CSECU-DSG at CheckThat! 2023: Transformer-based Fusion Approach for Multimodal and Multigenre Check-Worthiness

Abdul Aziz[1,*], Md. Akram Hossain[1] and Abu Nowshed Chy[1]

[1]*Department of Computer Science and Engineering, University of Chittagong, Chattogram-4331, Bangladesh*

Abstract

Check-worthiness is identifying verifiable factual claims present or not in content. It might be beneficial to automatically verify the political discourses, social media posts, and newspaper content. However, the multifaceted nature and hidden meaning of the content make it difficult to automatically identify the factual claims. To address these challenges, CheckThat! 2023 introduced a task to build automatic Check-worthiness classifiers in tweets with multimodal and multigenre settings. This paper presented our participation in CheckThat! 2023 Task 1. We perform fine-tuning on language-specific and vision pre-trained transformer models to extract the visual-contextualized or contextualized features representation for the multimodal and multigenre check-worthiness task. We add a BiLSTM layer on top of the contextual features and concatenate it with the other visual or contextualized features to get an enrich unified representation. Later, we employ a multi-sample dropout strategy to predict a more accurate class label. Experimental results show that our proposed method achieved competitive performance among the participants and obtained 1st place in the multimodal Arabic check-worthiness task.

Keywords

multimodal fact-checking, automatic fact-checking, multigenre check-worthiness, multimodal check-worthiness

1. Introduction

Nowadays, people frequently deliver their ideas, beliefs, visions, and breaking news using manifold social media platforms including Instagram, Reddit, Twitter, and Facebook based on their real-time behavior and useful features. Therefore, such platforms have increasingly evolved the skyrocketed means of discovering various information including public views, health situations, political mindsets, and customer choices. Disinformation is also shared on social media using these platforms. Thus, automated fact-checking is one of the most prominent tasks in recent years. Various works [1, 2, 3, 4, 5] have been introduced based on fact-checking in various formats including textual and multimodal. Most of the prior works utilise the traditional transformer-based approach [6, 7] for checking the worthiness of factual claims. Alam et al. [8] introduce a shared task at CheckThat! 2023 [9] to check-worthiness of tweets in both multimodal and multigenre (multiple languages) settings. To tackle this task we used a

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

[*] Corresponding author.

aziz.abdul.cu@gmail.com (A. Aziz); akram.hossain.cse.cu@gmail.com (. Md. A. Hossain); nowshed@cu.ac.bd (A. N. Chy)

transformer-based fusion approach with the BiLSTM module on top of the language model and the multi-sample dropout strategy. Our system ranked first in the multimodal Arabic task and achieved competitive performance in other tasks.

We organize the rest of the paper as follows: **Section** 2 describes our proposed system in the CheckThat! 2023 to automatically identify the worthiness of tweets, in **Section** 3, we present our proposed system design with parameter settings and conduct the results and performance analysis. Finally, we conclude with some future directions in **Section** 4.

2. Proposed Framework

Transformers models learn the necessary information about the relationship between words effectively. We employed the pre-trained transformers model with the BiLSTM module and a training strategy to identify the factual claim worthiness of tweets in multimodal and multigenre settings. The overview of our proposed transformer-based framework is depicted in Figure 1.

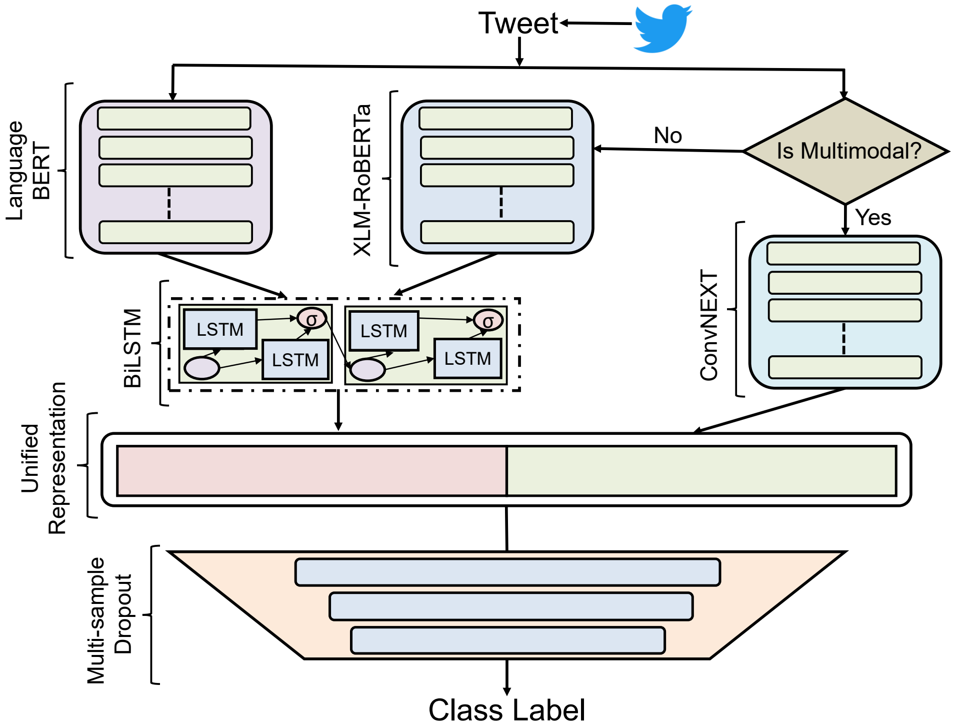

Figure 1: Overview diagram of our proposed method for multimodal and multigenre check-worthiness of tweets

Given a tweet containing a textual claim (and corresponding image for the multimodal tasks), we fed them into two transformer models. Our model identifies the data whether the task is multimodal or not, if multimodal (text and image inputs) our model uses language-specific BERT and ConvNEXT model to encode contextual-visual representation otherwise (only text input) our model uses language-specific BERT and the XLM-RoBERTa model to encode diverse contextual representation for multigenre tasks. We sum both models' contextual features for the multigenre task and fed them to a BiLSTM module. In multimodal tasks, we employ the BiLSTM module on top of the contextual representation only to handle the long-term contextual dependency present on tweets. Then, we concatenate visual-contextual features in a unified architecture for multimodal representation. These unified features vector fed to the multi-sample dropout strategy to speed up training and improve the robustness of our proposed model. We leverage the different dropout sample outputs using the arithmetic mean technique to get the final label of our proposed method.

2.1. Transformer Models

2.1.1. Twitter XLM-RoBERTa

Facebook AI launched the XLM-RoBERTa [10] as an upgrade to their initial XLM-100 model. It is a scaled cross-lingual sentence encoder. Using self-supervised training approaches, it offers state-of-the-art performances in cross-lingual understanding where a model is taught in one language and then applied to multiple languages with no additional training data. This model showed increased performance on numerous NLP applications. XLM-RoBERTa creates the possibility for a one-model-for-many-languages approach rather than a single model per language. Here, we use HuggingFace's implementation of the Twitter XLM-RoBERTa-base model [11] for multigenre tasks. It is composed of 12 layers (i.e. transformer block), the dimension of hidden size is 768 and the number of the self-attention head is 12. We use this model for all language-specific tasks as the 2nd model to capture diverse contextual representation along with the language-specific BERT model.

2.2. BERTweet

RoBERTa is an extension to the original BERT model which is named as a robustly optimized BERT pre-training approach. It focuses on the key hyper-parameters choices and removing the next sentence prediction (NSP) objective. Besides, it is training with much larger mini-batches and learning rates. BERTweet [12] is trained based on the RoBERTa pre-training procedure. BERTweet consists of 850M English Tweets. We use base sized BERTweet model for both multimodal and language-specific English tasks.

2.2.1. AraBERT

AraBERT [13] is a version of BERT that is gaining popularity for effective contextual representation of textual contents in various Arabic tasks including natural language inference in language-specific characteristics, named entity recognition in Arabic corpora. It is pre-trained on the Arabic language which used workpiece vocabulary. In our approach, we employ the huggingface implementation of the AraBERTV2 base-sized model for all Arabic tasks.

2.2.2. Spanish BERT

BERT stands for bidirectional encoder representations from transformers and is a new method of pre-training sentence representations which achieves state-of-the-art results on many NLP tasks including question-answering, and text classification. BETO [14] is a BERT model trained on a big Spanish corpus. For the Spanish task, we used BETO (Spanish BERT) a size similar to a BERT-base model that trained with the whole word masking technique.

2.2.3. ConvNEXT

ConvNEXT [15] is a pure convolutional model and builds inspired by the design of Vision Transformers (ViT). It is a hierarchical transformer model that reintroduced various convolution network priors, creating the transformer as a generic vision backbone and achieving notable performance on a wide variety of vision tasks. These motivate us to use the ConvNEXT vision transformer model in our proposed method. We employed the ConvNEXT base-sized model as the image encoder which trained on ImageNet-1k at resolution 224x224.

2.3. System Architecture

We jointly finetune two transformer models for all tasks. In the multimodal task, we used a language-specific pre-train BERT transformer model and a pre-train ConvNEXT vision transformer model for contextual and visual representation. We utilize the BiLSTM module on top of the contextual representation to learn the long-term contextual representation, respectively. Then, we concatenate these feature vectors in a unified architecture to get a visual-contextual representation. This feature vector is then fed to the multi-sample dropout strategy to get the final label of each multimodal task.

In multigenre tasks, we used a language-specific BERT model and Twitter XLM-RoBERTa for language-specific contextual representation. It helps the model to capture diverse contextual feature representations present in tweets. We sum two models' sequence outputs and fed them to a BiLSTM module to capture the effective contextual information. Later, we employ a multi-sample dropout strategy on top of the BiLSTM module output which predicts the final label for each task of multigenre tasks.

2.4. Training Strategies

Different training strategies improved the performance of the transformers model. In this paper, we use a training strategy named multi-sample dropout [16] technique. The multi-sample dropout technique improves the generalization ability and accelerates the training of the base model [16]. To improve the accuracy of the transformer-based trained network, we utilise the multi-sample dropout technique. We choose 3 number of the dropout samples based on the validation data performance. In multi-sample dropouts, we duplicate the features vector of the transformer model after the dropout layer, while sharing the weights among these duplicated fully connected layers. We calculate the loss for each sample and then the losses are averaged to obtain the final loss.

3. Experiment and Evaluation

3.1. Dataset Description

The organizers used a benchmark dataset [17] published in ECIR-2023 to evaluate the performance of the participants' systems at the CheckThat! 2023 shared task [8]. Organizers provide different datasets for multimodal Arabic, multimodal English, and language-specific dataset including English, Arabic and Spanish languages. Dataset texts/images are taken from English, Arabic and Spanish tweets. The dataset statistics of subtask 1A: check-worthiness of multimodal content and subtask 1B: check-worthiness of multigenre unimodal content are presented in Table 1.

Table 1
Dataset statistics of CheckThat! 2023 shared task 1 according to each label on the corresponding task.

Task	Label	Train	Dev	Dev-Test	Test
Subtask 1A: Check-Worthiness of multimodal content					
Multimodal-Arabic	Yes	776	113	220	203
	No	1,421	207	402	792
	Total	2,197	320	622	995
Multimodal-English	Yes	820	87	174	277
	No	1,536	184	374	459
	Total	2,356	271	548	736
Subtask 1B: Check-Worthiness of multigenre unimodal content					
Unimodal-Arabic	Yes	1,758	485	411	123
	No	4,301	789	682	377
	Total	6,059	1,274	1,093	500
Unimodal-English	Yes	4,058	1,355	238	108
	No	12,818	4,270	794	210
	Total	16,876	5,625	1,032	318
Unimodal-Spanish	Yes	2,208	299	704	509
	No	5,280	2,161	4,296	4,491
	Total	7,488	2,460	5,000	5,000

3.2. Preprocessing

To preprocess the given tweet text we employ various preprocessing techniques. In English tweets, we expand the contraction (e.g. "isn't", "can't", and "aren't") into their normalized form for effective representation. The URLs and special characters do not contain any causal indicative information which may be useful for this task, we discard them from the tweet texts. We utilize a publicly available Python library emot to demojize (i.e. convert emojis into text) all the available emoji in the tweet texts. Moreover, all the words, characters, and punctuation floodings are replaced with a single one. For example, "It's crooked she's she's guilty of a very very serious crime.” is becoming “It’s crooked she’s guilty of a very serious crime.” after removing punctuation flooding and consecutive words. We preprocess Arabic text using the publicly available AraBERT model’s ArabertPreprocessor [1] [13]. To preprocess Spanish tweets we utilise publicly available python text normalization library cucco [2] where we utilise a list of normalizations including *replace_urls, remove_extra_whitespaces, replace_emojis, replace_hyphens, replace_punctuation, replace_symbols.*

3.3. Experimental Settings

We now describe the details of our experimental settings and the hyper-parameter settings with the finetuning strategy that we have employed to design our proposed system for the CheckThat! 2023 shared task 1.

Parameter	Optimal Value
Learning rate	3e-5
Max-len	128
Number of epochs	5
Batch size	4
Manual seed	42
BiLSTM hidden state	256
Dropout	0.1, 0.2, 0.3

Table 2
Model settings for CheckThat! 2023 shared task 1 of our proposed method.

We finetune state-of-the-art Huggingface [18] transformer models including Twitter XLM-RoBERTa[3], AraBERT[4], BERTweet[5], Spanish BERT[6] and ConvNEXT[7] model for this task. We used all models as the base size in this work. We concatenate the training and development data during the model training phase. We used the CUDA-enabled GPU of the Google Colaboratory [19] platform and set the manual seed = 42 to generate reproducible results. We obtained the optimal parameter settings of our proposed model based on the performance of the development set which is articulated in Table 2. We use a multi-sample dropout training strategy on top of the unified representation of multimodal and multigenre tasks. To determine the optimal dropout values, we searched over the set {0.1, 0.2, · · ·, 0.9} and found the best dropout range was 0.1 to 0.3 based on our experimental results on the development set. We used the default settings for the other parameters.

[1] https://github.com/aub-mind/arabert
[2] https://pypi.org/project/cucco/
[3] https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base
[4] https://huggingface.co/aubmindlab/bert-base-arabertv2
[5] https://huggingface.co/vinai/bertweet-base
[6] https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased
[7] https://huggingface.co/facebook/convnext-base-224

3.4. Evaluation Measure

The CheckThat! 2023 shared task 1: check-worthiness in multimodal and multigenre content organizers employed a standard evaluation metric F1-score over the positive class as the primary evaluation metric to evaluate the participants' system on tweet data.

Team	Multimodal Arabic	Multimodal English
CSECU-DSG	**0.399**	0.628
Top performing team and baseline performance based on F1-score		
Fraunhofer SIT [20]	-	**0.712**
ZHAW-CAI [21]	-	0.708
marvinpeng	0.312	0.697
Z-Index [22]	0.301	0.495
Baseline	0.299	0.474

Table 3
Comparative results with other selected participants (Sub-task 1A).

Team	Arabic	English	Spanish
CSECU-DSG	0.662	0.834	0.599
Top performing team and baseline performance based on F1-score			
ES-VRAI [23]	**0.809**	0.843	0.627
OpenFact [24]	-	**0.898**	-
DSHacker [25]	0.633	0.819	**0.641**
Fraunhofer SIT [20]	-	0.878	-
Baseline	0.625	0.462	0.172

Table 4
Comparative performances with other selected participants (Sub-task 1B).

3.5. Results and Analysis

In this section, we analyze the performance of our proposed CSECU-DSG system in the CheckThat! 2023 worthiness identification of tweet shared task. The comparative performance of our proposed CSECU-DSG system on subtask 1A test data against other top-performing participants' systems is presented in Table 3. We have seen that our proposed method achieved a 0.399 score and ranked 1st in the multimodal Arabic task based on F1-score over the positive class. Moreover, our system achieved competitive performance on multimodal English task. This validates the effectiveness of our proposed method in multimodal check-worthiness tasks.

The comparative performance of our proposed CSECU-DSG system on subtask 1B language-specific test data including English, Arabic and Spanish against other top-performing participants' systems are presented in Table 4. In language-specific tasks, our method achieved relatively lower performance as we are not effectively tuning the hyperparameters of our method. However, our method achieved 3rd and 4th place in Spanish and Arabic check-worthiness tasks, respectively. This validates the potency and applicability of our proposed method in this task.

4. Conclusion and Future Directions

In this paper, we present an approach to automatically identify the worthiness of factual claims present in tweets in multimodal and multigenre settings using fine-tuned transformers models fusion architecture. We employ a BiLSTM module on top of the language model to handle the long-term dependencies present in tweets. Moreover, we use the multi-sample dropout training strategy to speed up training and get better generalization ability. Experimental results demonstrated the efficacy of our proposed transformer-based method, where the fusion of transformer variants with the BiLSTM module and multi-sample dropout prediction helped us to obtain competitive performance and ranked 1st in 1A Arabic and 3rd in 1B Spanish in the CheckThat! 2023 shared task.

Further research will be conducted on other SOTA transformers models with a unified architecture of two or more. However, the classes of the dataset are imbalanced, so the weighted average fusion strategy of different models may be exploiting better context for check-worthiness from multimodal and multigenre tweets effectively.

References

[1] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, Y. Choi, Truth of varying shades: Analyzing language in fake news and political fact-checking, in: Proceedings of the 2017 conference on empirical methods in natural language processing, 2017, pp. 2931–2937.

[2] P. Atanasova, A. Barron-Cedeno, T. Elsayed, R. Suwaileh, W. Zaghouani, S. Kyuchukov, G. D. S. Martino, P. Nakov, Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. task 1: Check-worthiness, arXiv preprint arXiv:1808.05542 (2018).

[3] P. Atanasova, P. Nakov, G. Karadzhov, M. Mohtarami, G. Da San Martino, Overview of the clef-2019 checkthat! lab: Automatic identification and verification of claims. task 1: Check-worthiness., CLEF (Working Notes) 2380 (2019).

[4] 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.

[5] Z. Guo, M. Schlichtkrull, A. Vlachos, A survey on automated fact-checking, Transactions of the Association for Computational Linguistics 10 (2022) 178–206.

[6] R. Vijjali, P. Potluri, S. Kumar, S. Teki, Two stage transformer model for covid-19 fake news detection and fact checking, arXiv preprint arXiv:2011.13253 (2020).

[7] E. Williams, P. Rodrigues, V. Novak, Accenture at checkthat! 2020: If you say so: Post-hoc fact-checking of claims using transformer-based models, arXiv preprint arXiv:2009.02431 (2020).

[8] F. and Alam, A. Barrón-Cedeño, G. S. Cheema, S. Hakimov, M. Hasanain, C. Li, R. Míguez, H. Mubarak, G. K. Shahi, W. Zaghouani, P. Nakov, Overview of the CLEF-2023 CheckThat! lab task 1 on check-worthiness in multimodal and multigenre content, in: Working Notes of CLEF 2023—Conference and Labs of the Evaluation Forum, CLEF '2023, Thessaloniki, Greece, 2023.

[9] A. Barrón-Cedeño, F. Alam, A. Galassi, G. Da San Martino, P. Nakov, , T. Elsayed, D. Azizov, T. Caselli, G. Cheema, F. Haouari, M. Hasanain, M. Kutlu, C. Li, F. Ruggeri, J. M. Struß, W. Zaghouani, Overview of the CLEF–2023 CheckThat! Lab checkworthiness, subjectivity, political bias, factuality, and authority of news articles and their source, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), 2023.

[10] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, arXiv preprint arXiv:1911.02116 (2019).

[11] F. Barbieri, L. Espinosa Anke, J. Camacho-Collados, Xlm-t: A multilingual language model toolkit for twitter, arXiv e-prints (2021) arXiv–2104.

[12] D. Q. Nguyen, T. Vu, A. T. Nguyen, BERTweet: A pre-trained language model for English Tweets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 9–14.

[13] W. Antoun, F. Baly, H. Hajj, Arabert: Transformer-based model for arabic language understanding, in: LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020, ????, p. 9.

[14] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[15] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, CoRR abs/2201.03545 (2022). URL: https://arxiv.org/abs/2201.03545. arXiv:2201.03545.

[16] H. Inoue, Multi-sample dropout for accelerated training and better generalization, arXiv preprint arXiv:1905.09788 (2019).

[17] A. Barrón-Cedeño, F. Alam, T. Caselli, G. Da San Martino, T. Elsayed, A. Galassi, F. Haouari, F. Ruggeri, J. M. Struss, R. N. Nandi, G. S. Cheema, D. Azizov, P. Nakov, The clef-2023 checkthat! lab: Checkworthiness, subjectivity, political bias, factuality, and authority, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2023, pp. 506–517.

[18] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., HuggingFace's Transformers: State-of-the-art Natural Language Processing, arXiv preprint arXiv:1910.03771 (2019).

[19] E. Bisong, E. Bisong, Google colaboratory, Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners (2019) 59–64.

[20] R. A. Frick, I. Vogel, J.-E. Choi, Fraunhofer SIT at CheckThat! 2023: Enhancing the detection of multimodal and multigenre check-worthiness using optical character recognition and model souping, ????

[21] P. von Däniken, J. Deriu, M. Cieliebak, Zhaw-cai at CheckThat! 2023: Ensembling using kernel averaging, ????

[22] P. Tarannum, M. A. Hasan, F. Alam, S. R. H. Noori, Z-Index at CheckThat! 2023: Unimodal and multimodal checkworthiness classification, ????

[23] H. T. Sadouk, F. Sebbak, H. E. Zekiri, Es-vrai at CheckThat! 2023: Analyzing checkworthiness in multimodal and multigenre contents through fusion and sampling approaches, ????

[24] M. Sawiński, K. Wecel, E. Ksieżniak, M. Stróżyna, W. Lewoniewski, P. Stolarski, W. Abramowicz, Openfact at CheckThat! 2023: Head-to-head gpt vs. bert - a comparative study of transformers language models for the detection of check-worthy claims, ????

[25] A. Modzelewski, W. Sosnowski, A. Wierzbicki, DSHacker at CheckThat! 2023: Check-Worthiness in Multigenre and Multilingual Content With GPT-3.5 Data Augmentation, ????