Improved identification of check-worthiness in social media data through multimodal analyses

Raphael Antonius Frick*, Martin Steinebach

Fraunhofer Institute for Secure Information Technology SIT | ATHENE — National Research Center for Applied Cybersecurity, Rheinstrasse 75, Darmstadt, 64295, Germany,
url=https://www.sit.fraunhofer.de/

Abstract

Combatting the spread of non-intentional and intentional false information on social media is challenging due to the vast amount of data that is shared each day. In order to still be able to retrieve credible information, assessing the check-worthiness of social media content can help to identify content that requires manual review. In this paper, we present a novel approach for detecting the check-worthiness in tweets. By incorporating the analysis of image content that is frequently shared along with social media posts, the proposed method, which consists of an analysis of the content, caption, and text obtained from optical character recognition, can outperform the current state-of-the-art recognition techniques with an F1 score of 0.7658 on the CheckThat! Lab 2023 benchmark dataset. Further experiments show, that by leveraging from multimodal information where applicable, the detection rate can be further improved.

Keywords

Check-Worthiness Estimation, Multimodality, Social Media, LMM

1. Introduction

Social media platforms have become a popular source of information for people around the world. However, the credibility of the information shared on these platforms is often questionable as most of the content shared originate from non-verified sources. A survey conducted by the Pew Research Center in September 2023[1] found that most participants either sometimes or often get their news from social media, making them a target for intentionally spread misinformation. The same is true for journalists, as user-generated content shared on social media is often the only source available. The dissemination of misinformation on social media can have serious consequences, such as the spread of false health information, political propaganda, and conspiracy theories [1]. On social media platforms such as X (Twitter), Facebook and TikTok, information can be shared in the form of text, images, videos and, in some cases, audio. Though originally focused on text, false information is now often conveyed also in other types of multimedia. For example, images are used to give the written post additional context, or the text is inserted into an image to avoid any form of blacklisting. With the rise of artificial intelligence, it has become increasingly accessible to the public to create new images or artificially forge them

[1] https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/

to depict a desired scene [2]. Existing methods for assessing credibility, which are often limited to text analysis, may therefore be inadequate for detecting multimodal misinformation [3]. In addition, data is constantly created and shared on the internet, making it impossible to manually review all of them. This has led to new challenges for estimating the credibility of information shared on social media.

While language models can be trained to identify fake news, their knowledge is often bound to the knowledge found within the dataset used for training. Therefore, they are often unable to recognize incorrect information that represents events that occurred after the model was trained [4]. Hence, the information still needs to be manually verified by a human expert. In order to drastically reduce the amount of data to be reviewed or to flag content that may contain incorrect information, the assessment of the check-worthiness of social media posts can be used. It serves either as a filter that is applied after data collection to prioritize the review of certain posts or in the analysis step to provide the classification result to the user. Hereby, check-worthy content is defined as content that contain factual claims that might be harmful and in interest of the general public [5]. Opinions and other types of subjectively written content are considered as non-check-worthy.

In this paper, we present a novel check-worthiness classification method that takes multi-modality into account to improve classification accuracy in cases where images are posted alongside the main text message. The method consists of three feature extraction networks that are used to extract information from the text and images, including the image description and the text embedded in the images, which are then fused to obtain a prediction. During evaluation on the CheckThat! Lab 2023 challenge dataset, the model has shown to improve upon the current state-of-the-art solution.

The main contributions can be summarized as follows:

- Proposal of a novel ensemble classification scheme for estimating the check-worthiness in tweets taking multimodality into account.
- It is the first model to combine textual features obtained from the tweets body with images descriptions retrieved by a large multimodal model and texts derived from a multilingual OCR analysis.
- The model is able to surpass the current state-of-the-art methods by a large margin and by providing an extensive ablation study, the effectiveness of the concept for improving the retrieval of credible information is showcased.

The remainder of the paper is structured as follows: Section 2 presents methods that have been proposed in the past to detect check-worthiness in tweets. Section 3 explains the proposed method and its components. The results on the evaluation dataset are discussed in section 4. The paper ends with a conclusion and an outlook for future work in section 5.

2. Related work

To detect multimodal disinformation on social media, several methods have been published in the past. The works of Wang et al. [3] and Singhal et al. [6] propose to analyze the text derived from a social media posting using a text transformer and the content of images posted with it

using a vision model. By this, visual and textual features are derived that are then fused in order to provide a classification decision. Their experiments reveal that taking additionally advantage of the analysis of media leads to improvements of accuracies up to 20%.

The detection of the check-worthiness in tweets has been part of the shared tasks of the CheckThat! Lab [7, 8, 9] at the CLEF conference for several years now. While the task originally only considered text as a modality across multiple languages, the task was extended in 2023 to include multimodal data by providing also the image data embedded in the tweets alongside the text for analysis. In each iteration, a dedicated labeled public dataset was published together with the challenge. Tests were conducted on a private dataset, which was released after the conclusion of the competition.

In 2021, the winning solution proposed by Martinez et al. [10] was based on a fine-tuned BertTweet language model [11]. Using grid search, they determined the parameters that provided the best performance on the development set. On the test dataset, the authors achieved a mean average precision (MAP) score of 0.224.

Savchev et al. [12] won the competition in 2022. They combined a fine-tuned RoBERTa [13] model with data augmentation and data preprocessing to slightly improve its performance. For each tweet, the links were substituted by a generic "@link" token. Further, the authors took advantage of back translation by translating the English tweets into French and then back into English. By this, they were able to achieve an F1-score of 0.698 surpassing the results of the follow-up [14] by 0.031 points obtained by an ensemble classification scheme consisting of ten weak classifiers.

In the first iteration of the task, which also took multimodality into account, Frick et al. [15] took first place in the competition. By taking the tweet text and the text embedded within the images into account during analysis, the proposed model was able to achieve an F1-score of 0.7297. While the model was able to improve its classification accuracy by conducting an OCR analysis on the embedded images, two major shortcomings were identified. Firstly, not every image featured visible text. Thus, in some cases no additional text could be extracted. Secondly, the texts were not always written in English, but also in other languages such as Chinese. However, the language model based on BERT offers no support for multilingualism, which may result in several misclassifications. This work aims to improve the problems identified.

3. Proposed approach

This paper presents a novel method that aims to combine various features from a tweet's embedded image content and its text to not only assess its check-worthiness but also improve classification accuracy. In particular, the proposed approach takes three features into account: the body text of the tweet, the text extracted from the embedded image as well as its description. The following provides a detailed explanation of the implementation of the individual components (Figure 1) of the classification scheme.

3.1. Dataset

For training and evaluating the proposed model, the dataset provided alongside the CheckThat! Lab 2023 Subtask 1A was used. The dataset consists of social media posts that were collected

Figure 1: Visualization of the individual processing steps within the proposed classification system
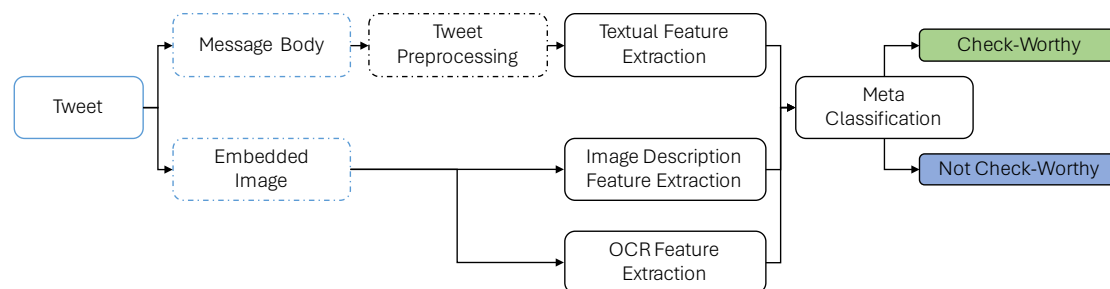
from Twitter with the help of its official API. Within the dataset, each entry contains the tweets body text and the image that was embedded into the tweet. The dataset is divided into four splits: a train split, a dev split, a dev-test split, and a test split. While labels for the train set, dev set, and dev-test set were provided upon release, the gold labels for the test split were not provided until after the competition concluded. The label distributions of each individual data set split as displayed in Table 1 suggest that the dataset suffers from class imbalance. Within each split, there were almost twice as many tweets not worthy of verification as tweets worthy of verification.

Table 1
Class distribution of the CheckThat! Lab 2023 task 1B English dataset

	Total	Yes	No
Train	2,356 / 100.00 %	820 / 34.80%	1,536 / 65.20%
Dev	271 / 100.00 %	87 / 32.10%	184 / 67.90%
Dev Test	548 / 100.00 %	174 / 31.75%	374 / 68.25%
Test	736 / 100.00 %	277 / 37.64%	459 / 62.36%
Sum	3,911 / 100.00 %	1,358 / 34.72%	2,553 / 65.28%

3.2. Textual feature extraction

For obtaining the textual features from given tweets, a fine-tuned RoBERTa model was used as a feature extractor. Prior to training and using the model, the tweets texts were preprocessed first. Here, URLs and user mentions were converted to generic tokens (HTTPURL, @USER), while emojis were converted into their respective descriptive tokens using the Python package pysentimiento [16]. Thus, the text 🚨*Coronavirus: China shuts down stock market till Feb 3 https://t.co/GvzFnhx9S8 https://t.co/M4AFZG1jbX* gets converted into *emoji police car light emoji Coronavirus: China shuts down stock market till Feb 3 HTTPURL HTTPURL* after preprocessing.

The processed data is then used to fine-tune a RoBERTa model (RoBERTa-base). For training, the train set of the 2023s CheckThat! Lab dataset was used, and its development split was used as a validation set for hyperparameter optimization. Training was carried out using Adam [17] as optimizer with an initial learning rate of 0.0004 was used. The batch size was set to 64 and while training was set to run for 10 epochs, the model converged after its third epoch of training

Figure 2: Description provided by ShareCaptioner for this image: *The image captures a breathtaking view of the Yellow Crane Tower, a renowned landmark in Wuhan, China. The tower, a multi-tiered structure, stands majestically against the backdrop of a vibrant sunset. Its red roof and gold accents gleam under the setting sun, reflecting the rich cultural heritage of the region. The tower is nestled amidst lush greenery, with trees encircling it, adding a touch of serenity to the scene. The perspective from which the photo is taken allows the tower to dominate the frame, its grandeur accentuated by the soft hues of the sunset.*

with a categorical-cross-entropy validation loss of 0.3830. Thus, early stopping was used as a mechanism to prevent overfitting on the train set.

## 3.3. Image description feature extraction

Image captioning models such as CLIP [18] can be used to generate a caption for a given image. With the rise of large language models such as ChatGPT and Llama [19], large language models are combined with image captioning models [20] to support additional tasks, such as reasoning, object detection and many more. In this work, we took advantage of ShareCaptioner [21] a state-of-the-art large multimodal model surpassing the performance of previous methods. Using the prompt *Analyze the image in a comprehensive and detailed manner and tell if the image is a photo, a text-image, a diagram, or an illustration* a detailed description from the given images was obtained. As it can be seen within Figure 2, the model can not only provide a description for the image, but is also able to correctly identify the object displayed in the photo.

Each of the image captions are then used to train another RoBERTa-based model. The model was trained in a similar way to the model for extracting text features. However, a learning rate of 0.0005 was chosen this time based on various experiments conducted on the development set. The final validation loss of the model was 0.4866, which was higher than the loss obtained when training the textual feature extractor. However, this is to be expected, as it is assumed that the written text posts mainly contain factual claims and images can often only be used to contextualize them.

Figure 3: Text identified by Tesseract: *WHEN IT'S FINALLY 2020 BUT YOU REALISE BOTH 1820 AND 1920 HAD MASSIVE PLAGUE OUTBREAKS*

Figure 4: Visualization of the fully connected neural network used to fuse the features obtained from multiple modalities
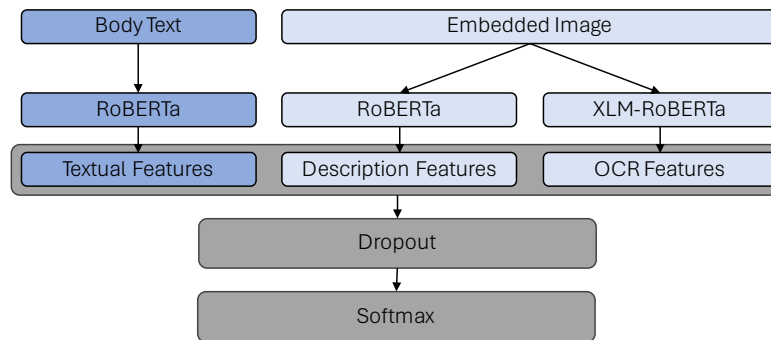
3.4. Embedded text features

To retrieve the text embedded within the images, an optical character recognition (OCR) model was used. Here, the models provided by tesseract V5 [22] were used, as they support multiple languages. For this purpose, the models supporting English, Chinese (simplified), Korean, Japanese and Korean were used after examining the languages on the pictures of the train set. An example OCR-analysis is displayed in Figure 3.

To support multilingualism by the feature extraction network, this time an XLM-RoBERTa-based model [23] was used in favor of the RoBERTa model as it was pre-trained on large amounts of texts containing multiple languages. The training concluded with a validation loss of 0.5734 on the validation split of the dataset.

Table 2
Results obtained on the CheckThat! Lab 2023 test dataset

	Accuracy	Precision	Recall	F1 Score
Text (no preprocessing)	0.7880	0.8457	0.5343	0.6549
Text (with preprocessing)	0.8043	**0.8519**	0.5812	0.6910
Image Description	0.6780	0.5637	0.6390	0.5990
Embedded Text	0.5122	0.4208	**0.7870**	0.5484
Text + Image Description	0.8234	0.8296	0.6679	0.7400
Text + Embedded Text	0.8234	0.8025	0.7040	0.7500
Image Description + Embedded Text	0.7120	0.6165	0.6209	0.6187
Text + Image Description + Embedded Text	**0.8288**	0.7893	0.7437	**0.7658**
Previous State-of-the-Art Solution [15]	0.8057	0.7659	0.6968	0.7297

3.5. Feature merging

After the individual features have been obtained by the respective feature extraction models, a meta classifier (Figure 4) was trained to fuse them together. For this, the classification head of the fine-tuned RoBERTa and XLM-RoBERTa models were removed first. In this way, feature embeddings can be extracted from each of the models. A fully connected neural network takes these embeddings as input and concatenates them. The concatenated values are then passed to a dropout layer with a dropout probability of 0.25 and then to a softmax layer that is responsible for outputting the prediction. For training, the dev-test split of the challenge dataset was used and the development split as the validation set. Moreover, the development set was also used to obtain an optimized threshold value by measuring the true positive and false positive rate at various thresholds.

4. Evaluation

4.1. CheckThat! Lab 2023 dataset

For evaluation, the test split of the CheckThat! Lab 2023 challenge dataset [2] was used. To demonstrate the effectiveness of the concept, each component (text, image description, embedded text) and the effects of tweet preprocessing were evaluated separately and then in combination. For better comparability with the meta classifier, the classification heads of all the models were retrained as described in Section 3.5. The obtained results are displayed in Table 2.

As can be seen, the results of the evaluation of the classifier, which only considers textual features, differ considerably depending on the application of the preprocessing. The scores across all the metrics were improved by preprocessing the tweets as proposed. However, it is also reveals that the model is not capable of surpassing the current state-of-the-art with textual information alone. The same also applies to the classifiers taking either solely the image descriptions or the embedded texts extracted from an OCR into account. This behavior was largely to be expected due to the validation losses determined during training. In addition, it can be assumed that the text of the tweets plays a greater role in the identification of the validation

---

[2]https://gitlab.com/checkthat_lab/clef2023-checkthat-lab/-/tree/main/task1

Table 3
Results obtained on the CheckThat! Lab 2021 test dataset

	MAP	MRR	RP	P@1	P@3	P@5	P@10	P@20	P@30
Text (with preprocessing)	0.2132	**1.000**	0.1579	**1.000**	**0.667**	**0.400**	0.300	0.150	0.140
Text + Image Description + Embedded Text	0.2153	**1.000**	2.105	**1.000**	0.333	**0.400**	**0.400**	**0.200**	**0.160**
Previous State-of-the-Art Solution[10]	**0.224**	**1.000**	**0.211**	**1.000**	**0.667**	**0.400**	0.300	**0.200**	**0.160**

worthiness than the associated images.

When combining two of the three modalities, a similar behavior can be identified. The classifier leveraging from the image description and the embedded text in the image performs worst on the test dataset, whereas any classifier that combines textual and features derived from the images is able to significantly enhance the classification results. Interestingly, combining the features provided by the text embedded in the images with the text of the tweet led to a better result than the combination of text and image description. One explanation for this could be that the text content within the tweet has some similarities with the text found using the OCR analysis; for example, similar terms such as Covid-19 are used in both representations. This means that both features can support each other better than, for example, just the visual description of the image. Compared to the solution proposed in [15], better performance was also achieved by introducing a multilingual model for analyzing the texts embedded in the image.

Combining all three features, however, resulted in the best results. With an accuracy of 0.8288 and an F1 score of 0.7658, it was able to improve upon the current state-of-the-art solution by a large margin.

4.2. CheckThat! Lab 2021 dataset

To assess, whether the model is also capable of identifying the check-worthiness of tweets it was not trained on, another evaluation was carried out on the test set of the CheckThat! Lab 2021 dataset [3]. While the dataset does not come with multimodal data, the dataset provides the output of the Twitter API. This is not the case for the 2022 dataset and was therefore not utilized. By this, it was possible to get access to the images that were posted alongside some of the posts. An image could be retrieved for 116 of the 350 posts. The model that was trained on the CheckThat! Lab 2023 dataset was applied without re-training on the visually enriched test set from 2021. The results are displayed in Table 3. As the results imply, the proposed model was not able to outperform the winning solution of the competition. One reason for this is, that by specifically training the model on the benchmark train dataset the model might be able to capture the given labels of the test set better. However, the results are still promising and indicate that by leveraging from multimodality where possible, the classification of check-worthiness can be slightly improved.

---

[3]https://gitlab.com/checkthat_lab/clef2021-checkthat-lab/-/tree/master/task1

5. Conclusion and future work

In this paper, we presented a novel system for detecting check-worthiness in social media data by leveraging from multimodality. The system that consists of the analysis of three features, textual features, image descriptions and features obtained from an OCR analysis, has shown promising results on the benchmark dataset of the CheckThat! Lab 2023. The model was able to improve upon the current state-of-the-art by 3.61%. Further, by leveraging from a multilingual model for the OCR analysis, the performance was improved by 1.03% upon the state-of-the-art. The cross-database evaluation also showed that the use of multimodal data can lead to an increase in performance wherever possible. Future work could revolve around improving the image captions so that they are more precise but shorter, as well as the overall performance of the OCR analysis. Moreover, additional means for providing explainability are to be investigated in the future.

References

[1] D. Allington, B. Duffy, S. Wessely, N. Dhavan, J. Rubin, Health-protective behaviour, social media usage and conspiracy belief during the covid-19 public health emergency, Psychological medicine 51 (2021) 1763–1769.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, 2022. `arXiv:2112.10752`.

[3] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, J. Gao, Eann: Event adversarial neural networks for multi-modal fake news detection, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 849–857. URL: https://doi.org/10.1145/3219819.3219903. doi:`10.1145/3219819.3219903`.

[4] H. Alkaissi, S. Mcfarlane, Artificial hallucinations in chatgpt: Implications in scientific writing, Cureus 15 (2023). doi:`10.7759/cureus.35179`.

[5] P. Nakov, G. D. S. Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, W. Mansour, B. Hamdan, Z. S. Ali, N. Babulkov, A. Nikolov, G. K. Shahi, J. M. Struß, T. Mandl, M. Kutlu, Y. S. Kartal, Overview of the clef–2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news, 2021. `arXiv:2109.12987`.

[6] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, S. Satoh, Spotfake: A multi-modal framework for fake news detection, in: 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), 2019, pp. 39–47. doi:`10.1109/BigMM.2019.00-44`.

[7] 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.

[8] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, C. Li, S. Shaar, et al., Overview of the clef-2022 checkthat! lab task 1 on

identifying relevant claims in tweets, in: 2022 Conference and Labs of the Evaluation Forum, CLEF 2022, CEUR Workshop Proceedings (CEUR-WS. org), 2022, pp. 368–392.

[9] 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).

[10] J. R. Martinez-Rico, J. Martinez-Romo, L. Araujo, Nlp&ir@ uned at checkthat! 2021: Check-worthiness estimation and fake news detection using transformer models., in: CLEF (Working Notes), 2021, pp. 545–557.

[11] D. Q. Nguyen, T. Vu, A. T. Nguyen, Bertweet: A pre-trained language model for english tweets, arXiv preprint arXiv:2005.10200 (2020).

[12] A. Savchev, Ai rational at checkthat! 2022: using transformer models for tweet classification, Working Notes of CLEF (2022).

[13] 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, arXiv preprint arXiv:1907.11692 (2019).

[14] N. Buliga, M. Raschip, Zorros at checkthat! 2022: Ensemble model for identifying relevant claims in tweets (2022).

[15] 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, Working Notes of CLEF (2023).

[16] J. M. Pérez, M. Rajngewerc, J. C. Giudici, D. A. Furman, F. Luque, L. A. Alemany, M. V. Martínez, pysentimiento: A python toolkit for opinion mining and social nlp tasks, 2023. `arXiv:2106.09462`.

[17] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2017. `arXiv:1412.6980`.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, 2021. `arXiv:2103.00020`.

[19] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. `arXiv:2302.13971`.

[20] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, S. Hoi, Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023, arXiv preprint arXiv:2305.06500 (????).

[21] L. Chen, J. Li, X. Dong, P. Zhang, C. He, J. Wang, F. Zhao, D. Lin, Sharegpt4v: Improving large multi-modal models with better captions, 2023. `arXiv:2311.12793`.

[22] R. Smith, An overview of the tesseract ocr engine, in: Ninth international conference on document analysis and recognition (ICDAR 2007), volume 2, IEEE, 2007, pp. 629–633.

[23] 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, 2019. `arXiv:1911.02116`.