

# Fraunhofer SIT at CheckThat! 2022: Semi-Supervised Ensemble Classification for Detecting Check-Worthy Tweets

Raphael Antonius Frick<sup>1</sup>, Inna Vogel<sup>1</sup> and Isabella Nunes Grieser<sup>1</sup>

<sup>1</sup>Fraunhofer Institute for Secure Information Technology SIT | ATHENE - National Research Center for Applied Cybersecurity  
Rheinstrasse 75, Darmstadt, 64295, Germany  
<https://www.sit.fraunhofer.de/>

## Abstract

During the corona pandemic misinformation has been increasingly spread on social media. Since the automatic verification of social media postings has shown to be challenging, there exists the need of classification systems, that can identify check-worthy posts in social media feeds. In this paper, the classification system used in the CLEF2022-CheckThat! Lab to detect check-worthiness in English tweets is presented. The proposed system, that took advantage of ensemble and semi-supervised learning showed promising results in the experimental evaluation. Further, first experiments were conducted with the novel framework *lambeq* to solve the classification problem using quantum natural language processing (QNLP), which were not part of the final model. The final model ranked fifth best in terms of the F1-score in the competition.

## Keywords

check-worthiness detection, Twitter, semi-supervised learning, GAN, QNLP

## 1. Introduction

Within several weeks after the outbreak of the coronavirus in 2019, fake news about the topic started to spread on the internet. The term fake news hereby refers to the targeted distribution of misleading or false information to influence a persons' opinion or behavior. Especially social media platforms and instant messengers have played an extensive role for the exchange and distribution of coronavirus related fake news. The information shared by humans manually or automatically through social bots on social media covered several topics, such as conspiracy theories about the origin of the outbreak and the existence of the virus, the negative effects of vaccines, as well as advertisements for untested or even harmful treatments. Since some of the misinformation can have an impact on a persons' life, there exists the need to detect whether a tweet or other social media postings contains false statements or claims.


However, even with the current advancements in natural language processing, the


---

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

✉ [raphael.frick@sit.fraunhofer.de](mailto:raphael.frick@sit.fraunhofer.de) (R. A. Frick); [inna.vogel@sit.fraunhofer.de](mailto:inna.vogel@sit.fraunhofer.de) (I. Vogel);

[isabella.grieser@sit.fraunhofer.de](mailto:isabella.grieser@sit.fraunhofer.de) (I. N. Grieser)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

automatic verification of social media postings remains an open issue, as they can not be classified with high accuracies [1]. The main objective of the subtask 1A on the CLEF-CheckThat! Lab [2, 3, 4] is to automatically determine the check-worthiness of tweets. The results of such an analysis could be used to create a set of tweets, that need to be further investigated by a human to verify their claims.

In this paper, a novel approach to detecting check-worthiness in tweets, that was developed in conjunction with the CheckThat! Lab, is presented. It combines multiple state-of-the-art transformer networks to an ensemble classification scheme and takes advantages of semi-supervised learning assisted by Generative Adversarial Networks (GAN). Furthermore, first experiments using Quantum Natural Language Processing (QNLP) and the framework lambeq were conducted. The proposed classification system ranked fifth place in the English 1A subtask.

The paper remainder is structured as follows: at first, an introduction to related work presented in the previous iteration of this competition is given in Section 2. Then, in Section 3 the task and the associated datasets are presented. In Section 4 an overview of the proposed approach is given along with explanations on the applied data augmentation and preprocessing. Section 5 showcases and discusses the experimental results achieved on the test set. The paper then concludes in Section 6 with an outlook for future work.

## 2. Related Work

The detection of check-worthiness in tweets has been a substantial part of the CheckThat! Lab over the past years [5, 4]. The lab held in 2021 already dealt with detecting the check-worthiness of tweets in the context of the pandemic. Back then, the best performing group proposed the usage of an ensemble classifier [6] consisting of BERT, ALBERT, RoBERTa, DistilBERT and Funnel-Transformer fine-tuned on a corpus featuring check-worthy and non-check-worthy tweets. The team that came in second [7] took advantage of heavy data augmentation, such as machine-translation and word substitutions. Their model was based on a fine-tuned Bertweet on the augmented data. The team that ranked third in the competition analyzed the semantic similarity estimation provided by a fine-tuned sentence transformer [8].

## 3. Task & Dataset Description

This year’s CheckThat! Lab task 1 [3] revolved around identifying relevant claims in tweets. The task itself was divided into several subtasks, with the objective of task 1A being to predict, whether a given tweet is worth fact-checking. While the task covered several languages, such as Arabic, Bulgarian, Dutch, English, Spanish and Turkish, the authors only participated in the English variant of the subtask.

**Table 1**

Class distribution of the CheckThat! Lab 2022 task 1A English dataset and of the CheckThat! Lab 2021 task 1A English dataset

	2022		2021	
	$\mathcal{L}_0$	$\mathcal{L}_1$	$\mathcal{L}_0$	$\mathcal{L}_1$
Train Split	1675	447	532	290
Validation Split	151	44	80	60
Validation-Test Split	445	129	331	19
Sum	2271	620	2271	620

### 3.1. 2022 Dataset

Along with the lab, a dataset was released. It consisted of tweets mentioning coronavirus centered topics collected in a time frame between 2020 and 2022, which were then manually labelled by humans. Tweets that were mapped with a label of  $\mathcal{L}_0$  represented tweets that were not check-worthy, while those labeled as  $\mathcal{L}_1$  were considered check-worthy by the annotators. The dataset was divided into four data splits, a train split, a validation split, a validation-test split and a test split. For all splits, except the test split, the annotations were released ahead of the lab evaluation phase. The label distribution showcased in Table 1 suggest that the dataset suffers from heavy class imbalance. Thus, this needs to be considered during training.

### 3.2. 2021 Dataset

Additionally, to the data from the current CheckThat! Lab, the dataset of the previous iteration of the lab [4] served as training data for the semi-supervised GANs and were used to train the meta classifier of the ensemble classifier. Some samples can be found in the dataset from 2022 and 2021. Thus, they were removed from the 2021 splits to avoid any duplicates. Similar to the 2022 dataset, this dataset was also heavily imbalanced regarding the class distribution (Table 1).

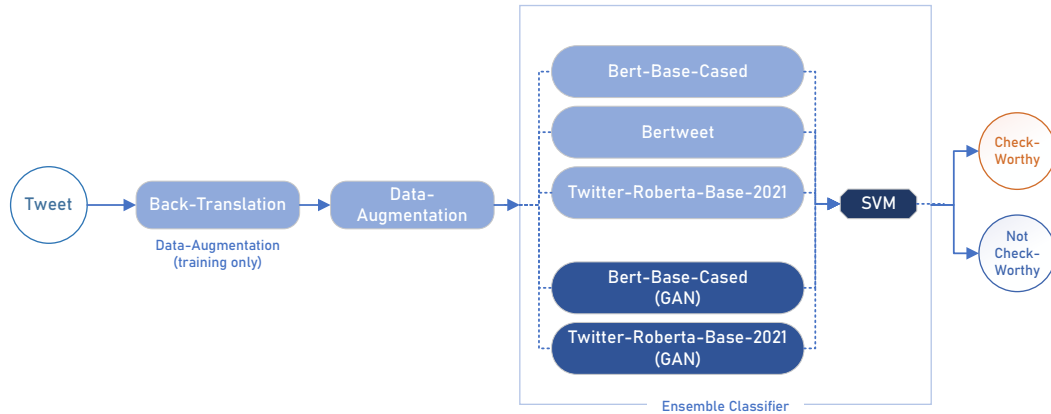
## 4. Methodology

In this section, the methods serving as candidates for the detection pipeline are presented. The overall concept used for the final submission is visualized in Figure 1. It consists of an ensemble classification scheme that takes advantage of data augmentation during training as well as several preprocessing steps.

### 4.1. Data Augmentation

As observed in Section 3, the training data suffers from data scarcity and class imbalance. To tackle the issue of data scarcity, the training data underwent a data augmentation process before applying any preprocessing.

**Figure 1:** Visualization of the check-worthiness detection pipeline used in the final submission



Back-translation was applied to generate new training samples. In comparison to synonym exchange, back-translation has the advantage that it introduces newly crafted sentences that syntactically differ a lot from the original sentence, but stay semantically very close to them. Thus, back-translation for the languages Spanish and Chinese was considered by first translating the tweet from English to the target language and then back to English using Google Translate. Chinese was considered for making the model more robust to samples that contain language errors. An example is showcased in Table 2. The utilization of back-translation resulted in an enriched dataset featuring three times the size of the original dataset.

**Table 2**  
Result of back-translation applied on a sample tweet

<b>Original</b>	Morning Headlines: Canada is marking 1 year of the global pandemic, the latest on vaccines in Ontario, and Metrolinx gets a payment update. <a href="https://t.co/EzHH8iArkl">https://t.co/EzHH8iArkl</a>
<b>Spanish</b>	Titulares de la mañana: Canadá cumple 1 año de la pandemia mundial, lo último sobre vacunas en Ontario, y Metrolinx recibe una actualización de pago. <a href="https://t.co/EzHH8iArkl">https://t.co/EzHH8iArkl</a>
<b>Back-Translated</b>	Morning headlines: Canada marks 1 year of global pandemic, vaccine update in Ontario, and Metrolinx gets paid update. <a href="https://t.co/EzHH8iArkl">https://t.co/EzHH8iArkl</a>

## 4.2. Preprocessing

Before feeding the tweets into the classifiers, they need to undergo several preprocessing steps. In comparison to news articles or speeches, sentences of social media postings often do not follow correct grammar and moreover often feature an extensive usage of slang

words. Furthermore, emojis are used to further express emotions and special control characters, such as @ and #, are used to identify user mentions or hashtags on Twitter.

For this purpose, the Python package pysentimiento[9] was utilized. It contains functions that allow exchanging emojis with their corresponding descriptive tokens. It also allows identifying URLs, user mentions as well as hashtags. Since URLs do not contain any information that drives the check-worthiness of a tweet, they were removed. User mentions were generalized by exchanging them with a @user-token. The result on a sample is displayed in Table 3.

To solve the issue of the imbalanced class distribution of the dataset, random under-sampling was utilized. This resulted in the removal of randomly selected samples assigned to the majority class until the overall class distribution in the datasets were balanced out.

**Table 3**

Result of the preprocessing applied on a sample tweet

<b>Original:</b>	<p>⚠ Over 75% of #Covid19 vaccinations carried out in just 10 countries          !! Vaccine monopolies are creating artificial scarcity.          We urge members of the @wto to support the #TRIPSwavier proposal and help boost global vaccine production.          Our statement → <a href="https://t.co/JYr7MMbSzR">https://t.co/JYr7MMbSzR</a> ← <a href="https://t.co/Cm5G8k4yJ7">https://t.co/Cm5G8k4yJ7</a></p>
<b>Preprocessed:</b>	<p>emoji warning emoji Over 75% of covid19 vaccinations carried out in just 10 countries ! Vaccine monopolies are creating artificial scarcity.          We urge members of the @USER to support the tripswaiver proposal and help boost global vaccine production.          Our statement emoji right arrow emoji emoji left arrow emoji</p>

### 4.3. Ensemble Classifier

The ensemble classifier consisted of five weak models, that were fine-tuned to solve the check-worthiness detection task. A meta-classifier is then used to conduct weighted decision-making. Three of the weak classifiers were trained in a supervised fashion, while two of them took additionally advantage of semi-supervised learning. The models that were chosen were determined based on the achieved F1-scores on the test-validation split. For supervised training, the following base-models were used:

- Bert-Base-Cased [10]: The BERT model was pretrained on a corpus consisting of 11,038 unpublished books and English Wikipedia articles. For the classifier, the case-sensitive model was considered, as sometimes emotions can convey through uppercase writing.
- Bertweet [11]: Bertweet is a language model based on RoBERTa, that was trained on a large-scale Twitter corpus featuring 845M tweets collected within a timeframe ranging from 2012 to 2019 as well as 5M tweets related to the COVID-19 pandemic. Thus, this model already has an extensive vocabulary of coronavirus related tokens.

- Twitter-Roberta-Base-2021 [12] Similar to Bertweet, Twitter-Roberta-Base is model based on RoBERTa that was trained on a twitter corpus. The corpus hereby consisted of 123.86M tweets gathered in 2021.

Semi-supervised learning was realized with GAN-Bert [13]. GAN-Bert is a generative adversarial network, that synthesizes a set of artificial vector embeddings, which allows fine-tuned models to achieve high accuracies when trained on scarce labeled data. It also enables the utilization of unlabeled samples in the training process. Hereby, the aforementioned Bert-Base-Cased [10] and Twitter-Roberta-Base-2021 [12] models served as base-models.

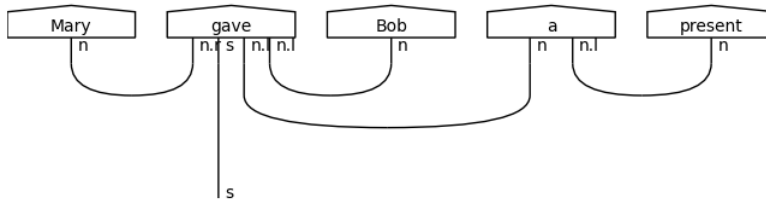
Each of the weak classifiers were trained on the train and validation split of the 2022 dataset using 10-fold cross-validation. This allows the models to be trained on as much data available as possible to identify optimal parameters as part of grid search. The models were trained for 15 epochs. To avoid overfitting, the use of model checkpoints discarding any model except the best performing one on the validation split has been considered. The learning rate was initially set to 0.00004, as lower learning rate resulted in bad convergence, while larger values worsened the validation scores. However, by using Adam as the optimizer, the learning rate is constantly adjusted during training.

During semi-supervised learning, the validation-test split of the 2022 dataset as well as the train, validation and test splits of the 2021 dataset were fed into the classifier as unlabeled data. Similar to the supervised learning classifiers, the GAN-Bert models were also trained using a 10-fold cross-validation. This results in a total of  $5 \times 10 = 50$  trained weak classifier. A meta classifier, here an SVM, takes the probabilities retrieved from the weak classifiers and outputs a combined classification decision. In comparison to averaging the probabilities, this has the advantage, that the single probabilities get weighted based on the classifiers individual performances.

#### 4.4. Quantum Natural Language Processing Modelling (QNLP)

A new method to deal with classification problems in the natural language processing field is quantum natural language processing (QNLP). One library that supports QNLP is lambeq [14].

**Figure 2:** Example of a string diagram annotated with pregroup grammar



In lambeq, sentences are represented as string diagrams (Figure 2), which can be further formulated as tensor networks. These describe the relationships between the different words in a sentence, which are defined according to the pregroup grammar

formalism. Lambeq supports the generation of two types of models: quantum-based models and classical models, which involves training classical tensor networks. For this competition, a classical model was trained to analyze the efficiency and usefulness of the framework.

There are three steps when creating a model in lambeq. First, the sentences need to be transformed into their corresponding sentence diagrams. This is done by using a parser, which calculates the syntax tree of the sentence and converts it to a sentence diagram. Then, a circuit is built based on the sentence diagrams. The form of the circuit differs depending on whether a quantum model or a classical model is built. At last, a model can be built based on the circuits of all sentence diagrams. This model can then be trained.

## 5. Experiments

Table 4 displays the classifications scores achieved by each model on the test set. The system used for the final submission consisted of an ensemble classification system featuring a meta classifier taking advantage of five weakly fine-tuned base-models. The QNLP-based model implemented using the framework lambeq posed several issues and thus was not included in the final submission.

**Table 4**

Results on the test set. Values displayed in bold indicate, which fine-tuned base-model was able to achieve the best score for a particular metric

	Accuracy		Precision		Recall		F1-Score	
	min	max	min	max	min	max	min	max
Bert-Base-Cased	<b>0.6980</b>	0.7517	<b>0.3778</b>	0.5263	0.0256	0.5128	0.0487	0.5194
Bertweet	0.6242	<b>0.7651</b>	0.3559	<b>0.600</b>	0.1538	0.6154	0.2449	0.5333
Twitter-Roberta-Base-2021	0.3557	0.6308	0.2857	0.3947	0.5641	0.9744	0.4418	0.5246
Bert-Base-Cased (GAN)	0.4228	0.6711	0.3089	0.4375	<b>0.7949</b>	0.9744	<b>0.4691</b>	<b>0.5882</b>
Twitter-Roberta-Base-2021 (GAN)	0.4026	0.6107	0.3015	0.3913	0.6923	<b>1.000</b>	0.4606	0.5496
Lambeq	-	0.7210	-	0.3570	-	0.3020	-	0.3270
Probability Averaging	-	0.6174	-	0.3929	-	0.8462	-	0.5366
Meta Classifier	-	0.6846	-	0.4394	-	0.7436	-	0.5524

### 5.1. Supervised Models

The models, that were trained in a supervised manner, consisted of a Bert-Base-Cased, Bertweet and Twitter-Roberta-Base-2021 model, by fine-tuning them on the train and validation split of the 2022 dataset using 10-fold cross-validation. The weak classifiers of the Bert-Base-Cased model were able to achieve the highest minimal accuracy across all

considered models. However, it did not perform well in terms of the achieved F1-score in comparison to the other models. The Bertweet model that was fine-tuned in supervision, maintained the highest accuracy and precision scores. It was also the best performing supervised classifier concerning the F1-score. The trained Twitter-Roberta-Base-2021 was able to achieve the highest recall score among the three. However, it did not perform well in terms of accuracy, precision and F1-score.

## 5.2. Semi-Supervised Models

In addition to the models that were trained in a supervised manner, experiments with GAN-Bert were conducted, which fine-tuned a Bert-Base-Cased and a Twitter-Roberta-Base-2021 model on labelled and unlabeled data. While accuracies achieved by both semi-supervised models were lower than their supervised counterparts, these models excelled in achieving by far the highest recall and F1-scores. Thus, this shows the effectiveness of semi-supervised learning for solving classification problems in natural language processing and especially in the detection of check-worthiness of tweets.

## 5.3. Ensemble Classification

For retrieving the final classification decision from the 50 weak classifiers, two types of ensemble classification were considered. The simplest approach is to average the probabilities gained from classifier and then estimation the class with the highest probability. This has the advantage that it does not need any additional data and does not increase the inference time. A major disadvantage, however, is that one bad performing weak classifier may have a massive impact on the overall classification result. Thus, a linear SVM was used, serving as a meta-classifier. While this requires additional data to train the meta-classifier, it conducts a weighted decision-making. This has the advantage that it is more robust to single classifiers that did not perform well. This advantage also is visible in the classification results on the test set, where the meta-classifier is able to achieve higher accuracy, precision and F1-scores than when taking solely advantage of probability averaging.

## 5.4. QNLP Model

For this competition, the usage of the new library lambeq posed some practical difficulties. First, the framework is not designed to handle multi-sentence inputs. This means that for tweets with multiple sentences, the data needs to be divided in its separate sentences. Furthermore, the default parser in lambeq, the BobcatParser, can not parse the tweets reliably in a structure which is necessary for the model creation and model training process. This makes the preprocessing and model creation process more difficult.

The results of the lambeq model were in total worse than the results of the final model. The results can be seen in Table 4. Thus, it was not included as part of the final classification system.



## 6. Conclusion

In this paper, a novel approach to detecting check-worthiness in tweets has been proposed. It takes advantage of state-of-the-art transformer models, such as BERT and Bertweet and combines supervised and semi-supervised learning by leveraging from ensemble learning. Moreover, there have been made first experiments on the utilization of QNLP to solve this classification problem. The evaluation had shown that QNLP and the framework lambeq still pose many issues. This might be connected to the fact, that both, the technique and the framework, are still fairly new at the time of writing. However, experimenting with semi-supervised learning, such as GAN-Bert, revealed promising results, as those models were able to achieve greater F1-scores than their counterparts learned in a supervised manner. Furthermore, taking advantage of a meta-classifier instead of averaging probabilities has shown to yield better results, as bad classifiers will not have a considerable influence on the final decision. Future work could revolve around taking advantage of other semi-supervised training techniques, such as virtual adversarial training.

## 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

- [1] G. K. Shahi, J. M. StruSS, T. Mandl, Overview of the CLEF-2021 CheckThat! lab task 3 on fake news detection, in: Working Notes of CLEF 2021—Conference and Labs of the Evaluation Forum, CLEF '2021, Bucharest, Romania (online), 2021. URL: <http://ceur-ws.org/Vol-2936/paper-30.pdf>.
- [2] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, J. M. Struß, T. Mandl, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, C. Li, S. Shaar, G. K. Shahi, H. Mubarak, A. Nikolov, N. Babulkov, Y. S. Kartal, J. Beltrán, M. Wiegand, M. Siegel, J. Köhler, Overview of the CLEF-2022 CheckThat! lab on fighting the COVID-19 infodemic and fake news detection, in: A. Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, F. Nicola (Eds.), Proceedings of the 13th International Conference of the CLEF Association: Information Access Evaluation meets Multilinguality, Multimodality, and Visualization, CLEF '2022, Bologna, Italy, 2022.
- [3] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, C. Li, S. Shaar, H. Mubarak, A. Nikolov, Y. S. Kartal, J. Beltrán, Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets, in: N. Faggioli, Guglielmo andd Ferro, A. Hanbury, M. Potthast

- (Eds.), Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.
- [4] S. Shaar, M. Hasanain, B. Hamdan, Z. S. Ali, F. Haouari, M. K. Alex Nikolov, F. A. Yavuz Selim Kartal, G. Da San Martino, A. Barrón-Cedeño, R. Míguez, T. Elsayed, P. Nakov, Overview of the CLEF-2021 CheckThat! lab task 1 on check-worthiness estimation in tweets and political debates, in: Working Notes of CLEF 2021—Conference and Labs of the Evaluation Forum, CLEF '2021, Bucharest, Romania (online), 2021. URL: <http://ceur-ws.org/Vol-2936/paper-28.pdf>.
  - [5] P. Nakov, G. Da San Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, W. Mansour, B. Hamdan, Z. S. Ali, N. Babulkov, A. Nikolov, G. K. Shahi, J. M. StruSS, T. Mandl, M. Kutlu, Y. S. Kartal, Overview of the CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news, in: Proceedings of the 12th International Conference of the CLEF Association: Information Access Evaluation Meets Multiliguality, Multimodality, and Visualization, CLEF '2021, Bucharest, Romania (online), 2021. URL: [https://link.springer.com/chapter/10.1007/978-3-030-85251-1\\_19](https://link.springer.com/chapter/10.1007/978-3-030-85251-1_19).
  - [6] J. R. Martínez-Rico, J. Martínez-Romo, L. Araujo, Nlp&ir@uned at checkthat! 2021: Check-worthiness estimation and fake news detection using transformer models, in: CLEF, 2021.
  - [7] X. Zhou, B. Wu, P. Fung, Fight for 4230 at checkthat! 2021: Domain-specific preprocessing and pretrained model for ranking claims by check-worthiness, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021, volume 2936 of CEUR Workshop Proceedings, CEUR-WS.org, 2021, pp. 681–692. URL: <http://ceur-ws.org/Vol-2936/paper-57.pdf>.
  - [8] I. B. Schlicht, A. F. M. de Paula, P. Rosso, UPV at checkthat! mitigating cultural differences for identifying multilingual check-worthy claims, CoRR abs/2109.09232 (2021) 2021. URL: <https://arxiv.org/abs/2109.09232>. arXiv:2109.09232.
  - [9] J. M. Pérez, J. C. Giudici, F. Luque, pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks, 2021. arXiv:2106.09462.
  - [10] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: <http://arxiv.org/abs/1810.04805>. arXiv:1810.04805.
  - [11] D. Q. Nguyen, T. Vu, A. T. Nguyen, BERTweet: A pre-trained language model for English Tweets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 9–14.
  - [12] F. Barbieri, J. Camacho-Collados, L. Espinosa-Anke, L. Neves, TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification, in: Proceedings of Findings of EMNLP, 2020.
  - [13] D. Croce, G. Castellucci, R. Basili, 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, Association for Computational Linguistics, Online, 2020, pp. 2114–2119. URL:

<https://www.aclweb.org/anthology/2020.acl-main.191>.

- [14] D. Kartsaklis, I. Fan, R. Yeung, A. Pearson, R. Lorenz, A. Toumi, G. de Felice, K. Meichanetzidis, S. Clark, B. Coecke, lambeq: An efficient high-level python library for quantum nlp, arXiv preprint arXiv:2110.04236 (2021).