AIMH at MULTI-Fake-DetectIVE: System Report

Giovanni Puccetti[1], Andrea Esuli[1]

[1]ISTI • Area della Ricerca CNR, via G. Moruzzi 1, 56124 Pisa, Italy

Abstract

This report describes our contribution to the EVALITA 2023 shared task MULTI-Fake-DetectIVE which involves the classification of news including textual and visual components. To experiment on this task we focus on textual data augmentation, extending the Italian text and the Images available in the training set using machine translation models and image captioning ones. To train using different set of input features, we use different transformer encoders for each variant of text (Italian, English) and modality (Image). For Task 1, among the models we test, we find that using the Italian text together with its translation improves the model performance while the captions don´t provide any improvement. We test the same architecture also on Task 2 although in this case we achieve less satisfactory results.

Keywords

MULTI-Fake-DetectIVE, Fake News, Multimodality

1. Introduction

Misinformation, intentional or not, is an ubiquitous phenomenon in social media. Whether due to malicious intent or scarce reviews, the number of outlets producing incorrect information is growing over time [1]. While the only true mean to protect one self from misinformation is careful review of trustworthy sources, the development of sound quantitative approaches for fake news detection is a worthy endeavour.

In this context there are works providing benchmark datasets for the very task of fake news detection in Twitter [2], however this is generally tackled in a unimodal setting where textual information is the only one examined. In this context, the MULTI-Fake-DetectIVE task [3], part of the EVALITA 2023 campaign [4] proposes to add multimodality, by challenging participants to classify fake news using both textual and visual features.

The task consists in classifying tweets reporting news about the war in Ukraine with both textual and visual content according to whether the reported news is true or fake. The task is subdivided into two subtasks:

- the first subtask is about detecting fake news by assigning a label among *Certainly False, Probably False, Probably True, Certainly True*;
- the second subtask is focused on detecting the agreement between text and image by assigning a label among *Misleading, Non Misleading, Unrelated*, which respectively indicate if the content of text and image support different interpretations, the same interpretation or are unrelated.

To perform the task we focus on exploring the effectiveness of augmenting the dataset by adding variants of the input extrapolated from both the existing text in Italian as well as the images leveraging the knowledge available in pre-trained models.

The idea of exploiting knowledge implicitly encoded in large pretrained models is used in several contexts with different goals, ranging from Neural Databases [5] to synthetic text detection [6].

The rest of the report is structured as follows: section 2 reports relevant literature, section 3 covers details of the dataset we found while preparing the models, section 4 is the System Description, section 5 outlines the results we obtained, and finally in section 6 we draw the conclusions of this work.

2. State of the Art

Recently, multimodal classification is tackled with visual language models such as OSCAR [7], VinVL [8] or with separate text and image encoding networks [9]. Built upon the idea of creating a shared representation space between text and images, developed in CLIP [10], several image captioning models have also been developed such as CoCa [11], we also try experimenting with these architecture for data augmentation. We could also use Multimodal Large Language Models for this same goal, i.e. augmenting data, some of the best performing ones are BLIP-2 [12] and Llava [13] these are too computationally costly and we avoid using them. Instead, to perform data augmentation across languages we employ Italian to English Neural Machine Translation models [14, 15].

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

giovanni.puccetti@isti.cnr.it (G. Puccetti);
andrea.esuli@isti.cnr.it (A. Esuli)

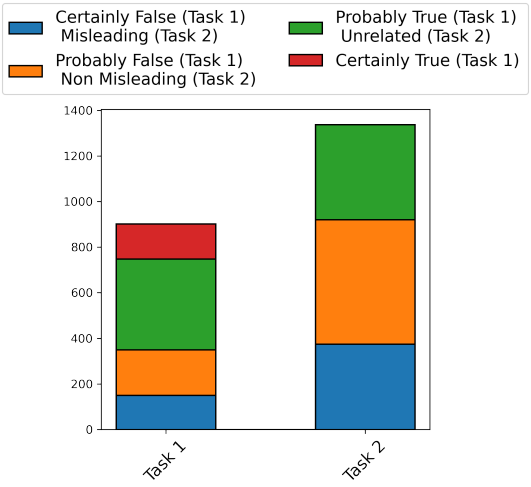

Figure 1: The labels distribution in the training dataset for Task 1 and 2.

3. Data

We perform an analysis of the dataset meant to understand if there are task specific preprocessing we have to apply to the data.

Figure 1 shows the distribution of labels in both tasks, we notice that both have heavily unbalanced distribution. The dataset of Task 1 Figure 1 shows how (likely) True news are the majority of samples, indeed, while ubiquitous in our everyday experience on the web, (likely) Fake news are still a minority of the total information shared.

Accordingly, for Task 2 Figure 1 shows that instances where Image and Text are heavily non aligned are also a minority.

While inspecting Task 1 training dataset, we observe non negligible data duplication, more specifically, there is 13.6% duplicated training samples, which we remove. On the contrary the dataset for Task 2 does not show any repetition.

4. Description of the System

In this Section we describe the methodology we developed to tackle the MULTI-Fake-DetectIVE task. We report the choices made and the steps that led us to them. In particular, we focus on data augmentation, for which we mainly adopt two systems working either on text or on images. Our architecture follows the one proposed by Gallo et al. [9].

We focus on data augmentation because the dataset is composed of Italian texts and since there aren't many models pre-trained specifically on this language we explore how well translating to English works. From here on, by *sample* we refer to a set of texts and images composing a single piece of news. Similarly, by *features* we indicate both texts or images.

To explore several data augmentation possibilities we build a unique pipeline that allows to add multiple pretrained models to process different input features schemes, based on different sets of texts and images.

Figure 2 outlines our architecture, using the same notation as in the figure, for each input sequence/image ($Feature_i$) in a sample, we use a pretrained model to embed it (N_i), then we add a linear layer ($Linear_i$) that maps all embeddings to the same dimension, finally we sum all such embeddings (entry-wise) to create a shared hidden state ($Hidden\ State$) and pass this vector through a linear layer ($Classification\ Head$) that maps it to a vector with length equal to the number of classes, 4 for task 1 and 3 for task 2. During training we optimize all parameters, including those of the pretrained models N_i.

4.1. Data Augmentation

The architecture we use allows us to seamlessly use as input any number of texts and images for each sample, in particular by adding extra features. We add features in two ways:

- We translate the textual documents to English using an open-source machine translation model [14, 15], in particular an Italian to English model[1];
- We caption the images using an image captioning model CoCa [11] fine tuned on the MSCOCO [16], we use an open source version [2].

Adding these extra inputs gives us the possibility to compose samples with different sets of features among *Italian Text, English Text, English Caption, Image.* We evaluate three sets of features:

- *English Text, Image*;
- *English Text, Italian Text, Image*;
- *English Text, English Caption, Image.*

4.2. Small Scale Ablation Study

All the models we test share the same high level architecture as shown in Figure 2, as mentioned above we use different pretrained transformer encoders to embed different modalities, sum all the embeddings entry wise after mapping them to the same dimension through a linear layer and finally with another linear layer we map to a vector with length the number of labels, finally we compute the usual Cross Entropy Loss for classification.

[1] www.huggingface.co/Helsinki-NLP/
[2] https://laion.ai/blog/coca/

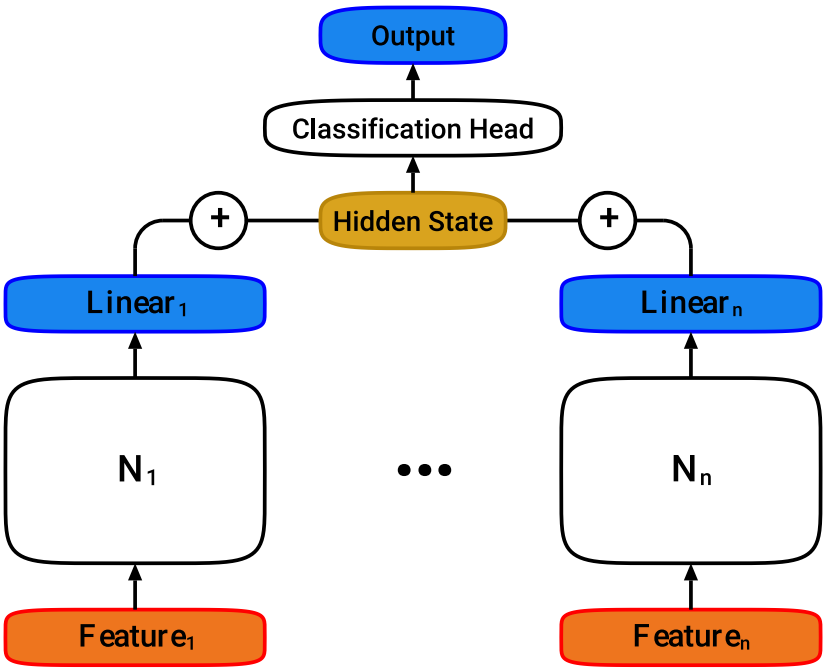

Figure 2: Neural Network architecture adaptable to several inputs schemes, N_i are pretrained transformers chosen specifically for each input augmented with a linear layer to map all output to the same dimension (512).

While summing the encoding of separate features, we multiply each of them by a coefficient, let us call it α_i (e.g. $\alpha_{eng-text}$ is the coefficient multiplying the embedding for the English translation), that modulates the relative importance of each feature. Similarly each feature has its own pre-trained encoder, we use the following ones:

- VIT [17] and in particular the vit-large-patch32-384 version[3] to encode images;
- RoBERTa large [18] to encode text in English, either the translated texts or the generated captions;
- a version of BERT-base pretrained on Italian[4] to encode all Italian text we use.

We perform all our validation test by splitting the training dataset in 80% training and 20% validation. The main architecture choices we make are, the shared size to which we map the embeddings output by each encoder, the α_i that multiply each of the embeddings before summing them, the classification head shape and the pre-trained models we use. Let us list how we chose each of them:

- For the vector size, we experiment with 512 and 1024, seeing that performance does not change depending on these two setting we use the smaller value, 512 in all our experiments.
- Concerning the α_i of each modality, we notice that $\alpha_{eng-text}$ is the most relevant one and after some tests, we choose the parameters as follows, $\alpha_{eng-text} = 1.0$ and all others equal to 0.1.
- The final ablation we have performed concerns the classification head, which eventually we choose to be a single linear layer with input size 512 and output size the number of labels, 4 for task 1 and 3 for task 2.
- Initially, we tried a different version with two Linear layers with *tanh* activation function in between and the hidden size of 2048, but this leads

[3] https://huggingface.co/google/vit-large-patch32-384
[4] https://github.com/dbmdz/berts

	Certainly Fake	Probably Fake	Probably Real	Certainly Real	weighted avg
support	16	52	106	21	195
precision	26.7	38.2	60.6	21.4	47.6
recall	25.0	25.0	75.5	14.3	51.3
f1-score	25.8	30.2	67.2	17.1	**48.6**

Table 1
Per class metrics using *English Text, Image* as features, our official task 1 submission (In **bold** the score used for the task).

	Misleading	Unrelated	Not Misleading	weighted avg
support	45	75	99	219
precision	28.3	46.8	47.6	43.3
recall	37.8	48.0	39.4	42.0
f1-score	32.4	47.4	43.1	**42.4**

Table 2
Per class metrics using *English Text, Image* as features, out official Task 2 submission (In **bold** the score used for the task).

to lower performance (although comparable) in all our experiments.

- Similarly, while choosing architecture we experimented with smaller versions of each transformer encoder, namely: (a) VIT with *patch16-224* instead of *patch32-384*; (b) roberta-base instead of roberta-large; (c) bert-base pretrained on English instead of Italian. However, while faster to train, switching any pretrained model to its smaller version reduced performance and therefore we opt for the larger ones when performing the grid search to choose our best model.

4.3. Hyper Parameter Selection

For Task 1 we perform a grid search using as features: *English Text, Image*.

We sweep over the following hyper parameters:

- learning rate: 1e-5, 2e-5, 3e-5, 5e-5;
- Max epochs: 3, 4, 5, 10, 20;
- Warmup steps: 0, 100;
- Batch size: 4, 8 (other values would not fit into our machine).

The best performance on our validation set is obtained with warmup 0, batch-size 8, epochs 4 and learning rate 1e-5 and therefore we use this set of hyper parameters when training with all groups of features[5].

Due to limitations in GPU memory, we clip all sequences to 256 length. We also tested a length of 400 using English Text and Images only, however this did not seem to affect performance[6].

Comparing the results obtained on our validation set when using different groups of features we eventually choose to only use the translated text together with the Images, as adding Italian didn't appear to provide significant improvements.

We tackle Task 2 keeping everything as we did in Task 1 switching training set.

5. Results

5.1. Task 1

Table 1 shows the performance of our approach on the first task. In bold we report the metric that has been used to evaluate our model, it reports how the class balance in the training set is reflected into per-class performance into the official test set (measured with the official evaluation script). Indeed the *Certainly Real* class is the most numerous in this case too as well as the one where our model is best performing. It is interesting to notice how the model performs better on the *Certainly False* class compared to the *Certainly Real* one despite the second being more populated, we speculate this is due to the similarity with the *Probably Real* class.

Although we chose a different method to submit to Task 1, we show that including the Italian text results into promising results on the official test set. Table 3 shows how this approach performs on the official test set and indeed it would improve over our submission.

Unlike adding the Italian text, using the captions does not results in performance improvements. Table 4 shows the performance obtained while adding the Captions for the images, generated by CoCa [11] and processed with a different roberta-large model.

5.2. Task 2

For task 2, we chose to keep all parameters as in Task 1.

[5]Performing a separate grid search for each feature group was not feasible.

[6]Other configurations using more features do not fit in memory for length of 400. All experiments are performed on NVIDIA GTX Quadro with 24GB of VRM.

	Cert. Fake	Prob. Fake	Prob. Real	Cert. Real	weight. avg
support	16	52	106	21	195.0
precision	50.0	41.2	62.9	20.0	51.4
recall	43.8	26.9	78.3	14.3	54.9
f1-score	46.7	32.6	69.7	16.7	**52.2**

Table 3
Per class metrics for Task 1 with features *English Text, Italian Text, Image* (In **bold** the score used for the task).

	Cert. Fake	Prob. Fake	Prob. Real	Cert. Real	weigh. avg
support	16	52	106	21	195
precision	12.5	46.2	59.6	20.0	47.9
recall	6.2	23.1	82.1	14.3	52.8
f1-score	8.3	30.8	69.0	16.7	**48.2**

Table 4
Per class metrics for Task 1 with features *English Text, English Captions, Image* (In **bold** the score used for the task).

Table 2 shows the scores we obtain on the official Test set for Task 2, it appears that the model we use can´t recognize the *Misleading* class for which f1-score is 32.4% while it does manage to achieve higher values on the remaining classes. We conclude that the relation between the Text and Image embeddings are not well captured by this model.

6. Conclusions

We have tackled the MULTI-Fake-DetectIVE task trying and improve performance with textual data augmentation techniques.

We show that our approach does provide some improvements and this is relevant as text-based data augmentation is a novel way to exploit the knowledge present within large pretrained models, made recently possible by pretrained models and has several application settings [19].

Moreover, in this report we show how using both Italian and English data at once, even though the English one is the translation of the Italian text, provides significant improvements in Task 1.

On the contrary, the lower performance of the model in Task 2 underlines how the relations between text and images are not well captured by our model and this offers the opportunity for further improvements.

A structural limitation of our approach is that, although we know that the dataset is composed of both tweets and articles and that the second document type is generally much longer than the tweets, we have not experimented with ways to use this longer context.

This too offers a promising future step, using longer context transformers when embedding text, while keeping our overall scheme of translating to English might give further improvements.

Indeed, given the scarcity of longer context transformers trained on Italian the English translation might be useful in this case as well.

Acknowledgments

This work is supported by the European Union under the scheme HORIZON-INFRA-2021-DEV-02-01 – Preparatory phase of new ESFRI research infrastructure projects, Grant Agreement n.101079043, "SoBigData RI PPP: SoBigData RI Preparatory Phase Project"

References

[1] X. Zhou, R. Zafarani, A survey of fake news: Fundamental theories, detection methods, and opportunities, ACM Comput. Surv. 53 (2020). URL: https://doi.org/10.1145/3395046. doi:10.1145/3395046.

[2] T. Fagni, F. Falchi, M. Gambini, A. Martella, M. Tesconi, Tweepfake: About detecting deepfake tweets, PLOS ONE 16 (2021) 1–16. URL: https://doi.org/10.1371/journal.pone.0251415. doi:10.1371/journal.pone.0251415.

[3] 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.

[4] 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.

[5] J. Thorne, M. Yazdani, M. Saeidi, F. Silvestri, S. Riedel, A. Halevy, From natural language processing to neural databases, Proc. VLDB Endow. 14 (2021) 1033–1039. URL: https://doi.org/10.14778/3447689.3447706. doi:10.14778/3447689.3447706.

[6] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, C. Finn, Detectgpt: Zero-shot machine-generated text detection using probability curvature, 2023. arXiv:2301.11305.

[7] X. Li, X. Yin, C. Li, X. Hu, P. Zhang, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, J. Gao, Oscar: Object-semantics aligned pre-training for vision-language tasks, ECCV 2020 (2020).

[8] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, J. Gao, Vinvl: Making visual representations matter in vision-language models, CVPR 2021 (2021).

[9] I. Gallo, A. Calefati, S. Nawaz, M. K. Janjua, Image and encoded text fusion for multi-modal classification, 2018 Digital Image Computing: Techniques and Applications (DICTA) (2018) 1–7.

[10] 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, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763. URL: https://proceedings.mlr.press/v139/radford21a.html.

[11] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, Y. Wu, Coca: Contrastive captioners are image-text foundation models, 2022. arXiv:2205.01917.

[12] J. Li, D. Li, S. Savarese, S. C. H. Hoi, BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models, CoRR abs/2301.12597 (2023). URL: https://doi.org/10.48550/arXiv.2301.12597. doi:10.48550/arXiv.2301.12597. arXiv:2301.12597.

[13] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, J. Gao, Llava-med: Training a large language-and-vision assistant for biomedicine in one day, CoRR abs/2306.00890 (2023). URL: https://doi.org/10.48550/arXiv.2306.00890. doi:10.48550/arXiv.2306.00890. arXiv:2306.00890.

[14] J. Tiedemann, Parallel data, tools and interfaces in OPUS, in: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, 2012.

[15] J. Tiedemann, S. Thottingal, OPUS-MT — Building open translation services for the World, in: Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT), Lisbon, Portugal, 2020.

[16] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Doll'a r, C. L. Zitnick, Microsoft COCO: common objects in context, CoRR abs/1405.0312 (2014). URL: http://arxiv.org/abs/1405.0312. arXiv:1405.0312.

[17] 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: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Ro{bert}a: A robustly optimized {bert} pretraining approach, 2020. URL: https://openreview.net/forum?id=SyxS0T4tvS.

[19] A. Mumuni, F. Mumuni, Data augmentation: A comprehensive survey of modern approaches, Array 16 (2022) 100258. URL: https://www.sciencedirect.com/science/article/pii/S2590005622000911. doi:https://doi.org/10.1016/j.array.2022.100258.