Analele Universității din București. Informatică
Vol. LXIII, p. 93-113
ISSN 3045-2392; e-ISSN 3045-2406

Bianca-Ștefania MUȘAT[1],
Cornelia CARAGEA[2],
Florentina HRISTEA[3]

DATA CARTOGRAPHY BASED AUGMENTATION TECHNIQUES FOR STANCE DETECTION

Abstract. Stance detection is the task of determining whether the information conveyed in a text is against, neutral, or in favor of a particular target. Since there is a plethora of targets upon which one can adopt a position, one common challenge of the stance detection task is the scarcity of annotations. Conversely, the emphasis on data quantity frequently entails a compromise in terms of the quality of the data. To address both challenges, we propose two data augmentation techniques that leverage training dynamics – the model behavior on individual instances during training – to identify and combine data instances with properties that differ, triggering, for example, the improvement of the generalization capabilities of the model or the enhancement of its optimization process. The first data augmentation method uses training dynamics to generate additional virtual samples during model training by interpolating existing annotated samples with characteristics that differ. The second data annotation approach is defined as a conditional masked language modeling task that generates additional samples by predicting the masked words of the input sentence, conditioned not only on its context but also on an auxiliary sentence sampled based on its characteristics. We empirically validated that fine-tuning a pre-trained language model on a subset of the training data, such

[1] Quant Risk Analyst at London Stock Exchange Group. E-mail: < bianca-stefania.musat@s.unibuc.ro >.

[2] Full Professor in the Department of Computer Science at the University of Illinois at Chicago, USA, and Adjunct Associate Professor at Kansas State University, USA. E-mail: < cornelia@uic.edu >.

[3] Full Professor Univ. Dr. in the Department of Computer Science, at the University of Bucharest, Romania. E-mail: < fhristea@fmi.unibuc.ro >.

Editura Universității din București
Bucharest University Press

that the instances that harm the training process are excluded, achieves better performance as compared to the same model fine-tuned on the entire training dataset. Moreover, in most cases, the performance of the existing augmentation approaches was also improved by using data with properties that differ during the annotation process, as opposed to random sampling.

Keywords: stance detection, data cartography, training dynamics, data augmentation

1. Introduction

Stance detection aims to automatically determine the standpoint taken by the author of a text towards a target of interest (Mohammad et al. 2017) and plays an important role in understanding how information is conveyed in everyday life. Stance detection has widespread applications, ranging from measuring public opinion towards social or political issues to identifying if the author of a text promotes a false idea or contradicts it as an integrated part of fake news detection systems. Most real-life stance detection systems are usually required to accommodate a wide range of targets, thus a common challenge of this task is represented by the scarcity of annotations. On the other hand, shifting the focus to the quantity of the data often comes with a cost regarding its quality.

To address both challenges, we first employ a technique called Data Cartography to characterize each instance of a dataset. Using this characterization, we remove the instances that harm the training process and then proceed to expand the dataset by using the different attributes of the remaining instances to intelligently augment the data.

The Data Cartography technique was first proposed in (Swayamdipta et al. 2020) and aims to characterize a dataset by analyzing the behavior of a model during training on each data instance (training dynamics). Specifically, it measures the variability and confidence of the model in the true class across multiple epochs, identifying three types of instances – easy-to-learn, ambiguous, and hard-to-learn – each having a different impact on the training process. The easy-to-learn instances are consistently labeled correctly by the model, having low variability and high confidence. On the opposing end, the hard-to-learn examples are defined by low confidence and variability, being often mislabeled by the classifier. Ambiguous examples exhibit high variability, as the model struggles to learn them.

Swayamdipta et al. (2020) tested the approach on several NLP tasks – natural language inference, question answering and commonsense reasoning – and found that ambiguous examples promote out-of-distribution generalization, the easy-to-learn examples contribute to model optimization, while the hard-to-learn instances often correspond to annotation errors. Inspired by these results, Park and Caragea (2022) leveraged the Data Cartography technique to improve the MixUp augmentation method (Zhang et al. 2018) which creates additional virtual examples during training by linearly interpolating the hidden representation of two randomly sampled data instances. In contrast, the method proposed by Park and Caragea (2022), called TDMixUp, interpolates examples with different data characteristics, specifically from the easy-to-learn and ambiguous categories, improving the results of the randomized MixUp strategy on six datasets corresponding to natural language inference, paraphrase detection and commonsense reasoning NLP tasks. Motivated by these results, we apply the TDMixUp method on stance detection, by leveraging training dynamics to characterize each data sample based on its contribution to the learning process. We equally divide the dataset into easy-to-learn, ambiguous, and hard-to-learn instances and then fine-tune a pre-trained language model on two of these sets (either easy-to-learn and ambiguous, or hard-to-learn and ambiguous), while also generating additional samples through interpolation between the two sets. In contrast with the approach proposed by Park and Caragea (2022), we do not remove the hard-to-learn examples, hypothesizing that, depending on the dataset, they might actually promote learning and out-of-distribution generalization.

Furthermore, we apply the Data Cartography technique on another augmentation method, called ASDA (Li and Caragea 2021), that was successfully employed in improving the stance detection task. The ASDA method is defined as a conditional masked language modeling (MLM) task that generates additional samples by predicting the masked words of the input sentence. Unlike other similar methods that condition the augmentation of a sentence only on its context and label (Wu et al. 2019), ASDA also uses an auxiliary sentence that provides additional context. This context encodes information regarding both the stance and the target of the input sentence, as well as an additional example sampled from the dataset that has the same target and stance. The results presented by Li and Caragea (2021) show that this approach leads to more diversified, target and stance-aware augmented sentences, compared to previous augmentation methods. We hypothesize that we can further improve the diversity of the augmented sentences, as well as the predictive performance of a model, by sampling the auxiliary sentence from a subset of the data that was characterized differently by the training dynamics compared to the input sentence.

2. Related Work

The topic of stance detection has been extensively studied in recent years, both as an independent task (Augenstein et al. 2016, Allaway and McKeown 2020, Glandt et al. 2021) and as part of larger systems, such as rumor veracity evaluation (Poddar et al. 2018, Ma et al. 2018) or fake news detection (Bhatt et al. 2018, Borges et al. 2019). Most stance detection approaches cover the in-target setup (Li, Zhao, and Caragea 2021, Glandt et al. 2021), where the targets present during the test stage have also been seen during training. Some studies, however, focus on cross-target stance detection (Xu et al. 2018, Zhang et al. 2020), where the aim is to generalize classifiers across targets, or zero-shot stance detection (Allaway and McKeown 2020, Zhang et al. 2023), where the test targets have not been seen during training. As the scarcity of annotations is a prevalent challenge in stance detection, many studies have focused on diminishing this issue through data augmentation (Li and Caragea 2021, Li and Yuan 2022, Zhang et al. 2023), multi-dataset learning (Li, Zhao, and Caragea 2021), or creating larger, more varied datasets (Zhang et al. 2023).

In this study, we explore two methods of improving the performance of both in-target and zero-shot stance detection through data augmentation. The first method is inspired by TDMixUp (Park and Caragea 2022), an augmentation method that has been successfully employed on several NLP tasks – natural language inference, paraphrase detection and commonsense reasoning. TDMixUp improves upon the MixUp strategy (Zhang et al. 2018), which extends the training distribution by linearly interpolating randomly selected input instances and their associated labels. Instead of randomized linear interpolation, TDMixUp employs the Data Cartography technique (Swayamdipta et al. 2020) to characterize each data instance and then combines instances with characteristics that are different to create more diverse new data samples.

The second augmentation approach extends the ASDA strategy proposed in (Li and Caragea 2021), by using Data Cartography to generate more informative examples. ASDA is formulated as a conditional masked language modeling task, where the masked words are conditioned on the context in which they appear, as well as on an auxiliary sentence. The latter contains the target and label information of the data instance and an additional example, randomly extracted from the same dataset, but having the same target and label. Similar to TDMixUp, our approach replaces the randomized selection of an additional example with a more informed selection process, that takes into consideration the characteristics of both the initial and additional example.

3. Data Cartography as a tool for data augmentation

The core of both data augmentation methods proposed in this paper is represented by the Data Cartography technique, first introduced in (Swayamdipta et al. 2020). The goal of this technique is to characterize a dataset, by analyzing the behavior of a model on individual examples during training – Training Dynamics. Similar to the original approach (Swayamdipta et al. 2020), we employ confidence and variability as the training dynamics used to characterize each data instance (x_i, y_i) over E training epochs. The confidence measure (μ_i) captures how confident in the true label the learner is, for a given example, and is defined as the mean model probability of the true label across epochs. The variability (σ_i) measures the spread of the model probability of the true label across epochs and is defined using the standard deviation:

$$\mu_i = \frac{\sum_{e=1}^{E} p_{\theta(e)}(y_i|x_i)}{E} \quad \sigma_i = \sqrt{\frac{\sum_{e=1}^{E}\left(p_{\theta(e)}(y_i|x_i) - \mu_i\right)^2}{E}}$$

The values of these statistics per sample are then used to map each data instance to one of the following three categories – easy-to-learn, ambiguous, and hard-to-learn – by equally splitting the dataset. The easy-to-learn category is characterized by high confidence and low variability and corresponds to those examples that are consistently labeled correctly by the model, while the hard-to-learn examples have low confidence and low variability, being usually mislabeled by the model. The ambiguous examples are the most challenging for the model, being represented by high variability. Furthermore, the confidence and variability metrics of each data instance can be used to construct Data Maps that help visualize the dataset with its three regions (easy-to-learn, ambiguous, hard-to-learn). Such examples can be seen in the Annex (

Figure 3 to

Figure 8).

4. Improving Stance Detection using Data Augmentation

We propose two augmentation methods for stance detection that make use of training dynamics to intelligently extract the data instances that will be used for augmentation. The first approach is similar to the TDMixUp (Park and Caragea 2022) proposal and implies interpolating examples from distinct regions identified using training dynamics. The second method extends ASDA (Li and Caragea 2021) by using training dynamics to create more diverse auxiliary sentences, that contain data from a different region than the sentence that is augmented.

4.1. TDMixUp

TDMixUp uses the same methodology of constructing virtual examples during training as in the original MixUp paper (Zhang et al. 2018), which extends the training distribution by including the prior knowledge that linear interpolations of embeddings should lead to linear interpolations of the associated labels:

$$\begin{cases} x = \lambda x_i + (1 - \lambda)x_j \\ y = \lambda y_i + (1 - \lambda)y_j \end{cases}$$

where x_i, x_j are raw input vectors, y_i, y_j are one-hot label encodings, and $\lambda \in [0,1]$ is sampled from a $\mathrm{Beta}(\alpha, \alpha)$ distribution and controls the strength of the interpolation through the hyper-parameter $\alpha \in (0, \infty)$. The MixUp approach aims to encourage a linear behavior of the model in between training examples, leading to stronger, more robust predictions.

In the original paper (Zhang et al. 2018), the two instances that are interpolated at some point are sampled randomly from the dataset. However, similar to the approach in (Park and Caragea 2022), we propose a method that interpolates examples from different regions, as identified using the Data Cartography technique. Specifically, using the MixUp strategy, we interpolate between the easy-to-learn and ambiguous sets, as well as between the hard-to-learn and ambiguous sets. The same model used to compute the training dynamics is retrained from scratch on the selected regions and on the interpolated samples generated during training using the MixUp method.

4.2. TDASDA

Inspired by the results obtained with the Auxiliary Sentence based Data Augmentation (ASDA) method (Li and Caragea 2021), we propose an adaptation of ASDA that makes use of the training dynamics when creating the auxiliary sentence. We refer to this proposed method as TDASDA (Training Dynamics ASDA). ASDA is an augmentation approach defined as a conditional masked language modeling task, which generates additional samples by predicting the masked words of the input sentences. As the model aims to generate new samples that are consistent with both the label and the target, an auxiliary sentence is concatenated to the example containing the masked words, in order to provide additional context. This auxiliary sentence has the same format as described in the original paper: "The authors of the following tweets are both [Label] [Target]. The first tweet is: [Additional Example]. The second tweet is:" (Li and Caragea 2021). However, instead of randomly sampling an additional example from the dataset, TDASDA further conditions the model to choose an example with characteristics that are different (derived from training dynamics) compared to those of the input sentence, but with the same stance and target. We hypothesize that this approach will increase the diversity of the augmented sentences generated by the model, by including samples with characteristics that differ in the context of the model.

5. Testing Setup

5.1. Datasets

Three stance detection datasets were used to test the performance of the approaches introduced in this paper: COVIDLies (Hossain et al. 2020), VAST (Allaway and McKeown 2020), and SemEval 2016 (Krejzl and Steinberger 2016).

The COVIDLies dataset was intended for misinformation detection and contains 62 claims extracted from a Wikipedia article about misconceptions related to the COVID-19 pandemic. The dataset contains 6591 tweets from March and April 2020, that have been mapped to misconceptions using BERTScore (Zhang et al. 2020). However, many of the misconceptions in COVIDLies have labeled examples only for the neutral class. Thus, we decided to construct an additional dataset, called Reduced COVIDLies that contains only those targets that have examples from at least two classes. The Reduced COVIDLies dataset contains 2110 tweets and 17 misconceptions.

VAried Stance Topics (VAST) is a large dataset created for zero-shot and few-shot stance detection. The dataset contains 18545 comments collected from The New York Times 'Room for Debate' section and 5634 topics, which are extracted from the debate topic or proposed by annotators. Some of these topics are mostly or only present in the testing and validation datasets to simulate a few-shot or zero-shot scenario. There are three validation and test datasets, one for zero-shot detection (the dataset contains completely new topics compared to the training data), one for few-shot (the dataset contains examples for the topics that have little representation in the training data), and a combined dataset.

SemEval 2016 is a stance detection dataset created for a shared task in the SemEval 2016 competition. The dataset consists of 4870 English tweets and 5 targets ("Atheism", "Feminist Movement", "Climate Change is a Real Concern", "Legalization of Abortion" and "Hillary Clinton"). The tweet-target pairs were manually annotated as either support, against, or neither. The latter label refers to both neutral examples and examples that contain no cue that can reveal the stance towards the given target.

5.2. Baselines and Parameter Tunning

To assess the performance of the two data augmentation methods, we compare them with the results of a fine-tuned language model on 100% of the training data, as well as on subsets of data having characteristics that differ (employing the Data Cartography method), as presented in Table 2 (see the Annex). The data augmentation methods are also benchmarked against their alternatives that use random sampling instead of training dynamics to generate additional data. BERT (Devlin et al. 2019) was used as the base language model for the VAST dataset, while Covid-Twitter-BERT (Müller et al. 2020), a BERT model pre-trained on a corpus of Tweets about COVID-19, was employed for the COVIDLies and SemEval 2016 datasets.

For the Data Cartography method and TDMixUp, we fine-tune the base models to predict the stance ("neutral", "in-favor", or "against") by appending a fully-connected layer to the hidden representation of the [CLS] token. An overview of the general model architecture used for these methods is presented in

Figure *1* (see the Annex). The model is fine-tuned for 4 epochs, using a batch size of 32, and the Adam optimizer with a learning rate of 2e-5 and no weight decay. The maximum sequence length is 256 for BERT and 128 for Covid-Twitter-BERT. In order for the model to learn to predict the stance based on both the input sentence and the target, the input sequence will contain both pieces of information, separated by the [SEP] token. Regarding TDMixUp, the hyper-parameter α from the Beta distribution, controlling the strength of the interpolation, is set to 0.4.

For the TDASDA approach we used the same base models as before on top of which we stacked a head for masked language modeling. A representation of the general model architecture used for this method is presented in

Figure 2 (see the Annex). We fine-tuned these models for 10 epochs, using a batch size of 16, the Adam optimizer with a learning rate of 1e-4, and the Sparse Categorical Cross Entropy loss function. The maximum sequence length is set to 400 for BERT and 256 for Covid-Twitter-BERT. The percentage of masked words in the input example is set to 15%, matching the percentage of tokens that were masked while training the base BERT model (Devlin et al. 2019). The labels provided to the model correspond to the tokens of the original sentence before any masking was done.

The results for all methods are averaged across 5 runs with random restarts. We evaluated our approaches using the macro averaged F1-score, in order to make sure that each class is given equal importance.

6. Results

The core of all methods proposed in this study is represented by the usage of training dynamics to characterize the instances in a dataset. To assess the effect of the Data Cartography method on the three datasets, we fine-tune a pre-trained language model on a subset of the initial dataset, characterized using training dynamics. Table 2 (see the Annex) describes all the subsets and their combinations that we used to test our approaches. Table 3 (see the Annex) presents the overall macro averaged F1-score of the Data Cartography technique and its benchmarks. The obtained results exhibit two main trends.

Firstly, we can see that the best results on VAST and SemEval are obtained using the combination of the easy-to-learn, ambiguous, and half of the hard-to-learn subsets. In order to better understand this result, we can look at the distribution of the instances between the three regions, represented in

Figure 3 to Figure 6 (see the Annex). In all these datasets, slightly more data seems to be condensed in the upper part of the graph, suggesting that some of the examples that have been characterized as hard-to-learn exhibit a behavior that is more characteristic of the ambiguous or easy-to-learn instances, thus, instead of harming the training process, they improve its generalization capabilities. This may explain why the combination of ambiguous, easy-to-learn, and half of the hard-to-learn sets leads to slightly better performance compared to the merger of the easy-to-learn and ambiguous sets that has been used so far (Swayamdipta et al. 2020, Park and Caragea 2022) and that represents the second best combination.

Secondly, by switching our attention to the results on Covid Lies, we can observe that the hard-to-learn instances seem to play a bigger role than before in aiding the learning process. Specifically, we see very good results obtained using the ambiguous and hard-to-learn sets, significantly better than those obtained by combining ambiguous with easy-to-learn data. By looking at the data distribution on the three regions (Figure 7 and

Figure *8* of the Annex), we identify an isolated cluster of very easy-to-learn instances that don't provide much information about the learning process of the model. Given the uneven distribution, the instances that led to these impressive results probably exhibit the behavior of easy-to-learn and ambiguous examples, promoting generalization and model convergence.

We hypothesize that the patterns presented above will extend to the data augmentation methods and present the results in. Some of the results obtained using solely the Data Cartography technique were included in the upper part of, to be used as benchmarks. We note that we included only the results of the models fine-tuned on two subsets of data with characteristics that differ, as the data augmentation methods further implemented in this paper, namely TDMixUp (Park and Caragea 2022) and ASDA (Li and Caragea 2021), were introduced with reference to combinations of two subsets only. However, we acknowledge the potential benefits of extending these data augmentation methods to combine three subsets of instances with characteristics that differ, given the aforementioned results obtained with the Data Cartography technique on all VAST and SemEval datasets (see Table 3 of the Annex). We leave this area of investigation and possible improvement for future studies.

The middle part of shows the macro averaged F1-score of the TDMixUp augmentation strategy, which interpolates examples from the ambiguous and easy-to-learn sets, as well as from the ambiguous and hard-to-learn pair. We also included the results of the randomized MixUp strategy.

The lower part of shows the macro averaged F1-score of the TDASDA augmentation strategy, which generated additional samples by predicting the masked words of the input sentence. The prediction of the masked words is conditioned on the context of the input sentence, its stance and target, as well as an additional example from the dataset that exhibits different characteristics compared to the input sentence. We tested the approach by sampling data from the following pairs of subsets: (ambiguous, easy-to-learn), and (ambiguous, hard-to-learn). We also show the results of TDASDA on randomly sampled data.

Table 1

The macro averaged F1-score results on all datasets for TDASDA (lower), TDMixUp (middle), and several Data Cartography benchmarks (upper)

Data Subset	VAST	VAST Zero	VAST Few	SemEval	CLies	CLies Red
100% train	0.6932	0.7116	**0.6961**	**0.7375**	**0.7433**	0.6261
66% train rand	0.6665	0.6978	0.6324	0.6357	0.6155	0.5506
amb + easy	**0.7157**	**0.7194**	0.6915	0.7229	0.7146	0.4975
amb + hard	0.6379	0.6600	0.6166	0.6884	0.7281	**0.6391**

TDMixUP:

Data Subset	VAST	VAST Zero	VAST Few	SemEval	CLies	CLies Red
33% + 33% train	0.6858	0.7001	0.6715	0.7190	0.6471	0.5758
amb + easy	**0.7134**	**0.7257**	**0.6926**	**0.7223**	0.6843	0.4923
amb + hard	0.6527	0.6671	0.6354	0.6442	**0.6869**	**0.6164**

TDASDA:

Data Subset	VAST	VAST Zero	VAST Few	SemEval	CLies	CLies Red
66% train	0.6900	0.7014	0.6692	**0.7402**	0.6578	0.5547
amb + easy	**0.7143**	**0.7248**	**0.7073**	0.7391	0.6730	0.4718
amb + hard	0.6780	0.6539	0.6177	0.6830	**0.7032**	**0.6145**

As hypothesized, by analyzing the results of TDMixUp and TDASDA, we can also identify two patterns, depending on which combination of instances gave better results.

Firstly, on all VAST datasets we can see that both TDMixUp and TDASDA data augmentation strategies consistently improve upon the results of the Data Cartography approach, when using the same input data. Moreover, the TDMixUp strategy that interpolates ambiguous and easy-to-learn instances significantly outperforms the randomized MixUp approach. Similarly, TDASDA on the ambiguous and easy-to-learn sets leads to better performance, when compared to the randomized ASDA. For the SemEval dataset, only the TDASDA approach led to better results compared to the Data Cartography method, but TDASDA on the ambiguous and easy-to-learn sets did not improve upon the randomized approach. Conversely, TDMixUp applied on SemEval did not improve the performance of Data Cartography, but TDMixUp on easy-to-learn and ambiguous sets did improve the randomized MixUp strategy.

Secondly, on both Covid Lies datasets, the best results obtained using either of the data augmentation techniques are achieved using the combination of hard-to-learn and ambiguous instances. However, these results do not improve upon the Data Cartography method. The only improvement of the data augmentation strategies upon the Data Cartography method for Covid Lies is achieved using randomized data.

7. Conclusions

We proposed two data augmentation techniques, TDMixUp and TDASDA, that aim to improve the Stance Detection task by leveraging training dynamics, namely the information extracted from the model behavior during training on each individual instance. Firstly, we characterized each example in the dataset using training dynamics – a technique called Data Cartography – and identified three groups of instances: easy-to-learn, ambiguous, and hard-to-learn. Then we hypothesized that, when fine-tuning a pre-trained language model only on certain groups, by removing instances that harm the training process, we may improve the predictive performance of the model. We empirically validated that the Data Cartography method achieves superior performance in terms of the macro averaged F1-score, usually by using a combination of easy-to-learn, ambiguous, and half of the hard-to-learn instances. On some datasets, better results were obtained using ambiguous and hard-to-learn examples, which can be explained by the unevenly distributed data between the three regions, making the ambiguous and hard-to-learn examples behave more closely to easy-to-learn and ambiguous data in these datasets.

Secondly, we validated the effect of the Data Cartography tool on two data augmentation methods. For these approaches, we used the same three groups of instances identified before. While TDMixUp interpolates instances from two different groups to create additional virtual examples during training, TDASDA is designed as a conditional masked language modeling task that generates additional data by predicting the masked words of an input sentence. The prediction of the masked words is conditioned on an auxiliary sentence that encodes the stance and the target of the input instance, as well as an additional example with characteristics that differ, but which displays the same stance towards the same target. Both augmentation approaches lead to similar conclusions: by combining instances with characteristics that are different, TDMixUp and TDASDA generally improve upon their random sampling alternatives.

Bibliography

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko (2016), "Stance and Sentiment in Tweets", in: *ACM Transactions on Internet Technology (TOIT)* 17, p. 1-23.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi (2020), "Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics", in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 9275-9293.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz (2018), "mixup: Beyond Empirical Risk Minimization", in: *Proceedings of the 6th International Conference on Learning Representations*.

Seo Yeon Park and Cornelia Caragea (2022), "A Data Cartography based MixUp for Pretrained Language Models", in: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 4244-4250.

Yingjie Li and Cornelia Caragea (2021), "Target-Aware Data Augmentation for Stance Detection", in: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 1850-1860.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu (2019), "Conditional BERT Contextual Augmentation", in: *Computational Science – ICCS 2019: 19th International Conference Proceedings*, Part IV, p. 84-95.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva (2016), "Stance Detection with Bidirectional Conditional Encoding", in: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 876-885.

Emily Allaway and Kathleen McKeown (2020), "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 (EMNLP)*, p. 8913-8931.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea (2021), "Stance Detection in COVID-19 Tweets", in: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers), p. 1596-1611.

Lahari Poddar, Wynne Hsu, Mong Li Lee, and Shruti Subramaniyam (2018), "Predicting Stances in Twitter Conversations for Detecting Veracity of Rumors: A Neural Approach", in: *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*, p. 65-72.

Jing Ma, Wei Gao, and Kam-Fai Wong (2018), "Detect Rumor and Stance Jointly by Neural Multi-Task Learning", in: *Companion Proceedings of the The Web Conference 2018*, p. 585-593.

Gaurav Bhatt, Aman Sharma, Shivam Sharma, Ankush Nagpal, Balasubramanian Raman, and Ankush Mittal (2018), "Combining Neural, Statistical and External Features for Fake News Stance Identification", in: *Companion Proceedings of the The Web Conference 2018*, p. 1353-1357.

Luıś Borges, Bruno Martins, and Pável Calado (2019), "Combining Similarity Features and Deep Representation Learning for Stance Detection in the Context of Checking Fake News", in: *Journal of Data and Information Quality (JDIQ)* 11.

Yingjie Li, Chenye Zhao, and Cornelia Caragea (2021), "Improving Stance Detection with Multi-Dataset Learning and Knowledge Distillation", in: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 6332-345.

Chang Xu, Cécile Paris, Surya Nepal, and Ross Sparks (2018), "Cross-Target Stance Classification with Self-Attention Networks", in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers), p. 778-783.

Bowen Zhang, Min Yang, Xutao Li, Yunming Ye, Xiaofei Xu, and Kuai Dai (2020), "Enhancing Cross-target Stance Detection with Transferable Semantic-Emotion Knowledge", in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 3188-3197.

Jiarui Zhang, Shaojuan Wu, Xiaowang Zhang, and Zhiyong Feng (2023), "Task-Specific Data Augmentation for Zero-Shot and Few-Shot Stance Detection", in: *Companion Proceedings of the ACM Web Conference 2023*, p. 160-163.

Yang Li and Jiawei Yuan (2022), "Generative Data Augmentation with Contrastive Learning for Zero-Shot Stance Detection", in: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 6985-6995.

Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh (2020), "COVIDLies: Detecting COVID-19 Misinformation on Social Media", in: *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Peter Krejzl and Josef Steinberger (2016), "UWB at SemEval-2016 Task 6: Stance Detection", in: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, p. 408-412.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019), "BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding", in: *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), p. 4171-4186.

Martin Müller, Marcel Salathé, and Per Egil Kummervold (2020), "COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter", in: *Frontiers in Artificial Intelligence 6 (2023)*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi (2020), "BERTScore: Evaluating Text Generation with BERT", in: *International Conference on Learning Representations 2020*.

All links were verified by the editors and found to be functioning before the publication of this text in 2024.

DECLARATION OF CONFLICTING INTERESTS

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

FUNDING

The authors received no financial support for the research, authorship, and/or publication of this article.

Annex

1. Datasets used for benchmarking

Table 2

Dataset splitting methodology

Dataset	Description
train 100%	the entire dataset
train 66%	66% of the dataset, randomly chosen
train 33%	33% of the dataset, randomly chosen
easy	the 33% most easy-to-learn examples
amb	the 33% most ambiguous examples
hard	the 33% most hard-to-learn examples
amb + easy	all dataset minus the 33% most hard-to-learn examples
amb + easy + 50% hard	all dataset minus the 16.5% most hard-to-learn examples
amb + hard	all dataset minus the 33% most easy-to-learn examples
amb + 50% hard	the 33% most ambiguous examples plus 16.5% least hard-to-learn examples

2. Data Cartography results

Table 3

Macro averaged F1-score results on all datasets and their subsets

Data Subset	VAST	VAST Zero	VAST Few	SemEval	CLies	CLies Red
100% train	0.6932	0.7116	**0.6961**	0.7375	**0.7433**	0.6261
33% train random	0.6802	0.6964	0.6388	0.6570	0.6368	0.6281
66% train random	0.6665	0.6978	0.6324	0.6357	0.6155	0.5506
easy	0.6980	0.6986	0.6525	0.7204	0.3207	0.3036
amb	0.6479	0.6878	0.6673	0.6585	0.6318	0.4394
hard	0.2585	0.2670	0.2658	0.4265	0.4511	0.5532
amb + easy	0.7157	0.7194	0.6915	0.7229	0.7146	0.4975
amb + easy + 50% hard	**0.7211**	**0.7313**	0.6936	**0.7552**	0.7068	0.4992
amb + hard	0.6379	0.6600	0.6166	0.6884	0.7281	**0.6391**
amb + 50% hard	0.7059	0.7045	0.6788	0.6994	0.6882	0.4609

3. Models' architectures

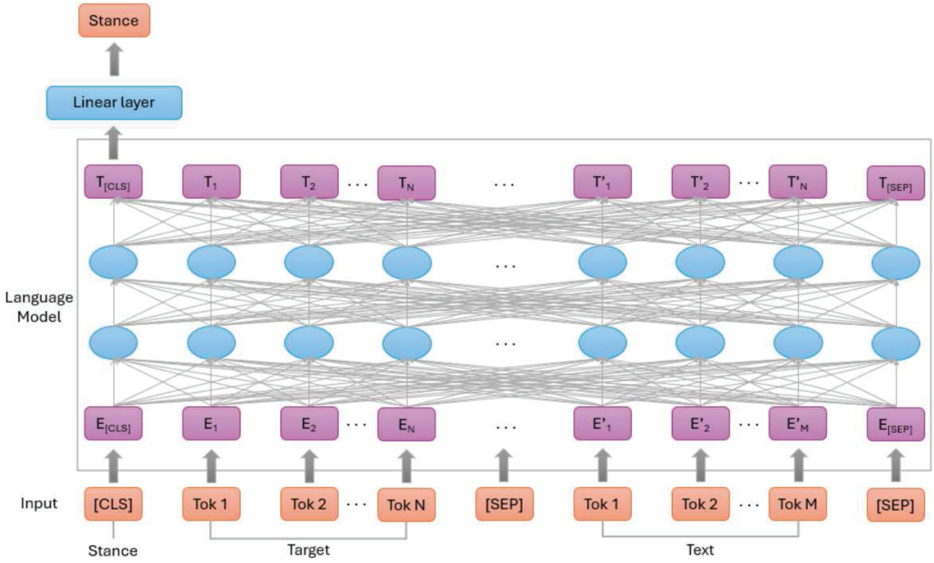

Figure 1. Language model architecture for classification

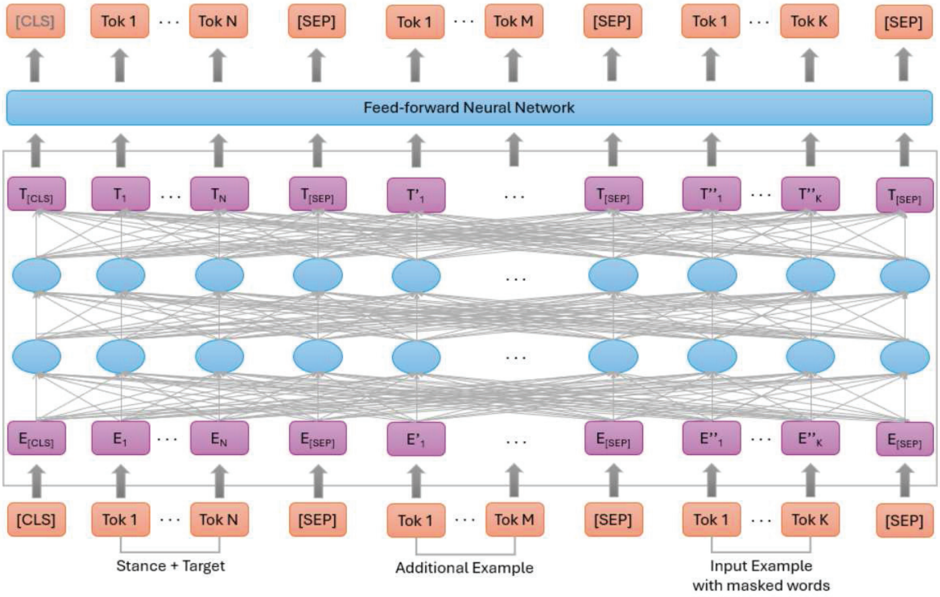

Figure 2. Language model architecture for conditional masked language modeling

4. Data Maps

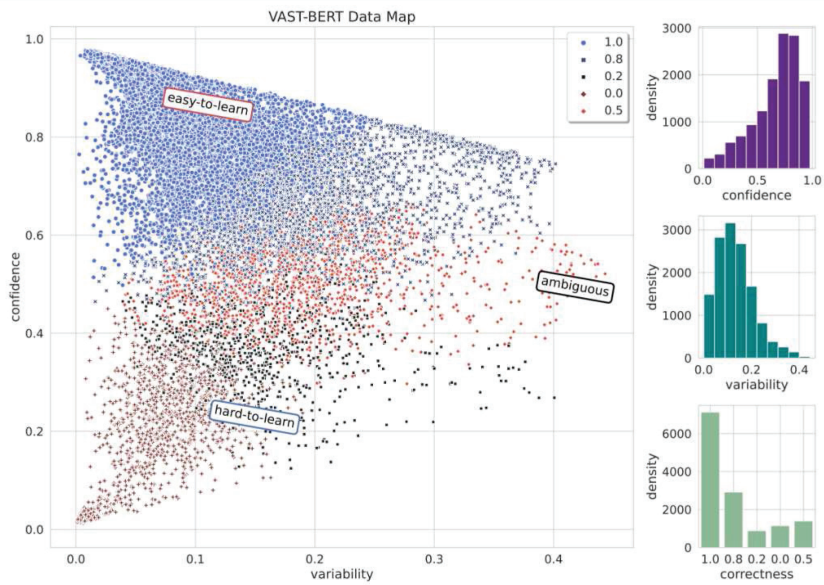

Figure 3. Data Map for the VAST Dataset, BERT model

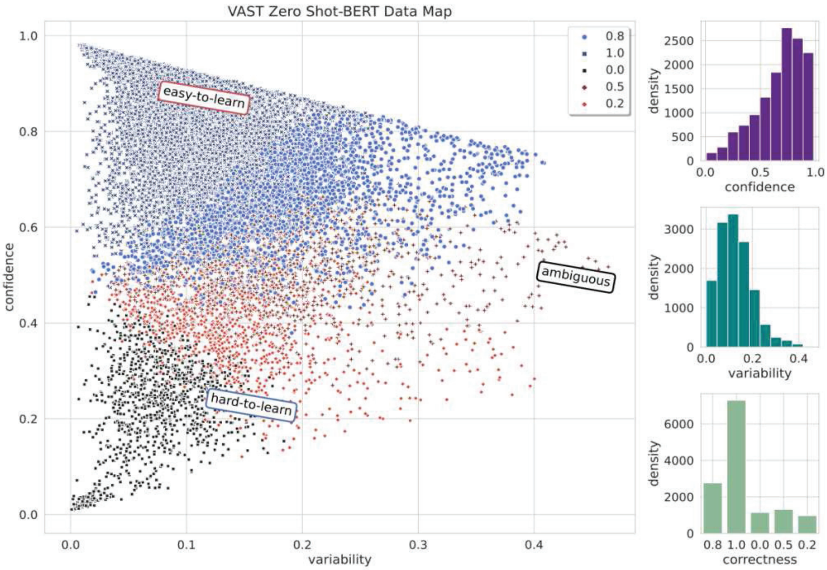

Figure 4. Data Map for the VAST Zero Shot Dataset, BERT model

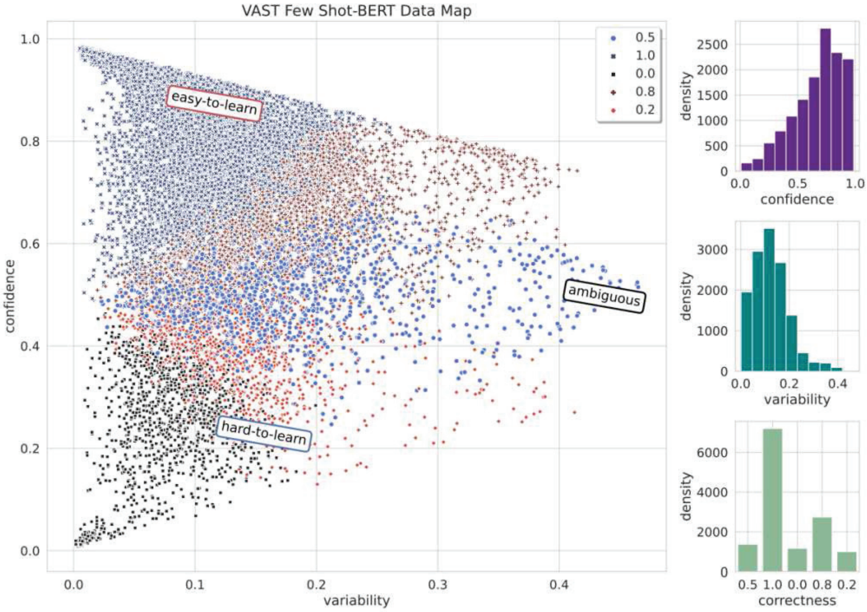

Figure 5. Data Map for the VAST Few Shot Dataset, BERT model

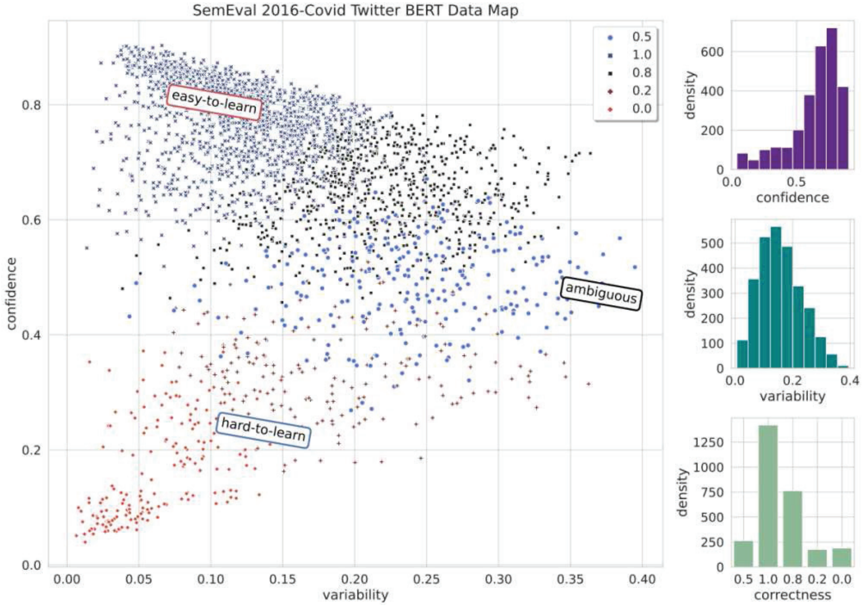

Figure 6. Data Map for the SemEval Dataset, Covid Twitter BERT model

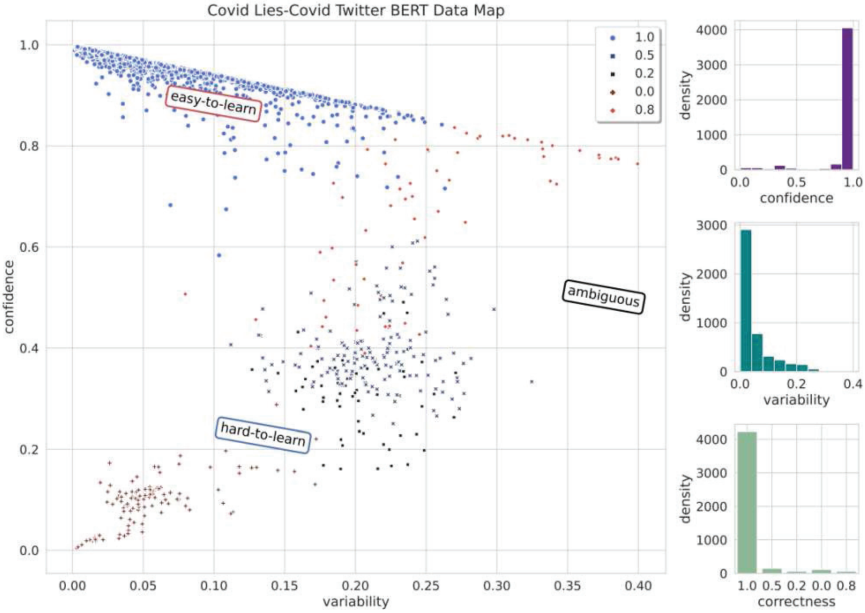

Figure 7. Data Map for the Covid Lies Dataset, Covid Twitter BERT model

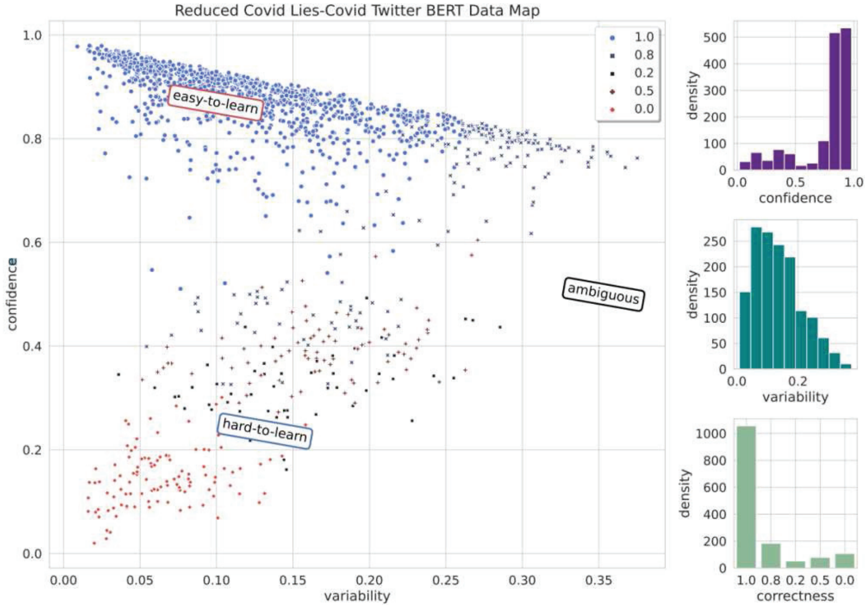

Figure 8. Data Map for the Reduced Covid Lies Dataset, Covid Twitter BERT model