MULTI-Fake-DetectiVE at EVALITA 2023: Overview of the MULTImodal Fake News Detection and VErification Task

Alessandro Bondielli[1,2,*], Pietro Dell'Oglio[3,4], Alessandro Lenci[2], Francesco Marcelloni[2], Lucia C. Passaro[2] and Marco Sabbatini[2]

[1]Department of Computer Science, University of Pisa, Italy
[2]Department of Philology, Literature and Linguistics, University of Pisa, Italy
[3]Department of Information Engineering, University of Florence, Italy
[3]Department of Information Engineering, University of Pisa, Italy

Abstract

This paper introduces the MULTI-Fake-DetectiVE shared task for the EVALITA 2023 campaign. The task was aimed at exploring multimodality within the realm of fake news and intended to address the problem from two perspectives, represented by the two sub-tasks. In sub-task 1, we aimed to evaluate the effectiveness of multimodal fake news detection systems. In sub-task 2, we sought to gain insights into the interplay between text and images, specifically how they mutually influence the interpretation of content in the context of distinguishing between fake and real news. Both perspectives were framed as classification problems.

The paper presents an overview of the task. In particular, we detail the key aspects of the task, including the creation of a new dataset for fake news detection in Italian, the evaluation methodology and criteria, the participant systems, and their results. In light of the obtained results, we argue that the problem is still open and propose some future directions.

Keywords

Fake News, Fake news detection, Multi-modality, Vision-Language models, Large Language Models

1. Introduction and Motivation

Recent years have seen a great increase in the online proliferation of disinformation and fake news [1]. This is especially true in the context of real-world events that are reported as breaking news. It is often the case that entities with malicious intents exploit breaking news to push their own agenda by distorting facts and intentionally publishing false or misleading information.

Distorted uses of online social media have been made mostly evident in the last few years by the first so-called *infodemic* following the COVID-19 pandemic [2], in what has been defined by several authors as a Post-Truth Era [3] dominated by emotions and pseudo-facts [4]. This phenomenon has grown further with the outbreak of the Russian war against Ukraine. Like in all conflicts, disinformation has become a powerful strategic weapon.

These issues have led over the years to the creation of numerous initiatives for independent fact-checking and fake news detection, and the topic has increased its relevance in the research community. The literature on issues related to fake news detection, disinformation, and fact-checking, is constantly growing despite the inherent challenges and many facets of the problem.

A large number of approaches and techniques have been proposed for content verification and fake news detection in a uni-modal setting. Most of the proposed approaches use either the actual content of the news (i.e., the text itself), its context (e.g., social network structures, temporal information), or a combination of both [5]. Most modern systems typically leverage transformer models with additional information [6].

It is clear that the easiest way to spread disinformation is in textual form. However, online outlets and social media allow for other modalities as well. Images for example can be leveraged in the context of disinformation and fake news in different ways: first, the inclusion of images in malicious content can be leveraged as a way to provide more credibility for the text containing the fake news; second, images could be described in such ways that their original content is misinterpreted by readers, leading again to disinformation; finally, they can be used in an attempt to increase the post attraction and get the fake news shared by as many social media users as possible.

We can argue that multimodal scenarios may be considered as closer in nature to real-world ones examin-

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
*Corresponding author.
alessandro.bondielli@unipi.it (A. Bondielli); pietro.delloglio@unifi.it (P. Dell'Oglio); alessandro.lenci@unipi.it (A. Lenci); francesco.marcelloni@unipi.it (F. Marcelloni); lucia.passaro@unipi.it (L. C. Passaro); marco.sabbatini@unipi.it (M. Sabbatini)
0000-0003-3426-6643 (A. Bondielli); 0000-0002-0793-5226 (P. Dell'Oglio); 0000-0001-5790-4308 (A. Lenci); 0000-0002-5895-876X (F. Marcelloni); 0000-0003-4934-5344 (L. C. Passaro); 0000-0002-8837-6592 (M. Sabbatini)

ing social media data. Nevertheless, multimodality has received relatively less attention over the years in this context [4]. This is rapidly changing, with a number of international multimodal shared tasks being organised for fake news and propaganda detection, fact-checking and related areas [7, 8, 9, 10]. Nevertheless, models combining multiple modalities for detecting fake news remain a major open challenge in the literature, as well as datasets including different modalities and different sources of fake news [4]. Moreover, we believe that a fundamental step towards a more nuanced understanding of the problem lies in actually understanding and modelling the interplay between the different modalities in generating disinformation.

In this context, we propose MULTI-Fake-DetectiVE[1] as part of the EVALITA 2023 Evaluation campaign [11]. The task is aimed at addressing both the textual and visual aspects of fake news on social media and online news outlets, from two key perspectives: we want to model fake news detection from a multimodal perspective, and we are interested in exploring how images and texts interact and influence each other in the context of real and fake news. Further, we contribute to this research area by creating a dataset of social media posts from Twitter and news articles regarding the Russian-Ukrainian war including fake and real news.

2. Definition of the task

MULTI-Fake-DetectiVE includes two sub-tasks. Both are formulated as multi-class classification problems. In the first sub-task, given a piece of content (i.e., a social media post or a news article) that includes both a visual and a textual component, the goal is to determine its likelihood of being a *real* or a *fake* news. In the second sub-task, given a text and an accompanying image, the goal is to decide whether their combination is aimed at *misleading* the interpretation of the reader about one or the other, or not. Note that for both sub-tasks we consider the visual component as all the images provided with a given textual content (i.e., news article or social media post). Thus, for example, if a tweet includes three images, and one of them is misleading, the expected label will be *misleading*.

In the following, we describe in detail both sub-tasks.

2.1. Sub-task 1: Multimodal Fake News Detection

The first sub-task is structured as a multi-class classification problem in a multimodal setting. The problem is defined as follows: given a piece of content $c = \langle t, v \rangle$ which includes a textual component t and a visual component v (i.e., one or more images), classify it into one of the following classes:

***Certainly Fake*:** news that is certain to be fake, whatever the context.

***Probably Fake*:** news that is likely to be fake, but may include some real information or at the very least be somewhat credible.

***Probably Real*:** news that is very credible but still retains some degree of uncertainty about the provided information.

***Certainly Real*:** news that is certain to be real and incontestable, whatever the context.

The classes refer to the informational content as a whole, and not to its single components. For example, a fake piece of news including a real image (e.g., in a misleading context) is still probably (or certainly) fake.

2.2. Sub-task 2: Cross-modal relations in Fake and Real News

The second sub-task is aimed at assessing how the two modalities (i.e., textual and visual) interact in the context of fake and real news. Our goal is to understand how images and texts in fake and real news can lead to misleading interpretations of the content pertaining to the other modality and to the whole news.

The sub-task is a three-class classification problem, and is defined as follows: given a piece of content $c = \langle t, v \rangle$ which includes a textual component t and a visual component v, decide whether their combination is:

***Misleading*:** one between the textual and visual components is used deceptively to lead to misinterpretation of the other.

***Not Misleading*:** the combination of the visual and textual component does NOT lead to misinterpretation of the news.

***Unrelated*:** the visual component is not related to the text component, or does not add information to the text component or does not change its interpretation in a meaningful way.

3. Dataset

The dataset for the shared task includes social media posts and news articles, containing both a textual and a visual component, concerning one or more real world events that are known to have been subject to the generation of fake news. In particular, the dataset focuses on

[1] https://sites.google.com/unipi.it/multi-fake-detective/home

the Ukrainian-Russian war, and includes data in a time span going from February 2022 to December 2022.

The dataset is composed of two sub-datasets, one for each sub-task. Each is further split into a training set and two different test sets. More specifically, the dataset for each sub task is divided as follows:

Training Set: the training data provided to participants. It includes data from February 2022 to September 2022.

Test Set (Official): the official test set used for evaluation. It includes data from October 2022 to December 2022.

Test Set (Additional): an additional batch of test data including data from the same time window as the training set.

The Official Test Set was developed to challenge participating systems to classify fake news and misleading content in a more real-world scenario (i.e., different time windows that might determine different data distributions). The Additional Test Set was instead aimed at giving us a clearer picture of how participating systems are resilient to changes in the context over time [12]. Note that the evaluation on the Additional Test Set was not mandatory.

The dataset is available for download on the website of the task.[2]

3.1. Data Collection and annotation

The dataset was collected and annotated via crowdsourcing following a multi-step process heavily inspired by the one proposed in [13]. First, we broadly collected Twitter data regarding the Ukrainian-Russian war in the chosen time span. To collect such data, we chose a set of keywords representative of the conflict, e.g., "Ucraina, Russia, Putin, Zelensky". In addition to this, we collected texts and images for news articles that were in the tweets. At this stage, the data were collected regardless of the sub-tasks.

Then, we exploited a manually collected set of verified fake news and misleading claims (henceforth referred to as *seed fake news and misleading claims*) to generate the dataset for each sub-task. We took into account different news outlets reporting on the fake news and independent fact-checking websites. These seed fake news and misleading claims were intended to serve a dual purpose. On the one hand, we used them to filter the original dataset by considering their similarity with data samples. This was done to ensure that: i) the resulting datasets would include only relevant elements (i.e., that actually refer to the Ukrainian-Russian war), and ii) the class distribution for both sub-tasks was not too skewed in favour of real news and not misleading claims, as it would have been in an uncontrolled scenario. On the other hand, the seed fake news and misleading claims served as context for the annotation process. Specifically, we used Prolific[3] to obtain labels for our dataset. For each sub-task, we provided annotators with the seed fake news and misleading claims as context, and asked them to label a few of the data samples. Each data sample was labelled by at least five different annotators. We collected human annotations and kept only data samples for which at least 3 out of the 5 annotators provided the same label.

Table 1
Dataset size for sub-task 1.

	C.F.	P.F.	P.R.	C.R.
Train	153	219	476	199
Test (official)	16	52	106	21
Test (additional)	27	58	101	32
Total	196	329	683	252

Datasets sizes and class distributions are reported in Tables 1 and 2 for, respectively, sub-tasks 1 and 2. In sub-task 1, inter-annotator agreement was calculated as the average Spearman correlation coefficient between annotator pairs, considering the ordered nature of the labels. The average correlation was 0.43 (σ = 0.04). In sub-task 2 we employed Fleiss' Kappa to measure the inter-annotator agreement since the labels were not inherently ordered. We obtained k = 0.25.

Participants were provided with a TSV file containing IDs, URLs, and numeric labels representing classes for sub-task 1 and sub-task 2. The label was excluded from the test set during the evaluation period. Participants had the option to download the data using their preferred method or utilise a provided download script. The script offered participants access to textual data, including metadata such as URLs, data type (e.g., tweet or article), and creation date if available, as well as associated images. Authorship information was not provided with the data. Note that while the datasets were treated separately for annotation, some data samples could be present in both sub-tasks. In such cases, the ID associated with the data point remained consistent across the two sub-tasks.

3.2. Copyright and Content Warning

The dataset includes tweets and news articles. The provided download script performs a coarse-grained anonymization of the data (e.g. by removing author information).

Upon download, users agree not to share the material they receive both during and after the competition. The

[2] https://sites.google.com/unipi.it/multi-fake-detective/data

[3] prolific.co

Table 2
Dataset size for sub-task 2.

	Misl.	Not Misl.	Unrel.
Train	373	546	417
Test (official)	45	75	99
Test (additional)	67	84	89
Total	**485**	**705**	**605**

data for the MULTI-Fake-Detective tasks is to be used for research purposes only. Note that by receiving the data users implicitly agree to Twitter Terms of Service, Privacy Policy, Developer Agreement, and Developer Policy for academic researchers.

We do not share responsibility for the contents of the dataset. Downloaded texts and images may include copyrighted material and sensitive contents. The downloaded data and the provided labels do not reflect in any way the social and political views of the task organisers.

4. Evaluation measures

Participants were allowed to present up to four different systems for predicting labels on the official test set, with one system marked as *primary*. Results for primary systems were used as basis for the final ranking. Specifically, the ranking was calculated based on the **weighted average F1-score** of the systems. The same evaluation procedure and criteria was applied to both sub-tasks. The evaluation procedure was conducted by means of an evaluation script (available to participants).

Note that due to restrictions in data distribution (see Section 3), not all participants may have had access to the exact same test dataset. For example, articles/tweets in the dataset may have been removed by the authors during the evaluation window. To ensure fair competition, we evaluated and ranked the systems only on the subsets of the test sets for which all the participants were able to provide a label.

4.1. Baseline models

Participating systems were evaluated against each other and against a set of baseline models.

Specifically, we proposed two different classification models, namely a Support Vector Machine (SVM) and a Multi-Layer Perceptron (MLP), with three different feature sets as the baseline models. As for the feature sets, we considered:

Text-only features extracted with a multilingual BERT model [14].

Image-only features extracted with ResNet-18 [15].

Multimodal features obtained by concatenating the text-only and image-only features.

All models were trained using the default parameters from scikit-learn[4].

To ensure fair reproducibility and comparisons, the baseline models and the evaluation scripts are available on the website of the task. [5]

5. Participating systems and results

A total of four teams participated in MULTI-Fake-DetectiVE. All four teams participated to sub-task 1 (Multimodal Fake News Detection), and two of them also participated to sub-task 2 (Cross-modal relations in Fake and Real News). The proposed approaches are quite different. We can distinguish between two truly multimodal approaches and two text-oriented ones. In the following, we broadly describe the core systems of participating teams.

PoliTo [16] participated to both sub-tasks with an approach focused on refining FND-CLIP [17], a fake news detection multimodal model based on CLIP [18]. Authors proposed several refinements to the original model via ad-hoc extensions including sentiment-based text encoding, image transformations in the frequency domain, and data augmentation via back translation. The final model for both sub-tasks is an ensemble that combines predictions for all the extensions.

AIMH [19] participated to both sub-tasks with a vision-text dual encoder approach. They used ViT to encode images and RoBERTa/BERT to encode texts. Authors experimented with different inputs for their model. They generated image captions and automatically translated Italian texts to English. They tested various input combinations and chose to use English texts and images as inputs for their final model.

ExtremITA [20] participated with a text-only approach aimed at solving all EVALITA tasks via prompt engineering of Large Language Models. The team proposed two Italian models, an encoder-decoder based on T5 [21] and an instruction-tuned decoder-only model based on LLaMA [22].

HIJLI-JU-CLEF [23] proposed a text-oriented model to solve sub-task 1. The model uses a pre-trained

[4] https://scikit-learn.org/stable/supervised_learning.html
[5] https://sites.google.com/unipi.it/multi-fake-detective/tasks-and-evaluation

Table 3
Sub-task 1 - Official Test Set.

Rank	TEAM-RUN	F1-Score
1	**Polito-P1**	**0.512**
2	extremITA-camoscio_lora	0.507
3	AIMH-MYPRIMARYRUN	0.488
4	*Baseline-SVM_TEXT*	0.479
5	*Baseline-SVM_MULTI*	0.463
6	*Baseline-MLP_TEXT*	0.448
7	*Baseline-MLP_IMAGE*	0.402
8	HIJLI-JU-CLEF-Multi	0.393
9	*Baseline-SVM_IMAGE*	0.386
10	*Baseline-MLP_MULTI*	0.374

Table 4
Sub-task 1 - Additional Test Set.

Rank	TEAM-RUN	F1-Score
1	**extremITA-camoscio_lora**	**0.464**
2	PoliTo	0.460
3	extremITA-it5	0.348

model applied to image captions and a pre-trained model to translate Italian texts into English. Finally, it uses translated texts and captions as inputs to either a BiLSTM or a Transformer-based model to classify the data.

Participating systems and baselines were evaluated and ranked according to the evaluation criteria described in Section 4. In the following, we present the results on each sub-task.

5.1. Results of Sub-task 1

All participating systems attempted to solve the Multimodal Fake News Detection sub-task. Tables 3 and 4 detail the obtained results of each system including baselines on the Official and Additional test sets, respectively.

As for the Official test set, the PoliTo ensemble model and the LLaMA-based ExtremITA model ranked first and second, with close results. The AIMH vision-text dual encoder model ranked third. All three models were able to outperform all baseline systems, albeit marginally. The best performing baselines are the text-based and multimodal SVM models. The other baseline models performed significantly worse. Finally, the HIJLI-JU-CLEF text-oriented system was able to outperform two out of the six baseline models proposed, ranking fourth among participants and eighth globally.

As for the Additional test set, the best performing model was the LLaMA-based ExtremITA, closely followed by the PoliTo ensemble approach. The T5-based ExtremITA model performed significantly worse. The AIMH and HIJLI-JU-CLEF systems did not participate in the additional evaluation.

Table 5
Sub-task 2 - Official Test Set.

Rank	TEAM-RUN	F1-Score
1	**Polito-P1**	**0.517**
2	*Baseline-MLP-TEXT*	0.506
3	*Baseline-SVM-TEXT*	0.482
4	*Baseline-MLP-MULTI*	0.461
5	*Baseline-SVM-MULTI*	0.442
6	*Baseline-SVM-IMAGE*	0.436
7	AIMH-MYPRIMARYSUB	0.421
8	*Baseline-MLP-IMAGE*	0.373

5.2. Results of Sub-task 2

Only the truly multimodal models participated in the Cross-modal relations in Fake and Real News sub-task. This is due to the fact that the task is inherently multimodal, and cannot be modelled properly with text-only models: the relationship between image and text features lies at the core of the task, and thus images have to be modelled in some capacity to face it effectively.

Table 5 shows the results obtained by each system, including baselines, on the Official test set. The PoliTo ensemble model ranked first, outperforming all baseline models, while the AIMH vision-text dual encoder model outperformed only the image-only MLP model, ranking seventh. Among baselines, surprisingly the best-performing ones are the text-only models, followed by the multimodal ones. We suspect that the text-only baseline performances are to be attributed to chance rather than to their effective modeling of the problem.

Only the PoliTo team participated in the Additional evaluation, obtaining a weighted average F1-Score of 0.61.

6. Discussion

We can draw some interesting insights from comparing the different proposed models both in terms of their architectures and obtained results.

General findings. First, we can argue that multimodal fake news detection and cross-modal analysis of images and texts in the context of fake news are two rather challenging tasks. As shown by agreement metrics, it was a challenging task for annotators as well (see Sec. 3). This is reflected also by the fact that even the best performing systems were not able to considerably improve over the baseline models results.

As for performances between the Official and Additional test sets, we saw a rather large discrepancy among tasks. We expected systems to perform better on data from the same time period. This appears to be true for sub-task 2, but not for sub-task 1. Note however that the only true comparison can be made on the PoliTo system, as it is the only one that participated to the Official and Additional evaluation for both tasks.

Finally, we must point out the weaknesses of the baseline models. While participating systems were able to perform consistently across sub-task and test sets, the performances of the baseline systems exhibit significant variability, with the relative rankings and performance disparities among models varying across tasks. This suggests that the baselines are unable to adequately model features of both modalities and to leverage them for the tasks.

The architectures of the systems. In sub-task 1, only two out of the four participating systems could be considered as truly multimodal, since they explicitly model image-level features (i.e., with an image encoder model). They are quite similar in principle: both leverage a dual encoder architecture [24] with a classification layer. A ViT image encoder was chosen by both approaches, albeit trained on different data. The text encoders employed in the AIMH system are RoBERTa (for English translations) and an Italian version of BERT for original texts. The PoliTo team uses the FND-CLIP text encoder, which is based on GPT instead. The popularity of CLIP-like Vision-Language models is evident due to their versatility and ease of adaptation for various scenarios and downstream tasks, including fake news detection.The main differences are the extensions (e.g., the inclusion of sentiment-aware text features and image transformations) proposed by the PoliTo team. Authors report performance increases with all proposed extensions, with the ensemble classifier performing best. The remaining two systems either disregard images due to the model architecture (ExtremITA) or consider their automatically generated caption (HIJLI-JU-CLEF), shifting the problem to a text-only space.

The role of textual content. The results of sub-task 1 do not directly highlight a clear advantage of a multimodal approach over a uni-modal one. Two out of the three models which outperform all the baselines are actually multimodal, with the PoliTo FND-CLIP-IT ensemble outperforming all the others. However, the runner-up was the text-only Italian LoRA based on LLaMA with near identical performances.

We can hypothesise that while modelling images in conjunction with text is actually helpful for determining whether a piece of content is a real or a fake news, a large part of the key information to answer the question lies within the textual content. If we assume this, it is easier to understand how model scale also plays a crucial role in performances. The only Large Langauge Model (LLM) presented can get close to the performances of a more refined and nuanced approach in a few-shot setting via prompting. Note that this may hold true regardless of the fact that Italian pre-trained LLMs are still not consistently outperforming all other approaches as their English counterparts due to their sheer size [25].

The importance of additional processing. Sub-task 2 was specifically developed to frame the problem as a multimodal one. Only the two truly multimodal systems participated. As previously discussed, both systems employ a similar architecture and are arguably comparable in terms of model sizes. Thus, we can hypothesize that the difference in performances both in sub-task 1 and 2 may be attributed mostly to the additional processing and extensions applied by the PoliTo system. We could further argue that the tasks are both very complex and nuanced, and additional forms of processing and/or features may provide important benefits in this scenario, rather than sole reliance on textual and visual features extracted from pre-trained models.

7. Conclusions and future directions

In this paper, we presented the MULTI-Fake-DetectiVE shared task for EVALITA 2023. The task was focused on multimodality in the context of fake news. We considered the problem from two perspectives: we wanted to assess fake news detection systems in a multimodal setting, and we wanted to understand how text and images influence each other in the interpretation of a piece of content in the context of fake and real news. We framed both as classification problems.

We saw an interesting degree of variety among proposed systems, which we categorized as truly multimodal or text-oriented. By analyzing the proposed approaches and their results, we can summarise our findings as follows. First, multimodal fake news detection is a very challenging task, especially when considering near real-world scenarios. Second, we saw that for similar vision-language models, both in terms of architecture and model scale, extending the boundaries of the problem by considering additional and/or alternative processing strategies, including affective-oriented features and image processing in the frequency domain, is highly beneficial. Third, we saw that model scale plays also an important role, with pre-trained LLMs approaching the performances of thoroughly fine-tuned systems.

Our findings suggest that the problem is still open and that moving forward it could be advantageous to jointly leverage the advantages of the best-performing approaches, for instance by focusing on Large pre-trained Vision-Language models augmented with additional features (e.g., by using available emotive resources [26]) via either fine-tuning or appropriate prompt tuning/engineering.

Due to the current challenges posed by deceptive or misleading content on social media, we believe that an effective understanding and modelling of such a complex problem may prove to be highly beneficial in contrasting online disinformation. In this regard, the MULTI-Fake-DetectiVE task, including the proposed approaches and the provided datasets, may serve the Italian NLP community as an initial stepping stone in addressing this issue for the Italian language.

Acknowledgments

This research was partially supported by the Italian Ministry of University and Research (MUR) in the framework of the PON 2014-2021 “Research and Innovation" resources – Innovation Action - DM MUR 1062/2021 - Title of the Research: “Modelli semantici multimodali per l’industria 4.0 e le digital humanities.”, of PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - “FAIR - Future Artificial Intelligence Research" - Spoke 1 “Human-centered AI", funded by the European Commission under the NextGeneration EU programme and of the CrossLab and FoReLab projects (Departments of Excellence).

References

[1] A. Bondielli, F. Marcelloni, A survey on fake news and rumour detection techniques, Information Sciences 497 (2019) 38–55.

[2] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, T. Chakraborty, Fighting an infodemic: Covid-19 fake news dataset, in: Combating Online Hostile Posts in Regional Languages during Emergency Situation, Springer International Publishing, 2021, pp. 21–29.

[3] S. Lewandowsky, U. K. Ecker, J. Cook, Beyond misinformation: Understanding and coping with the “post-truth” era, Journal of Applied Research in Memory and Cognition 6 (2017) 353–369. URL: https://www.sciencedirect.com/science/article/pii/S2211368117300700. doi:https://doi.org/10.1016/j.jarmac.2017.07.008.

[4] F. Alam, S. Cresci, T. Chakraborty, F. Silvestri, D. Dimitrov, G. D. S. Martino, S. Shaar, H. Firooz, P. Nakov, A survey on multimodal disinformation detection, in: Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 6625–6643. URL: https://aclanthology.org/2022.coling-1.576.

[5] L. Bozarth, C. Budak, Toward a better performance evaluation framework for fake news classification, Proceedings of the International AAAI Conference on Web and Social Media 14 (2020) 60–71. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/7279.

[6] L. C. Passaro, A. Bondielli, A. Lenci, F. Marcelloni, Unipi-nle at checkthat! 2020: Approaching fact checking from a sentence similarity perspective through the lens of transformers, in: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, 2020.

[7] G. Da San Martino, A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, P. Nakov, SemEval-2020 task 11: Detection of propaganda techniques in news articles, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, ICCL, Barcelona (online), 2020, pp. 1377–1414. URL: https://aclanthology.org/2020.semeval-1.186. doi:10.18653/v1/2020.semeval-1.186.

[8] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, in: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 2611–2624.

[9] D. Dimitrov, B. Bin Ali, S. Shaar, F. Alam, F. Silvestri, H. Firooz, P. Nakov, G. Da San Martino, SemEval-2021 task 6: Detection of persuasion techniques in texts and images, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), ACL, Online, 2021, pp. 70–98. URL: https://aclanthology.org/2021.semeval-1.7. doi:10.18653/v1/2021.semeval-1.7.

[10] F. 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 checkworthiness in multimodal and multigenre content, in: Working Notes of CLEF 2023—Conference and Labs of the Evaluation Forum, CLEF ’2023, Thessaloniki, Greece, 2023.

[11] 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.

[12] A. Cossu, T. Tuytelaars, A. Carta, L. Passaro,

V. Lomonaco, D. Bacciu, Continual pre-training mitigates forgetting in language and vision, arXiv preprint arXiv:2205.09357 (2022).
[13] L. C. Passaro, A. Bondielli, P. Dell'Oglio, A. Lenci, F. Marcelloni, In-context annotation of topic-oriented datasets of fake news: A case study on the notre-dame fire event, Information Sciences (2022).
[14] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[15] 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.
[16] L. D'Amico, D. Napolitano, L. Vaiani, L. Cagliero, Polito at multi-fake-detective: Improving fnd-clip for multimodal italian fake news detection, 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.
[17] Y. Zhou, Q. Ying, Z. Qian, S. Li, X. Zhang, Multimodal fake news detection via clip-guided learning, arXiv preprint arXiv:2205.14304 (2022).
[18] 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.
[19] G. Puccetti, A. Esuli, Aimh at multi-fake-detective: System report, 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.
[20] C. D. Hromei, D. Croce, V. Basile, R. Basili, Extremita at evalita2023: Multi-task sustainable scaling to large language models at its extreme, 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.
[21] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, The Journal of Machine Learning Research 21 (2020) 5485–5551.
[22] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).
[23] S. Sarkar, N. Tudu, D. Das, Hijli-ju-clef at multi-fake-detective: Multimodal fake news detection using deep learning approach, 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.
[24] Z. Gan, L. Li, C. Li, L. Wang, Z. Liu, J. Gao, Vision-language pre-training:: Basics, recent advances, and future trends, Foundations and Trends® in Computer Graphics and Vision 14 (2022) 163–352.
[25] V. Basile, Is EVALITA done? on the impact of prompting on the italian NLP evaluation campaign, in: Proceedings of the Sixth Workshop on Natural Language for Artificial Intelligence (NL4AI 2022), Udine, November 30th, 2022, volume 3287 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 127–140.
[26] L. C. Passaro, A. Lenci, Evaluating context selection strategies to build emotive vector space models, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA), Portorož, Slovenia, 2016, pp. 2185–2191.