Capitol Police to arrest undocumented immigrants at State of the Union.

Overview

- The emphasis of our study:
  - **The target of stance is a factual claim**.
  - The stance refers to a Twitter user’s belief or assertion regarding **the truthfulness of a factual claim**.
- An example of truthfulness stance:



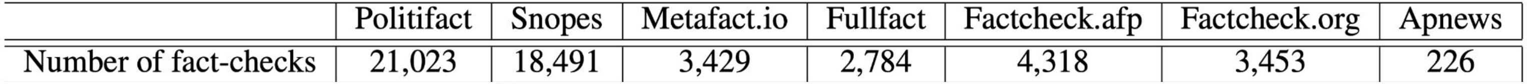
Claim: Rep. Paul Gosar asks Capitol Police to arrest illegal immigrants attending State of the Union.

Truthfulness Stance: Positive

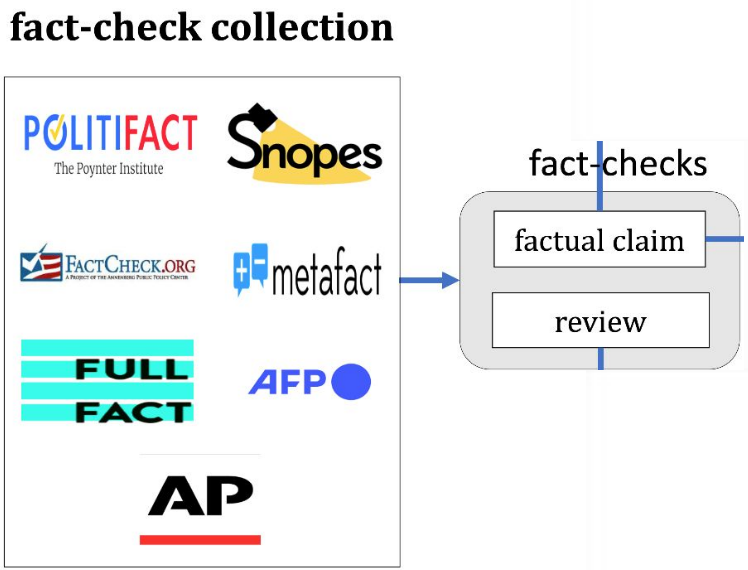
Tweet: They don't even try to hide their racism anymore! Remember his name at Election time! GOP Rep. Paul Gosar asks Capitol Police to arrest undocumented immigrants at State of the Union.

Data Collection

- We extracted factual claims and reviews from various trust-worthy websites.

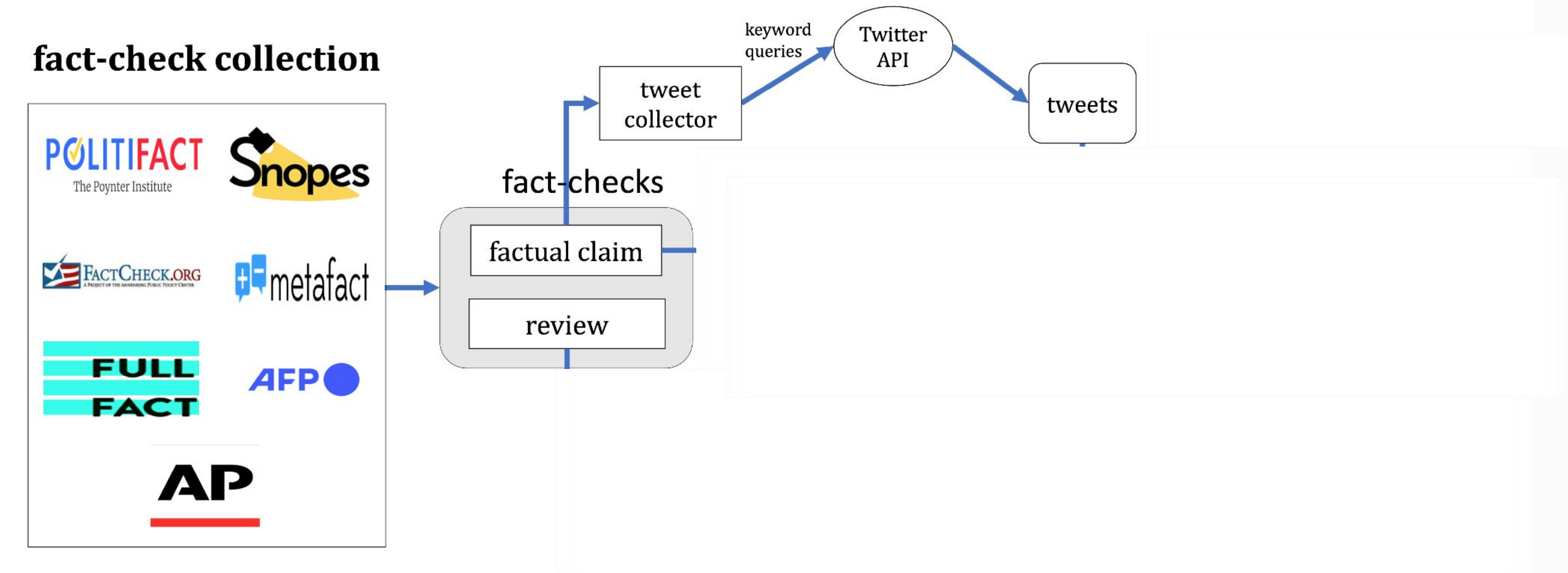


	Politifact	Snopes	Metafact.io	Fullfact	Factcheck.afp	Factcheck.org	Apnews
Number of fact-checks	21,023	18,491	3,429	2,784	4,318	3,453	226



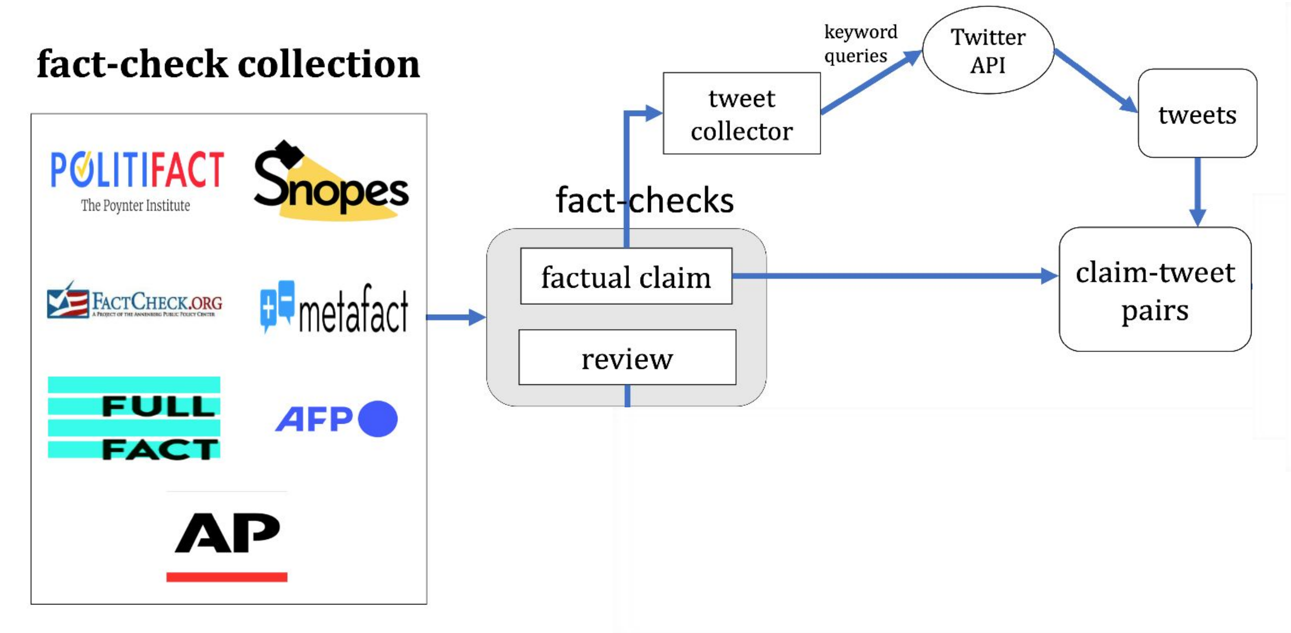
Data Collection

- We fetched claim-related English tweets using keyword queries to Twitter API.



Data Collection

- We annotated claim-tweet pairs by five options:
  Positive, Negative, Neutral, Unrelated, Invalid.
- The claim-tweet pairs are used for training stance detection model which take claim-tweet pair as input and generate the stance prediction.



Methodology

- Further pretraining language model
  - Language models are **large neural networks** that are used in a wide variety of natural language processing tasks.

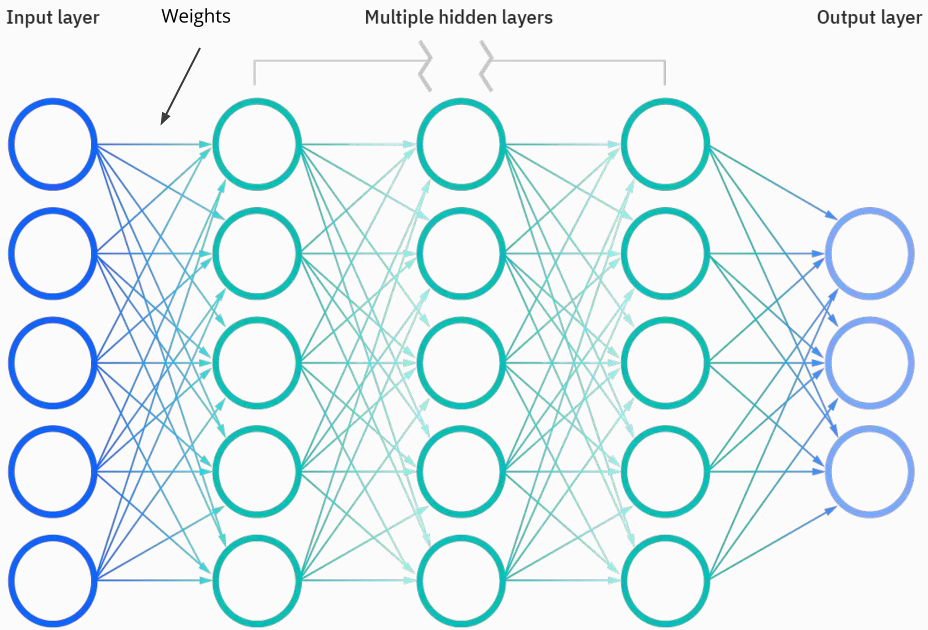


Image source: https://www.ibm.com/cloud/learn/neural-networks

Methodology

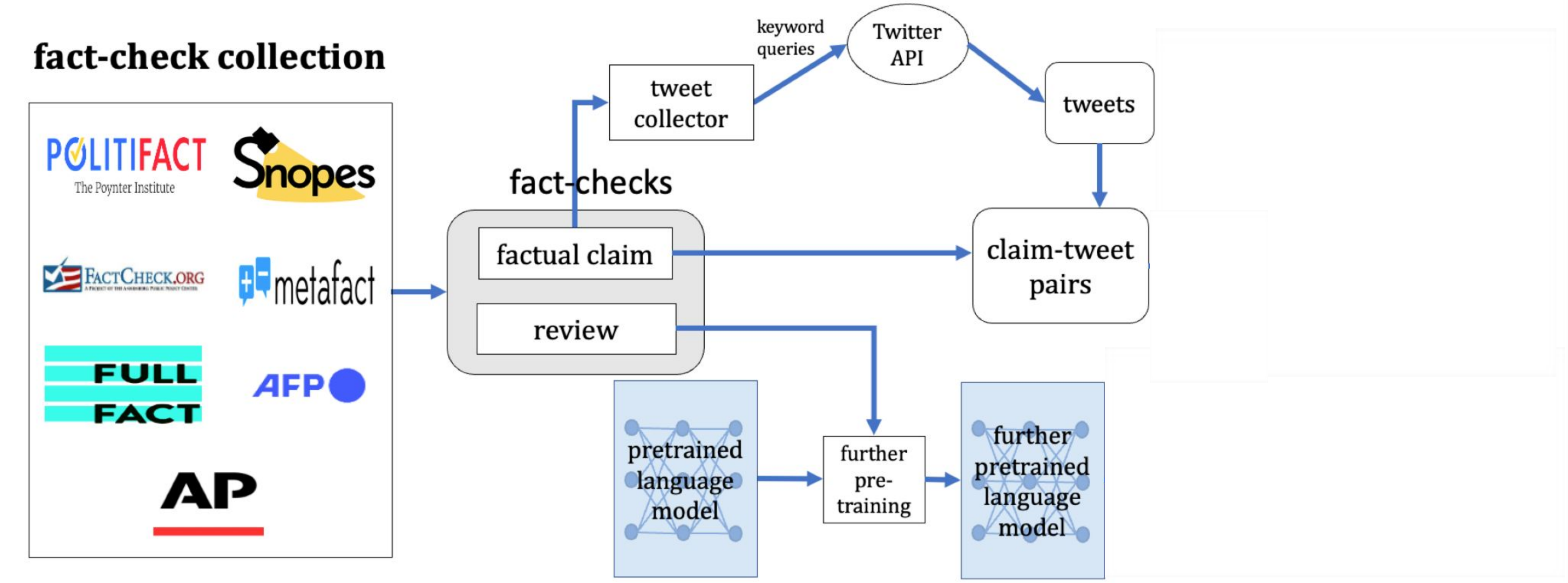
- Further pretraining language model
    - Language models are **large neural networks** that are used in a wide variety of natural language processing tasks.
    - Language models that have been trained with a large corpus still **don’t have domain knowledge about journalism**.

Methodology

- Further pretraining language model
  - Language models are **large neural networks** that are used in a wide variety of natural language processing tasks.
  - Language models that have been trained with a large corpus still **don’t have domain knowledge about journalism**.
  - Further pretrain language model on fact-check reviews can guide language models to **comprehend the semantics of factual claims.**

Methodology

- Further pretraining language model on fact-check reviews

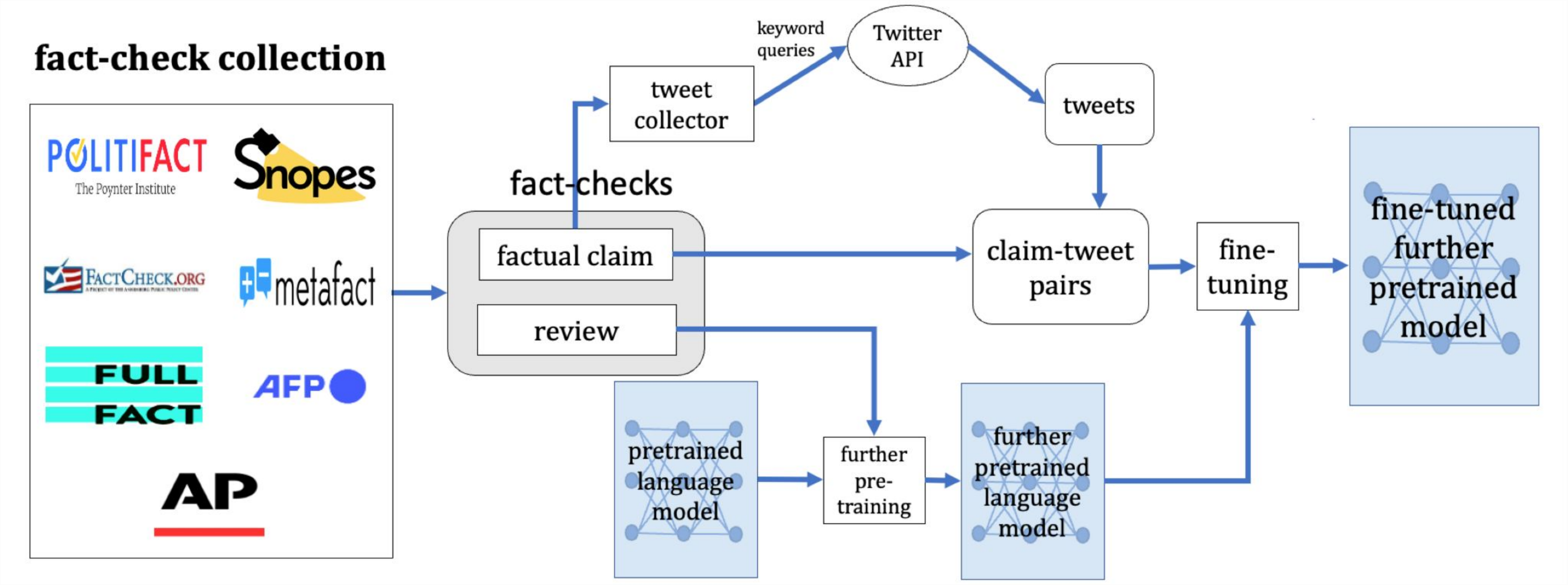


Methodology

- Fine-tune language model
  - Fine-tune refers to the procedure of **re-training a language model** on annotated claim-tweet pairs.
  - **The weights of the language model are updated** to account for the claim-tweet pairs' characteristics of the stance detection.

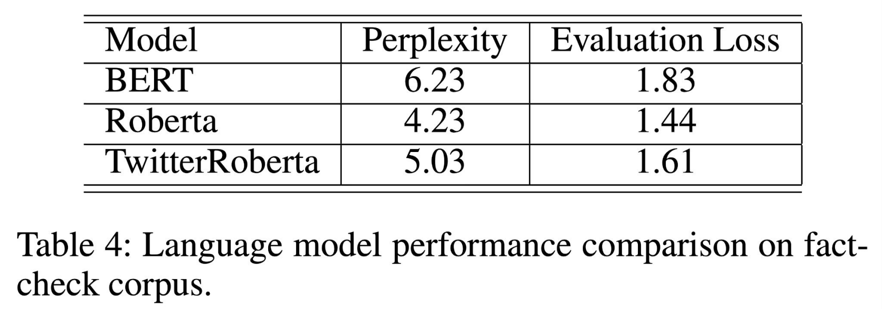
Methodology

- Fine-tune language model on annotated claim-tweet pairs



Model Evaluation: Further pretrain

- Roberta outperformed BERT and TwitterRoberta in predicting tokens in fact-checks by two evaluation metrics: perplexity (the lower the better) and loss (the lower the better).

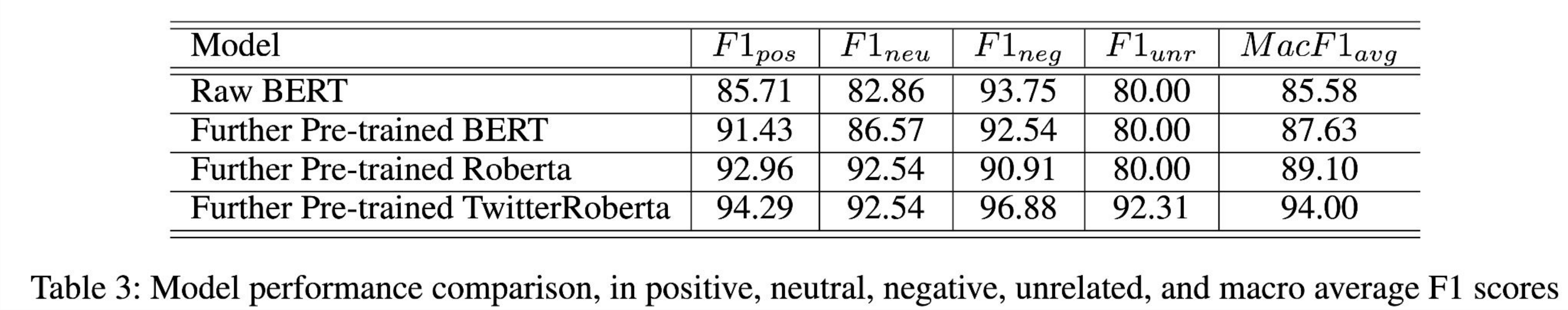


Model	Perplexity	Evaluation Loss
BERT	6.23	1.83
Roberta	4.23	1.44
TwitterRoberta	5.03	1.61

Table 4: Language model performance comparison on fact-check corpus.

Model Evaluation: Stance detection

- TwitterRoberta outperformed all other further pre-trained language models in predicting truthfulness stance by F1 score.

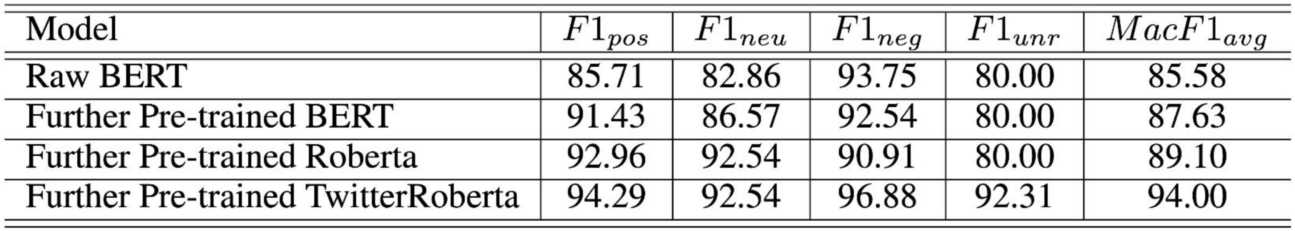


Model	$F1_{pos}$	$F1_{neu}$	$F1_{neg}$	$F1_{unr}$	$MacF1_{avg}$
Raw BERT	85.71	82.86	93.75	80.00	85.58
Further Pre-trained BERT	91.43	86.57	92.54	80.00	87.63
Further Pre-trained Roberta	92.96	92.54	90.91	80.00	89.10
Further Pre-trained TwitterRoberta	94.29	92.54	96.88	92.31	94.00

Table 3: Model performance comparison, in positive, neutral, negative, unrelated, and macro average F1 scores

Model Evaluation: Stance detection

- TwitterRoberta outperformed all other further pre-trained language models in predicting truthfulness stance by F1 score.
  - F1 score considers both
    - how many of the predictions are correct.
    - how many of the class samples are correctly predicted



Model	$F1_{pos}$	$F1_{neu}$	$F1_{neg}$	$F1_{unr}$	$MacF1_{avg}$
Raw BERT	85.71	82.86	93.75	80.00	85.58
Further Pre-trained BERT	91.43	86.57	92.54	80.00	87.63
Further Pre-trained Roberta	92.96	92.54	90.91	80.00	89.10
Further Pre-trained TwitterRoberta	94.29	92.54	96.88	92.31	94.00

Table 3: Model performance comparison, in positive, neutral, negative, unrelated, and macro average F1 scores

Conclusion

- We formulated the concept of truthfulness stance detection.
- We proposed a novel stance detection model which is based on additional pre-training of pre-trained language models using fact-checks.
- We created an API to allow for programmatic and large-scale usage of the tool.