Fired_from_NLP at CheckThat! 2024: Estimating the Check-Worthiness of Tweets Using a Fine-tuned Transformer-based Approach⋆

Md. Sajid Alam Chowdhury[1,*,†], Anik Mahmud Shanto[1,†], Mostak Mahmud Chowdhury[1,†], Hasan Murad[1] and Udoy Das[1]

[1] *Chittagong University of Engineering and Technology (CUET), Chittagong, Bangladesh*

Abstract

Due to immense usage and dependence on web-based and social media platforms, we nowadays come across a lot of information but all of them are not true. Thus, it is important to verify a statement before believing it. Therefore, Checking the validity of a statement has become a core research topic in Natural Language Processing (NLP) in both low-resource and resource-enriched languages. The CheckThat! Lab at CLEF 2024 has organized a shared task named Check-worthiness estimation (Task 1) where three datasets have been provided in the Arabic, English, and Dutch languages to determine whether a claim in a tweet and/or transcriptions is worth fact-checking. To perform the task, we have utilized several machine learning, deep learning, and transformer-based models to check which model performs best on the given datasets. Among all of these models, our proposed CW-BERT model has ranked 7^{th}, 10^{th}, and 12^{th}, scoring the F1 scores of 0.530, 0.543, and 0.745 in this task for the Arabic, English, and Dutch language respectively.

Keywords

Check-worthiness, Fact-checking, Tweets, Transcriptions, NLP, Transformer, Machine Learning, Deep Learning

1. Introduction

Social media platforms like Facebook and Twitter are now integral parts of our daily routines. These sites have revolutionized the way people communicate in the modern world. On the other hand, the rise and spreading of misinformation has become a matter of huge concern to the same extent. By spreading false news and misinformation using social media sites, public opinion and points of view can be changed. Therefore, as a vast amount of content is generated daily, it has become essential to identify which claims should be checked, whether they are facts or not, to allocate resources and stop the spread of misleading information effectively.

To overcome these challenges, The CheckThat! Lab [1] at CLEF 2024 has featured six distinct tasks described in the overview paper [2], each addressing different aspects of misinformation and content analysis on social media. Among these tasks, we have participated exclusively in Task 1: Check-Worthiness Estimation [3].

Verifying the accuracy of statements has emerged as a significant task within the field of Natural Language Processing. Like many other NLP tasks, this task has also been highly explored in high-resource languages like English [4]. Very few researches have been done in this domain in low-resource languages. Transformer-based approaches [5] have been used in all these research works. Before transformer-based approaches had begun to outperform all other approaches, machine learning (ML) algorithms [6] and deep learning (DL) techniques [7] have been employed for detecting if a tweet or statement should be verified or not.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

[*] Corresponding author.

[†] These authors contributed equally.

u1904064@student.cuet.ac.bd (Md. S. A. Chowdhury); u1904049@student.cuet.ac.bd (A. M. Shanto); u1904055@student.cuet.ac.bd (M. M. Chowdhury); hasanmurad@cuet.ac.bd (H. Murad); u1804109@student.cuet.ac.bd (U. Das)

The task of Check-Worthiness Estimation [3] at the CheckThat! Lab [1], part of CLEF 2024, seeks to bridge this gap by providing annotated datasets and encouraging the development of models that can accurately estimate the check-worthiness of texts. The introduction of multi-genre data further complicates the task, as tweets and transcriptions vary significantly in structure, style, and context. This task contains datasets in multiple languages (Arabic, English, and Dutch), making the task both linguistically and contextually challenging. Traditional approaches to check-worthiness estimation often rely on human annotators to determine if a claim is verifiable, potentially harmful, or otherwise significant. However, automating this process requires sophisticated natural language processing (NLP) techniques and models that are capable of handling diverse and complex data.

For this purpose, we have presented a detailed analysis of our approach to the Check-Worthiness Estimation task. We have also conducted a comparative analysis of several models, including Machine Learning models (Random Forest, SVM, XGBoost), Deep Learning models (LSTM, Bi-LSTM), and Transformer-based models (AraBERT [8] for the Arabic language, RobBERT [9] for the Dutch language, BERT-uncased [10] for the English language, and MultiLingual-BERT-uncased [10] for all three languages). We have named our approach of using the Multilingual-BERT-uncased model as CW-BERT.

We have used the respective datasets provided by the organizers for respective languages to train our respective models. After that, we have evaluated each of our models on the provided datasets. We have found that among all models, fine-tuned models based on transformers have obtained better results than all other models. The core contributions of our research are given below –

- We have developed a fine-tuned CW-BERT model specifically designed to assess the check-worthiness of claims in tweets and transcriptions across three languages: Arabic, English, and Dutch.
- We have conducted a comparative analysis among various models based on machine learning, deep learning, and transformer techniques to determine the most effective approach.

The implementation details of this task have been provided in this GitHub repository[1].

2. Related Works

The purpose of this work is to evaluate the value of fact-checking a claim made in a tweet and/or transcription. Previous works on this topic can be broadly categorized into machine learning, deep learning, and transformer-based approaches.

2.1. Previous Works Based on Machine Learning Approaches

An approach based on rule using feature engineering has been proposed [6]. The unsupervised approaches have been based on K-means clustering and the supervised approaches have been based on Cosine similarity, POS tags, and TF-IDF vectorization. An SVM-based model [11] has been proposed to predict whether a fact is worth checking. This approach focuses not only on sentence structure but also the context of the sentence.

2.2. Previous Works Based on Deep Learning Appoaches

Determining how trustworthy a claim is in the context of politics is a very important task. A deep learning approach with multiple tasks [12] has been used for predicting whether fact-checking should be given priority to a statement. A CNN-based deep learning model [7] has been suggested for obtaining semantic word embeddings while managing the complexity of natural language structures in diverse languages.

[1] https://github.com/Fired-from-NLP/CLEF-2024-check-worthiness-estimation

2.3. Previous Works Based on Transformer-based Approaches

Several researches have been done to estimate fact-checking worthiness in both low-resource and resource-enriched languages. This study suggests that NorBench [13] and NB-BERT-base [14] have been successfully employed for automated claim detection. Another research has proposed a transformer-based fine-grained technique to claim check worthiness [15].

3. Dataset

In our study, we have used respective datasets for three different languages (Arabic, Dutch, and English) provided by CLEF 2024 - CheckThat! Lab [1] for estimating the Check-worthiness of tweets and/or transcriptions. These datasets are categorized into two classes: ‘Yes’ (indicating the text is worth fact-checking) and ‘No’ (indicating otherwise). The datasets are divided into four sets: train, dev, test, and dev-test.

Table 1
Sample distribution across language

Language	Split	Label		Total
		Yes	No	
Arabic	Train	2243	5090	7333
	Dev	411	682	1093
	Dev-test	377	123	500
	Test	218	392	610
Dutch	Train	405	590	995
	Dev	102	150	252
	Dev-test	316	350	666
	Test	397	603	1000
English	Train	5413	17088	22501
	Dev	238	794	1032
	Dev-test	108	210	318
	Test	88	253	341

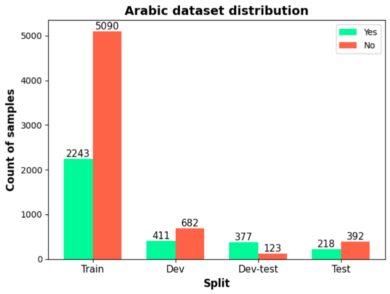

Figure 1: Distribution of Arabic Dataset Samples

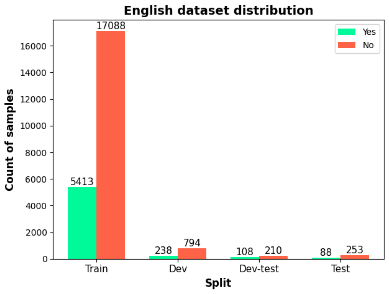

Figure 2: Distribution of English Dataset Samples

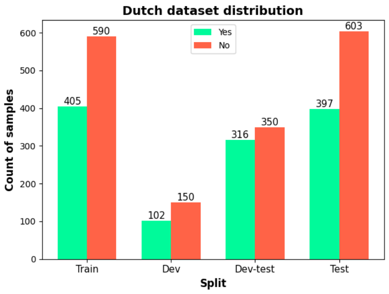

Figure 3: Distribution of Dutch Dataset Samples

For the English dataset, each split contains sentences comprising 8 to 9 words. The Arabic and Dutch datasets are 11 to 16 words and 7 to 9 words respectively. Figure 1, and Figure 2 show a significant imbalance in the English and Arabic datasets. The ‘Yes’ category, indicating check-worthy sentences, is substantially underrepresented compared to the ‘No’ category. Specifically, the English dataset contains 2243 samples labeled as ‘Yes’ and 5090 samples labeled as ‘No’. A similar imbalance is observed in the Arabic dataset, with 5413 ‘Yes’ samples and 17088 ‘No’ samples. Additionally, Table 1 reveals that the Dutch dataset demonstrates a smaller sample distribution across both categories. These datasets have been constructed from tweets and/or transcriptions, and reflect real-world text distributions.

4. Methodology

We have explained our methodology to develop models for check-worthiness estimation in this section. At first, we used some preprocessing strategies on the given datasets and then utilized a variety of deep learning and machine learning algorithms. Moreover, we have used different transformer models to develop the system. Figure 4 provides the summary of our methodology.

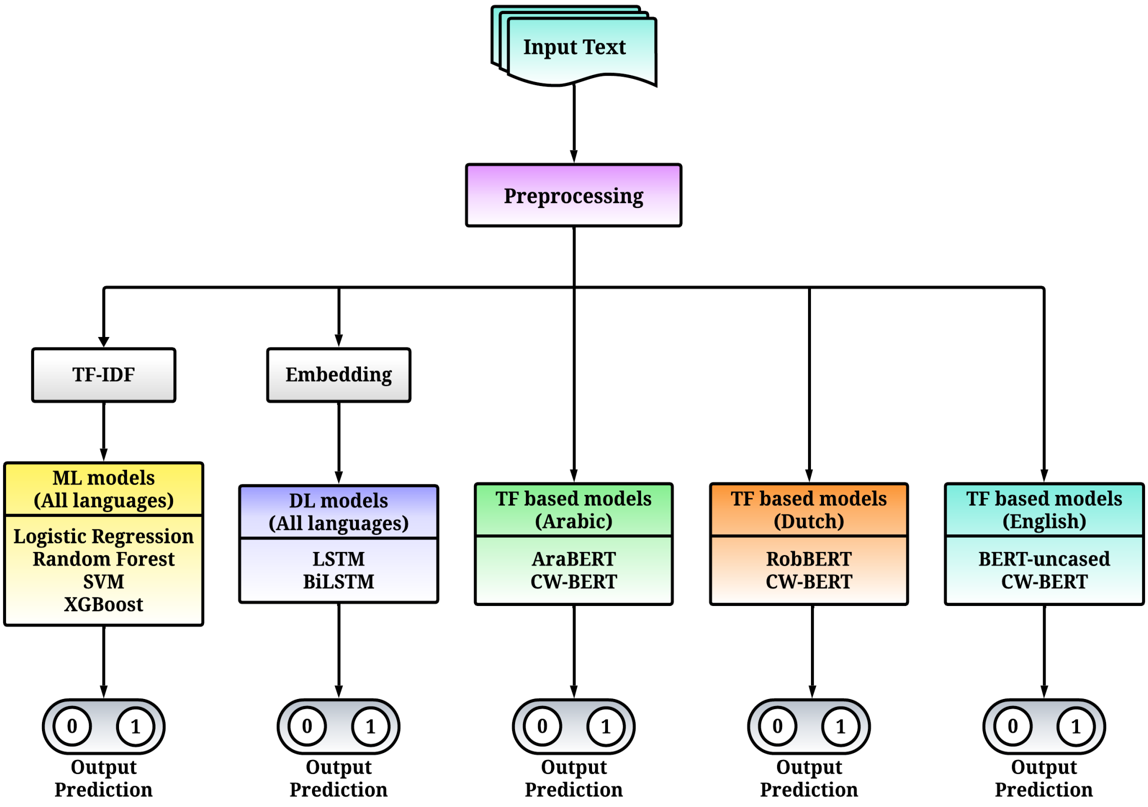

Figure 4: An outline of our approach

4.1. Preprocessing

Throughout the training and evaluation phases, we have applied several preprocessing steps. These include the removal of extraneous spaces and punctuation from the input text. However, as numerical text contributes to the overall meaning of the text we haven't removed them. For the final preprocessing step specific to transformer-based models, we have tokenized the sample text using the BERT tokenizer (bert-base-multilingual-uncased), added special tokens, and truncated/padded the sequences to a maximum length of 128 tokens.

4.2. Machine Learning Based Approaches

We have employed conventional machine learning algorithms like Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost to estimate check-worthiness. To determine the significance of each word within a text, we have employed the TF-IDF vectorizer. Using the modified training data, with a maximum of 1000 iterations, we have used a Logistic Regression model, a Random Forest model with 100 estimators and 42 as random state value, an SVM model with a linear kernel and a maximum number of iterations set to 1000. Additionally, we have developed an XGBoost model with 100 estimators and a max depth of 15, that builds an ensemble of decision trees repeatedly using gradient boosting.

4.3. Deep Learning Based Approaches

We have also utilized some deep learning-based models such as LSTM and BiLSTM to detect check-worthiness. First, we have defined an embedding layer to convert words into vectors. Then, we added a spatial dropout layer to prevent overfitting. Then we have defined an LSTM layer for the LSTM model and a bidirectional LSTM layer for the BiLSTM model to capture sequential dependencies, then a sigmoid activation function-based dense output layer for binary classification. We have used the Adam optimizer and the binary cross-entropy loss function to compile both the LSTM and the BiLSTM models.

4.4. Transformer-based Approaches

Transformer-based approaches are widely applied in numerous domains. For this reason, we have employed multiple transformer-based models for different languages in this task. We have used the AraBERT model [8], the RobBERT model [9], and the BERT-uncased model [10] for the Arabic, Dutch, and English language respectively.

Furthermore, we have fine-tuned the mBERT (bert-base-multilingual-uncased) model [10] for this task. We have named our model CW-BERT (Check-Worthiness-BERT) which is a generalized model for all three languages (Arabic, Dutch, and English).

Firstly, we have carried out a variety of preprocessing steps to prepare the data for training. Using the BERT tokenizer (bert-base-multilingual-uncased), we have tokenized each sample text to transform it into a format that can be entered into the BERT model. This involved padding/truncating sequences to a maximum length of 128 tokens and inserting special tokens. Then, we mapped the labels, which were 'Yes' for check-worthy claims and 'No' for non-check-worthy claims to binary values (1 for Yes, 0 for No). After that, we used padding for tokenized sequences to ensure uniform input length and converted the tokenized texts and labels PyTorch tensors for model training.

In our approach, we have chosen the mBERT model (bert-base-multilingual-uncased) for this task as it has been pre-trained on an extensive collection of multi-lingual data. We have specifically used the BertForSequenceClassification class, which is tailored for sequence classification tasks. The model has been fine-tuned using the training dataset using a carefully designed procedure. We have loaded the preprocessed training data into a DataLoader with a batch size of 32, enabling efficient mini-batch training. Then, in order to modify the learning rate during training, we used a linear learning rate scheduler with warm-up and the AdamW optimizer, which had a learning rate of $2e^{-5}$ and an epsilon of $1e^{-8}$. We have trained the model for several epochs, during which the forward pass computed the output of the model and loss for each batch, the backward pass computed gradients and updated model parameters. Finally, The learning rate was adjusted in accordance with the scheduler, and gradient clipping was used to stop exploding gradients.

5. Experimental Results

This section presents the experimental findings obtained during the training and evaluation stages of our proposed model and several other transformer-based, machine learning, and deep learning models for comparative study.

5.1. Environment Settings

A personal computer equipped with an Intel Core i7-9750H CPU running at 3.00 GHz and an NVIDIA GeForce GTX 2060 GPU was used to execute the simulation. Additionally, a Kaggle Notebook with a P100 GPU was used to guarantee the necessary processing capability.

5.2. Parameter Settings

Table 2 summarizes the parameter settings that we have used in different models.

Table 2
Parameter configurations for various models

Model	Learning Rate	Optimizer	Batch Size	Epsilon	Epochs
LSTM	$1e^{-3}$	Adam	64	-	15
BiLSTM	$1e^{-3}$	Adam	64	-	15
AraBERT	$2e^{-5}$	AdamW	32	$1e^{-8}$	12
RobBERT	$2e^{-5}$	AdamW	32	$1e^{-8}$	15
BERT-uncased	$2e^{-5}$	AdamW	32	$1e^{-8}$	10
CW-BERT	$2e^{-5}$	AdamW	32	$1e^{-8}$	10

5.3. Evaluation Metrics

As per the guidelines provided by the CLEF 2024 - CheckThat! Lab: Task 1 organizers, we have evaluated our models by calculating the F1 score on the test dataset. Equation 1 gives the mathematical description of the F1 score.

$$\textit{F1 Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{1}$$

where

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

True Positive, False Positive, and False Negative are represented by the symbols *TP*, *FP*, and *FN* respectively in equations 2 and 3.

5.4. Comparative Analysis

The CW-BERT model's performance on the Arabic, Dutch, and English test datasets is shown in Table 3.

Table 3
Performance of CW-BERT on the test dataset of different languages (Arabic, Dutch and English)

Language	**Macro Average**		
	Precision	**Recall**	**F1 Score**
Arabic	0.396	0.798	0.530
Dutch	0.491	0.607	0.543
English	0.822	0.682	0.745

We have conducted a comparative analysis of transformer-based, machine learning, and deep learning models, evaluating their performance using the F1 score. Table 4 shows how different models perform on the datasets of various languages by calculating the F1 score.

Table 4
Results across models on the test dataset of different languages (Arabic, Dutch, and English)

Category	**Model**	**Language**								
		Arabic			**Dutch**			**English**		
		P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
ML	Logistic Regression	0.285	0.070	0.107	0.418	0.249	0.312	0.703	0.295	0.416
	Random Forest	0.231	0.042	0.071	**0.628**	0.163	0.259	0.882	0.170	0.286
	SVM	0.397	0.409	0.403	0.427	0.403	0.415	0.439	0.455	0.447
	XGBoost	0.429	0.291	0.347	0.492	0.286	0.362	**0.889**	0.364	0.516
DL	LSTM	0.634	0.297	0.406	0.454	0.386	0.417	0.741	0.410	0.528
	BiLSTM	**0.671**	0.335	0.447	0.473	0.392	0.429	0.757	0.421	0.541
TF	AraBERT	0.384	0.733	0.504	-	-	-	-	-	-
	RobBERT	-	-	-	0.478	0.599	0.532	-	-	-
	BERT-uncased	-	-	-	-	-	-	0.809	0.637	0.713
	CW-BERT	0.396	**0.798**	**0.530**	0.491	**0.607**	**0.543**	0.822	**0.682**	**0.745**

When assessing models using the Arabic language test dataset, the SVM model has attained the best F1 score of 0.403 among machine learning models. Deep learning models, specifically LSTM and BiLSTM, have demonstrated enhanced efficacy. The BiLSTM model achieved an F1 score of 0.447, just surpassing the LSTM model's F1 score of 0.406. Transformer-based models have demonstrated superior performance in contrast to both deep learning and machine learning approaches. For instance, the AraBERT model has achieved an F1 score of 0.504, while our proposed CW-BERT model secured the highest F1 score of 0.530 among all models evaluated for the Arabic language, placing 7^{th} on the leaderboard.

In the context of the Dutch language, the SVM model once more demonstrated superior performance among machine learning models, achieving an F1 score of 0.415. Among the deep learning models, we have again seen improvement over the machine learning models as the LSTM model and the BiLSTM model have obtained the F1 score of 0.417 and 0.429 respectively. However, transformer-based models once again have demonstrated clear superiority over both deep learning and machine learning approaches. Specifically, the RobBERT model has secured the second-highest F1 score of 0.532, while the CW-BERT model attained the best F1 score of 0.543, ranking 12^{th} on the leaderboard.

Finally, for the English language, we have observed that the XGBoost model has surpassed the other models of machine learning with an F1 score of 0.516. Deep learning models have done little better than machine learning models. The BiLSTM model has scored an F1 score of 0.541 having a slight edge over the F1 score of the LSTM model of 0.528. However, the transformer-based approaches have achieved significant improvement with high margins contrasted with the deep learning and machine learning models. Although the BERT-uncased model has obtained an impressive F1 score of 0.713, our proposed CW-BERT model excels by obtaining the highest score among these models with a remarkable F1 score of 0.745, securing the 10^{th} position on the leaderboard.

5.5. Error Analysis

Table 4 shows that the CW-BERT model has outperformed all models in terms of estimating check-worthiness for all three languages. To get further insight into the model We have provided the confusion matrices of CW-BERT for the Arabic, Dutch, and English languages, presented in Figures 5, 6, and 7, in that order.

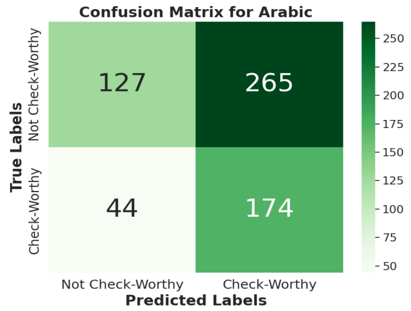

Figure 5: Confusion matrix of the CW-BERT model using the Arabic language test dataset

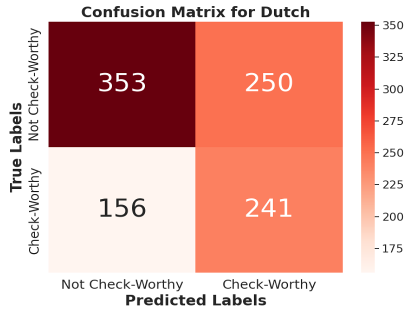

Figure 6: Confusion matrix of the CW-BERT model using the Dutch language test dataset

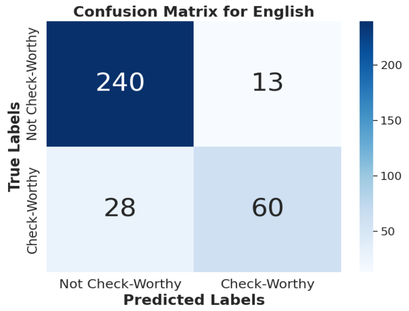

Figure 7: Confusion matrix of the CW-BERT model using the English language test dataset

From the confusion matrices, we have seen that the CW-BERT model for the English language has achieved the lowest True Positive Rate (TPR) of 79.8% even though the model is performing well in the Arabic language in terms of ensuring that most of the statements that are check-worthy are correctly identified. However, it has not performed so well in the Dutch and English languages as it has obtained very low TPR of 60.7% and 68.2% respectively, indicating that the model is missing some actual positive instances (check-worthy statements).

According to Figures 1, 2, and 3, it is evident that the datasets in all three datasets are highly imbalanced with more 'No' (Not Check-worthy) instances compared to 'Yes' (Check-worthy) instances. Because of this, the model is biased in favor of the majority class (No), which raises the number of false negatives. Besides that, we can see in Figure 3 that there are not enough training samples for the Dutch language which is also a reason for getting low TPR. Therefore, the model leads to some misclassifications.

6. Conclusion

Fact-checking is crucial for maintaining the integrity of information shared on social media. By focusing on influential tweets, fact-checkers can more effectively mitigate the spread of misinformation and contribute to a healthier information ecosystem. In this research, we have employed several machine learning, deep learning, and transformer-based approaches for detecting the fact-checking worthiness of tweets and/or transcriptions using the provided dataset in Task 1 (Check-Worthiness Estimation) organized by CheckThat! Lab under CLEF 2024. For three distinct languages, we have developed a single multilingual model. In the future, we plan to develop a more comprehensive multilingual model which will lead us to superior results. We intend to expand on this work in the future so that our model can efficiently handle data imbalance and apply large language models to the dataset.

References

[1] A. Barrón-Cedeño, F. Alam, T. Chakraborty, T. Elsayed, P. Nakov, P. Przybyła, J. M. Struß, F. Haouari, M. Hasanain, F. Ruggeri, X. Song, R. Suwaileh, The CLEF-2024 CheckThat! Lab: Check-worthiness, subjectivity, persuasion, roles, authorities, and adversarial robustness, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2024, pp. 449–458.

[2] A. Barrón-Cedeño, F. Alam, J. M. Struß, P. Nakov, T. Chakraborty, T. Elsayed, P. Przybyła, T. Caselli, G. Da San Martino, F. Haouari, C. Li, J. Piskorski, F. Ruggeri, X. Song, R. Suwaileh, Overview of the CLEF-2024 CheckThat! Lab: Check-worthiness, subjectivity, persuasion, roles, authorities and adversarial robustness, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. Di Nunzio, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[3] M. Hasanain, R. Suwaileh, S. Weering, C. Li, T. Caselli, W. Zaghouani, A. Barrón-Cedeño, P. Nakov, F. Alam, Overview of the CLEF-2024 CheckThat! lab task 1 on check-worthiness estimation of multigenre content, in: G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CLEF 2024, Grenoble, France, 2024.

[4] S. Shaar, A. Nikolov, N. Babulkov, F. Alam, A. Barrón-Cedeno, T. Elsayed, M. Hasanain, R. Suwaileh, F. Haouari, G. Da San Martino, et al., Overview of checkthat! 2020 english: Automatic identification and verification of claims in social media., CLEF (working notes) (2020).

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2023.

[6] S. Si, A. Datta, S. Naskar, A new approach to claim check-worthiness prediction and claim verification, in: P. Bhattacharyya, D. M. Sharma, R. Sangal (Eds.), ICON, NLPAI, 2020.

[7] H. Sinha, Sakshi, Y. Sharma, Text-convolutional neural networks for fake news detection in tweets, in: FICTA 2020, Volume 1, Springer, 2021.

[8] W. Antoun, F. Baly, H. M. Hajj, Arabert: Transformer-based model for arabic language understanding, CoRR (2020).

[9] P. Delobelle, T. Winters, B. Berendt, Robbert: a dutch roberta-based language model, CoRR (2020).

[10] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR (2018).

[11] P. Gencheva, P. Nakov, L. Màrquez, A. Barrón-Cedeño, I. Koychev, A context-aware approach for detecting worth-checking claims in political debates, in: R. Mitkov, G. Angelova (Eds.), RANLP, 2017.

[12] S. Vasileva, P. Atanasova, L. Màrquez, A. Barrón-Cedeño, P. Nakov, It takes nine to smell a rat: Neural multi-task learning for check-worthiness prediction, CoRR (2019).

[13] D. Samuel, A. Kutuzov, S. Touileb, E. Velldal, L. Øvrelid, E. Rønningstad, E. Sigdel, A. Palatkina, NorBench – a benchmark for Norwegian language models, in: T. Alumäe, M. Fishel (Eds.), NoDaLiDa, University of Tartu Library, 2023.

[14] P. E. Kummervold, J. De la Rosa, F. Wetjen, S. A. Brygfjeld, Operationalizing a national digital library: The case for a norwegian transformer model, in: NoDaLiDa, 2021.

[15] M. Sundriyal, M. S. Akhtar, T. Chakraborty, Leveraging social discourse to measure check-worthiness of claims for fact-checking, 2023.