Detecting Stance of Tweets Toward Truthfulness of Factual Claims

Zhengyuan Zhu, Zeyu Zhang, Foram Patel, Chengkai Li
The University of Texas at Arlington
Engineering Research Building, 414
500 UTA Blvd, Arlington, TX 76010
zhengyuan.zhu@mavs.uta.edu, zeyu.zhang@mavs.uta.edu, fxp3176@mavs.uta.edu, cli@uta.edu

Abstract

Journalists aim to understand misinformation on social media, especially in discerning the public's opinions toward the veracity of misinformation. For that, an algorithmic tool for *truthfulness stance detection* can be particularly useful. This paper introduce a deep learning model we developed for detecting the stance of tweets toward the truthfulness of factual claims. The models were constructed using a dataset curated and annotated in-house. While both the models and datasets warrant further development and refinement, preliminary experiments demonstrated promising results. The model is available through both an Application Programming Interface (API) and a demonstration website.

Introduction

Everyone can easily express their opinions on social media. Journalists are working hard on comprehending and mitigating misinformation on social media in order to reduce its harm. Such comprehension entails thoroughly discerning the public's opinions toward misinformation. This includes *stance detection*—determining whether a user supports or refutes a piece of misinformation based on their tweets. However, due to inherent personal bias and constraints on time and resources, it is virtually difficult for anyone to extensively comprehend tweets' stance, without the help of automation.

This paper introduces a deep learning model we developed for detecting the stance of tweets toward the truthfulness of factual claims. Prior research in related areas has focused on a piece of text's stance toward a subject (e.g., politicians in political debates [Lai et al. 2018], mergers and acquisitions [Conforti et al. 2020]), a topic (e.g., vaccination [Bechini et al. 2020], Brexit [Grčar et al. 2017], and other topics [Mohammad, Sobhani, and Kiritchenko 2017]), or a tweet on which the aforementioned text remarks [Derczynski et al. 2017, Gorrell et al. 2019]. The emphasis of our study is different in two aspects: The *target* of stance is a factual claim rather than another tweet, a topic, or a subject; the stance refers to a Twitter user's belief or assertion regarding the *truthfulness* of a factual claim, rather than the user's sentiment or emotion toward the claim.

Aiming at building a robust deep learning model for truthfulness stance detection, our work explores two core ideas. *First*, our model construction starts with widely-used pre-trained language models such as BERT [Devlin et al. 2018], Roberta [Liu et al. 2019], and TwitterRoberta [Barbieri et al. 2020], and we further pre-trained these models using contents of fact-checks. The premise is that in this way the further pre-trained models learn about knowledge related to truthfulness from fact-checks. Such knowledge shall become valuable in discerning truthfulness stance. *Second*, we fine-tuned these pre-trained models using a dataset of 600 pairs of tweets and factual claims (also from fact-checks). The dataset was collected and annotated in-house with ground-truth labels regarding the tweets' stance toward the corresponding factual claims' truthfulness. The results of evaluating our models show that they outperformed baseline models. A demonstration of our stance detection model is publicly available at https://idir.uta.edu/stance_detection.

C+J '22, June 9–11, 2022, New York, NY

Related Work

Stance detection is a classification task in natural language processing that aims to detect whether a piece of text is in favor of a given target or against it. Representative prior studies in this area can be contrasted with ours in two ways, as follows.

Sentiment vs. Truthfulness

Studies such as [Mohammad, Sobhani, and Kiritchenko 2017, Lai et al. 2018, Bechini et al. 2020, Grčar et al. 2017] seek to determine a piece of text's sentiment toward a target. The sentiment refers to subjective opinion and emotion in support of or against the target, expressed through the text. Hence we call it *sentiment stance detection*. On the contrary, our study focuses on *truthfulness stance detection* which examines whether the text's author believes the factual claim is true or not, which is orthogonal to sentiment. For instance, consider the factual claim *"Donald Trump once suggested that he would try to negotiate down the national debt."* and a related tweet *"Like he failed businesses he will crash country to try and negotiate debt. Meanwhile, good people are suffering."* Although the sentiment of the tweet is negative, it has positive stance regarding the truthfulness of the claim in that it shows the Twitter user believes the claim is credible.

In the literature, very few have studied truthfulness stance detection, with two exceptions [Derczynski et al. 2017, Con-

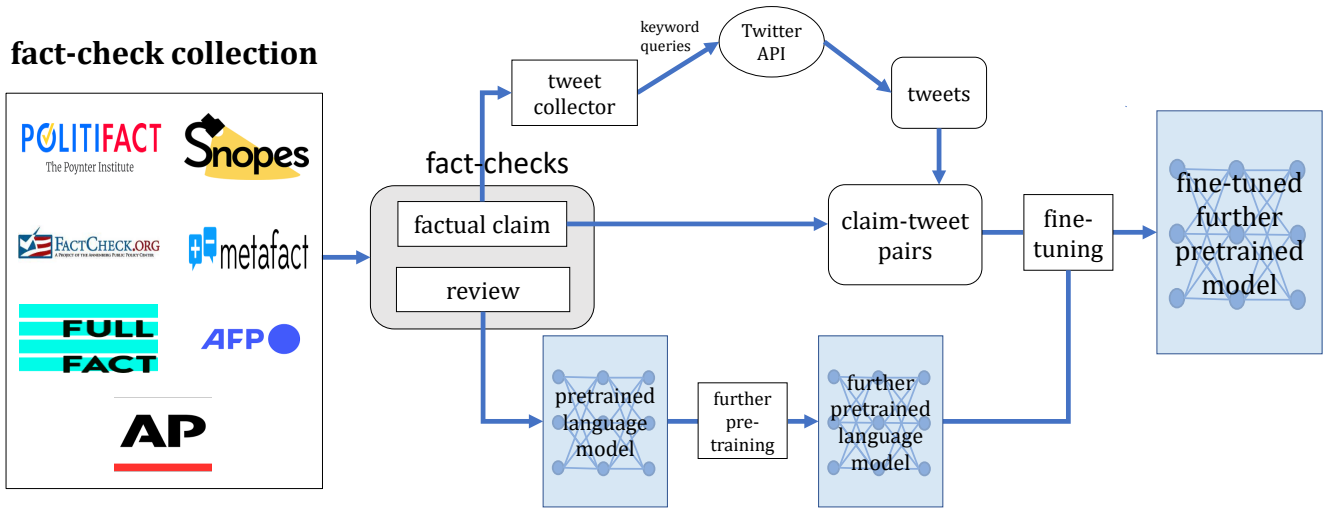

Figure 1: Overview of methodology

forti et al. 2020]. However, instead of stance toward factual claims, their focus is on stance toward tweets [Derczynski et al. 2017] or merger/acquisition events [Conforti et al. 2020].

Topic vs. Claim

As discussed in Introduction, most prior studies focused on stance towards subjects or topics. The exceptions are [Derczynski et al. 2017, Gorrell et al. 2019] in which the target of the stance can be a rumorous claim in the form of a tweet or a Reddit post. Our study focuses on general factual claims, which means the claim can be either a fact or misinformation.

Methodology

Problem Formulation

Our model aims to predict the truthfulness stance of a given tweet toward a factual claim. Formally, $t = [t_1, t_2, ..., t_n]$ denotes the textual content of a tweet where each t_i is a token, and $f = [f_1, f_2, ..., f_n]$ represents a factual claim where each f_i is also a token. Our model predicts the truthfulness stance $\hat{s} = F(t, f)$ of a tweet t towards a factual claim f, where $\hat{s} \in \{Positive, Neutral, Negative, Unrelated\}$.

Overview of our method

Figure 1 depicts the overall workflow of our method. It uses a collection of 53,724 fact-checks from several widely-used fact-checking websites. Each fact-check contains a factual claim and the corresponding review (i.e., detailed analysis of the claim). The component of *tweet collector* formulates queries using keywords from the claims and obtains tweets related to the claims by sending the queries to Twitter API. The claims and the corresponding tweets thus form our dataset of *tweet-claim pairs*, which are annotated in-house with ground-truth labels. The reviews of the fact-checks are used to carry out additional pre-training on pre-trained language models such as BERT [Devlin et al. 2018], Roberta [Liu et al. 2019], and TwitterRoberta [Barbieri et al. 2020]. Lastly, we created the final truthfulness stance detection models by fine-tuning the further pre-trained language models using the annotated tweet-claim pairs.

Further Pre-training on Fact-check Corpus

The purpose of performing further pre-training is to guide a model to comprehend the semantics of fact-checks so that it can better understand factual claims in the stance detection task. To the best of our knowledge, no prior studies have performed pre-training of language models using fact-checks. We applied masking tokens in fact-checks to further pre-train language models. [Devlin et al. 2018, Liu et al. 2019, Barbieri et al. 2020]. More specifically, given a document $D = [d_1, d_2, ..., d_n]$ where d_i is a token, we randomly select tokens based on a pre-defined probability p to replace the tokens with a special [MASK] token. Thus, we have the following token sequence input to a language model:

$$[CLS, d_1, d_2, [MASK], ..., d_{n-2}, [MASK], d_n]$$

The further pre-training is all about asking the language model to predict the selected tokens. In the hyper-parameters setting of the masked language model, we used a block size of $n = 512$, $p = 0.15$ as the probability to randomly mask tokens in the input, and 2 epochs in training models.

Data Collection

Our data collection consists of three main parts. We collected fact checks from various fact-checking websites. We also collected tweets using Twitter's API along with factual claim keywords as the queries. We manually annotated 533 tweet-claim pairs for model training and evaluation.

Fact-check Collection In order to build a stance detection model that has the ability to learn contextual knowledge from fact-checks, we collected data from seven well-known websites: Politifact, Snopes, Metafact.io, Fullfact,

	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
Claim	✓	✓	✓	✓	✗	✗	✓
Review	✓	✓	✓	✓	✓	✓	✓

Table 1: Number of fact-checks from each website

Stance	Positive	Neutral	Negative	Unrelated
Number of pairs	161	157	168	47

Table 2: The statistics of the annotated dataset.

Factcheck.afp, Factcheck.org, and Apnews. Table 1 provides the number of fact-checks collected from each website. The team "claim" refers to the factual claim vetted in a fact-check, whereas "review" is the fact-check body which contains the detailed analysis and background context. We collected a total of 53,724 fact-checks from all seven website, from which we randomly selected a subset of 5,764 fact-checks for this study. Note that Factcheck.afp and Factcheck.org do not provide explicit claims, while reviews from these two websites were still used in additional pre-training of models.

Tweet-claim Pairs Using the aforementioned 5,764 factual claims, we collected related English tweets using keyword queries to Twitter API. For each claim, we generated two queries: one using nouns and adjectives from the claim, and the other using nouns and verbs. For instance, consider the claim *"The economy was dead in the water when we got here. Virtually no jobs created."* The first query would be "economy dead water jobs" and the second query would be "economy water jobs got created".

From the tweets returned from Twitter API, we dropped retweets, quoted tweets, and replies, in order to avoid duplicates and unexpected contextual information. We also applied the longest contiguous matching subsequence (LCS) algorithm [Ratcliff and Metzener 1988] to filter out tweets that are almost identical to either already collected tweets or the factual claims. Such almost-identical tweet-claim pairs shall be handled as special cases in real-world applications based on empirical rules instead of using machine learning models. To ensure that the tweets are up to date, we only kept tweets posted within 100 days of the factual claim being made. Furthermore, we eliminated tweets shorter than 30 characters, since many of such extremely short tweets are not even comprehensible. By these measures, we obtained a total of 18,088 tweet-claim pairs.

Annotate Tweet-Claim Pairs For the dataset annotation process, an annotator was provided a tweet-claim pair and asked to evaluate whether the tweet author believes the factual claim is true or false based on the tweet content. More specifically, the annotator was asked to select one of the following five options: The tweet author believes the factual claim is true (Positive); The tweet author believes the factual claim is false (Negative); The tweet does not express positive or negative stance toward the veracity of the factual claim (Neutral); The tweet discusses a topic that is unrelated to the factual claim (Unrelated); The tweet-claim pair is problematic (Invalid), e.g., non-English or incomprehensible tweets or the claim is only a noun-phrase or a question. Out of the aforementioned 18,088 tweet-claim pairs, we annotated a total of 600 pairs. Each of these pairs is annotated by the same three experts from our team. We only kept those tweet-claim pairs that are labeled the same by at least two annotators. We also excluded the pairs on which the agreed-upon label is Invalid because such pairs do not contribute to constructing stance detection models. All these measures led to a total of 533 ground-truth pairs. The statistics of the annotated dataset are listed in Table 2.

Evaluation

To evaluate the effectiveness of our stance detection model, we conducted two experiments. We used three pre-trained language models, including BERT [Devlin et al. 2018], Roberta [Liu et al. 2019] and Twitter-Roberta [Barbieri et al. 2020], as backbones for evaluation. The rest of this section details the experiment setup and results.

Performance of Further Pre-training of Pre-trained Language Models

We evaluated how correctly language models can predict the tokens in fact-checks using the perplexity metric [Brown et al. 1992]:

$$PPL(X) = \exp\left\{-\frac{1}{t}\sum_{i}^{t}\log p_\theta\left(x_i \mid x_{<i}\right)\right\},$$

where $\log p_\theta\left(x_i \mid x_{<i}\right)$ denotes the log-likelihood of the i-th token conditioned on the preceding tokens $x_{<i}$ according to the language model. And the overall further pre-train objective is defined as minimizing the cross entropy loss.

Table 4 shows the language model performance comparison on fact-check corpus. It can be observed that the Roberta model achieved the smallest perplexity and evaluation loss. Hence, it outperformed other models in predicting masked tokens. In other words, it is more capable of learning the language features of fact-check.

Performance of Truthfulness Stance Detection

We compared the vanilla BERT model without further pre-pretraining (denoted as *Raw BERT*) and three other further pre-trained language models (denoted as *Further pre-trained BERT*, *Further pre-trained Roberta*, and *Further pre-trained TwitterRoberta*). We fine-tuned all these models and evaluated their performance, using the 533 annotated tweet-claim pairs. The performance metrics

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

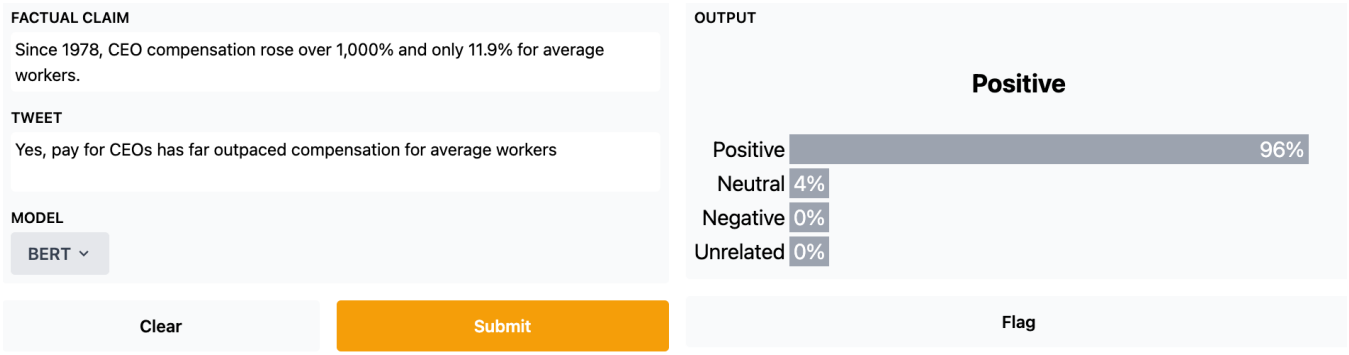

Figure 2: The user interface of the stance detection model

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.

are F1 scores for individual stance classes, denoted as $F1_{pos}, F1_{neg}, F1_{neu}, and F1_{unr}$, as well as the macro-average of F1 scores, denoted as $MacF1_{avg}$. The results are shown in Table 3. TwitterRoberta outperformed BERT and Roberta on every F1 score measure, although it did not achieve the best performance in the further pre-training stage. The reason could be that TwitterRoberta is better at handling tweets, whereas the models in the further pre-training stage dealt with fact-checks.

Stance Detection API

On top of the trained models, we created an application programming interface (API) using the gradio framework [Abid et al. 2019]. The API is available at https://idir.uta.edu/stance_detection/api/predict/ and a demonstration of the API is available at https://idir.uta.edu/stance_detection. Figure 2 illustrates the GUI of the demonstration website. The API takes a factual claim and a tweet from users as inputs, and it returns as output the tweet's truthfulness stance towards the factual claim. This demonstration accepts manual input. For using the stance detection tool on a large volume of tweet-claim pairs, requests can be posted in JSON format to the API.

Conclusion and Perspectives

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

On direction of our future work is to generate a larger dataset through more extensive data annotation. This could lead to more accurate and general models. In addition, we will work on developing a comprehensive truthfulness stance detection system, focusing on special cases that may not be effectively addressed by machine learning models, as well as separate machine learning tasks (e.g., detecting unrelated tweets given a factual claim, and auto-filtering problematic tweet-claim pairs).

The stance detection model and tool will become a key component in a misinformation surveillance system that we aim to build. Particularly, it can help understand how social media users respond to different types of misinformation, and how such insights may aid debunking misinformation. Several Twitter datasets have already been collected, all of which are related to major events such as the 2020 Presidential Election Fraud, the Cuomo Scandal, and Anti-Vaccine. Using our stance detection technique, we would be able to assess such events in great detail. We hope and expect to see our approach put into practice in the real world to help create a better online environment.

References

Abid, A.; Abdalla, A.; Abid, A.; Khan, D.; Alfozan, A.; and Zou, J. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569* .

Barbieri, F.; Camacho-Collados, J.; Neves, L.; and Espinosa-Anke, L. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421* .

Bechini, A.; Ducange, P.; Marcelloni, F.; and Renda, A. 2020. Stance analysis of twitter users: the case of the vaccination topic in italy. *IEEE Intelligent Systems* 36(5): 131–139.

Brown, P. F.; Della Pietra, S. A.; Della Pietra, V. J.; Lai, J. C.; and Mercer, R. L. 1992. An estimate of an upper bound for the entropy of English. *Computational Linguistics* 18(1): 31–40.

Conforti, C.; Berndt, J.; Pilehvar, M. T.; Giannitsarou, C.; Toxvaerd, F.; and Collier, N. 2020. Will-They-Won't-They: A Very Large Dataset for Stance Detection on Twitter. *arXiv preprint arXiv:2005.00388* .

Derczynski, L.; Bontcheva, K.; Liakata, M.; Procter, R.; Wong Sak Hoi, G.; and Zubiaga, A. 2017. SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 69–76. Vancouver, Canada: Association for Computational Linguistics. doi:10.18653/v1/S17-2006. URL https://www.aclweb.org/anthology/S17-2006.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Gorrell, G.; Kochkina, E.; Liakata, M.; Aker, A.; Zubiaga, A.; Bontcheva, K.; and Derczynski, L. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, 845–854.

Grčar, M.; Cherepnalkoski, D.; Mozetič, I.; and Kralj Novak, P. 2017. Stance and influence of Twitter users regarding the Brexit referendum. *Computational social networks* 4(1): 1–25.

Lai, M.; Patti, V.; Ruffo, G.; and Rosso, P. 2018. Stance evolution and twitter interactions in an italian political debate. In *International Conference on Applications of Natural Language to Information Systems*, 15–27. Springer.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* .

Mohammad, S. M.; Sobhani, P.; and Kiritchenko, S. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)* 17(3): 1–23.

Ratcliff, J. W.; and Metzener, D. E. 1988. Pattern-matching-the gestalt approach. *Dr Dobbs Journal* 13(7): 46.