SemanticCuetSync at CheckThat! 2024: Pre-trained Transformer-based Approach to Detect Check-Worthy Tweets

Notebook for the CheckThat! Lab at CLEF 2024

Symom Hossain Shohan[1,†], Md. Sajjad Hossain[1,†], Ashraful Islam Paran[1,†], Jawad Hossain[1], Shawly Ahsan[1] and Mohammed Moshiul Hoque[1,*]

[1]Chittagong University of Engineering and Technology, Chattogram - 4349, Bangladesh

Abstract

This paper presents an intelligent technique for classifying English, Arabic, and Dutch texts as checkworthy, harnessing the power of the BERT-based model. The study explores ten baseline models, including LR, MNB, SVM, CNN+LSTM, CNN+BiLSTM, BERT-Base-Uncased, RoBERTa, AraBERTv2, Dutch-RoBERTa, and Dutch-BERT, to address the shared task. The study also investigates an LLM using few-shots, such as SetFit, to identify checkworthy tweets or texts. Evaluation results unequivocally demonstrate the superiority of transformer-based models, with RoBERTa achieving the highest F1 scores of 75.82% for English tweets, Dehate-BERT scoring 52.55% for Arabic texts, and Dutch-BERT obtaining a maximum score of 58.42% for Dutch texts. Our team ranked 6th overall for English, 5th for Arabic, and 16th for Dutch in the shared task challenge.

Keywords

Natural Language Processing, Check-Worthiness, Fact-checking, Tweet-Verification, Transformers

1. Introduction

Checkworthy content refers to information that must be confirmed for accuracy, as it may have the potential to shape the opinions and decisions of others. The rise of social networks has led to an exponential growth in textual data on the internet, sometimes resulting in the spread of false claims that can be detrimental to society if left unaddressed. These claims can include political, religious, and health-related misinformation, which can cause discord in society. Fact-checking is a time-consuming task that requires extensive research, identification, verification, and expert analysis. Automating this entire process is a significant challenge, and the first step towards this goal is to determine whether the information is worth checking in the first place.

With the proliferation of communication and social media platforms, such as Facebook, Twitter, and Reddit, the dissemination of false information has become increasingly prevalent. A recent study has suggested that people struggle to differentiate facts from false news [1]. Intelligent technologies can be used to support human fact-checkers to identify claims worth fact-checking [2]. Many studies have been devoted to developing a fully automated system for fact-checking [3], [4], [5], [6], [7]. As social media data continues to expand daily, it is impractical to monitor everything efficiently by human experts. Therefore, developing an automatic system has emerged as the ultimate solution to this problem. This work proposes a solution to classify English, Arabic, and Dutch texts or tweets as checkworthy, harnessing the power of BERT-based approaches. The critical contributions of this study are:

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

*Corresponding author.

†These authors contributed equally.

u1904048@student.cuet.ac.bd (S. H. Shohan); u1904031@student.cuet.ac.bd (Md. S. Hossain); u1904029@student.cuet.ac.bd (A. I. Paran); u1704039@student.cuet.ac.bd (J. Hossain); u1704057@student.cuet.ac.bd (S. Ahsan); moshiul_240@cuet.ac.bd (M. M. Hoque)

0009-0004-0834-2037 (S. H. Shohan); 0009-0008-8670-8857 (Md. S. Hossain); 0009-0001-4795-3816 (A. I. Paran); 0009-0006-6051-8989 (J. Hossain); 0009-0003-9940-9681 (S. Ahsan); 0000-0001-8806-708X (M. M. Hoque)

- Introducing a fine-tuned transformer-based model to classify checkworthy texts for three languages (English, Arabic, and Dutch).
- Exploring various machine learning (LR, SVM, and MNB), deep learning (CNN, CNN + LSTM, and CNN + BiLSTM), and transformer-based models for finding a suitable method for detecting checkworthy texts in multiple languages.

2. Related Work

Inaccurate news is quickly spreading throughout social media. Checking the authenticity of any post that surfaces on social media becomes crucial. Intelligent fact-checking systems have emerged as a significant area of research to tackle this problem. Several domains allow for the detection of trustworthiness, such as digital scam [8], the healthcare sector [9], politics [10], and many more fields. An overview of Task 1 in the fourth edition of the CheckThat! The lab was provided by Shaar et al. [11]. Their job was anticipating which tweets involving politics and COVID-19 needed to be verified. Williams et al. [12] presented a transformer-based solution with data augmentation for this problem, and it received an mAP (mean average precision) of 0.66 in the Arabic language. Checkworthiness in multimodal [13] is another popular research area these days; in addition to unimodal, Sadouk et al. [14] proposed a multimodal transformer-based model (BERT+ResNet50) to identify checkworthiness in English which recorded F1 score of 0.71 and transformer based model (MarBERT) with downsampling recorded F1 score of 0.61 in Arabic for image dataset. Meanwhile, Ivanov et al. [15] proposed audio datasets from past political debates and ensemble techniques for detecting checkworthiness. Their audio model (wav2vec2.0) received a mAP of 0.34 when extra noise was eliminated. Ensembles using BERT and an audio model outperformed BERT alone, with a mAP of 0.38.

This work addresses a significant gap in the existing literature by comprehensively comparing machine learning (ML), deep learning (DL), and transformer-based solutions. In addition, it investigates the use of few-shot models like SetFit for determining check worthiness in Dutch, Arabic, and English. This study improves the understanding of the various models' performance in these distinct languages.

3. Dataset and Task Description

The dataset consists of tweets or texts in English, Arabic, and Dutch languages, along with their corresponding labels ('Yes' for texts worth checking, 'No' otherwise). Table 1 shows the distribution of train, dev, dev-test, and test sets. We trained all models using the training set and evaluated the model's performance based on the test set. CLEF 2024 - CheckThat! Lab [16, 17, 18] consists of six tasks

Table 1
Dataset statistics for Task-1, where TW stands for Total words and UW stands for Unique words.

Language	Train	Dev	Dev-Test	Test	Total	TW	UW
English	22501	1032	318	341	24192	432903	11605
Arabic	7333	1093	500	610	9536	251619	50001
Dutch	995	252	666	1000	2913	36062	8240
Total	**30829**	**2377**	**1484**	**1951**	**36641**	**720584**	**69846**

[19, 20, 21, 22, 23]. We participated in task-1 of this shared task. Task-1 [19] focuses on assessing whether a claim in a tweet or transcription requires further investigation for fact-checking. The traditional approach for such decisions involves human experts, either professional fact-checkers or annotators, who evaluate the claim based on various criteria. Table 2 illustrates an example of training data for the different languages.

Table 2
Task-1 sample with text and label.

Ex	Text	Label
1	I'd like to mention one thing.	No
2	I'm proud of the fact that violent crime is down in the State of Texas.	Yes
3	فتح الاجواء من هذه الليلة ألف مبروك لكل الخليجيين (Opening the skies from tonight. Congratulations to all Gulf people)	No
4	يا قدسنا إسمعيني حبّك جاري في شرايييني ❤️ #القدس_عاصمة_فلسطين_الأبدية (O Jerusalem, listen to me, your love is running through my veins ❤️ #Jerusalem_The_Eternal_Capital_of_Palestine)	No
5	Aantal restaurants in #groningen nu op slot. #blijfthuis (Number of restaurants in #Groningen now closed. #stay at home)	Yes
6	Ik heb deze hele dag nog geen 1-aprilgrap gezien of gehoord. #houdafstand (He will be here again after April 1, 2017. #houdafstand)	No

4. System Overview

This task exploited various ML, DL, and transformer-based approaches across all three languages. ML-based techniques used include linear regression (LR), support vector machine (SVM), and multinomial naive Bayes (MNB). DL-based techniques involve CNN, CNN+long short-term memory (LSTM), and CNN+bidirectional LSTM (BiLSTM). Lastly, various BERT-based transformers are fine-tuned for each language for the given task. Figure 1 illustrates the schematic process of checkworthy text detection.

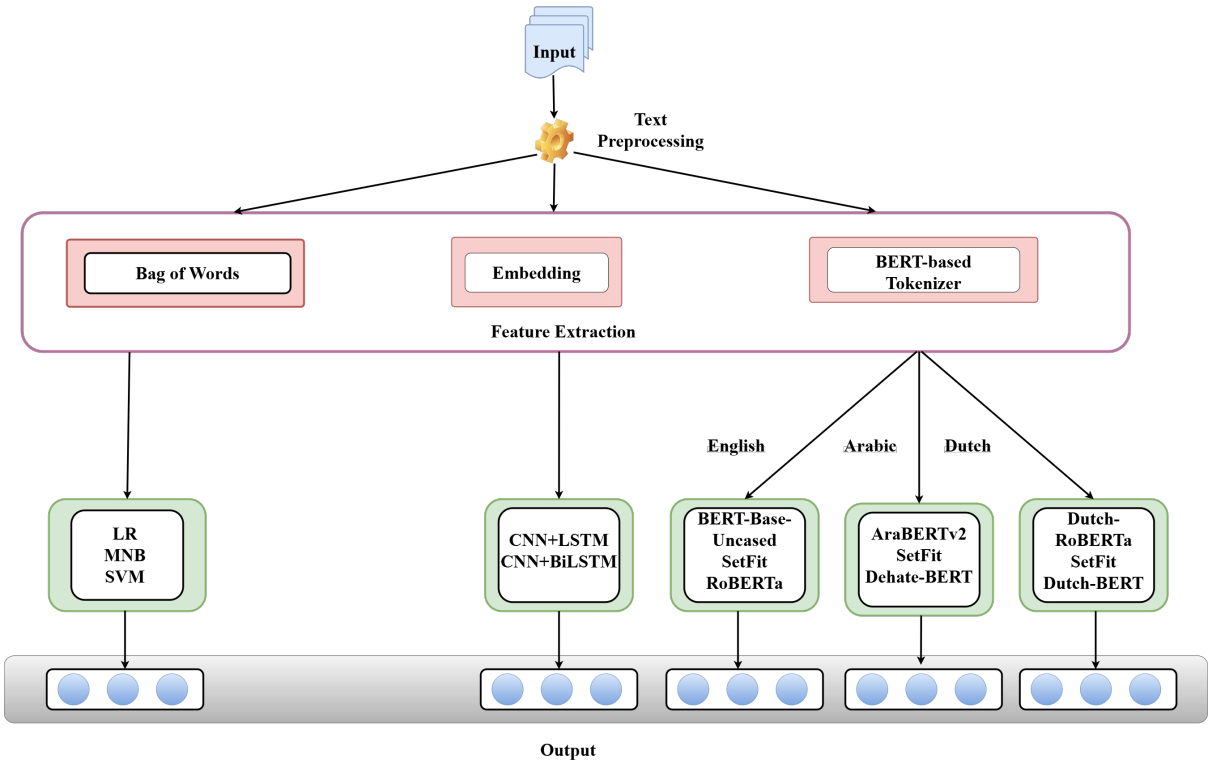

Figure 1: Schematic process for check-worthy text detection.

Textual Feature Extraction: Textual feature extraction is one of the essential steps in natural

language processing, which involves transforming raw textual data into numerical representations. This numerical representation aids the models in understanding and processing textual data. A Count Vectorizer is used in the ML models examined in this work. It is a widely used technique for textual feature extraction that transforms text data into a matrix of token counts. In DL models, tokenization and padding combine to convert raw texts into structured numerical data. These numerical representations are then passed through an embedding layer, which captures more advanced features such as semantic relationships. This study uses the embedding layer instead of Word2Vec [24] or GloVe [25] to allow the model to learn task-specific embedding during training. Finally, BERT-based tokenizers are employed for transformer-based models to exploit the BERT architecture.

ML Models: Various ML models are examined in this work, such as LR, SVM, MNB, KNN, and RF. All the hyperparameter settings for these models are illustrated in Table 3.

Table 3
Parameters of the employed ML models.

Classifier	Parameters	Value
LR	solver	lbfgs
	max_iter	20000
MNB	alpha	1.0
	fit-prior	False
SVM	kernel	linear
	gamma	auto

CNN: This work employed a CNN model comprising an embedding layer with an output dimension of 200. The model features two Conv1D layers with 64 and 128 filters, respectively. Both layers used a kernel size of 2 and ReLU activation. For downsampling, the model incorporates a GlobalMaxPooling1D layer. Subsequently, a dense layer with 128 units and ReLU activation is followed by a dropout layer with a rate of 0.5 to prevent overfitting. The output layer has a single unit with sigmoid activation. The model utilizes the '*binary_crossentropy*' loss function and '*Nadam*' optimizer and has trained with a batch size of 32 for three epochs.

CNN+LSTM: The CNN+LSTM model used in this work has almost the same architecture as the CNN model, incorporating a single LSTM layer comprising 64 units and a dropout rate of 0.2 for sequence modeling. Furthermore, the dense layer included in this design features 64 units and utilizes the ReLU activation function. The remaining hyperparameter configurations are consistent with those employed in the CNN model.

CNN+BiLSTM: This model has an architecture similar to the CNN+LSTM model but replaces LSTM with a Bidirectional LSTM.

Transformer models for English: This study fine-tuned three transformer-based models for a specified task in the English dataset. The models employed were BERT-Base-Uncased [26], SetFit [27], and RoBERTa [28]. The necessary text preprocessing steps were followed before feeding the data into the transformers. These text preprocessing steps include lowercasing, emoji removal, stop word removal, stemming, contraction expansion, simple Unicode spelling correction, and HTML tag removal. For stop word removal, the NLTK stopwords list is used. The main agenda of the text preprocessing steps was to reduce the noise in the dataset and focus on meaningful words. The BERT-Base-Uncased used in this task is a pre-trained transformer model with exceptional performance across various natural language processing (NLP) tasks. This model demonstrated satisfactory performance on the specified task. On the other hand, SetFit leverages pre-trained transformers with limited labeled data. We explored the potential of this few-shot learning framework for the given task. SetFit does not require manual prompts for classification, in contrast to LLMs. Finally, RoBERTa, another optimized version of BERT, is used here for the specified task and outperforms other models.

Transformer models for Arabic: This study also exploited three transformer-based models and fine-tuned them in the Arabic dataset. Models used for the Arabic dataset were AraBERTV2 [29], SetFit (Few-shot) [27], and Dehate-BERT [30]. Similar to the English dataset, some text-preprocessing

steps were also performed. Again, the preprocessing steps are lowercasing, emoji removal, stop word removal, stemming, contraction expansion, simple spelling correction using Unicode, HTML tag removal, punctuation removal, URL removal, whitespace removal, and number removal. Stemming was performed here using ArabicLightStemmer. Finally, normalization was used to convert similar characters to a standard form.

AraBERTV2 is the improved version of AraBERT, which leverages the BERT architecture. This model was trained on a sizeable Arabic dataset and has demonstrated effectiveness in various downstream NLP tasks, including sentiment analysis, NER, and Arabic question answering. Dehate-BERT is a pre-trained transformer model primarily designed for hate speech detection, and it outperformed all other models in Arabic in this specific task.

Transformer models for Dutch: This study investigated Dutch RoBERTa [31], SetFit, and Dutch-BERT in the Dutch dataset. Rather than undertaking extensive text preprocessing, this work limited its processing to removing non-Dutch characters from the texts. Table 4 illustrates the hyperparameters of transformer-based models.

Table 4
Hyperparameters for transformer-based models, where LR, WD, WS, and EP stand for learning rate, weight decay, warmup steps, and number of epochs, respectively.

Models	LR	WD	WS	EP
AraBERTv2	$3e^{-5}$	0.01	500	3
Dehate-BERT	$4e^{-5}$	0.01	500	2
BERT-Base-Uncased	$3e^{-5}$	0.01	500	2
RoBERTa	$3e^{-5}$	0.01	500	2
Dutch-RoBERTa	$3e^{-5}$	0.01	500	3
Dutch-BERT	$5e^{-5}$	0.01	500	5

5. Results and Analysis

Table 5 illustrates an in-depth analysis of the performance of ML, DL, and transformer-based models in English, Arabic, and Dutch on the test set.

In evaluating English language data, logistic regression (LR) achieved a precision of 75.00%, recall of 37.50%, and F1-score of 50.00%. However, support vector machine (SVM) emerged as the top-performing ML model, with a precision of 68.97%, recall of 45.45%, and the highest F1-score of 54.79%. Multinomial Naive Bayes (MNB) showed a precision of 64.81%, recall of 39.77%, and F1-score of 49.29%. Among DL models, CNN+BiLSTM demonstrated the best performance with a precision of 66.10%, recall of 44.32%, and F1-score of 53.06%. Among Transformers, RoBERTa showcased remarkable performance, achieving a precision of 89.23%, recall of 65.91%, and the highest F1-score of 75.82%. The SetFit model achieved a precision of 52.34%, recall of 63.64%, and an F1-score of 57.44%. This model shows comparable performance to well-known models such as SVM and LR.

Arabic language evaluations revealed LR achieves a precision of 38.52%, recall of 21.55%, and F1-score of 27.65%. However, SVM followed closely with a precision of 40.57%, recall of 32.57%, and F1-score of 36.13%. MNB showed a precision of 36.00%, recall of 24.77%, and the highest F1-score of 29.35%. Among DL models, CNN+BiLSTM showed the best performance with a precision of 34.27%, recall of 27.98%, and F1-score of 30.81%. Among Transformers, Dehate-BERT emerged as the top-performing model with a precision of 40.24%, recall of 75.69%, and the highest F1-score of 52.55%. The SetFit model attained a precision of 37.75%, recall of 69.26%, and an F1-score of 48.86%. Although it does not match the performance of the leading transformer-based model, Dehate-BERT, it still demonstrates potential in tackling Arabic language classification issues.

For Dutch language evaluations, LR achieves a precision of 50.98%, recall of 32.75%, and F1-score of 39.88%. SVM showed a precision of 43.86%, recall of 37.78%, and F1-score of 40.60%. MNB attained a precision of 46.91%, recall of 9.57%, and F1-score of 15.89%. Among DL models, CNN+BiLSTM

Table 5
Performance of the employed models on the test set.

Language	Method	Classifier	Pr(%)	Re(%)	Ac(%)	F1(%)
English	ML Models	LR	75.00	37.50	80.65	50.00
		SVM	**68.97**	**45.45**	**80.65**	**54.79**
		MNB	64.81	39.77	78.88	49.29
	DL Models	CNN+LSTM	74.47	39.77	80.94	51.85
		CNN+BiLSTM	**66.10**	**44.32**	**79.77**	**53.06**
	Transformers	BERT-Base-Uncased	84.85	63.64	87.68	72.73
		SetFit	52.34	63.64	75.66	57.44
		RoBERTa	**89.23**	**65.91**	**89.15**	**75.82**
Arabic	ML Models	**LR**	**38.52**	**21.55**	**59.67**	**27.65**
		SVM	40.57	32.57	58.85	36.13
		MNB	36.00	24.77	57.38	29.35
	DL Models	CNN+LSTM	33.16	28.44	53.93	30.62
		CNN+BiLSTM	**34.27**	**27.98**	**55.08**	**30.81**
	Transformers	AraBERTV2	40.40	55.05	54.92	46.60
		SetFit	37.75	69.26	48.20	48.86
		Dehate-BERT	**40.24**	**75.69**	**51.15**	**52.55**
Dutch	ML Models	**LR**	**50.98**	**32.75**	60.80	39.88
		SVM	43.86	37.78	56.10	40.60
		MNB	46.91	9.57	59.80	15.89
	DL Models	CNN	33.33	0.25	60.20	0.50
		CNN+BiLSTM	**52.06**	**44.58**	**61.70**	**48.03**
	Transformers	Dutch-RoBERTa	6.21	60.33	0.32	11.26
		SetFit	45.31	58.44	55.50	51.05
		Dutch-BERT	**48.40**	**73.80**	**52.40**	**58.42**

demonstrated the best performance with a precision of 52.06%, recall of 44.58%, and F1-score of 48.03%. In the transformer category, Dutch-BERT outperformed others with a precision of 48.40%, recall of 73.80%, and the highest F1-score of 58.42%.

The SetFit model exhibited a precision of 45.31%, recall of 58.44%, and an F1-score of 51.05%. It shows competitive results compared to transformer-based models such as Dutch-BERT, indicating its possible usefulness in Dutch language classification tasks.

In general, transformer-based models surpass both ML and DL models across various languages, demonstrating the effectiveness of pre-trained language models in numerous natural language processing applications. Additionally, within each category, specific models show superior performance, emphasizing the necessity of choosing the appropriate model based on the specific task and language.

5.1. Error Analysis

A comprehensive quantitative and qualitative error analysis is conducted to provide detailed insights into the proposed model's performance.

Quantitative Analysis

Figure 2 illustrates the confusion matrix of the best-performing models across English, Arabic, and Dutch.

From a total of 341 test cases in English, RoBERTa demonstrates strong performance in identifying the positive class, with 247 True Positives and only 6 False Positives. This indicates a high precision, meaning the model is highly accurate when it predicts "Yes." Additionally, with 57 True Negatives, it correctly identifies many negative instances. However, there are 31 False Negatives, which indicates that some positive instances are being missed. RoBERTa shows a balanced approach with notable proficiency in minimizing incorrect optimistic predictions, resulting in a high F1 score of 75.82% over

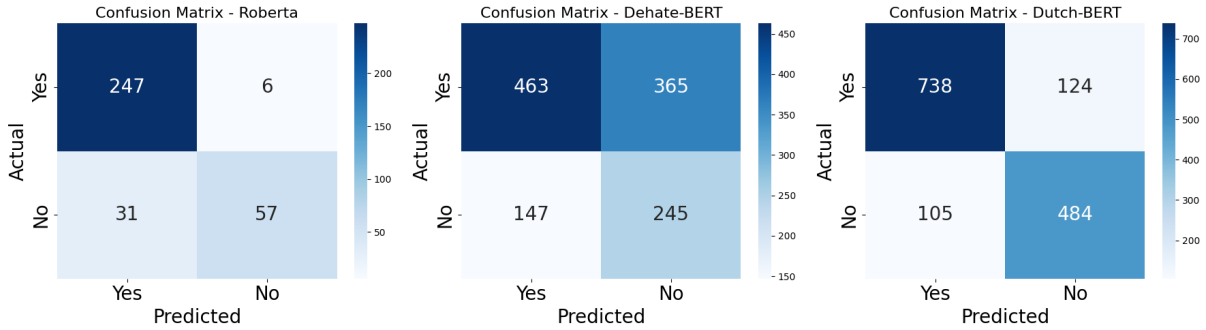

Figure 2: Confusion matrix of RoBERTa, Dehate-BERT and Dutch-BERT model.

positive samples.

From a total of 610 test cases in Arabic, Dehate-BERT shows a different pattern in its confusion matrix. With 215 True Positives and 120 True Negatives, the model accurately identifies many instances from both classes. However, the model has many False Positives (177) and False Negatives (98). This suggests that while Dehate-BERT can identify positive instances, it incorrectly classifies many negative instances as positive, leading to a lower precision. Additionally, the relatively high count of False Negatives indicates room for improvement in recall, highlighting the need for better distinction between the two classes.

Dutch-BERT performs moderately from 1000 test cases in Dutch, with 446 True Positives and 214 True Negatives. The model, however, has 157 False Positives and 183 False Negatives. This indicates that while Dutch-BERT can identify positive instances reasonably well, it struggles with precision and recall. The high number of False Positives suggests a tendency to overpredict the positive class, and the significant count of False Negatives shows it also misses many positive instances. Consequently, Dutch-BERT's overall performance is balanced but shows substantial room for improvement in minimizing misclassifications to enhance its F1 score.

Qualitative Analysis

Table 6 presents some actual labels (AL) and predicted labels (PL) of the developed models.

Table 6
Few predictions with actual and predicted label.

Ex	Text	AL	PL
1	They said they were just going to get inspectors in.	Yes	No
2	And from that point on, I've voted to -- I moved to bring those troops home.	Yes	Yes
3	بدأت مشاهد المذبحة الكبرى بالوصول ليلة البارحة قطع الاحتلال الإسرائيلي الإنترنت بالكامل وبدأ بقصف هستيري حتى أن المتحدث باسم جيشه صرح بأنه "يهاجم غزة بقوة عظيمة "وسائل الإعلام التقليدية غير قادرة على نقل الصورة بالوضوح الذي تنقله هواتف الناشطين الآن بدأت مشاهد مذبحة البارحة تصل⋯ (Scenes of the great massacre began to arrive last night. The Israeli occupation cut off the entire Internet and began hysterical bombing. Even its army spokesman declared that it was "attacking Gaza with great force."…)	Yes	Yes
4	De SGP in de Eerste Kamer wil de nieuwe #stikstof wetgeving controversieel verklaren nu het kabinet is gevallen. #landbouw (The SGP in the Senate wants to declare the new \#nitrogen legislation controversial now that the cabinet has fallen. \#Agriculture)	Yes	No
5	Het was verwacht, maar nu dus ook een feit: 2020 was niet alleen in België maar ook globaal en in Europa het warmste jaar sinds de metingen... #klimaatverstoring (It was expected, but now it is also a fact: 2020 was not only in Belgium but also globally and in Europe the warmest year since measurements... \#climatedisruption)	Yes	Yes

It is clear that the models accurately predicted the labels for examples 2, 3, and 5 but made errors with examples 1 and 4. For the first example, the sentence's intent is ambiguous, leading to an incorrect label prediction by the model. In the case of example 4, although the sentence is checkworthy, the model mislabeled it due to inadequate training data in the Dutch language, which hindered proper learning.

6. Conclusion

This work investigated the various ML, DL, and transformer-based models for identifying checkworthy tweets or texts in English, Arabic, and Dutch. The results indicate that transformer-based models shine in this task and exhibit exceptional capability in detecting checkworthy text. Specifically, RoBERTa excels in English, Dehate-BERT for Arabic, and Dutch-BERT for Dutch, achieving the highest F1 scores of 75.82%, 52.55%, and 58.42%, respectively. The study recommends that further advancements be made by increasing the training data and incorporating advanced LLMs and GPT models.

References

[1] F. Olan, U. Jayawickrama, E. O. Arakpogun, J. Suklan, S. Liu, Fake news on social media: the impact on society, Information Systems Frontiers 26 (2024) 443–458.

[2] P. Nakov, D. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barrón-Cedeño, P. Papotti, S. Shaar, G. D. S. Martino, Automated fact-checking for assisting human fact-checkers, arXiv preprint arXiv:2103.07769 (2021).

[3] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, J. Han, A survey on truth discovery, ACM Sigkdd Explorations Newsletter 17 (2016) 1–16.

[4] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective, ACM SIGKDD explorations newsletter 19 (2017) 22–36.

[5] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, et al., The science of fake news, Science 359 (2018) 1094–1096.

[6] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, science 359 (2018) 1146–1151.

[7] F. Xu, V. S. Sheng, M. Wang, A unified perspective for disinformation detection and truth discovery in social sensing: a survey, ACM Computing Surveys (CSUR) 55 (2021) 1–33.

[8] X. Chen, R. Chandramouli, K. P. Subbalakshmi, Scam detection in twitter, in: Data Mining for Service, Springer, 2014, pp. 133–150.

[9] S. D. Gollapalli, M. Du, S.-K. Ng, Identifying checkworthy cure claims on twitter, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 4015–4019.

[10] A. Patwari, D. Goldwasser, S. Bagchi, Tathya: A multi-classifier system for detecting check-worthy statements in political debates, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 2259–2262.

[11] S. Shaar, M. Hasanain, B. Hamdan, Z. S. Ali, F. Haouari, A. Nikolov, M. Kutlu, Y. S. Kartal, F. Alam, G. Da San Martino, et al., Overview of the clef-2021 checkthat! lab task 1 on check-worthiness estimation in tweets and political debates., in: CLEF (working notes), 2021, pp. 369–392.

[12] E. Williams, P. Rodrigues, S. Tran, Accenture at checkthat! 2021: interesting claim identification and ranking with contextually sensitive lexical training data augmentation, arXiv preprint arXiv:2107.05684 (2021).

[13] F. Alam, A. Barrón-Cedeño, G. S. Cheema, S. Hakimov, M. Hasanain, C. Li, R. Míguez, H. Mubarak, G. K. Shahi, W. Zaghouani, et al., Overview of the clef-2023 checkthat! lab task 1 on check-worthiness in multimodal and multigenre content, Working Notes of CLEF (2023).

[14] H. T. Sadouk, F. Sebbak, H. E. Zekiri, Es-vrai at checkthat! 2023: Analyzing checkworthiness in multimodal and multigenre (2023).

[15] P. Ivanov, I. Koychev, M. Hardalov, P. Nakov, Detecting check-worthy claims in political debates,

speeches, and interviews using audio data, in: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2024, pp. 12011–12015.
[16] 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.
[17] 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.
[18] 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.
[19] 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: [18], 2024.
[20] J. M. Struß, F. Ruggeri, A. Barrón-Cedeño, F. Alam, D. Dimitrov, A. Galassi, G. Pachov, I. Koychev, P. Nakov, M. Siegel, M. Wiegand, M. Hasanain, R. Suwaileh, W. Zaghouani, Overview of the CLEF-2024 CheckThat! lab task 2 on subjectivity in news articles, in: [18], 2024.
[21] J. Piskorski, N. Stefanovitch, F. Alam, R. Campos, D. Dimitrov, A. Jorge, S. Pollak, N. Ribin, Z. Fijavž, M. Hasanain, N. Guimarães, A. F. Pacheco, E. Sartori, P. Silvano, A. V. Zwitter, I. Koychev, N. Yu, P. Nakov, G. Da San Martino, Overview of the CLEF-2024 CheckThat! lab task 3 on persuasion techniques, in: [18], 2024.
[22] F. Haouari, T. Elsayed, R. Suwaileh, Overview of the CLEF-2024 CheckThat! Lab Task 5 on Rumor Verification using Evidence from Authorities, in: [18], 2024.
[23] P. Przybyła, B. Wu, A. Shvets, Y. Mu, K. C. Sheang, X. Song, H. Saggion, Overview of the CLEF-2024 CheckThat! lab task 6 on robustness of credibility assessment with adversarial examples (incrediblae), in: [18], 2024.
[24] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems 26 (2013).
[25] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.
[26] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.
[27] L. Tunstall, N. Reimers, U. E. S. Jo, L. Bates, D. Korat, M. Wasserblat, O. Pereg, Efficient few-shot learning without prompts, arXiv preprint arXiv:2209.11055 (2022).
[28] 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, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.
[29] W. Antoun, F. Baly, H. Hajj, Arabert: Transformer-based model for arabic language understanding, arXiv preprint arXiv:2003.00104 (2020).
[30] S. S. Aluru, B. Mathew, P. Saha, A. Mukherjee, Deep learning models for multilingual hate speech detection, arXiv preprint arXiv:2004.06465 (2020).
[31] P. Delobelle, T. Winters, B. Berendt, RobBERT: a Dutch RoBERTa-based Language Model, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 3255–3265. URL: https://www.aclweb.org/anthology/2020.findings-emnlp.292. doi:10.18653/v1/2020.findings-emnlp.292.