Multilingual Models for Checkworthy Social Media Posts Detection

Anonymous ACL submission

Abstract

This work presents an extensive study of transformer-based NLP models, dedicated to detection of social media posts containing verifiable factual claims and harmful claims. The document summarizes activities carried out during the pipeline execution, which led to the design of the NLP models for post detection. These activities included dataset collection, dataset pre-processing, architecture selection, setting up the experiments, model training (fine-tuning), model testing and implementation. Comprehensive analysis of various models was conducted. Special attention was focused on multilingual models, which are capable of processing English social media posts and simultaneously posts of low-resource languages, like: Polish, Czech, Slovak and Bulgarian. The obtained results were validated with state-of-the-art models and the comparison proved the robustness of the created models. The work's novelty consists in multi-label, multilingual classification models, which can efficiently perform simultaneous detection of harmful posts and of social media posts containing verifiable factual claims.

1 Introduction

The application of NLP methods along with the development of transformer architectures reaches broader subject domains. Nowadays, a lot of works focus on exploring the possibility to deploy NLP techniques in the combat against misinformation, fake news or propaganda. Although there is a consensus in the community, that the concept of completely replacing a human in the process of detecting fake news is currently rejected, the transformer architectures are being studied as solutions that can significantly optimize and improve the work of human fact checkers. With the development of electronic communication, particularly social media, and the simultaneous increase in awareness and also responsibility of governmental, social and opinion-forming institutions, the need to create applications and tools to combat fake news and disinformation is growing.

The ambiguity and diversity of natural languages, the high intensity of irony and satire in social media texts and posts, and the presence of a cultural context make the task of detecting fake news a challenging and relatively time-consuming process. The process requires the involvement of fact checkers, i.e. experts with specific knowledge in this domain area. In order to relieve the fact checkers, overwhelmed with information and to facilitate their work, NLP and DL (deep learning) methods are applied. The paper is dedicated to AI models facilitating the work of human fact checkers, it presents models for the detection of verifiable factual claims and harmful claims.

Detection of checkworthy claims represents the first stage in the fake news detection process (Cheema et al., 2022). Verifiable factual claims are posts that state a definition, mention a quantity in the present or the past, make a verifiable prediction of the future, reference laws, procedures, rules of operation, discuss images or videos, and make statements on correlation or causation (Nakov et al., 2022). Harmful claims are defined as offensive and/or hateful content on social media that can harm an individual, organization, and society (Nakov et al., 2022). It is assumed that automatic checkworthy claim detection can significantly support the work of human fact checkers.

The paper presents a series of experiments consisting in training (fine-tuning) existing transformer architectures in order to perform checkworthy claim detection tasks, which are also classification tasks. The following architectures were fine-tuned: DistilBERT (Sanh et al., 2019), BERT-large (Devlin et al., 2019), BERT-base (Devlin et al., 2019), XLM-RoBERTa-base (Conneau et al., 2019), XLM-RoBERTa-large (Conneau et al., 2019). The study applied the Flair tool (Akbik et al.,

2019) and the cloud computing platform Google Colaboratory (Google). An important aspect of the work was to create models that would be characterized not only by a high level of performance but that would also be easy to implement, use and maintain. Apart from the analysis of the typical metrics of NLP models for classification tasks (like accuracy, recall, f1-score) the inference times were also analysed. The research also focused on obtaining models that would support multilingual texts, in particular for low-resource languages such as Slovak, Czech and Polish.

The paper presents the entire pipeline starting from data selection and pre-processing, through the selection and fine-tuning of architectures, to model testing and additional validation in real life applications. Three types of models were obtained: models for detecting verifiable factual claims, models for detecting harmful claims, and models that simultaneously detect verifiable factual claims and harmful claims. The verification of the results confirms the reliability and usefulness of the proposed solutions. The data used in the training process was on multiple subjects, which makes the models effective at detecting checkworthy claims in posts with multiple topics.

The key contributions of the paper are:

- Creating models based on the XLM-RoBERTa-large and the XLM-RoBERTa-base architectures for simultaneous detection of verifiable factual claims and harmful claims in multilingual posts; tests were carried out for the following languages: English, Arabic, Bulgarian, Dutch, Turkish, Polish, Slovak, Czech.

- Selecting and composing multi-subject and multilingual data collections containing verifiable factual claims allowing for the creation of above mentioned models.

- Conducting a detailed inference time analysis of models implemented using (a) only CPU units; (b) also GPU units. The obtained times show that the models inference is much faster than that of humans.

2 Related Works

With the emergence of transformer architectures in the NLP domain, solutions focusing on automated fake news detection began to appear. Such attempts are presented in (Kula et al., 2021), where the BERT architecture (Devlin et al., 2019) and other derivative architectures like RoBERTa (Liu et al., 2019), DistilBERT (Sanh et al., 2019) or even the autoregressive DistilGPT2 (Sanh et al., 2019) and XLNet (Yang et al., 2019) were applied to classifying news articles (mainly on political and social topics) into articles that contain/do not contain fake news. Transformer-based models were also applied in the fight against COVID-19 misinformation in (Glazkova et al., 2020; Li et al., 2021; Koloski et al., 2021), where they achieved the best results.

The task of fact checking is also gaining prominence in research. The first stage of the fact-checking task pipeline is the detection of claims in posts. Fact-checking methods not only allow to determine whether a given post is true or false (binary classification), but also to verify and potentially demystify the disinformation contained in the content, e.g. by highlighting verifiable claims in the text and linking to sources relevant for fact checkers.

The task of claim detection using various transformer architectures is discussed in articles (Gupta et al., 2021; Stammbach et al., 2022; Reddy et al., 2022; Nakov et al., 2022). Fine-tuned architectures of DistilBERT (Sanh et al., 2019), RoBERTa (Liu et al., 2019), and ClimateBERT (Webersinke et al., 2021) were applied to detect environmental claims using a dataset dedicated specifically to the environmental domain (Stammbach et al., 2022). The CLEF2022 (Nakov et al., 2022) competition was devoted to the detection of verifiable factual claims and harmful claims, in which the best results were achieved by methods based on BERT (Eyuboglu et al.), XLM-RoBERTa (Savchev), GPT-3 (Agresti et al.). In its turn, a claim-spotting system, called claimBuster (Hassan et al., 2017) was used to detect sentences containing claims in news articles about COVID-19 (Reddy et al., 2022). Models for detecting claims in any type of an online text are presented in (Gupta et al., 2021).

3 Overview of the Models' Architectures

In order to create an optimal model for detection of social media posts with verifiable factual claims and harmful claims, a number of experiments were carried out using architectures based on Transformers (DistilBERT, BERT-large, BERT-base, XLM-RoBERTa-base, XLM-RoBERTa-large). Standard

```
(pooler): RobertaPooler(
      (dense): Linear(in_features=768, out_features=768, bias=True)
      (activation): Tanh()
    )
  )
)
(decoder): Linear(in_features=768, out_features=2, bias=True)
(loss_function): CrossEntropyLoss()
```

Figure 1: Top layers of the XLM-RoBERTa-base for the one-class model.

```
(pooler): RobertaPooler(
      (dense): Linear(in_features=768, out_features=768, bias=True)
      (activation): Tanh()
    )
  )
)
(decoder): Linear(in_features=768, out_features=4, bias=True)
(loss_function): BCEWithLogitsLoss()
```

Figure 2: Top layers of the XLM-RoBERTa-base for the multilingual, multi-label model.

versions of the already mentioned architectures were applied and no additional changes were made to their hyperparameters (such as the typical number of layers or the number of self-attention heads). The models were trained as classifiers: the last (top) layer in each model is a linear dense layer. Following the standard PyTorch approach, the final activation is built into the loss function: a softmax layer (the `CrossEntropyLoss()` loss function) for the single-label case and a sigmoid layer (the `BCEWithLogitsLoss()` loss function) for the multi-label case (Foundation; Paszke et al., 2017). The top layers are also presented in Figure 1 and Figure 2.

As a result of experiments and an analysis of requirements for the claim detection tasks, the multilingual XLM-RoBERTa architecture was chosen to obtain the final models, due to the multilingual character of the detection task. The RobertaLayer of the XLM-RoBERTa-base architecture is shown in Figure 3.

4 Datasets

Data is a crucial component in the training of transformer-based architectures for downstream tasks. Appropriate selection of training data is a necessary precondition for the correct performance of the resulting models. Due to the very specific and unique tasks considered here, the availability of appropriate labeled data was very limited. The following datasets were applied in this work: CLEF2022 (task 1B and 1C) (Nakov et al., 2022), CLEF2021 (task 1B) (Shaar et al., 2021),

```
(0): RobertaLayer(
      (attention): RobertaAttention(
        (self): RobertaSelfAttention(
          (query): Linear(in_features=768, out_features=768, bias=True)
          (key): Linear(in_features=768, out_features=768, bias=True)
          (value): Linear(in_features=768, out_features=768, bias=True)
          (dropout): Dropout(p=0.1, inplace=False)
        )
        (output): RobertaSelfOutput(
          (dense): Linear(in_features=768, out_features=768, bias=True)
          (LayerNorm): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
          (dropout): Dropout(p=0.1, inplace=False)
        )
      )
      (intermediate): RobertaIntermediate(
        (dense): Linear(in_features=768, out_features=3072, bias=True)
      )
      (output): RobertaOutput(
        (dense): Linear(in_features=3072, out_features=768, bias=True)
        (LayerNorm): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
        (dropout): Dropout(p=0.1, inplace=False)
```

Figure 3: The RobertaLayer within the XLM-RoBERTa-base architecture.

LESA2021 (noisy and semi-noisy datasets) (Gupta et al., 2021), Monant dataset (Srba et al., 2019, 2022).

Dataset CLEF2022 (task 1B) contains Twitter posts, dedicated to the topic of COVID-19, which are binary-labeled as containing/not containing verifiable factual claims. The dataset contains five different languages (English, Turkish, Dutch, Arabic, Bulgarian) and they are highly unbalanced, there are many more items labeled as 1 (containing verifiable factual claims) than 0 (without verifiable factual claims). Table 1 shows the number posts in each category across the included languages.

Dataset CLEF2022 (task 1C) is very similar to CLEF2022 (task 1B) – it contains posts almost identical to those in CLEF2022 (task 1B), but labeled for containing/not containing posts with harmful claims. This dataset is also unbalanced: it contains significantly more 0 (not harmful) posts than 1 (harmful) posts.

Dataset CLEF2021 (task 1B) contains data on political debates, the content of the collection refers to a variety of topics that are the subject of current political debates. The samples were labeled as containing/not containing fact-check-worthy verifiable factual claims. The dataset is characterized by a large imbalance, it contains more non-check-worthy elements that do not contain verifiable factual claims.

The LESA2021 dataset is a collection of data on a variety of topics collected from Twitter (COVID-19 topic) and six publicly available benchmark datasets (Gupta et al., 2021).

Finally, the Monant dataset is a unique collection of data, containing posts and verified factual claims paired with them. The topics of the posts

Dataset language	Nr. of posts with verifiable factual claims	Nr. of posts without verifiable factual claims
English	3,040	1,753
Turkish	2,480	1,331
Dutch	1,861	2,162
Arabic	4,121	2,093
Bulgarian	2,697	1,329
Total	**14,199**	**8,668**

Table 1: Amount of verifiable and not verifiable factual claims in the CLEF2022 (task 1B) dataset (Nakov et al., 2022).

are diverse, and the collection is also multilingual.

The presented datasets were used to create several collections: an overview of these is given in Table 2. Each collection was split into train/validation/test folds (using (pandas development team, 2020; Akbik et al., 2019)); the sizes of the splits are shown in Table 3. In the remaining part of this section, we are going to describe all the individual collections in more detail.

Collection 1 Collection 1 is based entirely on the CLEF2022, task 1B dataset, contains all items and all five original languages. The task is to classify posts as containing / not containing verifiable factual claims. The validation (dev) subset is a combination of dev.csv and dev_test.csv from the CLEF2022 (task 1B) dataset, the train and test sets are train.csv and test_GOLD.csv from the the CLEF2022 (task 1B) dataset, respectively.

Collection 2 Collection 2 is based entirely on the CLEF2022, task 1C dataset and the original five languages. The task is to classify posts as harmful or not. The collection consists of three subsets: the train, validation and test sets of 14018, 5124 and 3649 posts respectively. A version of this collection with only English posts was also created, which we refer to as collection 2tr.

Collection 3 Collection 3 is the same as collection 1, but with all non-English (i.e. Turkish, Dutch, Arabic and Bulgarian) posts automatically translated into English using Google Translate. The division of collection 3 into folds is as follows: 14032 (train), 5137 (dev), 3698 (test). The implemented splitting resulted from the subsets of the source dataset CLEF2022 (task 1B), which consists of four subsets and files (train.csv, dev.csv + dev_test.csv as dev and test_GOLD.csv) (Nakov et al., 2022).

Collection 4 Collection 4 is a combination of the CLEF2022, task 1B with selected data from the Monant dataset (Srba et al., 2019, 2022) and with selected data from CLEF2021 (task 1B). Concerning the Monant dataset, the selection of elements for collection 4 was based on the selection of posts scripted in the Latin alphabet or in Bulgarian or Arabic language. All posts selected from the Monant dataset were labeled as 1 (containing verifiable factual claims). The selection of posts from the CLEF2021 (task 1B) dataset consisted in selecting only posts labeled as 0 (posts without verifiable factual claims).

Collection 4 also aggregated posts from CLEF2021 (task 1B) by combining three posts into one post. The purpose of this operation was to avoid bias related to the length of posts. CLEF2021 (task 1B) contains very short posts relative to other items in collection 4, hence the requirement of the aggregation.

Collection 4 also contains selected semi-noisy items from LESA2021 (Gupta et al., 2021), labelled as 0 (posts without verifiable factual claims), along with posts translated from the CLEF2022 (task 1B) collection into Slovak, Czech and Polish. Finally, collection 4 also includes the entire collection 2 with items labeled as harmful vs. non-harmful posts. In summary, collection 4 contains four labels (verifiable factual claims, non-verifiable factual claims, harmful posts, non-harmful posts). 49,265, 10,148 and 3,705 items are used for training, validation and testing respectively.

All described collections have been pre-processed in accordance with the requirements of the conducted experiments. Two methods of pre-processing were applied. Pre-processing method 1 consisted in eliminating punctuation, www addresses, e-mail addresses, white spaces, emoticons, newlines, empty lines. This method of pre-processing was adopted as the fundamental one, because it was recognized that the crucial element in the implementation of the task of detecting posts with verifiable factual claims and with harmful content is the analysis of the content itself, and not of the form. Moreover the models were designed to detect content in social media posts that generally contain a severe level of noise. Noisy posts are often contain incorrect grammar and punctuation, hence the elimination of punctuation in the pre-processing phase is justified.

Coll. #	Datasets	Languages	Claim Detection Task	In
1	CLEF2022:1B	en, tr, nl, ar, bg	verifiable factual	exp2
2	CLEF2022:1C	en, tr, nl, ar, bg	harmful	exp3
2tr	CLEF2022:1C	en + (tr, nl, ar, bg) → en	harmful	exp4
3 (1tr)	CLEF2022:1B	en + (tr, nl, ar, bg) → en	verifiable factual	exp1
4	CLEF2022:1B; Coll. 2; Subset of: Monant; CLEF2021:1B; LESA2021	en, tr, nl, ar, bg + CLEF2022:1B → sk, cz, pl	verifiable factual + harmful	exp5

Table 2: Collections and data splits applied during the experiments conducting.

Coll. #	Train	Validation	Test
1	14032	5137	3698
2	14018	5124	3649
2tr	14018	5124	3649
3 (1tr)	14032	5137	3698
4	49265	10148	3705

Table 3: The data splits (by collection).

Pre-processing method 2 is an extension of pre-processing 1, consisting in the elimination of posts, which are not scripted in Bulgarian, Arabic or Latin alphabet, and in elimination of posts shorter than 15 characters and longer than 500 characters. Posts shorter than 30 characters, without including the digits were also eliminated due to the observation, that there are relatively short posts containing mainly digits. This pre-processing framework was imposed as an attempt to eliminate very noisy posts that convey no or relatively little content, and therefore do not contain verifiable factual claims.

The removal of posts exceeding 500 characters is based on a statistical analysis of the datasets, which revealed that the average length of a post in the Monant dataset is three times longer than the same in the CLEF2021 (task 1B) dataset.

Duplicates, hashtags and social media handles were also removed in order to eliminate details such as user account names and keywords from the posts. The motivation behind this was the observation that posts using a non-Latin alphabet very often contained hashtags and proper names scripted in the Latin alphabet. Therefore, the filtering of non-Latin posts resulted in a relatively large number of posts containing only hashtags and social media handles, and this led to an increase in the collection's noise level.

5 Setup and Execution of Experiments

The data preparation process as well as the execution of the experiments themselves were performed remotely in the Google Colaboratory cloud platform (Google), with the use of the Flair tool (version 0.7) (Akbik et al., 2019) and the pandas library (pandas development team, 2020). The actual instance used for the experiments contained the Tesla T4 card with 16 GB RAM, and the Intel Xeon CPU @2.00 GHz with 12.68 GB RAM.

In order to identify the optimal model for detecting posts with claims, five experiments were carried out. Two of them concerned the detection of posts containing verifiable factual claims (experiment 1 and experiment 2), and the next two were related to the detection of posts containing harmful claims (experiment 3 and experiment 4). The final experiment (experiment 5) concerned the multi-label model, which simultaneously performed the task of detecting posts containing verifiable factual claims and the task of detecting harmful claims.

In all experiments the data were split into train/validation/test folds. The sizes of the folds are presented in Table 3. The train set was used for training and the validation (a.k.a. dev) set was used to determine validation accuracy at each epoch. During training, both the last model (trained for the most epochs) and the best model (in terms of validation accuracy) were checkpointed. The best models were then taken and tested on the withheld test set. Tests on the last models were also conducted.

Table 2 gives an overview of which collections were used in which experiments. The collections themselves are described in sec. 4. In this section, we are therefore going to concentrate on other aspects of the experiments such as the hyperparameters and the training process.

Experiment 1 Experiment 1 uses collection 3, i.e. the version of collection 1 with non-English content

Name of the hyperparameter	Hyperparameter value
learning rate	3e-05
batch size	32
anneal factor	0.5
patience	3
max number of epochs	5/10*
mini batch chunk size	1**

Table 4: Training hyperparameters' values; * 10 epochs were applied for the XLM-RoBERTa-large architecture; * mini batch chunk size was set only for the XLM-RoBERTa-large architecture due to the GPU memory limitations.

machine-translated to English, where the task is to detect posts that contain verifiable factual claims. In experiment 1, only the BERT-large-uncased architecture was applied, the main hyperparameters of the experiment are presented in Table 4.

Experiment 2 Experiment 2 is related to multilingual architectures and original, non-English-translated posts. Collection 1 was used in the experiment. The following architectures were trained: XLM-RoBERTa-base, XLM-RoBERTa-large, multilingual DistilBERT-base, multilingual BERT-base. The same hyperparameters were applied as for experiment 1 (Table 4).

Experiment 3 Experiment 3 addresses detection of posts containing harmful claims based on collection 2, the collection containing five different languages. To select the optimal model, the following architectures were fine-tuned: XLM-RoBERTa-base, XLM-RoBERTa-large.

Experiment 4 Experiment 4 is analogous to experiment 3, except that it uses collection 2tr (i.e. the version of collection 2 with non-English posts machine-translated into English) and it uses a model designed for English – fine-tuning was conducted on DistilBERT-base.

Experiment 5 Experiment 5 was crucial for the conducted work and its scope and shape was the result of the other four experiments. The focus of the experiment was to create a single model capable of detecting posts containing verifiable factual claims and harmful claims, which can also detect posts not related to COVID-19 topics. All the experiments presented heretofore were based on collections related to the topics of COVID-19. To generate a multi-thematic, universal model, collection 4 was used.

6 Results

The models were compared in terms of metrics (recall, f1-score, accuracy), typically used in NLP models testing in classification tasks. Models were also compared in terms of the requirements during the implementation phase of the models in hands-on applications and in terms of performance of the models in the inference process.

The main metric that was taken into account was recall. The recall is used when the model's failure to detect the sought phenomenon (in our case, a specific post) results in significant negative effects. It was considered that it is a much worse case scenario for fact-checkers to overlook a post with verifiable factual claims and harmful claims than to unnecessarily check a post that does not contain verifiable factual claims or harmful claims. The analysis showed that the best results for recall were obtained for models based on the BERT-large-uncased (recall=0.7938) and the XLM-RoBERTa-large (recall=0.7724) architecture in case of verifiable factual claims detection, Table 5. In case of harmful claims detection the best models were the XLM-RoBERTa-large (recall=0.3765) and the XLM-RoBERTa-base (recall=0.3466).

Since the discrepancies between the mentioned models for individual detection tasks were relatively small in relation to recall, the second criterion was analysed, i.e. the ease of implementation and use of the solution. The XLM-RoBERTa models are multilingual and, unlike the BERT-large-uncased, do not require English-only texts. A comparative analysis of inference time was also performed and the XLM-RoBERTa-base model needed up to 3.5 times less time for the inference than the XLM-RoBERTa-large. Therefore, the XLM-RoBERTa-base model was considered the best performing model out of models mentioned in the previous paragraph. The inference time analyses were the subject of the experiment 2. In the analysis, apart from comparing models, comparisons between GPU and CPU computing platforms were conducted, Table 6. The results of the inference time analysis showed that the use of AI models jointly with GPU cards allows for a significant acceleration of the process of claims detection in posts in relation to the time needed by a human. A human for the task of identifying

Model	Accuracy	Recall	Test dataset
XLM-RoBERTa-large 10 epochs	0.7558	0.7724	5 original languages
BERT-large-uncased	0.7388	0.7938	EN + translations into EN
XLM-RoBERTa-base	0.7520	0.6976	5 original languages
distilBERT-base-multilingual-cased	0.7223	0.6811	5 original languages
BERT-base-multilingual-cased	0.7239	0.6597	5 original languages

Table 5: Task of verifiable factual claims detection. Comparison between generated multilingual and English based language models. Results of experiment 1 and experiment 2.

sentences with claims needs about 30 seconds per sentence (Reddy et al., 2022). The obtained models at the same amount of time classify about 1,000 or even more than 2,000 posts (in the case of XLM-RoBERTa-base), which results several thousand times acceleration with comparison to the work done by a human.

The credibility of the presented models was verified by comparing them with the results from the CLEF2022 (Nakov et al., 2022) paper. In Table 7 results for 5 different languages are included, the models were tested on the same testing datasets as the models from the CLEF2022 article (in terms of accuracy). In comparison to the best models from CLEF2022, task 1B (Nakov et al., 2022), the proposed models showed better accuracy in two cases (BG and AR), similar accuracy (the difference is max. 2.7 points) in two cases (EN and NL) and worse (the difference is more than 6 points) accuracy in one case (TR). Considering the above, as well as the fact that the generated XLM-RoBERTa-large and XLM-RoBERTa-base models are multilingual, and the CLEF2022 paper models were aimed at processing only one specific language, the generated models were considered reliable, and the obtained results as SOTA for Bulgarian and Arabic.

Table 8 compares models for detection of verifiable factual claims with the model that simultaneously detects verifiable factual claims and harmful claims (multi-label XLM-RoBERTa-base). Based on the results collected in the table, it can be observed that the multi-label XLM-RoBERTa-base model, despite being dedicated to two different tasks, achieves better results than models dedicated only to the task of verifiable factual claims detection. This leads to the conclusion that the XLM-RoBERTa-base multi-label model is a reliable and credible model in the task of detecting posts with verifiable factual claims. In Table 8 results for low-resource languages are also presented, test sets for these languages (SK, CZ, PL) were created as translations of the original CLEF2022 (task 1B) datasets.

Regarding the detection of harmful claims, the XLM-RoBERTa-base is a multi-label model, which is the same multi-label model placed in Table 8, the model achieves slightly worse results than models dedicated only to the task of harmful claims detection (models with one class), Table 9. In Table 9 the results from the literature (CLEF2022, task 1C (Nakov et al., 2022)) were compared with the results achieved by the generated models. The f1-score metric was compared on the original CLEF2022 test sets (task 1C), for a total of 5 original languages considered jointly; for English + the 4 languages translated into English; and for the individual original languages (EN, TR, AR), which were considered separately.

7 Conclusion

The paper has shown how fine-tuned, pre-trained languages models can be used to accurately and simultaneously detect verifiable factual claims and harmful claims in posts (in a multi-label setup), and how the models perform during the inference process. A number of models were trained and compared, starting with unilingual (EN) models, through multilingual models, and ending with multilingual, multi-label models.

Furthermore it was studied how the multilingual models perform with low-resource languages like Slovak, Czech, and Polish in the task of verifiable factual claims detection. The presented results confirm that the generated models are credible and efficient for the task of detecting verifiable factual claims and harmful claims in posts. The reported results are in all cases based on evaluation on the withheld test sets. The trained models can be successfully used as tools supporting manual fact-checking processes conducted by humans. The future work will be dedicated to study the impact of particular data on multi-label models.

Nr. of items	Model			
	XLM-RoBERTa -large GPU **[s]**	XLM-RoBERTa -large CPU **[s]**	XLM-RoBERTa -base GPU **[s]**	XLM-RoBERTa -base CPU **[s]**
100	3.13	110.23	1.99	31.26
1000	32.97	1174.14	13.71	337.86
2000	61.48	2339.22	26.67	680.62

Table 6: The inference time in seconds of the XLM-RoBERTa-large and XLM-RoBERTa-base models, comparison between GPU computational platform (Tesla T4 & Intel(R) Xeon(R) CPU @ 2.00GHz) and CPU computational platform (AMD EPYC 7B12). Analysis made for the social media posts from the Monant dataset (Srba et al., 2019, 2022).

Language of the test dataset	**Accuracy**		
	best results of CLEF2022	XLM-RoBERTa-large 10 epochs	XLM-RoBERTa-base
EN	0.761	0.7331	0.7371
BG	0.839	0.8176	0.8511
NL	0.736	0.7239	0.7172
TR	0.801	0.7383	0.7539
AR	0.570	0.7861	0.7660

Table 7: Accuracy comparison between generated multilingual models for verifiable factual claims detection (experiment 2) and results from the CLEF2022 (task 1B) paper (Nakov et al., 2022). Tests done for 5 original languages from the CLEF2022 dataset.

Model	Recall	Test dataset
XLM-RoBERTa-large 10 epochs	0.7724	5 original languages
XLM-RoBERTa-base	0.6976	5 original languages
multi-label XLM-RoBERTa-base	0.8024	5 original languages
multi-label XLM-RoBERTa-base	0.8249	translations into SK
multi-label XLM-RoBERTa-base	0.8212	translations into CZ
multi-label XLM-RoBERTa-base	0.8282	translations into PL

Table 8: Task of verifiable factual claims detection, metric recall for the positive class. Comparison between multi-label and multilingual language models. Results of experiment 2 and experiment 5. Tests done for 5 original languages from the CLEF2022 dataset and for translations into Slovak, Czech and Polish.

Model	f1-score	Test dataset
multi-label XLM-RoBERTa-base	0.3741	5 original languages
XLM-RoBERTa-base	0.4032	5 original languages
distilBERT-base-uncased	0.3587	EN + translations into EN
model of Zorros (Nakov et al., 2022)	0.397	original EN
model of ARC-NLP (Nakov et al., 2022)	0.366	original TR
XLM-RoBERTa-base	0.5609	original AR
model of iCompass (Nakov et al., 2022)	0.557	original AR

Table 9: Task of harmful claims detection, metric f1-score for the positive class. Comparison between genereted models and results from the CLEF2022 (task 1C) paper (Nakov et al., 2022). Results of experiment 3 and experiment 5. Tests done for 5 original languages, for translations and singular languages from the CLEF2022 dataset.

References

Stefano Agresti, S. Amin Hashemian, and Mark J. Carman. Polimi-flatearthers at checkthat! 2022: Gpt-3 applied to claim detection. pages 422–427.

Alan Akbik. *flair flairNLP*. https://github.com/flairNLP/flair, Accessed= May 23, 2023.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *NAACL-HLT (Demonstrations)*, pages 54–59. Association for Computational Linguistics.

Gullal Singh Cheema, Sherzod Hakimov, Abdul Sittar, Eric Müller-Budack, Christian Otto, and Ralph Ewerth. 2022. MM-claims: A dataset for multimodal claim detection in social media. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 962–979, Seattle, United States. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ahmet Bahadir Eyuboglu, Mustafa Bora Arslan, Ekrem Sonmezer, and Mucahid Kutlu. Tobb etu at checkthat! 2022: Detecting attention-worthy and harmful tweets and check-worthy claims. pages 478–491.

The PyTorch Foundation. *PYTORCH DOCUMENTATION*. https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html, Accessed= February 23, 2023.

Anna Glazkova, Maksim Glazkov, and Timofey Trifonov. 2020. g2tmn at constraint@aaai2021: Exploiting CT-BERT and ensembling learning for COVID-19 fake news detection. *CoRR*, abs/2012.11967.

Google. *Welcome to Colaboratory*. https://colab.research.google.com/, Accessed= February 27, 2023.

Shreya Gupta, Parantak Singh, Megha Sundriyal, Md. Shad Akhtar, and Tanmoy Chakraborty. 2021. LESA: Linguistic encapsulation and semantic amalgamation based generalised claim detection from online content. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3178–3188, Online. Association for Computational Linguistics.

Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, Vikas Sable, Chengkai Li, and Mark Tremayne. 2017. Claimbuster: The first-ever end-to-end fact-checking system. *Proc. VLDB Endow.*, 10(12):1945–1948.

Boshko Koloski, Timen Stepisnik Perdih, Senja Pollak, and Blaz Skrlj. 2021. Identification of COVID-19 related fake news via neural stacking. *CoRR*, abs/2101.03988.

Sebastian Kula, Rafał Kozik, Michał Choraś, and Michał Woźniak. 2021. Transformer based models in fake news detection. In *Computational Science – ICCS 2021*, pages 28–38, Cham. Springer International Publishing.

Xiangyang Li, Yu Xia, Xiang Long, Zheng Li, and Sujian Li. 2021. Exploring text-transformers in AAAI 2021 shared task: COVID-19 fake news detection in english. *CoRR*, abs/2101.02359.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *CoRR*, abs/1907.11692.

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Rubén Míguez, Tommaso Caselli, Mücahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, Hamdy Mubarak, Alex Nikolov, and Yavuz Selim Kartal. 2022. Overview of the CLEF-2022 checkthat! lab task 1 on identifying relevant claims in tweets. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, volume 3180 of *CEUR Workshop Proceedings*, pages 368–392. CEUR-WS.org.

The pandas development team. 2020. pandas-dev/pandas: Pandas.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *NIPS-W*.

Revanth Gangi Reddy, Sai Chetan, Zhenhailong Wang, Yi R. Fung, Kathryn Conger, Ahmed Elsayed, Martha Palmer, Preslav Nakov, Eduard Hovy, Kevin Small, and Heng Ji. 2022. Newsclaims: A new benchmark for claim detection from news with attribute knowledge.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Aleksandar Savchev. Ai rational at checkthat! 2022: Using transformer models for tweet classification. pages 656–659.

Shaden Shaar, Maram Hasanain, Bayan Hamdan, Zien Sheikh Ali, Fatima Haouari, Alex Nikolov, Mücahid Kutlu, Yavuz Selim Kartal, Firoj Alam, Giovanni Da San Martino, Alberto Barrón-Cedeño, Rubén Míguez, Javier Beltrán, Tamer Elsayed, and Preslav Nakov. 2021. Overview of the CLEF-2021 checkthat! lab task 1 on check-worthiness estimation in tweets and political debates. In *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*, pages 369–392. CEUR-WS.org.

Ivan Srba, Robert Moro, Jakub Simko, Jakub Sevcech, Daniela Chuda, Pavol Navrat, and Maria Bielikova. 2019. Monant: Universal and extensible platform for monitoring, detection and mitigation of antisocial behavior. In *Proceedings of Workshop on Reducing Online Misinformation Exposure (ROME 2019)*, pages 1–7.

Ivan Srba, Branislav Pecher, Tomlein Matus, Robert Moro, Elena Stefancova, Jakub Simko, and Maria Bielikova. 2022. Monant medical misinformation dataset: Mapping articles to fact-checked claims. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*, New York, NY, USA. Association for Computing Machinery.

Dominik Stammbach, Nicolas Webersinke, Julia Anna Bingler, Mathias Kraus, and Markus Leippold. 2022. A dataset for detecting real-world environmental claims.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2021. Climatebert: A pretrained language model for climate-related text. *CoRR*, abs/2110.12010.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

A Work Limitations

The limitations apply primarily to the obtained models, they were trained, fine-tuned mainly on noisy and on semi-noisy data, and then tested only on noisy data. The models are therefore intended for this type of data, they can be used to detect verifiable factual and harmful claims in other types of texts, but the reliability of the models in this scope has not been tested due to the work concerns social media posts. The test sets derive from CLEF2022 (Nakov et al., 2022) and are dedicated to the COVID-19 topic, no testing was performed for data on other topics. The final models proposed are XLM-RoBERTa architecture models, i.e. they are multilingual models, but testing was carried out for 8 languages (AR, BG, NL, EN, TR, PL, SK, CZ).

B Artifacts licenses

The main tool used in the work was Flair library. This is including the usage and the modification of the code available at the tool web-page. Flair is based on the MIT license and is free of charge and allows using and also tool modification (Akbik et al., 2019), (Akbik). The version of Google Colaboratory, used in the work is free of charge and it access rules are described in the Google Terms of Service (Google). Pandas library is free software based on BSD 3-clause license (pandas development team, 2020). Datasets used in the work are free for general research use. The Monant dataset is not allowed to be re-shared, the rest datasets are publicly available.

The main purpose of the study was research work and the obtained models are intended for research purposes only. It is in full compliance with the requirements of the creators of datasets and tools used in the work.

C Additional Details of Experiments

Model	Training time [minutes]
multilabel	
XLM-RoBERTa-base	180
vfc XLM-RoBERTa-large	
10 epochs	227
vfc XLM-RoBERTa-base	55
vfc BERT-large-uncased	97
vfc distilBERT-base-multilingual-cased	30
vfc BERT-base-multilingual-cased	49
harm XLM-RoBERTa-base	55
harm distilBERT-base-uncase	17

Table 10: Training times comparison for different models, vfc - verifiable factual claim detection model, harm - harmful claims detection model.

The work uses version 0.7 of the Flair library, and the code prepared is based on available Flair documents and tutorials (Akbik). The code is used

Architecture	Number of parameters
BERT base	110
BERT large	336
DistilBERT	66
XLM-RoBERTa large	355
XLM-RoBERTa-base	125

Table 11: Architectures size comparison regarding the number of parameters (Sanh et al., 2019), (Devlin et al., 2019), (Conneau et al., 2019).

to launch training and then display the test results (Akbik). The discrepancies in the code from the Flair tutorials and the code used in the work relate to: different values and parameter settings, the choice of a different architecture, as well as the adaptation of the code to the requirements of generating a multi-label model. Essential parts of the code are available in the anonymized repository.[1]

[1] https://anonymous.4open.science/r/multilingual_checkworthy-F06D