PoliTo at MULTI-Fake-DetectiVE: Improving FND-CLIP for Multimodal Italian Fake News Detection

Lorenzo D'Amico[1], Davide Napolitano[1], Lorenzo Vaiani[1] and Luca Cagliero[1]

[1]Politecnico di Torino, Turin, Italy

Abstract

The MULTI-Fake-DetectiVE challenge addresses the automatic detection of Italian fake news in a multimodal setting, where both textual and visual components contribute as potential sources of fake content. This paper describes the PoliTO approach to the tasks of fake news detection and analysis of the modality contributions. Our solution turns out to be the best performer on both tasks. It leverages the established FND-CLIP multimodal architecture and proposes ad hoc extensions including sentiment-based text encoding, image transformation in the frequency domain, and data augmentation via back-translation. Thanks to its effectiveness in combining visual and textual content, our solution contributes to fighting the spread of disinformation in the Italian news flow.

Keywords

Fake News Detection, Multimodal Learning

1. Introduction

In recent years, the proliferation of fake news and disinformation in online platforms has become a significant challenge, impacting public discourse, political landscapes, and social dynamics [1]. This phenomenon has been particularly evident during real-world events, where misinformation spreads rapidly, often leading to harmful consequences.

With the proliferation of multimodal data sources, detecting fake news content has become more and more challenging as fake content can be hidden in either visual and textual news elements. Studying the interplay between these modalities is crucial for understanding how misinformation is crafted and disseminated and how it can be effectively detected.

The MULTI-Fake-DetectiVE challenge [2] proposed at EVALITA 2023 [3] focuses on overcoming the limitations of existing approaches in coping with multimodal Italian news content. It addresses the automatic detection of Italian fake news in a multimodal setting, where both textual and visual components potentially contribute as sources of fake content. The challenge has the twofold aim of accurately discriminating between real and fake news content and investigating the influence of visual and textual components on each other's interpretation.

In this work, we present the PoliTO approach to both MULTI-Fake-DetectiVE tasks. Our solution outperforms all the baselines and competitors in both tasks, strengthening the state-of-the-art multimodal fake news detectors in a challenging and previously underexplored scenario. Furthermore, it releases new open-source resources to the community to fight disinformation in the Italian flow of news articles.

We propose FND-CLIP-IT, an extension of the state-of-the-art FND-CLIP multimodal architecture. We explore the integration of a sentiment-based text encoder, a data augmentation stage based on back-translation, an image transformation module based on Discrete Fourier Transform, as well as different approaches to combine and weigh the input embeddings.

The remainder of this paper is organized as follows. In Section 2 we review the literature on fake news detection, considering both text-only and multimodal approaches. Section 3 briefly describes the dataset, task, and metrics used in the challenge. In Section 4 we describe the methodology, primarily focusing on the proposed FND-CLIP extensions. Section 5 presents the experimental setup and the obtained results. Finally, Section 6 draws the conclusions and discusses the main limitations and future directions.

2. Related Work

NLP-based approaches. Early approaches focused on linguistic features, such as lexical and syntactic patterns, to distinguish between real and fake news. However, with the advancement of deep learning researchers have increasingly turned to more sophisticated methods such as recurrent neural networks [4], convolutional neural networks [5], and transformer models [6] to capture semantic and contextual information for improved

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7–8, Parma, IT

lor.damico@studenti.polito.it (L. D'Amico); davide.napolitano@polito.it (D. Napolitano); lorenzo.vaiani@polito.it (L. Vaiani); luca.cagliero@polito.it (L. Cagliero)

0000-0001-9077-4103 (D. Napolitano); 0000-0002-3605-1577 (L. Vaiani); 0000-0002-7185-5247 (L. Cagliero)

detection accuracy. In this work, we mainly rely on state-of-the-art transformers pretrained on Italian textual data to effectively extract information from the news textual component.

Multimodal Approaches. Incorporating multimodal information such as text and images has shown to be promising to improve the accuracy of fake news detection systems [7]. Recently, the adoption of multimodal architectures and transformers has shown to be particularly effective in capturing the semantic relationships among different modalities for fake news detection, e.g., CB-Fake [8], CAFE [9] and TTEC [10].

FND-CLIP, proposed by [11], is among the most recently proposed multimodal architectures for fake news detection. It relies on the established CLIP model [12] to measure the cross-modal similarity and guide the mapping and fusion of the input features. The architecture develops along three main streams: a textual one, which extracts information using BERT and CLIP, a visual one which extracts features from the images using ResNet and CLIP, and a multimodal one which combines the features extracted using CLIP from both modalities. FND-CLIP suffers from the following limitations:

- The natural language encoder neglects the polarity of the input text, which is known to be relevant to fake news detection [13].
- Fake news examples are likely to be undersampled in real training data. Hence, the classification model may suffer from class imbalance effects.
- Multimodal fake news often contains tampered visual content. Tampered images are more likely to be detected in the frequency domain space. However, FND-CLIP does not consider any frequency-based image descriptor [14].

Our research endeavors to address the aforesaid limitations by proposing FND-CLIP-IT, i.e., an improved version FND-CLIP suited to multimodal Italian fake news detection.

3. Task and Dataset Description

3.1. Tasks Description

Task 1: Multimodal Fake News Detection. The problem can be formulated as a multi-class classification task, where the input content $c = \langle t, v \rangle$, consisting of a textual component t and a visual component v, can be classified as follows: *Certainly Fake* (CF), when the news is very likely to be fake, regardless of the context in which it is presented; *Probably Fake* (PF), when the news is likely to be fake but may contain some real information or exhibit a certain level of credibility; *Probably Real* (PR), when the news is highly credible but retains some degree of uncertainty regarding the information provided; *Certainly Real* (CR), when the news is most certain to be real and indisputable, regardless of the context. It is worth noticing that these labels pertain to the overall informational content and should not be assigned based solely on the individual components.

Task 2: Analysis of Cross-Modal Relations in Fake and Real News. The purpose is to examine the relationship between the textual and visual modalities within the context of fake and real news. The primary objective is to gain insights into how images and texts in fake and real news can potentially lead to misleading interpretations of the content, both within each modality and as a whole. The task can be formulated as a three-class classification problem. Given a multimodal piece of content c, the goal is to automatically assign one of the following categories to c: *Misleading* (M), when either the image or the text is misleading in terms of interpreting the information conveyed by the other modality or the content as a whole; *Not Misleading* (NM), when both image and text are related to each other, providing support to the overall information presented, and are not intended to mislead; *Unrelated* (U), when the image and the text are not related to each other.

Evaluation Metrics. For both tasks, the evaluation metrics are accuracy, average per-class precision and recall, and macro- and weighted- F1 score. Weighted-F1 score has been selected as the reference metric for ranking participants.

3.2. Dataset Description

Both task-specific datasets consist of a collection of Twitter posts and newspaper articles describing one or more real events. For Task 1 the training set contains 908 distinct labeled samples[1]. The labels in the training data are distributed as follows: CF 16.4%, PF 22.0%, PR 44.4%, CR 17.2%. Around 80.0% of the samples are tweets, whereas the remaining ones are news articles. The test set consists of 193 samples following roughly the same per-class and per-type distributions as in the training data. For Task 2, the training set contains 1309 distinct samples and the per-class distribution is M 26.9%, U 40.6%, NM 31.5%. 66.0% of the samples are tweets, whereas the remaining ones are news articles. The test set contains 219 samples. Compared to the training data, the per-type sample distribution is slightly more biased towards tweets (75.0%) and Non-Misleading content (45.2%).

[1] available at the time of writing, June 2023

4. Methodology

Here we present FND-CLIP-IT, an improved version of FND-CLIP suited to the MULTI-Fake-DetectiVE challenge. Our solution is rooted in the original FND-CLIP model [11] and a set of unimodal language and visual encoders described below.

Unimodal language baselines. We utilize the following models tailored to the Italian language: BERT-IT[2], GilBERTo[3], BART-IT[4] [15].

Since the input text can be longer than the maximum model size, we adopt a hierarchical approach: the text is divided into chunks of fixed length, where each one is fed to the transformer encoder, and then the final representation is obtained by averaging all the [CLS] tokens.

Unimodal visual baselines. We exploit two established models, i.e., ViT [16] and ResNet-152 [17]. Since more pictures can be associated with the same sample, at inference time we separately evaluate all the images and the final prediction is the average of all obtained output logits.

Multimodal baselines. To leverage visual and textual content at the same time, we rely on (i) the standard FND-CLIP [11] architecture, adapted to handle Italian text rather than English, (ii) CLIP [12], and (iii) a late fusion approach combining BERT-IT and ResNet-152.

4.1. FND-CLIP-IT

FND-CLIP-IT extends the state-of-the-art FND-CLIP architecture to address the current limitations of fake news detection approaches. By incorporating the proposed extensions, the overall efficacy and robustness of FND-CLIP-IT shows significant improvements compared to the baseline versions. A detailed description of the proposed extensions, hereafter denoted by *A*, *B*, *C*, *D*, and *E* for the sake of brevity, is given below.

A. **Sentiment-based textual representation**: To consider the polarity of the input text for fake news detection [13], we enrich the textual representation by adding a sentiment-based encoding to the existing text encoders. Specifically, we use the Italian-BERT model finetuned on a sentiment analysis task[5]. We also consider the following variants of sentiment-based textual representation:
   - $\mathbf{A}_1$. A concatenation of the sentiment-based embedding to the initial original representation, on top of which we apply the textual projection head.
   - $\mathbf{A}_2$. A separate stream of information with a dedicated projection head.

B. **DFT-based additional stream**: we convert the image from the spatial domain to the frequency domain by applying Discrete Fourier Transform (DFT). The purpose is to detect tampered images, which likely occur in multimodal fake news [14]. We encode both real and imaginary parts using a dedicated VGG19 [18]. The obtained representations are then concatenated to generate a parallel stream of information that will be then combined with the others before applying the final FND-CLIP classifier.

C. **Embedding concatenation**: instead of summing the embedding of each stream we concatenate them. Concatenation has already been proven to be an effective way of combining multimodal information [19]. The rationale behind it is that by keeping more fine-grained pieces of information the classification head, adapted to handle the new encoding, can capture the most discriminating source features in a more effective way.

D. **Class rebalancing through data augmentation**: since the dataset is quite imbalanced across the classes, we re-balance the data distribution by penalizing the most frequent class. In particular, we generate new samples of the minority classes by applying a textual augmentation based on back-translation [20], which already proved to be beneficial in both multimodal [21] and fake news detection [22] tasks. The auxiliary language adopted is English, and the translation models used are provided by Helsinki-NLP[6].

E. **Additional Squeeze and Excitation Layers**: Similar to FND-CLIP, we employ a squeeze-and-excitation operation [23] to weigh the input embedding streams. The purpose of a squeeze-and-excitation block is to adaptively recalibrate channel-wise feature responses by explicitly modeling interdependencies between channels. Unlike [11], where they weigh differently the textual and visual streams, we also adopt squeeze-and-excitation within each modality to weigh the relevance of each encoder. The key idea is to give more importance to discriminating modality-specific embeddings.

Beyond considering each FND-CLIP extension separately, we also build both models that combine the pro-

[2] https://huggingface.co/dbmdz/bert-base-italian-xxl-cased

[3] https://huggingface.co/idb-ita/gilberto-uncased-from-camembert

[4] https://huggingface.co/morenolq/bart-it

[5] https://huggingface.co/neuraly/bert-base-italian-cased-sentiment

[6] https://huggingface.co/Helsinki-NLP

Model	Accuracy	Precision	Recall	F1-macro	F1-weighted
BERT-IT	0.554	0.558	0.470	0.491	0.531
GilBERTo	0.522	0.510	0.462	0.473	0.514
BART-IT	0.495	0.447	0.428	0.429	0.472
ResNet-152	0.451	0.406	0.373	0.380	0.436
ViT	0.402	0.336	0.322	0.324	0.388
BERT-IT+ResNet-152	0.516	0.516	0.423	0.433	0.479
CLIP-IT	0.538	0.523	0.478	0.484	0.520
FND-CLIP-IT	0.560	0.551	0.492	0.503	0.537
FND-CLIP-IT$_{A_1}$	0.565	0.558	0.525	0.530	0.552
FND-CLIP-IT$_{A_2}$	0.560	0.541	0.521	0.526	0.550
FND-CLIP-IT$_{B}$	0.587	0.587	0.506	0.525	0.565
FND-CLIP-IT$_{C}$	0.576	0.578	0.521	0.539	0.568
FND-CLIP-IT$_{D}$	0.565	0.540	0.517	0.524	0.555
FND-CLIP-IT$_{E}$	0.576	0.563	0.558	0.560	0.574
FND-CLIP-IT$^{*}_{A_2,B}$	0.587	0.578	0.548	0.560	0.581
FND-CLIP-IT$^{*}_{A_1,C,D}$	<u>0.609</u>	0.595	<u>0.593</u>	<u>0.593</u>	<u>0.606</u>
FND-CLIP-IT$^{*}_{A_1,D,E}$	0.598	<u>0.603</u>	0.569	0.582	0.596
FND-CLIP-IT$_{A_1,B,C,D,E}$	0.576	0.568	0.529	0.542	0.572
ENSEMBLE	**0.658**	**0.661**	**0.621**	**0.637**	**0.653**

Table 1
Results obtained on the validation set from our baselines, FND-CLIP-IT variants and ensemble models. The best results are highlighted in bold, while the best results of a single model are underlined. Models marked with * are involved in the reported ensemble.

posed extensions *A-E* in different ways and ensemble methods that combine best-performing individual models. To this end, we use a weighted average of individual logits for each class.

5. Experimental Results

5.1. Setup

The models were fine-tuned for a maximum of 80 epochs, using a batch size of 16, a learning rate of 1e-3, an AdamW optimizer with a weight decay of 0.001 and a linear scheduler. All baseline models were trained using a cross-entropy-loss, while all FND-CLIP-IT variants were trained with both cross-entropy and focal losses.

5.2. Results

Table 1 presents the results of the baselines (upper part) and our proposed solutions (lower part), obtained on the Task 1 validation set. Significantly, the outcomes reveal an intriguing pattern wherein text-only models exhibit superior performance when compared to image-only models, underscoring the paramount importance of textual information within the context of the task. Notably, the multimodal CLIP baseline demonstrates results comparable to the text-only model. At the same time, FND-CLIP-IT architecture attains performance marginally better than the BERT-IT model. Furthermore, our diverse extensions of the FND-CLIP-IT framework, when applied individually, yield notable improvements over the original implementation. In addition, select combinations of these variants produce even more promising outcomes. Although both focal loss and cross-entropy were evaluated, we chose to report only the results obtained with focal loss, due to their overall superior performance compared to cross-entropy. It is worth noting, however, that the combination of all variants does not surpass the performance of specific combinations, indicating a potential susceptibility to overfitting. Furthermore, an intriguing observation emerges with the implementation of an ensemble model that leverages the best-performing combinations. This ensemble model outperforms the individual models, further accentuating the benefits of employing ensemble techniques to enhance overall performance.

5.3. Competition

We employed our ensemble method to evaluate the performance of our FND-CLIP-IT variants on the test samples. The test results are presented in Table 2. The upper part of Table 2 shows the outcomes obtained for Task 1. Although these results are worse than the performance achieved on our validation set, they surpass all other baselines and competitors.

Furthermore, we fine-tuned the same ensemble of models for Task 2 by replacing the classification head last layer. The bottom of Table 2 reports the achieved results for Task 2. Remarkably, our approach outperforms both the baseline and the competitors.

	Team Run	F1-weighted
Task 1	PoliTo-P1	**0.512**
	extremITA-camoscio_lora	0.507
	AIMH-MYPRIMARYRUN	0.488
	Baseline-SVM-TEXT	0.479
	HIJLI-JU-CLEF-Multi	0.393
Task 2	PoliTo-P1	**0.517**
	Baseline-MLP-TEXT	0.506
	AIMH-MYPRIMARYRUN	0.421

Table 2
Official MULTI-Fake-DetectiVE results. For the official baselines, we report only the result of the best-performing approach.

By leveraging our best ensemble method, we have demonstrated the robustness and versatility of our FND-CLIP-IT variants across both Task 1 and Task 2, surpassing existing approaches in terms of performance and effectiveness.

6. Conclusion and Future Directions

In this study, we introduced the FND-CLIP-IT architecture exploring several variants for fake news detection in a multimodal setting. Our findings demonstrate the effectiveness of these variants, with notable improvements observed over the original implementation. Furthermore, selected combinations of these variants and model ensembling yield even more promising results.

In the future, we plan to continue refining and optimizing the proposed variants to further enhance their performance. Additionally, investigating the interpretability of the FND-CLIP-IT architecture and understanding its decision-making process will be an interesting direction for future research.

Acknowledgments

This study was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PNRR M4C2, INVESTIMENTO 1.3 D.D. 1555 11/10/2022, PE00000013). This study was carried out within the MICS (Made in Italy – Circular and Sustainable) Extended Partnership and received funding from the European Union Next-GenerationEU (PNRR M4C2, INVESTIMENTO 1.3 D.D. 1551.11-10-2022, PE00000004). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them. The research leading to these results has been partly funded by the SmartData@PoliTO center for Big Data and Machine Learning technologies. Computational resources were provided by HPC@POLITO, a project of Academic Computing within the Department of Control and Computer Engineering at the Politecnico di Torino[7].

References

[1] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, Science (2018).

[2] A. Bondielli, P. Dell'Oglio, A. Lenci, F. Marcelloni, L. C. Passaro, M. Sabbatini, Multi-fake-detective at evalita 2023: Overview of the multimodal fake news detection and verification task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[3] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[4] C. Iwendi, S. Mohan, S. khan, E. Ibeke, A. Ahmadian, T. Ciano, Covid-19 fake news sentiment analysis, Computers and Electrical Engineering (2022).

[5] K. Ma, C. Tang, W. Zhang, B. Cui, K. Ji, Z. Chen, A. Abraham, Dc-cnn: Dual-channel convolutional neural networks with attention-pooling for fake news detection, Applied Intelligence (2023).

[6] H. Jwa, D. Oh, K. Park, J. M. Kang, H. Lim, exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert), Applied Sciences 9 (2019).

[7] I. Segura-Bedmar, S. Alonso-Bartolome, Multimodal fake news detection, Information 13 (2022).

[8] B. Palani, S. Elango, V. Viswanathan K, Cb-fake: A multimodal deep learning framework for automatic fake news detection using capsule neural network and bert, Multimedia Tools and Applications (2022).

[9] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, L. Shang, Cross-modal ambiguity learning for multimodal fake news detection, in: Proc. of the ACM Web Conference 2022, 2022, pp. 2897–2905.

[10] J. Hua, X. Cui, X. Li, K. Tang, P. Zhu, Multimodal fake news detection through data augmentation-based contrastive learning, Applied Soft Computing 136 (2023) 110125.

[11] Y. Zhou, Q. Ying, Z. Qian, S. Li, X. Zhang, Multimodal fake news detection via clip-guided learning, arXiv preprint arXiv:2205.14304 (2022).

[7] https://www.hpc.polito.it/

[12] 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.
[13] M. A. Alonso, D. Vilares, C. Gómez-Rodríguez, J. Vilares, Sentiment analysis for fake news detection, Electronics 10 (2021).
[14] J. Jing, H. Wu, J. Sun, X. Fang, H. Zhang, Multimodal fake news detection via progressive fusion networks, Information Processing & Management (2023).
[15] M. La Quatra, L. Cagliero, Bart-it: An efficient sequence-to-sequence model for italian text summarization, Future Internet 15 (2023).
[16] 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, in: 9th International Conference on Learning Representations, ICLR 2021, 2021.
[17] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. of the IEEE conference on computer vision and pattern recognition, 2016.
[18] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proc. of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, 2015.
[19] L. Vaiani, M. La Quatra, L. Cagliero, P. Garza, Leveraging multimodal content for podcast summarization, in: Proc. of the 37th ACM/SIGAPP Symposium on Applied Computing, 2022.
[20] S. Edunov, M. Ott, M. Auli, D. Grangier, Understanding back-translation at scale, in: Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing, ACL, Brussels, Belgium, 2018.
[21] L. Vaiani, L. Cagliero, P. Garza, PoliTo at SemEval-2023 task 1: Clip-based visual-word sense disambiguation based on back-translation, in: Proc. of the 17th International Workshop on Semantic Evaluation (SemEval-2023), ACL, Toronto, Canada, 2023.
[22] M. Amjad, G. Sidorov, A. Zhila, Data augmentation using machine translation for fake news detection in the urdu language, in: Proc. of the 12th language resources and evaluation conference, 2020, pp. 2537–2542.
[23] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proc. of the IEEE conference on computer vision and pattern recognition, 2018.