Zero-Shot Stance Detection on Tweets

Keywords: Stance Detection, Large Language Models, Zero-Shot, NLP, Tweets

The stance or opinion of an individual towards a target (policy, issue, product, individual, etc.) informs fields of research like social science, psychology, cognitive science, linguistics, and network science. For instance, in the space of contradicting web search results to the effectiveness of a medicine to treat a disease, stance detection can help summarise the results to indicate its efficacy.

While surveying individuals is a traditional method to detect stance, it falls short in scalability and cost. Surveys can also produce biased results due to the tendency that people to report more socially acceptable positions even in anonymized surveys. This may not only provide access to larger populations but can also produce more ecologically valid results. However, inferring the stance of individuals using social media text is a challenging problem given factors like incoherency on social media posts, abbreviated writing, spelling errors, presence of hyperlinks, hashtags, and usernames.

Many existing methods followed the supervised learning paradigm—identifying linguistic features from a curated *sentence-target-stance* dataset to predict the stance of a given sentence towards a particular pre-selected target—an expensive process [1]. These methods, in general, struggle to infer stance towards targets not seen during training.

The rise of large language models (LLMs) like BERT—self-supervised models pre-trained on web-based text corpora with the objective of performing simpler tasks like next-word prediction or mask-filling [2] —birthed the paradigm of *finetuning* LLMs. This paradigm further trained the pre-trained LLMs on task-specific training data. Finetuned LLMs for stance detection currently show state of the art (SOTA) performance (Fig. 1b, BERT) on the SemEval 2016 Task 6 dataset curated to detect stance ("favor", "against", or "none") towards pre-selected targets—"Atheism (AT)", "Hillary Clinton (HC) ", "Legalization of Abortion (LA)", "Climate Change is a real concern (CC)", "Feministic Movement (FM) " (Task 6A), and "Donald Trump (DT)" (Task 6B)—from related tweets [3]. This paradigm eliminated the need for extensive feature engineering, and could handle unseen targets (as it is likely to have been seen in pre-training). However it requires large training data and computing resources.

More recently, the generative family of LLMs like GPT, T5, etc. enabled a completely different approach—in-context inference. Rather than finetuning LLMs, input examples (Fig. 1a, orange) are shown as a guide for it to produce the desired *kind* of outputs of a downstream task. The performance of this approach is competitive with the supervised learning paradigm, requires very few labeled data (approx. 10-20 examples), and is also not overly sensitive to the accuracy of labeling [4]. An extreme case of in-context inference is zero-shot inference (ZSI) in which no input examples are used to guide the model (Fig. 1a, black) —a highly desirable, naturalistic, but challenging setting. In ZSI, the LLM is treated as a human, and can directly be asked a question. For instance, given a social media post "Vaccination is safe." the LLM can be asked "What is the stance of 'Vaccination is safe' towards vaccines?" to which an ideal LLM will reply "Favor." This response is based on the *understanding* of words gained by LLMs during pre-training which is *transferred* over to a downstream task like stance detection. In addition to requiring no training data ZSI, works on significantly fewer computing resources, and user-friendly platforms like hugging face allow practitioners to use it with as little as three lines of code. The focus of this work is to evaluate how well ZSI performs on stance detection, particularly on the SemEval 2016 Task 6 dataset on which many previous models have been evaluated.

We use FlanT5, an open-source instruction-tuned model by Google [5]. Instruction-tuned models are the latest breed of LLMs created, among other reasons, to make LLMs more adherent to human instructions (Fig. 1a, boldface black). The performance of FlanT5 on the stance detection task in a zero-shot setting, with and without pre-processing of tweets is shown in Fig. 1b. Similar to the implementation in finetuned BERT [6], pre-processing of tweets involves case-folding, expanding abbreviations, and splitting hashtags. The numbers are the F scores for each target. AMF1 is the macro-average of the F score for 'favor' and 'against' classes in Task 6A. For FlanT5, the F score is obtained by averaging the scores across 10 instructions which are paraphrases to "Please answer the question with True or False." outputted by ChatGPT, similar in spirit to the approach in Zhou 2022 [7].

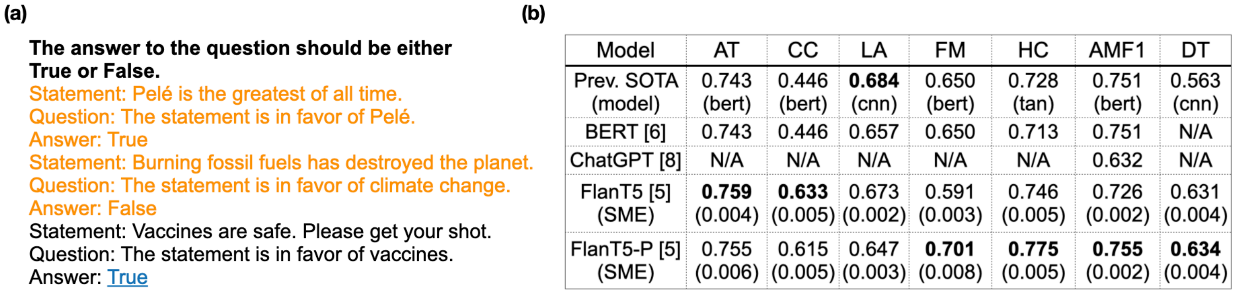
The results show that zero-shot stance detection using FlanT5 outperforms finetuned BERT—the previous SOTA. More impressively, for the target "Donald Trump" (Task 6B) for which no training data is available, the previous SOTA AMF1 is significantly lesser than that of FlanT5 which does not need training data in a zero-shot setting (Fig 1b, rightmost column). FlanT5 also outperforms ChatGPT in a zero-shot setting and this *may* be because ChatGPT is a decoder-only LLM particularly useful as a conversational system while FlanT5 is an encoder-decoder LLM which is more beneficial for NLP tasks.

To ensure that stance detection is a truly zero-shot task for FlanT5, a preliminary inspection of the training data of FlanT5 revealed no signature of the SemEval 2016 Task 6 data. The same cannot be said about ChatGPT since its training data is not public and recent work inferred that ChatGPT may have seen the SemEval 2016 Task 6 dataset during its training [8].

In sum, the benefits of in-context learning, particularly zero-shot inference combined with its impressive performance shown on the stance detection task expose the potential of instruction-tuned open-source LLMs like FlanT5 for data-driven social media research.

References

1. Abeer AlDayel and Walid Magdy. "Stance detection on social media: State of the art and trends". *Information Processing & Management 58.4* (2021)
2. Qiu, Xipeng, et al. "Pre-trained models for natural language processing: A survey." *Science China Technological Sciences* 63.10 (2020)
3. Saif Mohammad et al. "Semeval-2016 task 6: Detecting stance in tweets". *Proceedings of the 10th international workshop on semantic evaluation* (2016)
4. Min, Sewon, et al. "Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?." *arXiv preprint arXiv:2202.12837* (2022).
5. Chung, Hyung Won, et al. "Scaling instruction-finetuned language models." *arXiv preprint arXiv:2210.11416* (2022)
6. Schiller, Benjamin, Johannes Daxenberger, and Iryna Gurevych. "Stance detection benchmark: How robust is your stance detection?." *KI-Künstliche Intelligenz* (2021)
7. Zhou, Yongchao, et al. "Large language models are human-level prompt engineers." *arXiv preprint arXiv:2211.01910* (2022).
8. Kocoń, Jan, et al. "ChatGPT: Jack of all trades, master of none." *arXiv preprint arXiv:2302.10724* (2023).



(a)

The answer to the question should be either True or False.
Statement: Pelé is the greatest of all time.
Question: The statement is in favor of Pelé.
Answer: True
Statement: Burning fossil fuels has destroyed the planet.
Question: The statement is in favor of climate change.
Answer: False
Statement: Vaccines are safe. Please get your shot.
Question: The statement is in favor of vaccines.
Answer: True

(b)

Model	AT	CC	LA	FM	HC	AMF1	DT
Prev. SOTA (model)	0.743 (bert)	0.446 (bert)	**0.684** (cnn)	0.650 (bert)	0.728 (tan)	0.751 (bert)	0.563 (cnn)
BERT [6]	0.743	0.446	0.657	0.650	0.713	0.751	N/A
ChatGPT [8]	N/A	N/A	N/A	N/A	N/A	0.632	N/A
FlanT5 [5] (SME)	**0.759** (0.004)	**0.633** (0.005)	0.673 (0.002)	0.591 (0.003)	0.746 (0.005)	0.726 (0.002)	0.631 (0.004)
FlanT5-P [5] (SME)	0.755 (0.006)	0.615 (0.005)	0.647 (0.003)	**0.701** (0.008)	**0.775** (0.005)	**0.755** (0.002)	**0.634** (0.004)

Figure 1. FlanT5, in a zero-shot setting, outperforms finetuned LLMs and previous models in the task of stance detection. (a) An example of in-context inference for stance detection on tweets ("Statement" indicates each tweet). The black and orange text can be the input to an LLM and the blue underlined text is the output of an ideal LLM. In a zero-shot setting, the focus of this work, the examples which guide the model (orange text) are not present. Boldface black text represents the instruction to an instruction-tuned LLM like FlanT5. This helps the model produce the right kind of output (say, True/False), especially in the absence of guiding examples (b) Performance (F scores) of various models on the stance detection task on the test sets of SemEval 2016 Task 6A and 6B. Task 6A comprises of targets for which training tweets are available—AT, CC, LA, FM, and HC. BERT, CNN, TAN models are all finetuned/trained on this training dataset. In contrast, Task 6B—DT—has no training data. The last two rows indicate the performance of FlanT5 in a **zero-shot setting** (has not used any task specific training data) with and without pre-processing (P) of tweets. AMF1 is the macro-average F score calculated by pooling together the tweets in the test sets of Task 6A. SME is the standard mean error on FlanT5 across the 10 instruction paraphrases. Boldface numbers indicates new SOTA.