Fraunhofer SIT@SMM4H'22: Learning to Predict Stances and Premises in Tweets related to COVID-19 Health Orders Using Generative Models

Raphael Antonius Frick and **Martin Steinebach**
Fraunhofer Institute for Secure Information Technology SIT |
ATHENE — National Research Center for Applied Cybersecurity
Rheinstrasse 75, Darmstadt, 64295, Germany
{raphael.frick, martin.steinebach}@sit.fraunhofer.de

Abstract

This paper describes the system used to predict stances towards health orders and to detect premises in Tweets as part of the Social Media Mining for Health 2022 (#SMM4H) shared task. It takes advantage of GPT-2 to generate new labeled data samples which are used together with pre-labeled and unlabeled data to fine-tune an ensemble of GAN-BERT models. First experiments on the validation set yielded good results, although it also revealed that the proposed architecture is more suited for sentiment analysis. The system achieved a score of 0.4258 for the stance and 0.3581 for the premise detection on the test set.

1 Introduction

During the pandemic many countries, such as the US, China, and Germany, introduced health mandates to cope against the fast spread of the coronavirus. In this time, social networks like Facebook and Twitter enabled users to share their opinion regarding the implemented health orders. These social media posts can be used as part of argument mining to find reasons for and against those. Further, it also allows analyzing whether any opposing opinions are based on false information and intentionally spread disinformation.

In Task 2. of the Social Media Mining for Health 2022 (#SMM4H) shared task (Weissenbacher et al., 2022), one objective was to develop a classification system that can predict the stance towards imposed health mandates during the pandemic, such as *face masks*, *school closures* and *stay at home orders*, in Tweets. Another was to detect whether a Tweet contained statements that could be used for reasoning.

In this paper, the proposed approach of our team, *Fraunhofer SIT*, is described. It takes advantage of two generative models to generate new labeled samples as well as a generative adversarial network for classification and to also fully utilize unlabeled samples during training. These are then incorporated into an ensemble classification scheme to improve the classification accuracy.

The paper is organized as follows: in Section 2 the proposed system developed for the shared task is presented. Section 3 showcases and discusses the experimental results achieved by the proposed components on the validation set. The paper then concludes in Section 4.

2 Proposed Approach

The main component of our classification system to estimate stances and to identify premises in Tweets was built around an ensemble of fine-tuned GAN-BERT (Croce et al., 2020) models. *GAN-BERT* is a generative adversarial network (Goodfellow et al., 2014) that allows to fine-tune models on labeled and unlabeled data. Its generator network produces artificially crafted embeddings, while the discriminator tries to distinguish them from real embeddings extracted from transformer models. Here, *BERT uncased*, *BERT cased* (Devlin et al., 2019) and *ALBERT v2* (Lan et al., 2019) served as transformer models due to their performance on the validation set. In the case of real samples, the discriminator also tries to predict their classes. The class-labels hereby consisted of L_{FAVOR} and $L_{AGAINST}$ for the stance estimation task and L_0 and L_1 for the premise detection task. The prediction probabilities of all the three fine-tuned GAN-BERT models were averaged to provide a unified classification decision.

To provide additional labeled training data and to prevent data scarcity having an impact on the classification accuracy, new samples were generated through *GPT-2* (Radford et al., 2019). To ensure that the generated data represent a particular class, one text-generation model was trained on data of a specific class. The data then underwent several preprocessing steps, where emojis were converted to their descriptive terms, URLs and user mentions

Proceedings of the 29th International Conference on Computational Linguistic, pages 111–113
October 12–17, 2022.

	Stance				Premise			
	$\mathrm{F1}_{mask}$	$\mathrm{F1}_{school}$	$\mathrm{F1}_{home}$	$\mathrm{F1}_{avg}$	$\mathrm{F1}_{mask}$	$\mathrm{F1}_{school}$	$\mathrm{F1}_{home}$	$\mathrm{F1}_{avg}$
bert-uncased	0.8845	0.8663	0.7397	0.8302	0.6994	**0.8235**	0.6415	0.7215
bert-uncased + PP	0.8348	0.8492	0.7500	0.8113	0.7322	0.7676	**0.7360**	**0.7453**
GAN + bert-uncased + PP	0.8974	0.8587	0.7879	0.8480	0.6993	0.8075	0.6604	0.7224
GAN + bert-uncased + PP +U	0.8945	0.8488	0.8235	0.8556	0.6107	0.7397	0.5934	0.6479
GAN + bert-cased + PP +U	0.8622	0.8353	0.8000	0.8325	0.6565	0.7200	0.5773	0.6513
GAN + albert-v2 + PP +U	0.8561	0.7836	0.6437	0.7615	0.7034	0.7590	0.6729	0.7118
GAN + bert-uncased + PP + U + GPT2	**0.9030**	**0.8667**	0.7887	0.8528	0.7226	0.7738	0.6606	0.7189
GAN + bert-cased + PP + U + GPT2	0.8496	0.8383	**0.8308**	0.8396	0.7143	0.7329	0.6909	0.7127
GAN + albert-v2 + PP + U + GPT2	0.8462	0.8457	0.7826	0.8248	0.7044	0.7297	0.6604	0.6982
GAN + ensemble + PP + U + GPT2	0.8945	0.8605	0.8182	**0.8577**	**0.7451**	0.7898	0.6916	0.7422

Table 1: F1-scores achieved by each model configuration on the validation set of each task. *PP* refers to preprocessed data and *U* to the usage of unlabeled data from Glandt et al. (2021) during training.

were exchanged with generic tokens.

3 Experimental Results

The classification system was implemented using Python. For data preprocessing, the *emoji*[1] package as well as libraries provided by Python were used.

The GPT-2 models were trained on the preprocessed train split of the dataset provided along the shared task (Davydova and Tutubalina, 2022) using the *aitextgen* framework[2]. To avoid overfitting, the models were only trained for 3500 steps with a learning rate of $1e^{-3}$. The words *the*, *face masks*, *school closures*, *stay at home orders*, *Corona* and *COVID19* served as prefixes for the text-generation. Further, various temperatures were considered, such as 0.8, 1.0 and 1.2. For each configuration consisting of a label, prefix and temperature, 50 samples were generated and occurring duplicates were removed. The following examples showcase Tweets generated by our trained model:

> **Stance — Favor:** COVID19 @user What a frightening lot of people. No one is wearing a mask. and related ways of managing it. It's so disappointing.
>
> **Stance — Against:** The death figures are inflated. A major percentage of deaths are those who died with Covid at home, and not from it.

For training the GAN-BERT models in a semi-supervised manner, data from the *Stance Detection in COVID-19 Tweets Dataset* (Glandt et al., 2021) served as unlabeled data. The sequence length for the embeddings was set to 280 and the models were trained using a batch size of 16 and learning rates of $5e^{-5}$ for both the generator and the discriminator. As optimizer *adam* (Kingma and Ba, 2015) was used to take advantage of adaptive learning rates and for regularization, a dropout rate of 0.3 was chosen. To further avoid overfitting, the training procedure took advantage of model checkpoints to keep only the best performing models while the models were trained for 20 epochs.

Table 1 showcases the F1-scores achieved by various model configurations. A *BERT uncased* model fine-tuned using *GAN-BERT* achieved overall higher F1-scores when trying to estimate stances than the same model trained traditionally with or without the application of preprocessing. When trying to classify premises using GAN-BERT, the performance on the validation set decreased. One reason for this could be, that the stance towards a particular topic can often be expressed using a single term (e.g., #WearAMask), while for premises, the structure, and wording of the entire sentence is of high importance. The generator of GAN-BERT may not be able to replicate sentence structures as accurately as wordings. Similar effects are also to be expected when using GPT-2 to craft new samples, as the synthesis is prone to grammatical errors. However, synthesizing new samples as well as averaging the classification probabilities of multiple transformer networks helped with increasing the classification accuracies. On the test set a score of 0.4258 for the stance and a score of 0.3581 for the premise detection was achieved.

4 Conclusion

In this paper, a classification system to estimate stances of health orders and to detect premises in Tweets is proposed. The system incorporates two types of generative models to allow training on unlabeled data and to generate new labeled data. The F1-scores achieved by the proposed model on the validation set yielded promising first results. However, it also showed, that the classification architecture is more suited for sentiment estimation rather than premise detection.

[1] https://github.com/carpedm20/emoji/

[2] https://github.com/minimaxir/aitextgen

Acknowledgements

This work was supported by the German Federal Ministry of Education and Research (BMBF) and the and the Hessian Ministry of Higher Education, Research, Science, and the Arts within their joint support of "ATHENE – Disinformation on Corona" and "Lernlabor Cybersicherheit".

References

Danilo Croce, Giuseppe Castellucci, and Roberto Basili. 2020. GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2119, Online. Association for Computational Linguistics.

Vera Davydova and Elena Tutubalina. 2022. Smm4h 2022 task 2: Dataset for stance and premise detection in tweets about health mandates related to covid-19. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages –.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

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)*, pages 1596–1611, Online. Association for Computational Linguistics.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Davy Weissenbacher, Ari Z. Klein, Luis Gascó, Darryl Estrada-Zavala, Martin Krallinger, Yuting Guo, Yao Ge, Abeed Sarker, Ana Lucia Schmidt, Raul Rodriguez-Esteban, Mathias Leddin, Arjun Magge, Juan M. Banda, Vera Davydova, Elena Tutubalina, and Graciela Gonzalez-Hernandez. 2022. Overview of the seventh social media mining for health applications #smm4h shared tasks at coling 2022. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages –.