Zero-Shot Stance Detection using Contextual Data Generation with LLMs

Ghazaleh Mahmoudi, Babak Behkamkia, Sauleh Eetemadi

School of Computer Engineering, Iran University of Science and Technology, Iran
{gh_mahmoudi, babak_behkamkia}@comp.iust.ac.ir, sauleh@iust.ac.ir

Abstract

Stance detection, the classification of attitudes expressed in a text towards a specific topic, is vital for applications like fake news detection and opinion mining. However, the scarcity of labeled data remains a challenge for this task. To address this problem, we propose **Dy**namic **Mo**del **Adap**tation with Contextual Data Generation (DyMoAdapt) that combines Few-Shot Learning and Large Language Models. In this approach, we aim to fine-tune an existing model at test time. We achieve this by generating new topic-specific data using GPT-3. This method could enhance performance by allowing the adaptation of the model to new topics. However, the results did not increase as we expected. Furthermore, we introduce the **M**ulti **G**enerated **T**opic VAST (MGT-VAST) dataset, which extends VAST using GPT-3. In this dataset, each context is associated with multiple topics, allowing the model to understand the relationship between contexts and various potential topics.

Introduction

With the growth of social media platforms, an increasing number of individuals turn to platforms such as Twitter for news consumption. Consequently, automatically identifying the opinions expressed in news articles and by people regarding specific topics (Stance Detection) has become a pressing issue in the field of Natural Language Processing (NLP) (Kaushal, Saha, and Ganguly 2021). Initially, some researchers assumed that the topics encountered during the testing phase would align with those seen during training (Mohammad et al. 2016).

However, in reality, the number of topics is virtually limitless. Unfortunately, collecting a comprehensive dataset that encompasses all possible topics is impractical. As a result, researchers have explored alternative approaches such as zero-shot and few-shot learning, as well as the extraction of topic-invariant features (Allaway, Srikanth, and McKeown 2021). Nonetheless, none of these methods can outperform real data examples that explicitly convey the stance of a given post regarding a specific topic.

In recent years, large language models (LLMs), like GPT-3 (Brown et al. 2020a) and LLaMA (Touvron et al. 2023), have brought a revolution to the NLP field. These models have been trained on large amounts of data and can understand and generate human-like text. Hence, LLMs hold significant potential for generating synthetic data that closely resembles real data.

By combining the concept of using LLMs with few-shot learning, we introduce a novel method for dataset generation. Additionally, to overcome the challenge of unseen topics in the testing phase, we suggest a novel approach. Our main contributions can be divided into two parts.

- We introduced MGT-VAST, a new dataset generated by GPT-3 from VAST (Allaway and McKeown 2020a), where each context is paired with multiple topics. The underlying idea is to assist the model in comprehending the relationship between a context and various possible topics. In comparison to the VAST Dataset, MGT-VAST contains a greater number of unique topics.
- We propose the DyMoAdapt approach using LLMs in the test phase to enhance the performance of existing models. In this approach, we additionally fine-tune the model in the test phase with data generated by GPT-3 according to the given topic. After fine-tuning, the model would be ready for the actual test data and perform better because it has seen similar examples of a particular topic.

Related work

Deep learning methods have proven to be effective in solving NLP problems, especially Stance Detection; however, they perform poorly when the training dataset is limited. Recently, few-shot learning has been proposed as a solution to this problem. This learning paradigm aims to generalize to new tasks with limited training data (zero or few labeled examples) using prior knowledge. The lack of labeled samples makes the estimation of the loss value during model training more challenging, which is the key issue of few-shot learning (Hossain et al. 2022).

Allaway and McKeown (2020b) is the first publication in the few-shot stance detection field. This research presents a new dataset called VAST. They collected this data according to the gap with existing datasets that contain a limited number of topics. They also introduced a novel deep learning approach that focuses on generalization when only a limited amount of data is available for each topic.

Liu et al. (2021) uses the VAST dataset and introduces a new model to improve its generalization capabilities. An enhanced general knowledge module was introduced to exploit semantic and structural level information. In this model, knowledge is limited to the relationships between documents and topics.

Another study introduced a zero-shot model called **TO**pic-**AD**versarial Network (TOAD), which employs adversarial learning (Allaway, Srikanth, and McKeown 2021). They used domain-transfer ideas (Ganin and Lempitsky 2015) to produce topic-invariant representations, allowing the model to generalize to unseen topics.

Vamvas and Sennrich (2020) proposed a zero-shot model for generalizing across languages, in contrast to the previous work that focused on generalization across topics. They collected a dataset containing more than 150 political questions and 67k comments written by candidates. The comments comprise a mixture of German, French, and Italian. They fine-tuned a multilingual BERT model for stance detection.

Methodology

In this section, we first explain the concept of MGT-Vast data generation. Following that, we provide a detailed description of DyMoAdapt, the novel approach used during the test phase to address unseen topics.

MGT-VAST Dataset

In order to create the MGT-VAST dataset, we used GPT-3 to generate topics that are either in favor or opposing the existing posts in the VAST dataset. By using the prompt shown in Figure 1 The generated topics have lengths ranging from 2 to 4 words. Through this way, for each post, we have multiple topics, allowing the model to learn the relationship between each topic and the corresponding post. We have provided statistics for the MGT-VAST dataset, as shown in Table 1 and Table 4.

	Train	Test
# Examples	4986	2305
# Examples with Agree label	2516	1204
# Examples with Disagree label	2470	1101
# Unique Post	1233	563
# Unique Topics	4877	2293
# of Words in Topic	17655	8141
# of Unique Words in Topic	5890	3501
Average # of words per Topic	3.51	3.53

Table 1: MGT-VAST dataset Statistic

DyMoAdapt Approach

The intuition behind our proposed approach is to enhance the performance of existing models. This method operates as follows: For each new topic encountered during the test phase, we generate 2k new data points using GPT-3 (k samples in favor and k with opposing labels). We set k to 3 because of the limitations of the GPT-3 API. Figure 2 shows the prompt used to generate a synthetic post for the given topic. Afterward, the model is fine-tuned on the generated data. Then, with the fine-tuned model, predictions are made for the original test data.

	Train	Dev	Test
SemEval-T6	4870	-	1956
VAST	13477	2062	3006
MGT-VAST	4986	-	2305

Table 2: This table illustrates the number of instances in each proportion of SemEval2016-T6, VAST, and MGT-VAST (our generated dataset).

> List the most potential topics of the given post with their labels.
> a label can be ”agree” or ”disagree” only. try to find both labels.
> summarize each topic in 2 to 4 words.
> return a JSON in which topics are keys and labels are values:
> example: {{topic here: label here}}
> post: “‘{post}’”

Figure 1: The prompt for dataset generation via GPT-3, which gets a post as input and gives all the possible topics for each post.

> Your task is to generate a human written post. Do not mention that you are an intelligent assistant.
> The generated post must discuss the given topic at some point in itself.
> The generated post must have a stance toward this topic. the stance could be ”agree” or ”disagree”. the post should be at most 2 paragraphs. just return the post.
> topic: {topic}
> stance: {label}
> Here is an example post, but we do not know the stance of it toward the topic. you can learn from its structure.
> post: “‘{post}“‘

Figure 2: The prompt for synthetic post generation via GPT-3 in DyMoAdapt approach.

Model	VAST (%)				SEM2016-T6 (%)					
	Pro	Con	Neut	All	DT	HC	FM	LA	A	CC
BERT	**82.3**	65.3	16.68	37.2	33.5	**57.6**	62.1	**58.2**	55.0	58.6
TOAD	42.6	36.7	**43.8**	**41.0**	49.5	51.2	54.1	46.2	46.1	30.9
GPT-3	63.5	69.1	38.9	40.4	**53.7**	57.5	**62.3**	47.9	24.6	59.3
DyMoAdapt (3 labels)	58.8	**88.8**	6.4	35.8	27.0	38.2	21.2	29.3	**58.9**	**74.6**

Table 3: This table compares several baseline models with our pipeline. We report $F1_{macro}$ for all of our experiments (the average of F1 on pro and con). It is important to note for BERT's experiments we only used a simple approach by just using the main text as input for this classification task.

Experiments and Results

Our experiments consist of two main parts: first, the evaluation of the MGT-VAST dataset, and second, the investigation of the DyMoAdapt approach.

MGT-VAST Dataset Evaluation

For evaluating the MGT-VAST dataset, we chose three different models. Below, we describe the selected models and the experimental procedure:

i BERT: We selected BERT (Devlin et al. 2019), which is a transformer-based model. The input is provided in the format ”post [SEP] topic”.

ii GPT-3: LLMs that have recently been introduced and have demonstrated high performance in various tasks. In this experiment, we used the prompt shown in Figure 3 to instruct GPT-3 (Brown et al. 2020b) to determine the stance for the input post along with its topic. We utilized GPT3.5.Turbo and the OpenAI API.

iii TOAD: another selected model is TOAD (Allaway, Srikanth, and McKeown 2021), which utilizes Bicond LSTM(Augenstein et al. 2016) and adversarial learning.

Topic	Frequency
Charter schools	6
dual citizenship	5
Illegal Immigration	4
Declawing cats	4
Immigration	4

Table 4: Most Frequent Topic in MGT-VAST (Train).

The results obtained are displayed in Table 5. The analysis of the results on the MGT-Vast dataset demonstrates that the models have achieved promising outcomes. In particular, the BERT and GPT-3 models have outperformed TOAD in terms of $F1_{macro}$. It is worth noting that the results of BERT and GPT-3 are quite comparable, and there is no significant superiority of one over the other in a meaningful sense. As expected, considering that a part of the MGT-VAST dataset has been generated using LLMs, transformer-based models have better performance.

Model	Stance Label		
	Agree	Disagree	All
BERT	**68.5**	**81.3**	**60.0**
GPT-3	68.4	70.6	59.9
TOAD	56.8	47.5	52.2

Table 5: $F1_{macro}$ score of models on MGT-VAST dataset.

DyMoAdapt Approach Evaluation

We chose BERT, GPT-3, and TOAD models to evaluate DyMoAdapt using the VAST(Allaway and McKeown 2020b) and SemEval2016-T6(Mohammad et al. 2016) datasets. The procedure for obtaining the results of each model is explained as follows:

i BERT: we fined-tune BERT with a linear classification at the last layer on the training set for 10 epochs with a learning rate of 1e-5. Then, we evaluated the model on the test set of each dataset. The results are split through each topic (e.g., DT, HC, FM, LA, A, CC) for SemEval2016-T6 and for each stance label (e.g., Pro, Con, Neut) for VAST.

ii GPT-3: we aimed to perform the stance classification task using prompts (Figure3) with GPT3.5.turbo without fine-tuning due to our limited access to the OpenAI API.

iii TOAD: the results of the TOAD model are from Allaway, Srikanth, and McKeown (2021) research.

iv DyMoAdapt (3 labels): this experiment is similar to the BERT section in terms of training, with the main difference being in the testing phase. In the test phase, for each input, We asked GPT-3 to generate more posts according to the given topic using a prompt (Figure 2). Then, we fine-tuned BERT on the data generated by GPT-3 before performing the final predictions. This process is repeated for all samples in the test set.

post	Topics	Label
Without government to ensure their behavior, companies will attempt to make a profit even to the DETRIMENT of the society that supports the business. We have seen this in the environment, in finances, in their treatment of workers and customers. Enough.	Role of government Corporate behaviour Profit motive Environmental impact Worker treatment Customer treatment Social responsibility	Agree Disagree Disagree Agree Agree Disagree Agree
I have two serious issues with plug-in cars, local and national. Locally, NH has the highest electric rates in the country (thanks, Seabrook). Nationally, plugging in is a huge waste of energy. To get a gallon of gas worth of electricity out of a wall socket, we need to put at least three gallons into the generator. Every time energy changes state, there is a loss: fuel to heat, heat to steam, steam to mechanical energy, mechanical to electrical, and in the car, electrical to battery, battery to mechanical, mechanical to tires, tires to motion. Add that to transmission losses, and you end up with a lot of waste instead of savings. Plug-in cars will add to the problem, not solve it.	plugging in is a waste of energy plug-in cars will add to the problem high electric rates	Agree Disagree Disagree

Table 6: In the first two rows, we demonstrate the generated topics and their stances towards the given text, while the last two instances are examples of generated posts related to a given combination of topic and label. We utilized GPT-3 to generate these posts, topics, and labels.

In general, the DyMoAdapt method can be a suitable alternative for use in real-time applications and with real data compared to other methods, including GPT and TOAD. The results obtained are presented in Table 3.

What is the stance of the post which is delimited by triple backticks toward the given topic?
your answer should be agree, disagree or neutral. keep your answer 1 word long.
please double-check your answer before responding and be sure about it.
topic: {topic}
post: “‘{post}“‘

Figure 3: The prompt for stance detection using GPT-3

Discussion

The data generated by GPT-3, which receives the neutral label, usually doesn’t meet acceptable quality standards. Therefore, the performance of DyMoAdapt with three labels is weaker than DyMoAdapt with two labels. While the detection of the neutral label has consistently posed challenges for models, it is imperative to explore alternative methods for generating neutral data.

Conclusion

In this work, we proposed a new idea for the stance detection pipeline during the test phase called DyMoAdaptt, which almost improves performance on unseen topics. Using BERT with DyMoAdapt, achieved an average improvement of 24% in F1 score across DT, HC, A, and CC topics in SEMEval2016-T6. However, it is important to note a corresponding reduction in performance for other labels. Furthermore, we introduced the MGT-VAST dataset, which contains more than one topic for each post sample, generated using LLMs. The most important advantage of this dataset is the possibility of generating more samples.

For future work, Transformer-based models with attention layers can be employed to gain a deeper understanding of the relationship between topics and posts in the NLP domain. Moreover, other data augmentation methods, such as EDA can be used in DyMoAdapt to generate additional data and compare their results with the current approach.

Limitations

We generated topics with stances (”agree” or ”disagree”) towards a given text because we used GPT-3 to generate data, and GPT-3 couldn’t generate an acceptable quality text with a neutral stance toward a topic. However, the stance detection task has three labels (”pro”, ”con”, and ”neutral”). Thus, our proposed dataset and pipeline don’t perform well on instances with the ”Neutral” label. Moreover, our experiments include a small portion of the datasets because we had very limited access to GPT-3.

References

Allaway, E.; and McKeown, K. 2020a. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Allaway, E.; and McKeown, K. 2020b. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8913–8931. Online: Association for Computational Linguistics.

Allaway, E.; Srikanth, M.; and McKeown, K. 2021. Adversarial Learning for Zero-Shot Stance Detection on Social Media. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tur, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4756–4767. Online: Association for Computational Linguistics.

Augenstein, I.; Rocktäschel, T.; Vlachos, A.; and Bontcheva, K. 2016. Stance Detection with Bidirectional Conditional Encoding. In Su, J.; Duh, K.; and Carreras, X., eds., *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 876–885. Austin, Texas: Association for Computational Linguistics.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020a. Language Models Are Few-Shot Learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020b. Language Models are Few-Shot Learners. arXiv:2005.14165.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised Domain Adaptation by Backpropagation. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, 1180–1189. JMLR.org.

Hossain, M. E.; Kabir, M. A.; Zheng, L.; Swain, D. L.; McGrath, S.; and Medway, J. 2022. A systematic review of machine learning techniques for cattle identification: Datasets, methods and future directions. *Artificial Intelligence in Agriculture*, 6: 138–155.

Kaushal, A.; Saha, A.; and Ganguly, N. 2021. tWT–WT: A dataset to assert the role of target entities for detecting stance of tweets. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3879–3889.

Liu, R.; Lin, Z.; Tan, Y.; and Wang, W. 2021. Enhancing Zero-shot and Few-shot Stance Detection with Commonsense Knowledge Graph. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 3152–3157. Online: Association for Computational Linguistics.

Mohammad, S.; Kiritchenko, S.; Sobhani, P.; Zhu, X.; and Cherry, C. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, 31–41.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Vamvas, J.; and Sennrich, R. 2020. X-Stance: A Multilingual Multi-Target Dataset for Stance Detection. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*. Zurich, Switzerland.