ES-VRAI at CheckThat! 2023: Analyzing Checkworthiness in Multimodal and Multigenre Contents through Fusion and Sampling Approaches

Hamza Tarik Sadouk[1,*], Faouzi Sebbak[1] and Hussem Eddine Zekiri[1]

[1]*Ecole Militaire Polytechnique, PO Box 17, 16111 Bordj El Bahri, Algiers, Algeria*

Abstract

This paper describes our participation in the CLEF2023 CheckThat! Lab [1], specifically focusing on Task 1, which addresses Checkworthiness in Multimodal and Unimodal Contents [2]. The task involves determining the worthiness of fact-checking a claim in a tweet. Traditionally, this decision relies on professional fact-checkers or human annotators who consider auxiliary questions like verifiability and harmfulness. Task 1 comprises two subtasks: Subtask 1A focuses on assessing the Checkworthiness of tweets containing both text and images, offered in Arabic and English, while Subtask 1B involves assessing the Checkworthiness of text snippets from tweets or debate/speech transcriptions available in Arabic, English, and Spanish. For subtask 1B, we proposed different methods based on pre-trained transformer models and sampling techniques that helped us achieve the first position in Arabic, the second position in Spanish, and the fifth position in English. For subtask 1A, we presented various multimodal fusion strategies utilizing pre-trained language and vision models. Although we obtained a commendable third-place result in the English Multimodal dataset, unfortunately, we did not secure a position on the leaderboard.

1. Introduction

Selecting claims for verification is a complicated task. Fact-checking is a time-intensive process, and it can be challenging to determine if a claim is genuine or misleading. Fact-checkers must weigh the potential harm caused by misleading claims, such as risks to health, democracy, and emergencies, against the effort needed to verify them. Additionally, fact-checkers strive to maintain impartiality, so their tools mustn't introduce unfair biases.

In some countries, reliable official statistics are not published by governments, making sure claims related to statistics are nearly impossible to validate. While simple algorithms can often identify viral content, assessing the "Checkworthiness" of a claim is more complicated. For instance, breaking news stories can be both popular and accurate. Due to the limited resources of fact-checking organizations, many check-worthy claims go unchecked. Historical lists of

checked or unchecked claims are unreliable indicators for determining which similar claims should be fact-checked.

Claims can be found across various sources, such as news websites, social media platforms (text, audio, or video), and broadcast media. Fact-checkers use diverse technologies to monitor these sources, including news alerts, automatic speech recognition, and translation tools, which rely on underlying AI technologies.

2. Related Work

Fact-checkers face a deluge of claims and must determine which are worth investigating, leading to the development of AI solutions such as those in the CLEF CheckThat! lab 2018-2022 [3][4][5][6][7] and within dedicated fact-checking organizations like Full Fact [8]. The problem is often approached as a ranking task, where systems assign Checkworthiness scores to increase transparency and help fact-checkers prioritize or filter claims. Fact-checkers can then provide feedback on the scores' accuracy, which can be used to refine the system.

ClaimBuster [9] is the first Checkworthiness detection system employed by fact-checkers in the Duke Reporters' Lab project [10]. The system was trained on a manually annotated dataset to differentiate between non-factual sentences, unimportant factual claims, and check-worthy factual claims using features such as sentiment, named entities, part-of-speech tags, words, and claim length. The dataset used in this study included historical US election debate transcripts from 1960 to 2012, covering 30 debates and 28,029 transcribed sentences. Each sentence was labeled with the speaker (candidate or moderator) [9]. The ClaimBuster system utilized an SVM classifier and a variety of features, such as sentiment, TF-IDF word representations, part-of-speech (POS) tags, and named entities. It provided a Checkworthiness ranking based on the SVM prediction scores but did not attempt to replicate the Checkworthiness decisions of any specific fact-checking organization. CNN and PolitiFact later evaluated the system.

Pepa Gencheva et al. [11] focused on US 2016 Presidential Campaign debates and used existing annotations from nine respected fact-checking organizations (PolitiFact, FactCheck, ABC, CNN, NPR, NYT, Chicago Tribune, The Guardian, and Washington Post). Their model considered the context of each sentence, including its position within a more extended intervention by one of the actors, and predicted (i) whether any fact-checking organizations would select the target sentence and (ii) whether a specific organization would select it.

Konstantinovskiy et al. [12] developed a more detailed schema and dataset for Checkworthiness annotation of TV shows. Gencheva et al. [13] created a dataset of political debates by observing which sentences were fact-checked and modeled the sentence structure and claim context. This dataset was used in the ClaimRank system [14] and extended for multitask learning from nine fact-checking organizations [15]. Further extensions were used in the CLEF CheckThat! lab, with participants developing models based on pre-trained transformers like BERT and RoBERTa [16, 17, 18]. The task was also modeled using positive unlabeled learning [19].

Social media companies are also working to combat misinformation and disinformation on their platforms. Facebook developed a proprietary tool to identify claims for fact-checking [20]. They use user flags indicating potentially false posts and features from reply content to predict

Table 1
Multimodal dataset description for subtask 1A

Language	Class	Train	Dev	Dev-Test	Total
English	No	**1,635**	182	345	2,162
	Yes	**664**	74	180	918
	Total	2,299	256	525	3,080
Arabic	No	**1,421**	207	402	2,030
	Yes	**776**	113	220	1,109
	Total	2,197	320	622	3,139

Table 2
Multigenre dataset description for subtask 1B

Language	Class	Train	Dev	Dev-Test	Total
Arabic	No	**4,301**	789	682	5,772
	Yes	**1,758**	485	411	2,654
	Total	6,059	1,274	1,093	8,426
English	No	**12,818**	4,270	794	17,882
	Yes	**4,058**	1,355	238	5,651
	Total	16,876	5,625	1,032	23,533
Spanish	No	**5,280**	2,161	4,296	11,737
	Yes	**2,208**	299	704	3,211
	Total	7,488	2,460	5,000	14,948

if a post contains false information, updating the model based on fact-checker feedback.

3. Data Overview

In this section, we discuss the different datasets related to the subtasks (subtask 1A: Multimodal, and subtask 1B: Multigenre) provided as part of the Checkthat 2023 competition [21, 2].

3.1. Datasets Description

Checkworthiness consists of determining whether a given tweet is worth fact-checking. This task is multimodal (text+image) provided in English and Arabic (see table 1), and multigenre (text only) provided in Arabic, English, and Spanish (see table 2).

3.2. Class imbalance

Observation of Table 1 and Table 2 reveals the presence of class imbalance. In order to address this particular concern, we employ a set of techniques [22] as described subsequently.

3.2.1. Class weight

Class weighting is a technique in machine learning that assigns different weights to address imbalanced data. It emphasizes underrepresented classes during training, improving perfor-

mance in terms of precision, recall, and F1-Score, making it valuable for real-world scenarios with important minority classes or higher costs for misclassification.

3.2.2. Upsampling

Upsampling is a method used to tackle imbalanced data by increasing the number of samples in the minority class. It can improve performance for underrepresented classes but has drawbacks like overfitting due to duplicated or similar instances, leading to limited generalization on unseen data.

3.2.3. Downsampling

Downsampling is a method used to handle imbalanced data by decreasing the number of instances in the majority class. It aims to achieve balance by randomly removing samples from the majority class. However, downsampling has limitations, including the loss of valuable information and smaller dataset size, which may not be ideal for training complex models.

4. Subtask 1A: Checkworthiness-multimodal

4.1. Method

This subtask consists of checking the worthiness of a given tweet by considering both text and associated image + its OCR text. Our methodology involves the integration of text and image features through the utilization of diverse techniques, encompassing pre-trained models, classifiers, and fusion strategies. This amalgamation allows for the effective combination and synthesis of textual and visual information, leading to a comprehensive and enriched representation of the data.

4.1.1. Pre-trained Models

We used several pre-trained models to extract features from both text and images:

- **BERT**: Bidirectional Encoder Representations from Transformers is a popular pre-trained language model that effectively captures contextual information in a text [23].
- **DistilBERT**: A distilled version of BERT that retains most of its performance while being smaller and faster [24].
- **CLIP-ViT16 and CLIP-RN50**: These models are part of the Contrastive Language Image Pretraining (CLIP) framework [25], which learns joint image and text representations. ViT16 is based on the Vision Transformer architecture [26], and RN50 is based on a ResNet-50 architecture [27].
- **ViT**: short for Vision Transformer [26]. It adapts the Transformer architecture, originally designed for natural language processing, to process images by dividing them into non-overlapping patches and treating them as tokens.
- **ResNet**: short for Residual Network, is a family of deep convolutional neural networks introduced by He et al. in 2015 [27]. It employs residual learning to address the vanishing gradient problem that arises in deep architectures.

Table 3
Classifier Parameters

Model Name	Parameters / Late Fusion Strategy
RF1	Random Forest (1000 estimators)+class-weight
RF2	Random Forest (100 estimators)+class-weight
RF3	Random Forest (10 estimators)+class-weight
GB1	Gradient Boosting (100 estimators)
GB2	Gradient Boosting (1000 estimators)
MLP1	MLP(hidden_layer_sizes=(100, 50))
MLP2	MLP(hidden_layer_sizes=(1000, 50))
MLP3	MLP(hidden_layer_sizes=(1000, 500))
XGB1	XGBoost (100 estimators)
DNN1	DNN(4 hidden layers, 20 epochs)
DNN2	DNN(6 hidden layers, 20 epochs)
DNN3	DNN(10 hidden layers, 20 epochs)
SVM1	SVM(linear kernel)
SVM2	SVM(rbf kernel, C=5, gamma='scale')
SVM3	SVM(rbf kernel, C=8, gamma='scale')

4.1.2. Classifiers

After extracting features from the text and images, we combined these features using various classifiers. Table 3 shows the used classifiers with their parameters:

- **Random Forest (RF)**: An ensemble method that builds multiple decision trees and combines their outputs [28].
- **Gradient Boosting (GB)**: A boosting technique that builds a series of weak learners, iteratively improving on their performance by focusing on misclassified examples [29].
- **Multi-Layer Perceptron (MLP)**: A feedforward artificial neural network with one or more hidden layers [30].
- **XGBoost**: An optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable [31].
- **Deep Neural Network (DNN)**: A neural network with multiple hidden layers, allowing for more complex feature representations [32].
- **Support Vector Machine (SVM)**: A classifier that finds the optimal hyperplane to separate different classes in a high-dimensional feature space [33].

4.1.3. Fusion Techniques

We used two main fusion techniques [34] to combine the features extracted from text and images:

- **Early Fusion**: In this approach, features from both modalities are combined before being fed into the classifier. This allows the model to learn joint representations of the input data and capture the interactions between the modalities more effectively.

Table 4

Late fusion models and strategies

Model Name	combined models	Late Fusion Strategy
LF1	DistilBERT+VIT	averaging
LF2	DistilBERT+VIT	weighted averaging
LF3	DistilBERT+VIT	trainable fusion layer
LF4	DistilBERT+VIT	maximum probability

- **Late Fusion**: In this approach, the modalities are processed separately, and their outputs are combined afterward. This allows each modality to be modeled independently, focusing on the most discriminative features of each modality without being affected by noise from the other modality.

Late fusion methods employed in the experiments are summarized in Table 4. They include averaging, weighted averaging, trainable fusion layer, majority voting, and maximum probability.

4.2. Results and discussion

We evaluated the performance of different models and fusion techniques on the Dev-Test English dataset of subtask 1A using the metrics of accuracy, F1-Score, precision, and recall. By analyzing these results, we aimed to identify the most effective strategies for combining text and image features in this specific case. From table 5, we have the following observations:

- The top-performing model, BERT+ResNet50 with early fusion, achieves a balanced F1-Score of 0.7160, precision of 0.8056, and recall of 0.6444. Late fusion models, like DistilBERT+ViT transformer (LF3), show promise with an F1-Score of 0.7029, precision of 0.8271, and recall of 0.6111. The early fusion model with CLIP-RN50 and 10 hidden layers DNN achieves the highest recall of 0.7944 but has lower precision (0.6059). Early fusion captures modality interactions for higher F1-Scores and recall, while late fusion focuses on discriminative features for higher precision. These observations highlight the trade-offs and strengths of early and late fusion approaches in integrating multimodal information.
- During the evaluation cycle, we employed **BERT+ResNet50 with a classification layer and early fusion**, which yielded an impressive F1-score of **0.704**. It is important to highlight that processing multimodal data demands significant resources. Unfortunately, due to this limitation, we were unable to produce the required results within the evaluation cycle deadline. Consequently, despite achieving a commendable third-place result, we did not secure a position on the leaderboard.

5. Subtask 1B: Checkworthiness-multigenre

5.1. Method

This subtask consists of checking the worthiness of a given tweet by considering text only in three languages: Arabic, English, and Spanish. The designated objective was accomplished

Table 5
English Multimodal models performance comparison

Model	Method	Accuracy	F1-Score	Precision	Recall
clip-vit16+RF1	Early Fusion	0.7752	0.5597	0.8523	0.4167
clip-vit16+RF2	Early Fusion	0.7733	0.5576	0.8427	0.4167
clip-vit16+RF3	Early Fusion	0.7390	0.4669	0.7792	0.3333
clip-vit16+GB1	Early Fusion	0.7943	0.6276	0.8273	0.5056
clip-vit16+GB1+PCA	Early Fusion	0.7714	0.6026	0.7459	0.5056
clip-vit16+GB2	Early Fusion	0.8038	0.6555	0.8235	0.5444
clip-vit16+MLP1	Early Fusion	0.7962	0.6646	0.7626	0.5889
clip-vit16+MLP2	Early Fusion	0.8114	0.6877	0.7956	0.6056
clip-vit16+MLP3)	Early Fusion	0.8133	0.6957	0.7887	0.6222
clip-vit16+XGB1	Early Fusion	0.7810	0.6048	0.7928	0.4889
clip-rn50+RF1	Early Fusion	0.7714	0.5652	0.8125	0.4333
clip-rn50+MLP2	Early Fusion	0.8038	0.6709	0.7895	0.5833
clip-rn50+MLP3	Early Fusion	0.8000	0.6645	0.7820	0.5778
clip-rn50+MLP4	Early Fusion	0.8038	0.6791	0.7730	0.6056
clip-rn50+GB1	Early Fusion	0.7981	0.6443	0.8136	0.5333
clip-rn50+DNN1	Early Fusion	0.8133	0.7135	0.7531	0.6778
clip-rn50+DNN2	Early Fusion	0.8133	0.7118	0.7562	0.6722
clip-rn50+DNN3	Early Fusion	0.7524	0.6875	0.6059	0.7944
clip-rn50+1	Early Fusion	0.8095	0.6689	0.8279	0.5611
clip-rn50+SVM2	Early Fusion	0.8133	0.6859	0.8106	0.5944
clip-rn50+SVM3	Early Fusion	0.8152	0.6901	0.8120	0.6000
BERT+ResNet50	Early Fusion	0.8248	**0.7160**	0.8056	0.6444
DistilBERT+VIT	Early Fusion	0.7848	0.6271	0.7724	0.5278
LF1	Late Fusion	0.8019	0.6232	0.8958	0.4778
LF2	Late Fusion	0.8171	0.6800	0.8500	0.5667
LF3	Late Fusion	0.8229	0.7029	0.8271	0.6111
LF4	Late Fusion	0.8019	0.6232	0.8958	0.4778

through the process of fine-tuning pre-trained language models in combination with sampling techniques to address class imbalance.

5.2. Language models

5.2.1. English models

- **RoBERTa** [35] is an enhanced version of BERT, trained on a larger dataset without the next sentence prediction task, resulting in improved performance.
- **XLNet** [36] is an advanced NLP model that combines masked language modeling and autoregression techniques, introducing variations to the transformer structure.
- **ALBERT** [37] is a memory-efficient variant of BERT that reduces parameters and improves speed while incorporating a Sentence Order Prediction (SOP) loss during training.
- **BigBird** [38] is a transformer model optimized for long sequences, using sparse attention and a novel sentence-level training objective.

- **GPT-2** [39] is a powerful language model trained on massive text data, capable of performing various NLP tasks and applied in real-world applications.

5.2.2. Arabic models

- **Arabert** [40], a BERT-based language model, is optimized for Arabic NLP tasks, demonstrating effectiveness in sentiment analysis, and text classification.
- **ARBERT** [41] is a large-scale pre-trained masked language model for Modern Standard Arabic, based on BERT-base architecture, with 163 million parameters.
- **MARBERT** [41] is an Arabic language model pre-trained on a diverse Twitter dataset specifically designed to handle Arabic dialect variations.
- **ARAELECTRA** [42] is an Arabic language representation model based on ELECTRA, optimized for Arabic reading comprehension tasks.

5.2.3. Spanish models

- **BETO** [43] is a BERT model trained on a large Spanish language dataset using Whole Word Masking, similar in size to BERT-Base.
- **BERTIN** [44] is a series of BERT-based models for Spanish, trained from scratch on the Spanish portion of mC4 using Flax.

5.2.4. Multilinguall models

- **BERT-base-multilingual** [23] is a pretrained language model for cross-lingual transfer learning, trained on a diverse corpus of 104 languages.
- **XLM-RoBERTa** [45] is a cross-lingual language model based on RoBERTa, pre-trained on multilingual datasets and outperforming previous models.
- **GPT-3** [46] is an advanced language model by OpenAI, known for its impressive text generation and few-shot learning capabilities.

5.3. Results and discussion

We evaluated the performance of different models and sampling techniques on the Dev-Test Arabic, English, and Spanish dataset of subtask 1B using the metrics of accuracy, F1-Score, precision, and recall.

5.3.1. Arabic

Upon analyzing table 6, the following observations has been made:

- AraBERT achieves the highest accuracy of 0.6917, but considering class imbalance, accuracy might not be the most suitable metric. The MarBERT + Downsampling model performs best according to the F1-Score (0.6053), showcasing a good balance between precision and recall. MarBERT + Downsampling exhibits a high recall of 0.8887, effectively identifying a significant portion of Check-worthy tweets, albeit with a precision of

Table 6

Check-worthiness multigenre Arabic models evaluation

Model	Accuracy	F1-Score	Precision	Recall
MarBERT+Upsampling	0.6279	0.5730	0.5088	0.6557
MarBERT+Downsampling	0.5589	**0.6053**	0.4590	0.8887
MarBERT	0.6397	0.3634	0.5551	0.2701
ArBERT+Upsampling	0.5974	0.4500	0.4627	0.4380
ArBERT+Downsampling	0.5706	0.6039	0.4654	0.8598
ArBERT	0.6805	0.5666	0.5859	0.5485
AraBERT	0.6917	0.5415	0.6149	0.4845
AraBERT+Downsampling	0.6734	0.5873	0.5595	0.6181
AraBERT+Oversampling	0.6587	0.5224	0.5512	0.4962
XLM-RoBERTa	0.6487	0.3356	0.5807	0.2365
XLM-RoBERTa+Downsampling	0.5837	0.5427	0.4622	0.6575
XLM-RoBERTa+Oversampling	0.6285	0.4116	0.5093	0.3455
BERT-multilingual	0.6313	0.3958	0.5156	0.3217
AraElectra	0.5873	0.5472	0.5156	0.3271

0.4590, indicating potential false positives. In contrast, AraBERT demonstrates relatively balanced precision (0.6149) and recall (0.4845), offering a better compromise between identifying Check-worthy tweets and reducing false positives.

- In the evaluation cycle, we utilized **MarBERT + Downsampling** after the Test dataset release. We achieved first place with a remarkable F1-score of **0.809**, surpassing the second-place team by a substantial margin with their F1-score of 0.733, indicating a significant lead on our part.

5.3.2. English

After examining the data presented in Table 7, the following observations have been identified:

- The BERT, XLNet, BERTweet, BERT Multilingual, RoBERTa, DistilBERT, and BigBird models are fine-tuned on the dataset without modifications. Although these models exhibit high accuracy and F1-Scores, the imbalanced nature of the dataset limits their performance. Implementing Downsampling leads to a decrease in accuracy and F1-Score for most models, suggesting information loss. Oversampling has minimal impact on performance. Class-weight slightly decreases the F1-Score for BERTweet and DistilBERT. GPT-2 demonstrates high accuracy but relatively lower F1-Score, potentially due to its text generation focus. GPT-3 Curie excels with the highest accuracy (0.9815) and F1-Score (0.9603) among all models. GPT-3 Ada achieves comparable performance with XLNet and BERTweet, displaying high accuracy (0.9757) and F1-Score (0.9478). Overall, GPT-3 Curie achieves the highest F1-Score (0.9603), followed by BERTweet (0.9540) and GPT-3 Ada (0.9478), performing well on the imbalanced dataset.
- During the evaluation cycle, we utilized BERTweet despite it being the second-best performer in the Dev-Test dataset. Unfortunately, we couldn't use GPT-3 due to limited access and time constraints imposed by the competition. We secured fifth place by employing **BERTweet**, achieving an F1-score of **0.843**.

Table 7
Check-worthiness multigenre English models evaluation

Model	Accuracy	F1-Score	Precision	Recall
BERT	0.9748	0.9454	0.9452	0.9452
BERT + Downsampling	0.9651	0.9277	0.8885	0.9705
BERT + Oversampling	0.9651	0.9277	0.8885	0.9705
XLNet	0.9758	0.9478	0.9419	0.9538
XLNet + Downsampling	0.9554	0.9105	0.8478	0.9831
XLNet + Oversampling	0.9719	0.9409	0.9130	0.9705
BERTweet	0.9787	**0.9540**	0.9500	0.9580
BERTweet	0.9767	0.9504	0.9351	0.9663
BERTweet + Oversampling	0.9680	0.9339	0.8929	0.9790
BERTweet + Downsampling	0.9545	0.9084	0.8473	0.9790
BERT Multilingual	0.9729	0.9414	0.9375	0.9452
BERT Multilingual + Downsampling	0.9438	0.8885	0.8190	0.9705
RoBERTa	0.9765	0.9461	0.9344	0.9580
RoBERTa + Downsampling	0.9535	0.9073	0.8393	0.9874
XLM-RoBERTa + Downsampling	0.9535	0.9066	0.8441	0.9790
DistilBERT	0.9705	0.9342	0.9441	0.9244
DistilBERT + Class-weight	0.9680	0.9328	0.9051	0.9621
DistilBERT + Oversampling	0.9564	0.9116	0.8561	0.9748
DistilBERT + Downsampling	0.9273	0.8609	0.7708	0.9748
BigBird	0.9748	0.9458	0.9381	0.9538
BigBird + Downsampling	0.9428	0.8876	0.8120	0.9790
BigBird + Oversampling	0.9767	0.9508	0.9280	0.9748
GPT-2	0.9800	0.9200	0.9300	0.9100
GPT-3 Ada	0.9757	0.9478	0.9419	0.9537
GPT-3 Curie	0.9815	0.9603	0.9543	0.9663

5.3.3. Spanish

Upon analyzing the information provided in Table 8, the following observations have been made:

- The BERT-based models, specifically BERT-Spanish, BERTin-RoBERTa-Spanish, and XLM-RoBERTa, demonstrate strong performance on the imbalanced dataset. Notably, XLM-RoBERTa achieves the highest F1-Score and accuracy without using sampling techniques, indicating its suitability for handling the imbalanced dataset. The impact of Downsampling and Oversampling varies across different models, with slight decreases in F1-Scores observed for BERT-Spanish and XLM-RoBERTa, while BERT-multilingual benefits from both Downsampling and Oversampling. This highlights the dependence of sampling technique effectiveness on the specific model and dataset.
- In the evaluation cycle, we opted for **XLM-RoBERTa**, resulting in us securing second place with an F1-score of **0.627**. Notably, the first-place position was attained with a slightly higher F1-score of 0.641.

Table 8
Check-worthiness multigenre Spanish models evaluation

Model	Accuracy	F1-Score	Precision	Recall
BERT-Spanish	0.9138	0.6897	0.6991	0.6804
BERT-Spanish + Downsampling	0.8890	0.6630	0.5790	0.7756
BERT-Spanish + Oversampling	0.9074	0.6623	0.6806	0.6449
BERTin-RoBERTa-Spanish	0.9200	0.6700	0.7400	0.6200
BERTin-RoBERTa-Spanish + Downsampling	0.8800	0.6500	0.5500	0.8000
BERTin-RoBERTa-Spanish + Oversampling	0.9200	0.6700	0.7400	0.6100
XLM-RoBERTa	0.9130	**0.6943**	0.6871	0.7017
XLM-RoBERTa + Downsampling	0.8736	0.6445	0.5335	0.8139
XLM-RoBERTa + Oversampling	0.8940	0.6803	0.5912	0.8011
BERT-multilingual	0.6313	0.3958	0.5155	0.3212
BERT-multilingual + Downsampling	0.8600	0.6187	0.5018	0.8068
BERT-multilingual + Oversampling	0.8906	0.6500	0.5914	0.7216

6. Conclusion

When considering both multimodal and multigenre data, several important findings emerged. For multimodal data, the best-performing model was BERT+ResNet50 with early fusion, achieving an F1-Score of 0.7160 and demonstrating a balanced precision and recall. Late fusion models, like DistilBERT+ViT transformer with trainable fusion layer, achieved a slightly lower F1-Score of 0.7029. The model with the highest recall was CLIP-RN50 with ten hidden layers DNN (early fusion) at 0.7944 but with a lower precision of 0.6059. Early fusion models excelled at capturing modality interactions, resulting in higher F1-Scores and recall, while late fusion models prioritized discriminative features for improved precision. Our employed BERT+ResNet50 with early fusion model achieved an impressive F1-Score of 0.704. Still, due to resource limitations, it couldn't meet the evaluation cycle deadline, resulting in a third-place result without securing a position on the leaderboard.

When examining performance in multigenre data, English dataset models consistently outperformed Spanish and Arabic counterparts. English achieved the highest F1-Score of 0.9603 (GPT-3 Curie), while Spanish and Arabic reached 0.6943 (XLM-RoBERTa) and 0.6053 (MarBERT+Downsampling), respectively. Resampling techniques notably impacted Spanish and Arabic models, particularly enhancing the Arabic F1-Score from 0.3634 to 0.6053 using Downsampling with MarBERT. Arabic exhibited more significant performance variation, potentially due to its complexity and variations in model quality. Selecting appropriate pre-trained models and resampling techniques, such as BERT-Spanish for Spanish and MarBERT for Arabic, proved crucial. A thorough experimentation with models and processes is necessary to achieve optimal language-specific performance, emphasizing the need to adapt strategies based on language characteristics for maximum effectiveness across diverse languages and tasks.

Finally, we must point out that the field of fake news detection and information authenticity verification is continuously progressing, encompassing various aspects beyond checkworthiness estimation, particularly in the context of social media platforms. In light of this, the present study establishes a foundational framework for such a verification pipeline, incorporating Language

Models, Sampling techniques, and Fusion strategies to enhance the efficacy of checkworthiness assessment in scenarios involving multimodal and multilingual content, which commonly arise in social media environments. Furthermore, the pipeline can incorporate additional mechanisms for information authentication, such as considering the Stance of credible information sources and employing Subjectivity detection to differentiate objective facts from subjective opinions.

References

[1] 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.

[2] 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.

[3] P. Nakov, A. Barrón-Cedeño, T. Elsayed, R. Suwaileh, L. Màrquez, W. Zaghouani, P. Atanasova, S. Kyuchukov, G. Da San Martino, Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims, in: CLEF, 2018.

[4] T. Elsayed, P. Nakov, L. Màrquez, M. Hasanain, R. Suwaileh, P. Atanasova, Overview of the clef-2019 checkthat!: Automatic identification and verification of claims, in: CLEF, 2019.

[5] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, C. Li, S. Shaar, H. Mubarak, A. Nikolov, Y. S. Kartal, J. Beltrán, Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets, in: Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[6] P. Nakov, T. Elsayed, M. Hasanain, R. Suwaileh, P. Atanasova, A. Barrón-Cedeño, Overview of the clef-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news, in: CLEF, 2021.

[7] A. Barrón-Cedeño, T. Elsayed, P. Nakov, M. Hasanain, R. Suwaileh, P. Atanasova, W. Salloum, Overview of checkthat! 2020: Automatic identification and verification of claims in social media, in: CLEF, 2020.

[8] Full fact, https://fullfact.org/, ???? Accessed: 2022-04-02.

[9] N. Hassan, G. Zhang, F. Arslan, C. Caragea, M. Tremayne, B. Adair, C. Yang, Claimbuster: The first-ever system for automated detection of check-worthy factual claims in political debates, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, ACM, 2017.

[10] Duke reporters' lab, https://reporterslab.org/, ???? Accessed: 2022-04-02.

[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: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, 2017, pp. 267–276.

[12] L. Konstantinovskiy, I. Augenstein, Check-worthiness detection as positive unlabelled learning, arXiv preprint arXiv:2101.09585 (2021).

[13] P. Gencheva, P. Nakov, L. Màrquez, A. Barrón-Cedeño, A context-aware approach for detecting worth-checking claims in political debates, in: Proceedings of the International Conference Recent Advances in Natural Language Processing, ACL, 2017.

[14] I. Jaradat, P. Gencheva, P. Nakov, A. Barrón-Cedeño, L. Màrquez, Claimrank: Detecting check-worthy claims in arabic and english, in: Proceedings of the 27th International Conference on Computational Linguistics, ACL, 2018.

[15] M. Vasileva, P. Gencheva, P. Nakov, L. Màrquez, Multitask learning for check-worthiness detection, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL, 2019.

[16] M. Hasanain, T. Elsayed, Automatic arabic claim identification, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL, 2020.

[17] A. Nikolov, A. Barrón-Cedeño, R. Suwaileh, M. Hasanain, T. Elsayed, P. Nakov, Adapting bert for check-worthiness detection in english and arabic, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL, 2020.

[18] P. Williams, G. Murdock, Z. Karim, Check-worthiness detection in english and arabic using transformer models, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL, 2020.

[19] J. Wright, I. Augenstein, Claim detection in online debates as positive unlabeled learning, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL, 2020.

[20] About fact-checking in facebook and instagram, https://www.facebook.com/help/publisher/182222309230722, 2022. Accessed: 2023-05-23.

[21] 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, et al., The clef-2023 checkthat! lab: Checkworthiness, subjectivity, political bias, factuality, and authority, in: Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III, Springer, 2023, pp. 506–517.

[22] H. He, E. A. Garcia, Learning from imbalanced data, IEEE Transactions on knowledge and data engineering 21 (2009) 1263–1284.

[23] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2019).

[24] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 5–10.

[25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[27] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[28] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.

[29] J. H. Friedman, Greedy function approximation: a gradient boosting machine, Annals of statistics (2001) 1189–1232.

[30] D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning internal representations by error propagation, Technical Report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[31] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.

[32] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, nature 521 (2015) 436–444.

[33] C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (1995) 273–297.

[34] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE transactions on pattern analysis and machine intelligence 41 (2018) 423–443.

[35] 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, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, 2019, pp. 216–221.

[36] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, in: Advances in neural information processing systems, 2019, pp. 5753–5763.

[37] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, Albert: A lite bert for self-supervised learning of language representations, arXiv preprint arXiv:1909.11942 (2020).

[38] M. Zaheer, G. Guruganesh, A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, et al., Big bird: Transformers for longer sequences, in: International Conference on Learning Representations, 2021.

[39] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI Blog (2019).

[40] W. Antoun, F. Baly, H. Hajj, Arabert: Transformer-based model for arabic language understanding, arXiv preprint arXiv:2003.00104 (2020).

[41] M. Abdul-Mageed, A. Elmadany, E. M. B. Nagoudi, Arbert & marbert: deep bidirectional transformers for arabic, arXiv preprint arXiv:2101.01785 (2020).

[42] W. Antoun, F. Baly, H. Hajj, Araelectra: Pre-training text discriminators for arabic language understanding, arXiv preprint arXiv:2012.15516 (2020).

[43] 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.

[44] J. D. la Rosa y Eduardo G. Ponferrada y Manu Romero y Paulo Villegas y Pablo González de Prado Salas y María Grandury, Bertin: Efficient pre-training of a spanish language model using perplexity sampling, Procesamiento del Lenguaje Natural 68 (2022) 13–23. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6403.

[45] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, et al., Unsupervised cross-lingual representation learning at scale, arXiv preprint arXiv:1911.02116 (2020).

[46] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, arXiv preprint arXiv:2005.14165 (2020).