Zorros at CheckThat! 2022: Ensemble Model for Identifying Relevant Claims in Tweets

Nicu Buliga[1,2], Madalina Raschip[1]

[1]Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iasi, Iasi, Romania
[2]Bitdefender, Iasi, Romania

Abstract

This paper describes the system used by Zorros team in the CLEF2022 CheckThat! Lab for Task 1 on identifying relevant claims in tweets. Task 1 was divided into four subtasks, which try to detect if the tweets are worth fact-checking (1A), contain verifiable factual claims (1B), are harmful to society (1C) and are attention-worthy (1D). For each subtask, we proposed different models based on pre-trained transformer models that helped us achieve the first position for subtasks 1C and 1D, the second position for subtask 1A, and the fifth position for subtask 1B.

Keywords

check-worthiness, COVID-19, transformer models, ensemble

1. Introduction

Social media platforms like Twitter play a major role in facilitating human communication and socialization by sharing different information, thus it is widely used by almost everyone who interacts with technology. Despite all the advantages, it has some dark sides, like fake news which spread faster than authentic ones and has increased considerably with this pandemic situation. A lot of false information spread online during the COVID-19 disease outbreak, from discrediting the threat of COVID-19 to conspiracy theories about vaccines. Misinformation on social media about COVID-19 is linked to vaccine hesitancy [1]. Therefore, the process of automatically identifying fake news is a very crucial and hard challenge for social media platforms, because even humans can not distinguish between fake and authentic news accurately. The CLEF2022 CheckThat! lab [2] is a good and relevant initiative for the current times since there is an urgent need for solutions to combat misinformation.

In this paper, we present an approach used for solving Tasks 1 (English) on Identifying Relevant Claims in Tweets [3] of CLEF 2022 CheckThat! Lab [2] [4]. Task 1 was divided into four subtasks: 1A - Check-worthiness of tweets, 1B - Verifiable factual claims detection, 1C - Harmful tweet detection and 1D - Attention-worthy tweet detection. The approach used for each subtask is based on transformers. The key innovation of transformers is the introduction of a self-attention mechanism. The computation for each item is independent of all the others, which means that it can be easily parallelized. This parallelism enabled transformers to be trained on large general purpose corpora, leading to pre-trained transformer models like BERT

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
nicu.buliga2000@gmail.com (N. Buliga); madalina.raschip@uaic.ro (M. Raschip)

[5] and T5. Those pre-trained models can be used to transfer knowledge to other NLP tasks, leading to significant improvements. Different pre-trained BERT [5] and RoBERTa [6] models are used in ensemble models to solve the subtasks 1A-1D.

The rest of the paper is organized as follows: section 2 - Related Work, section 3 - Problem Description, section 4 - Methodology, where the models used are described, section 5 - Evaluation and Results from the competition and section 6 - Conclusion and Future Work.

2. Related Work

Fake news detection has gained a lot of attention in the last years, especially during the worldwide COVID-19 pandemic, because of the misinformation spreading about COVID-19 on social media and there is a need of platforms to prevent it. Thus, many systems have been tested and used for detecting fake news, but they can not classify information accurately, because of their inability to fully understand the data.

Some of the most recent deep learning models used for fake news detection tasks are described below. The paper [7] presents a hybrid Neural Network architecture that combines the capabilities of CNN and LSTM. Two different dimensionality reduction approaches, Principle Component Analysis and Chi-Square are used in order to reduce the dimensionality of the feature vectors before passing them to the classifier. To develop the idea, the authors acquired a dataset from the Fake News Challenges website which has four types of stances: agree, disagree, discuss, and unrelated. Their results show a 20% improvement in the F1-Score. Another model is described in [8]. The authors used a transformer model for fake news classification of a specific domain dataset, and included human justification and metadata for added performance. They have used multiple BERT models with shared weights between them to handle various inputs.

Related research areas are check-worthiness and credibility assessment of tweets. There are many approaches in literature for identifying check-worthiness in social media, starting from working with features like TF-IDF representations [9] until more recently used word embeddings from transformers. For example, the paper [9] analyzes different classifiers and uses different features like TF-IDF representations, part of speech tags, sentiment scores, and entity types to detect check-worthy factual claims in presidential debates.

CLEF has been organizing CheckThat! Labs since 2018. The best model [10] from 2018 for check-worthiness in political claims used a multilayer perceptron and many features like averaged word embeddings and bag-of-words representations. In the last two editions, with the emergence of the COVID-19 pandemic, the competition considered the task of check-worthiness of tweets about COVID-19. Also, transformer-based models have begun to be used often. In the last year, for the check-worthiness of tweets task the best approach [11] used several pre-trained transformer models, BERTweet [12] giving the best results on the development set.

Some recent works for credibility assessment are presented below. In [13], a semi-supervised ranking model is described to automatically evaluate the credibility of a tweet. In [14], a large multilingual dataset for fact checking is introduced along with several automated fact checking models based on transformers.

3. Problem Description

The Task 1 of the competition was separated into four subtasks, all of them related to the COVID-19 topic. The first three tasks are binary classification tasks, and the last one is a multiclass classification task. Every subtask had its own dataset, splitted into train, dev and test sets as shown in Table 1.

Table 1
Dataset Details

Task	Data		
	Train	Dev	Test
1A	2122	195	574
1B	3324	307	911
1C	3323	307	910
1D	3321	306	909

The label distributions are shown in Figure 1 for subtasks 1A, 1B, 1C and in Figure 2 for subtask 1D. We can clearly observe that for subtasks 1A, 1C and 1D the data are very unbalanced, so we have to use different techniques for balancing the data.

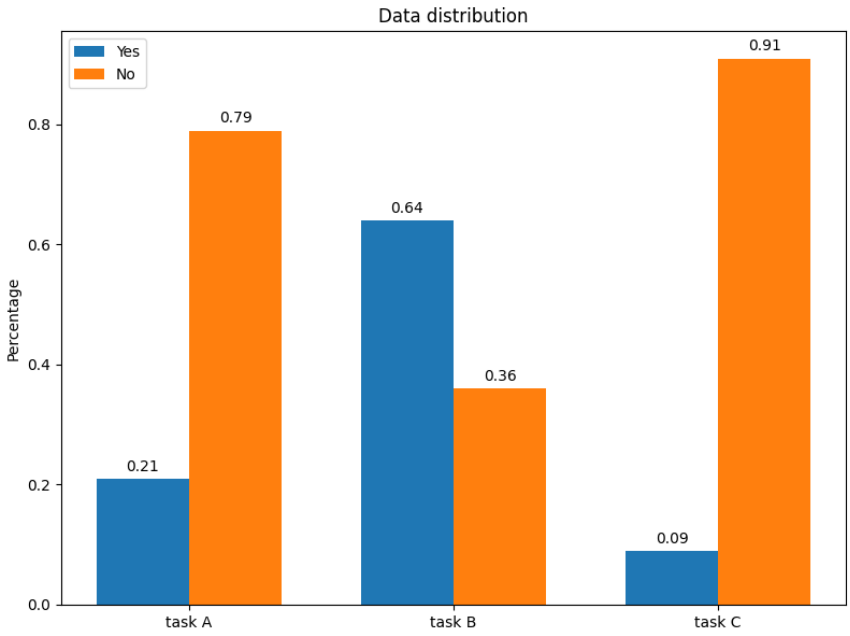

Figure 1: Data distribution for subtasks 1A, 1B, 1C

The description of the subtasks is given below:

- **Subtask 1A: Check-worthiness of tweets**
  Given a tweet, the task is to predict whether it is worth fact-checking, so it has two labels:

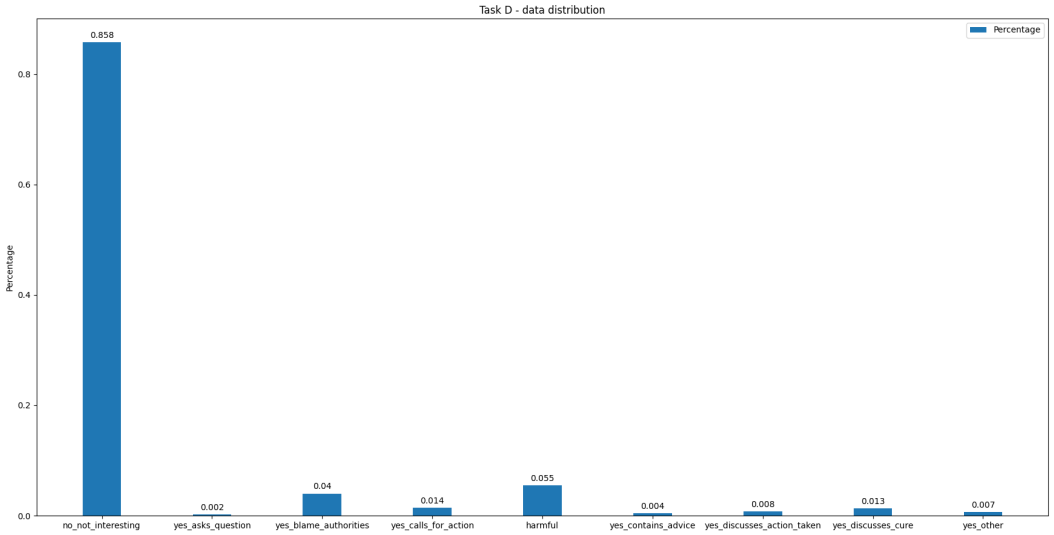

Figure 2: Data distribution for subtask 1D

Yes and No.

- **Subtask 1B: Verifiable factual claims detection**
  For this subtask, we have to predict whether a tweet contains a verifiable factual claim.
- **Subtask 1C: Harmful tweet detection**
  We have to predict whether a tweet is harmful to society.
- **Subtask 1D: Attention-worthy tweet detection**
  This subtask has nine class labels and we have to predict whether a tweet should get the attention of policy makers and why. The labels are:
  - No, not interesting
  - Yes, asks question
  - Yes, blame authorities
  - Yes, calls for action
  - Yes, harmful
  - Yes, contains advice
  - Yes, discusses action taken
  - Yes, discusses cure
  - Yes, other

4. Methodology

The proposed approach for solving the subtasks 1A-1D consists of four main steps: text preprocessing, tokenization for transformer based models, selection of the model architecture and the construction of the ensemble model. Each of them is described below.

4.1. Text Preprocessing

Given the small dataset for each task and the irrelevance of some tokens in the tweets for training our model, we performed the following modifications on raw tweets:

1. Lowercased the text of the tweet
2. Replaced all shorts, like *don't* to their normal form, *do not*
3. Removed all URLs, TAGs and non alphanumeric characters
4. Removed all stand-alone numbers

4.2. Tokenization

To pass the tweets to the pre-trained models like BERT and RoBERTa, we need to convert every tweet into a list of tokens based on the vocabulary of the model and then find the tokens ids and the attention masks. For this step, we used the respective tokenizer of each model because it already knows the accepted structure of the model and can easily convert the tweets.

The sentences are grouped into batches, so the tweets in the same batch must have the same number of tokens. To satisfy this condition, we used the parameters of the tokenizer by specifying a maximum length based on tokens list length distribution (100 was the best length in experiments). Therefore, shorter tweets will be padded with the same predefined token and longer tweets will be truncated.

4.3. Model Architecture

We have used only the encoder block of pre-trained BERT and RoBERTa models as the core of the final model, which can offer really good performance on NLP tasks. On top of that, we need a classification header for predicting, because until now we have only tweets embeddings. We need a binary classification model for subtasks 1A, 1B, 1C and a slightly different model for subtask 1D, because it has more than two labels. So, for the first three subtasks, we added an output neuron with the sigmoid as an activation function on top of the pre-trained model with a threshold at 0.5, giving us the probability of a tweet being in class *YES*. The loss function is Binary Cross Entropy and the optimizer is Adam with a weight decay of 0.01 and a learning rate of $2 \cdot 10^{-5}$. The classification head has a dropout layer with a rate of 0.5. Dropout was applied to avoid over-fitting. The model architecture is given in Figure 3.

For the final subtask 1D, we have as a classification header nine neurons with softmax as an activation function and Cross Entropy as a loss function. The predicted output gives us a probability distribution over all classes. The optimizer respects the same parameters from the first model, and the dropout layer also has a rate of 0.5. The model architecture for subtask 1D is given in Figure 4.

We fine-tuned these models for every subtask on the train set consisting of a concatenation between the *train set* and the *test set*. For testing models performance we have used the *dev set*. For training, we chose 20 epochs with data grouped into batches of 16 tweets. To deal with the unbalanced datasets, we have used weights for classes in the loss functions, essentially assigning a higher weight to the loss of the minor classes.

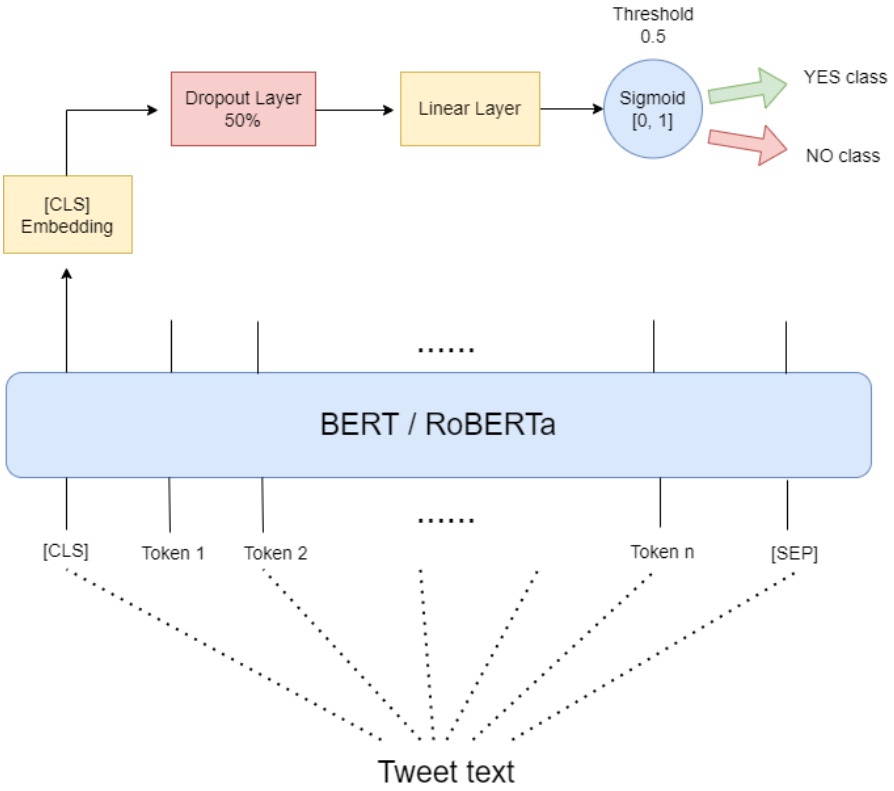

Figure 3: Model Architecture for Binary Classification

4.4. Fine-tuned BERT and RoBERTa Models

We fine-tuned the following ten models for every subtask:

4.4.1. TweetEval based models

TweetEval [15] is a pre-trained RoBERTa-base model, further trained on $\sim 58M$ tweets, randomly collected. The result of this step is a Twitter-domain adapted version of RoBERTa.

We also used a selection of five **TweetEval** models, each fine tuned for a specific task: Irony Detection[16], Offensive Language Identification[17], Emotion Recognition[18], Hate Speech Detection[19], Sentiment Analysis[20].

4.4.2. BERTweet models

BERTweet [12] is the first public large-scale language model pre-trained for English Tweets. BERTweet is trained based on the RoBERTa pre-training procedure. The corpus used to pre-train BERTweet consists of $850M$ English Tweets ($16B$ word tokens $\sim 80GB$), containing $845M$ Tweets streamed from $01/2012$ to $08/2019$ and $5M$ Tweets related to the COVID-19 pandemic.

We used three versions of this model:

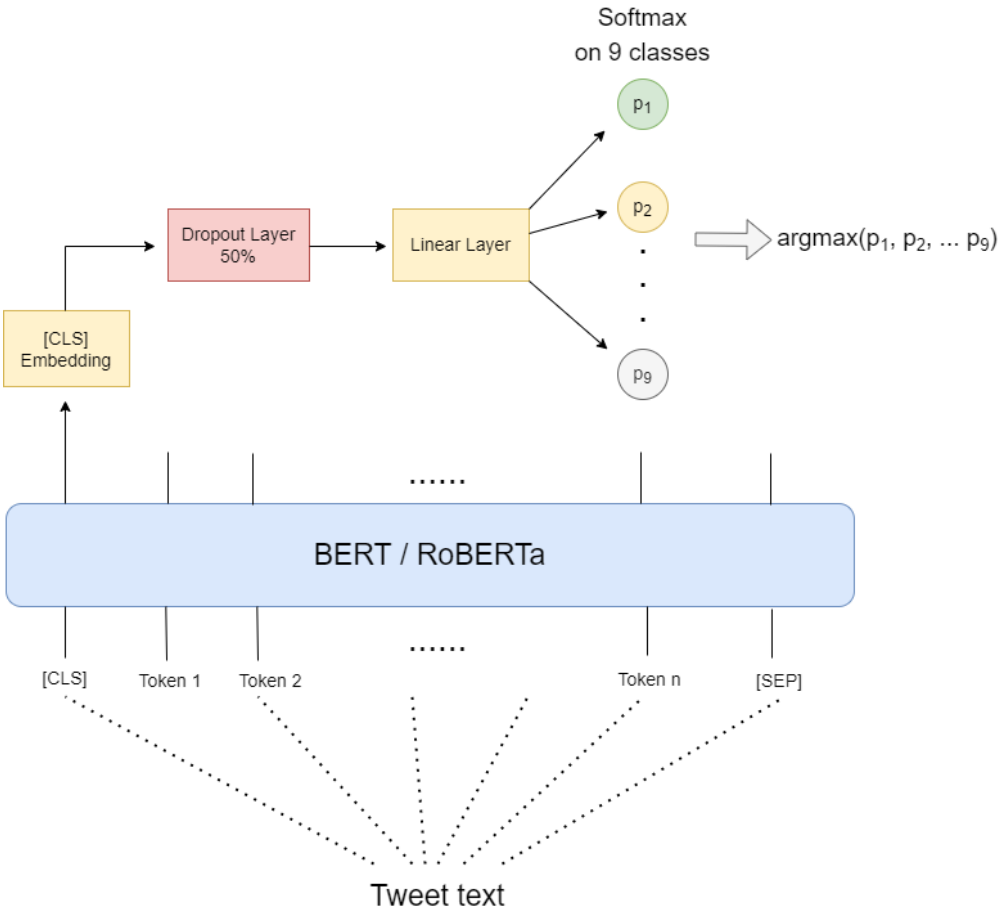

Figure 4: Model Architecture for Multi-class Classification

- BERTweet Large, the large version of the model, pre-trained on $873M$ English tweets (cased)
- BERTweet COVID-19 Base Cased, the base version of the model, additionally trained on $23M$ COVID-19 English tweets (cased)
- BERTweet COVID-19 Base Uncased, the base version of the model, additionally trained on $23M$ COVID-19 English tweets (uncased)

4.4.3. COVID-Twitter-BERT v2

It is a BERT-large-uncased model, pre-trained on a corpus of messages from Twitter about COVID-19. This model is identical to COVID-Twitter-BERT v1 [21] but was trained on more data, resulting in higher downstream performance.

The first version of the model was trained on $160M$ tweets collected between January 12 and April 16, 2020, containing at least one of the keywords "wuhan", "ncov", "coronavirus", "covid", or "sars-cov-2". These tweets were filtered and preprocessed to reach a final sample of $22.5M$

tweets, containing $40.7M$ sentences and $633M$ tokens, which were used for training.

4.5. Ensemble Model

For subtasks 1A, 1B and 1C we used an ensemble model to predict the labels. The structure of the ensemble model is composed of ten BERT and RoBERTa pre-trained and fine-tuned models described above, with another classification header on top of them, with one neuron and sigmoid as activation function, predicting the probability of a tweet being in class *YES*. Essentially, we took the predictions from every fine-tuned model without the sigmoidal activation function and feed them to the classification header. Therefore every tweet will have a new feature vector of length 10, where an element is a raw prediction from one model.

The model structure can be seen in Figure 5.

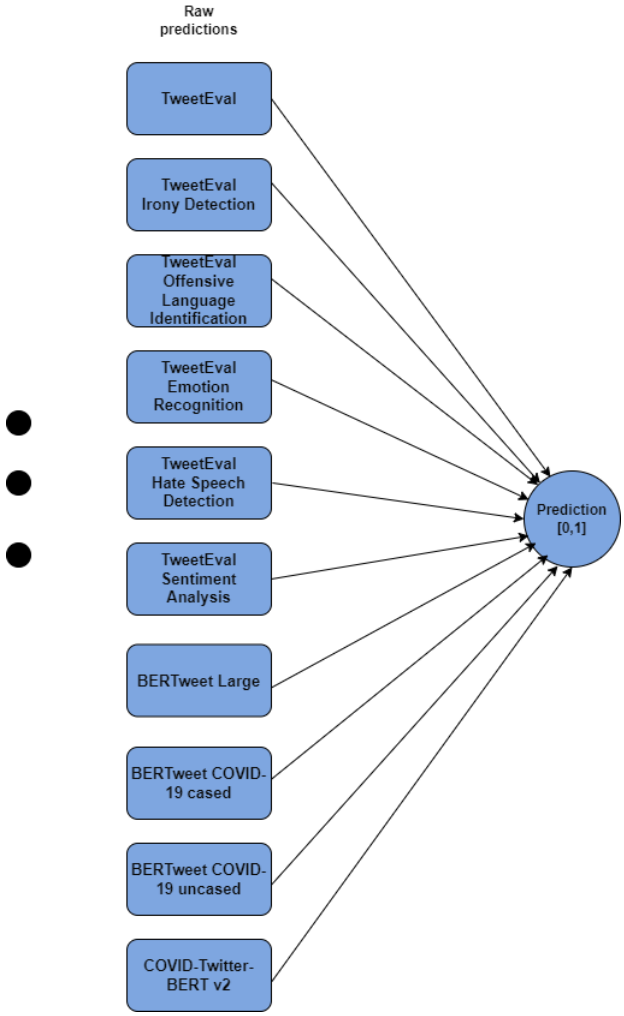

Figure 5: Ensemble Model Structure

Table 2
Subtask 1A results

Model	**Positive Class F1 Score**	Accuracy	Macro F1 Score	Macro Precision	Macro Recall
BERTweet Large	0.554	0.590	0.587	0.684	0.714
BERTweet COVID-19 Base Cased	0.579	0.698	0.671	0.678	0.729
BERTweet COVID-19 Base Uncased	0.568	0.724	0.683	0.676	0.714
TweetEval	0.601	0.698	0.679	0.696	0.754
TweetEval Irony	0.562	0.718	0.677	0.671	0.709
TweetEval Offensive	0.636	0.731	0.711	0.720	0.785
TweetEval Emotion	0.607	0.731	0.701	0.699	0.752
TweetEval Hate	0.598	0.684	0.669	0.696	0.753
TweetEval Sentiment	0.574	0.711	0.678	0.676	0.721
COVID Twitter BERT v2	0.609	**0.785**	0.730	0.724	0.738
Big Ensemble Model	**0.667**	0.771	**0.746**	**0.740**	**0.804**

This new model was trained for 50 epochs with a batch size of 8 and a learning rate of $2 \cdot 10^{-4}$, checking its performance at every epoch and saving the model with the best performance.

For task 1D, we did not use an ensemble model but the model with the best performance out of all ten, which was COVID-Twitter-BERT v2 [21].

5. Evaluation and Results

Task 1 is a classification task. Classification algorithms can be evaluated using several metrics including accuracy, precision, recall, and F1-score. For the subtasks 1A and 1C, the organizers used the F1 measure with respect to the positive class (minor class), for subtask 1B - accuracy and for 1D - weighted F1 score. We describe next the results obtained for every subtask on the last test set offered by the organizers and used for the contest. Tables 2-5 contain the results for all metrics; the column in bold from each table is the metric used by the organizers to establish the winners.

The results for subtask 1A can be found in Table 2. Here we have tested ten transformer-based models and an ensemble model built from these. We can observe that the ensemble model has the best performance for every metric except accuracy, which is not that relevant when the data is very unbalanced. Therefore, this is the model we used for the contest.

For subtask 1B, we tested only five TweetEval derived models and an ensemble model made from these, because of the increased computation time. The results are given in Table 3. The ensemble model yield the best results for this subtask, except for the macro recall, which is only 0.01 smaller than the maximum value of this metric. Therefore, we used the ensemble model for this subtask in the competition.

The results for subtask 1C are illustrated in Table 4. The tested models are the same as those from subtask 1B. Here the ensemble has the best results for two out of five metrics, including the F1 score for positive class, which was used as evaluation for the contest.

For subtask 1D, we did not use an ensemble. We tested four models due to increased computation time. The results are given in Table 5. Here the decision was simple, because COVID

Table 3
Subtask 1B results

Model	Accuracy	Macro F1 Score	Macro Precision	Macro Recall
TweetEval Irony	0.709	0.683	0.703	0.679
TweetEval Offensive	0.703	0.674	0.697	0.661
TweetEval Emotion	0.689	0.658	0.681	0.656
TweetEval Hate	0.691	0.677	0.680	0.674
TweetEval Sentiment	0.703	0.683	0.693	**0.680**
Small Ensemble Model	**0.709**	**0.683**	**0.703**	0.679

Table 4
Subtask 1C results

Model	Positive Class F1 Score	Accuracy	Macro F1 Score	Macro Precision	Macro Recall
TweetEval Irony	0.347	0.625	0.542	0.569	0.625
TweetEval Offensive	0.396	0.711	0.602	0.597	0.662
TweetEval Emotion	0.392	0.753	**0.618**	0.608	0.650
TweetEval Hate	0.337	**0.796**	0.608	**0.612**	0.605
TweetEval Sentiment	0.370	0.729	0.598	0.592	0.636
Small Ensemble Model	**0.397**	0.685	0.592	0.599	**0.671**

Table 5
Subtask 1D results

Model	Weighted F1 Score	Accuracy	Weighted Precision	Weighted Recall
COVID Twitter BERT v2	**0.725**	**0.721**	0.735	**0.721**
BERTweet Large	0.716	0.713	0.724	0.713
TweetEval Base	0.706	0.713	0.702	0.713
BERTweet COVID-19 Base Uncased	0.724	0.717	**0.737**	0.717

Twitter BERT v2 had the best results on three out of four metrics, including the one used for evaluation.

6. Conclusions and Future Work

In this paper, we present the results obtained in Task 1 of the CLEF 2022 CheckThat! lab. Different models based on pre-trained BERT encoders and fine-tuned BERT models offered really good results. For the first three subtasks we used ensemble models, while for subtask 1D a fine-tuned BERT model was enough. We obtained first place for subtasks 1C and 1D, second place for subtask 1A and fifth place for subtask 1B. The experimental results show the power of the existing pre-trained models, no longer being necessary to retrain a model from scratch, saving a lot of time on computation.

For future work, we will focus on learning more features from a sentence, making different

combinations from BERT layers, like pooling the last four layers, or concatenating them to obtain a higher understanding of the semantic in a sentence.

Acknowledgement

This paper is partially supported by the Competitiveness Operational Programme Romania under project number SMIS 124759 - RaaS-IS (Research as a Service Iasi).

References

[1] F. Pierri, B. L. Perry, M. R. DeVerna, K.-C. Yang, A. Flammini, F. Menczer, J. Bryden, Online misinformation is linked to early covid-19 vaccination hesitancy and refusal, Scientific reports 12 (2022) 1–7.

[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, The clef-2022 checkthat! lab on fighting the covid-19 infodemic and fake news detection, in: M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, V. Setty (Eds.), Advances in Information Retrieval, Springer International Publishing, Cham, 2022, pp. 416–428.

[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: Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[4] 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: Proceedings of the 13th International Conference of the CLEF Association: Information Access Evaluation meets Multilinguality, Multimodality, and Visualization, CLEF '2022, Bologna, Italy, 2022.

[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[7] M. Umer, Z. Imtiaz, S. Ullah, A. Mehmood, G. S. Choi, B.-W. On, Fake news stance detection using deep learning architecture (cnn-lstm), IEEE Access 8 (2020) 156695–156706. doi:10.1109/ACCESS.2020.3019735.

[8] D. Mehta, A. Dwivedi, A. Patra, M. Anand Kumar, A transformer-based architecture for fake news classification, Social Network Analysis and Mining 11 (2021) 39. URL: https://doi.org/10.1007/s13278-021-00738-y. doi:10.1007/s13278-021-00738-y.

[9] N. Hassan, C. Li, M. Tremayne, Detecting check-worthy factual claims in presidential

debates, in: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, 2015, pp. 1835–1838.
[10] C. Zuo, A. Karakas, R. Banerjee, A hybrid recognition system for check-worthy claims using heuristics and supervised learning, in: CEUR workshop proceedings (Vol. 2125), 2018.
[11] J. R. Martinez-Rico, J. Martinez-Romo, L. Araujo, Nlpir@uned at checkthat! 2021: check-worthiness estimation and fake news detection using transformer models, in: CEUR workshop proceedings (Vol. 2936), 2021.
[12] 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.
[13] A. Gupta, P. Kumaraguru, C. Castillo, P. Meier, Tweetcred: Real-time credibility assessment of content on twitter, in: International conference on social informatics, 2014, pp. 228–243.
[14] A. Gupta, V. Srikumar, X-fact: A new benchmark dataset for multilingual fact checking, in: arXiv preprint arXiv:2106.09248, 2021.
[15] 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.
[16] C. Van Hee, E. Lefever, V. Hoste, Semeval-2018 task 3: Irony detection in english tweets, in: Proceedings of The 12th International Workshop on Semantic Evaluation, 2018, pp. 39–50.
[17] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval), in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 75–86.
[18] S. Mohammad, F. Bravo-Marquez, M. Salameh, S. Kiritchenko, Semeval-2018 task 1: Affect in tweets, in: Proceedings of the 12th international workshop on semantic evaluation, 2018, pp. 1–17.
[19] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63. URL: https://www.aclweb.org/anthology/S19-2007. doi:10.18653/v1/S19-2007.
[20] S. Rosenthal, N. Farra, P. Nakov, Semeval-2017 task 4: Sentiment analysis in twitter, in: Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017), 2017, pp. 502–518.
[21] M. Müller, M. Salathé, P. E. Kummervold, Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter, arXiv preprint arXiv:2005.07503 (2020).