On the Categorization of Corporate Multimodal Disinformation with Large Language Models

Ana-Maria Bucur[1,2,*], Sónia Gonçalves[3] and Paolo Rosso[2,4]

[1]Interdisciplinary School of Doctoral Studies, University of Bucharest, Romania
[2]PRHLT Research Center, Universitat Politècnica de València, Spain
[3]Universidad de Sevilla, Spain
[4]ValgrAI Valencian Graduate School and Research Network of Artificial Intelligence, Spain

Abstract

Disinformation is becoming more prevalent in the corporate sphere, especially as brands choose to promote their products through influencers or micro-celebrities who are perceived as reliable and impartial, but may facilitate false information. The spread of disinformation can have negative economic impacts on companies and brands, which can even affect their reputation. Artificial Intelligence can help detect false information and has become increasingly important in combating disinformation. The current work addresses the problem of characterizing multimodal disinformation targeting corporations and provides a collection of content that spreads disinformation in digital media. The content was manually annotated with information about the target (Organization, Brand, or Other) and the source (Corporate, Advertising, or Other) of the false content. We conduct comprehensive experiments to evaluate the effectiveness of state-of-the-art Unimodal and Multimodal Large Language Models in identifying the source and target of the content.

Keywords

Corporate Multimodal Disinformation, Multimodal Large Language Models, Spanish

1. Introduction and Related Work

According to [1], the concept of disinformation refers to a deliberate and organized attempt to confuse or manipulate people by providing dishonest information. In the corporate sphere, disinformation is gaining more ground. It is orchestrated to persuade audiences and hold great appeal for advertisers who promote their dissemination as a lure "because it fits more easily into people's prejudices" [2]. The issue can become even more dangerous when we consider that more and more brands choose to promote their products through influencers or micro-celebrities, which can facilitate false information [3]. These opinion leaders are perceived with high levels of reliability and impartiality, allowing them to recommend products and services on various social media platforms and generate word of mouth that brands leverage for their commercialization [4].

The spread of disinformation can be a risk to companies and brands and cause a negative economic impact [5] that can even affect their reputation. Disinformation that can impact a company's reputation may stem from political, financial, emotional, or internal motivations, such as discontented employees [6]. Therefore, it is important for organizations to manage trusting relationships with the public. Organizations can become victims of individuals and advanced technologies with the intention to damage their reputation for twisted purposes [7] through the use of deepfakes, a new form of fake news that threatens companies, organizations, and brands [8, 9, 10]. As the reputation of organizations can be affected by the spread of disinformation, to protect the corporate image, communication officers need to be aware of strategies to combat it, such as fact-checking. Artificial Intelligence has enabled the implementation of automated approaches capable of detecting false information [11, 12], also from a multimodal perspective [13, 14, 15, 16, 17, 18].

Proceedings of the 1st Workshop on COuntering Disinformation with Artificial Intelligence (CODAI), co-located with the 27th European Conference on Artificial Intelligence (ECAI), pages 29–39, October 20, 2024, Santiago de Compostela, Spain
*Corresponding author.
ana-maria.bucur@drd.unibuc.ro (A. Bucur); songomgon2@alum.us.es (S. Gonçalves); prosso@dsic.upv.es (P. Rosso)
0000-0003-2433-8877 (A. Bucur); 0000-0002-5579-7761 (S. Gonçalves); 0000-0002-8922-1242 (P. Rosso)

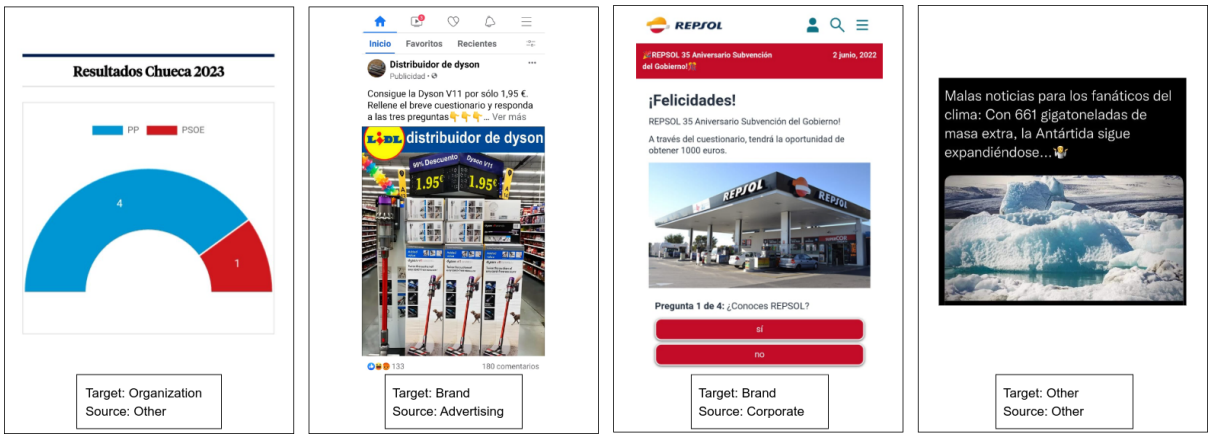

Figure 1: Selected examples of false content. The data is diverse, containing screenshots from social media, websites, etc. Translated text, first image: “Results for Chueca”. Translated text, second image: “Get Dyson V11 for only 1,95 euros. Fill in the short questionnaire and respond to the three questions...”. Translated text, third image: “Congratulations! Repsol 35th anniversary government subsidy! Through the questionnaire, you will have the opportunity to obtain 1000 euros.”. Translated text, fourth image: “Bad news for the climate fanatics: with 661 gigatons of extra mass, Antarctica continues to expand...”.

Unlike general disinformation, which can target individuals, events, or broad societal issues, corporate disinformation often has direct financial implications and can damage trust in brands and organizations. Recognizing the unique characteristics and potential impacts of such disinformation, our work aims to deepen the understanding of what are the actors targeted by corporate disinformation and the sources spreading it. By classifying the target of the false content, we can identify whether the affected entity is an organization or a brand. Furthermore, identifying the source will enable affected entities to take action and develop appropriate responses to counter the disinformation being spread about them.

As there are many previous works on multimodal fake content detection [18, 14, 13, 16, 17], we aim to characterize content that has been already fact-checked and confirmed as false. To the best of our knowledge, this is the first time that the problem of multimodal disinformation targeting corporations has been addressed automatically. For this purpose, a collection of multimodal content in Spanish that was already fact-checked is collected and annotated by expert annotators with information about the target and source of the content (Figure 1). Our dataset consists of 534 samples, together with annotations for the target (Organization, Brand, or Other) and the source (Corporate, Advertising, or Other) spreading disinformation. The false content can be targeted at an Organization, such as a company, institution, or an individual representing them. It can also target a Brand or a person associated with it. Alternatively, disinformation can be classified as Other, meaning it is not aimed at an organization or brand but contains misleading information intended to deceive the general population. Furthermore, false content can originate from various sources. It may stem from a Corporate origin, where a corporate entity is responsible for spreading disinformation, rather than just an individual. Alternatively, it could be a result of persuasive Advertising, typically in the form of paid posts on social media. Lastly, false content may originate from Other sources, such as online users disseminating misleading information.

In this paper, we address the problem of characterizing multimodal disinformation targeting corporations. Our work makes the following contributions:

- A collection of multimodal false content (visual and textual information in Spanish) that spread disinformation in digital media on corporations is compiled and annotated with information about the source and target of the false content;
- Comprehensive experiments are conducted to evaluate the effectiveness of state-of-the-art Unimodal and Multimodal Large Language Models (LLMs) in characterizing false content.

2. Data Collection

The dataset used in this work is obtained from the IBERIFIER repository[1], which includes online content that has been fact-checked and verified[2]. IBERIFIER is a project that aims to fight disinformation in digital media in Spain and Portugal, in which data from various fact-checking websites is collected and analyzed. In our research, we specifically focus on false content in Spanish that was verified by EFE Verifica[3] and Maldita.es[4], as these organizations contributed the most content to the IBERIFIER database. Our dataset consists solely of posts that were confirmed by these fact-checking entities to contain false information. This limits the dataset size, as obtaining fact-checked data is challenging. Our dataset contains 496 samples from Maldita.es and 38 samples from EFE Verifica, with multimodal data represented through both visual and textual information in Spanish. By deliberately focusing on posts that have been verified to contain disinformation, we can more effectively evaluate the performance of pre-trained visual transformer models and LLMs in characterizing deceptive information. This dataset allows us to study and understand how these models identify the different targets and sources spreading disinformation. The dataset is an essential resource for studying the effectiveness of LLMs in classifying false content from visual and textual cues found in images.

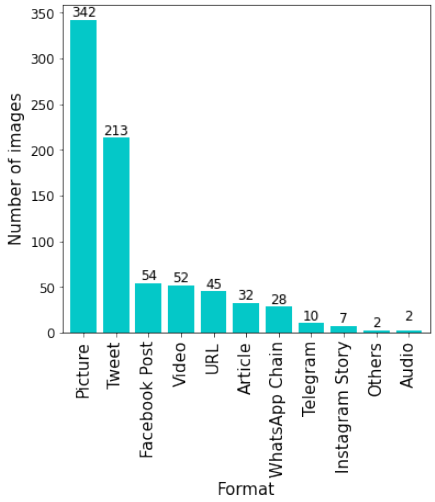

Figure 2: The format of the false content found in the collected data: pictures, screenshots from social media platforms, from different websites, or news articles.

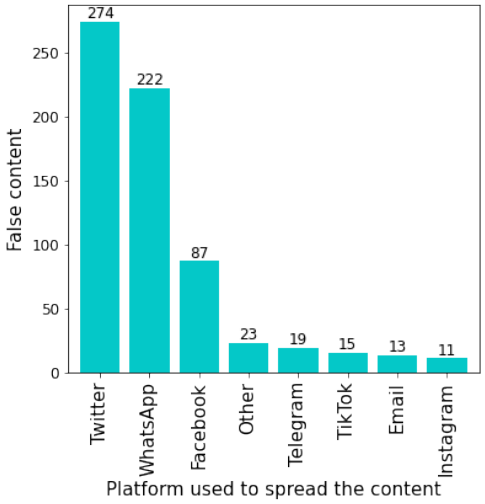

Figure 3: Platforms used to spread the false content. Most of the content was shared on social media platforms and WhatsApp.

For each of the collected images, we also retrieved information about the format of the content and the platform used to spread it using the IBERIFIER API. In Figure 2, we present the various formats of false content. The most common type of false content is represented by pictures, followed by screenshots from social media. Figure 3 shows the platforms used to spread the disinformation content. The data suggests that social media platforms like Twitter, Facebook, TikTok, and Instagram are the primary channels used to spread false content. However, we found that a considerable amount of false information is also shared through messaging apps like WhatsApp.

Two expert annotators have labeled each instance of false content with information about the target and source. The **target** of the disinformation can be an **Organization** (either a company, an institution, or a person representing it), a **Brand** (or a person representing it), or it can be **Other**, meaning that it is not targeted towards an organization or a brand, and it contains false information intending to mislead

[1] https://iberifier.eu/
[2] https://iberifier.eu/factchecks/
[3] https://verifica.efe.com/
[4] https://maldita.es/

the general population about various topics, such as climate change, immigrants, conspiracy theories, local news. With regard to the different **sources** of false content (i.e. the origin of the content), the content can be of **Corporate** origin (usually, there is an entire corporate entity behind the spread of disinformation, not just an individual), persuasive **Advertising** (usually paid posts on social media), or **Other** - usually false content spread by other users. The Other class also contains false content in which the identity of the spreader does not appear in or cannot be inferred from the image/text (see Figure 1, 1st and 4th example). We obtained a strong agreement between the two annotators (Cohen's κ 0.90). The disagreements between them have been resolved by a senior researcher in the field. The final dataset contains 347 samples targeting an organization, 87 targeting a brand, and 100 targeting other entities. Regarding the sources of the false content, the dataset is comprised of 52 Corporate, 4 Advertising, and 478 Other sources.

We showcase 4 examples from the collected data in Figure 1. The dataset includes different types of disinformation found in digital media, which makes it difficult to identify the source and target spreading the content. The first example shows an image with a figure representing the electoral results from the Chueca neighborhood of Madrid. However, the image is spreading disinformation because the results are actually from a municipality in Toledo with the same name. This is a classic example of how disinformation can be spread by manipulating images and providing false information. The source of the content was classified as *Other* because the origin of the information is unknown, it does not appear in the text or the image. On the other hand, the target is *Organization* because the disinformation publication affects one or more organizations, in this case, political parties (People's Party (PP)) and Spanish Socialist Workers' Party (PSOE)).

The second example is a sponsored post from Facebook, asking individuals to complete a brief questionnaire for the chance to purchase a discounted vacuum cleaner. However, this image represents a classic phishing post where individuals are persuaded to share their banking information with malicious entities. This example illustrates how social media platforms can be used to spread phishing scams that can deceive unsuspecting users. The source of the content was categorized as *Advertising* due to the information originating from a clearly identified advertising publication (sponsored content), indicating that the advertising is conducted on a social network through payment. Conversely, the target is identified as *Brand* because the disinformation publication impacts brands, specifically Dyson and Lidl.

The third example is a screenshot from a website that claims to be of Repsol S.A., an energy and petrochemical company from Spain. However, the website is not the real website of the company, and it is used for phishing. Malicious actors are using the website to trick users into sharing their personal data. The content was categorized as *Corporate* because the web page appears to be created by a corporate entity rather than an individual. On the other hand, the target is *Brand*, as it targets Repsol.

In the fourth example, we present a screenshot from social media that is not targeted towards a corporate entity or a brand, and it was labeled as *Other* - trying to mislead the general population. The source of the content was labeled as *Other*, with no information about the source provided in the text or image.

3. Methodology

We perform experiments in zero-shot or few-shot settings to evaluate the effectiveness of state-of-the-art visual transformer models and LLMs in characterizing false content within multimodal data.

3.1. Pre-trained Visual Transformer Models

Pre-trained visual transformer models, such as CLIP [19], have shown great performance on downstream tasks without additional training, obtaining competitive results with a supervised baseline. CLIP was pre-trained in a self-supervised manner on a large collection of image-text pairs with a contrastive learning objective. The model was trained to maximize similarity between pairs of the same class and minimize similarity between pairs of different classes. CLIP extracts embeddings by processing the

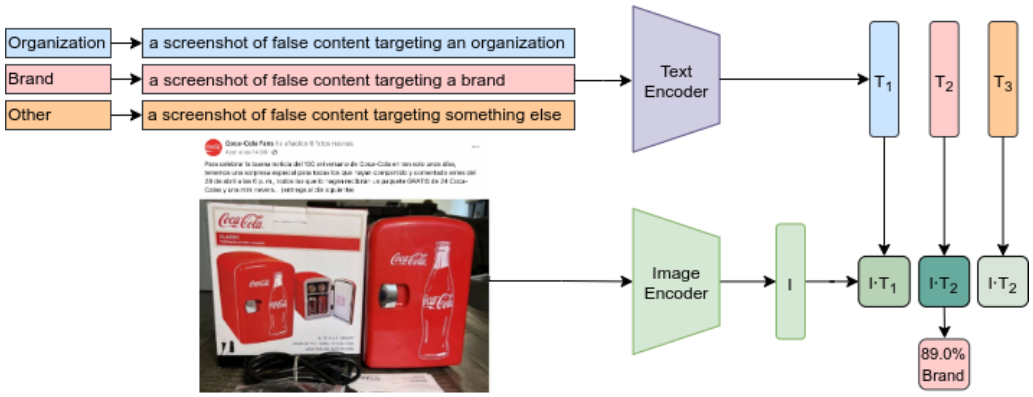

Figure 4: Zero-Shot Classification pipeline for state-of-the-art visual transformer models: CLIP, OpenCLIP, MetaCLIP, SigLIP. Images and class names/descriptions are passed through frozen encoder models, and the final prediction is represented by the text that is most similar to a given image.

image and text through a visual and textual encoder, respectively. The embeddings are then mapped to a shared space where similarities between image-text pairs can be computed. Pre-training allows CLIP to represent images and text with similar content closer in the embedding space while unrelated image-text pairs are represented further apart. In this way, the model can compute the relationship between a given image and its corresponding textual description.

We are exploring the effectiveness of using CLIP and similar models [20, 21] for zero-shot classification. To achieve this, we investigate how well the models can predict the target and the source of online disinformation. The zero-shot classification pipeline is presented in Figure 4. The process involves passing images and texts, in our case, the names/descriptions of the categories, through frozen visual and textual encoder models. The similarity between the image and each category name/description is computed, and the category with the highest similarity score is selected as the final prediction. We conducted our experiments in two settings: by providing the class names as labels and by providing a short definition/description of the content we expect to find for each class. The two types of label names, short and long, are shown in Figure 4. For target classification, we first experimented with short label names such as Organization, Brand, and Other. We also experimented with longer names, such as "a screenshot of false information targeting an organization (a company or an institution)", etc. Inspired by recent works highlighting the importance of the definitions of the concepts [22], we added more information to the text describing the categories. For the source classification, we followed a similar approach and experimented with both the short label names, such as Corporate, Advertising, and Other, and longer variants.

In our experiments, we have tested the abilities of various pre-trained transformer models like CLIP [19], OpenCLIP [23], MetaCLIP [20], SigLIP [21]. CLIP and OpenCLIP [23] have identical vision transformer architecture, but OpenCLIP was trained on the open-source dataset LAION-2B [24], whereas CLIP was trained on a private dataset of image-text pairs. MetaCLIP [20] uses the same architecture and training regime as above, but the authors ensure that only high-quality image-text pairs are used for pre-training. SigLIP [21] replaces the softmax-based contrastive loss from CLIP with a sigmoid loss. We experiment with different variants of the models, either base, large, or huge, if available.

3.2. Large Language Models

With the great success of leveraging LLMs in various vision and language tasks [25, 26, 27, 28], we also choose to test their abilities in characterizing multimodal disinformation shared in digital media. We experiment with two LLMs that have shown good results in language tasks, LLaMa-2 [27], and Mistral [25]. LLaMa is a competitive model, with good results over a suite of benchmarks related to commonsense reasoning, word knowledge, reading comprehension, etc. [27]. Mistral is another LLM

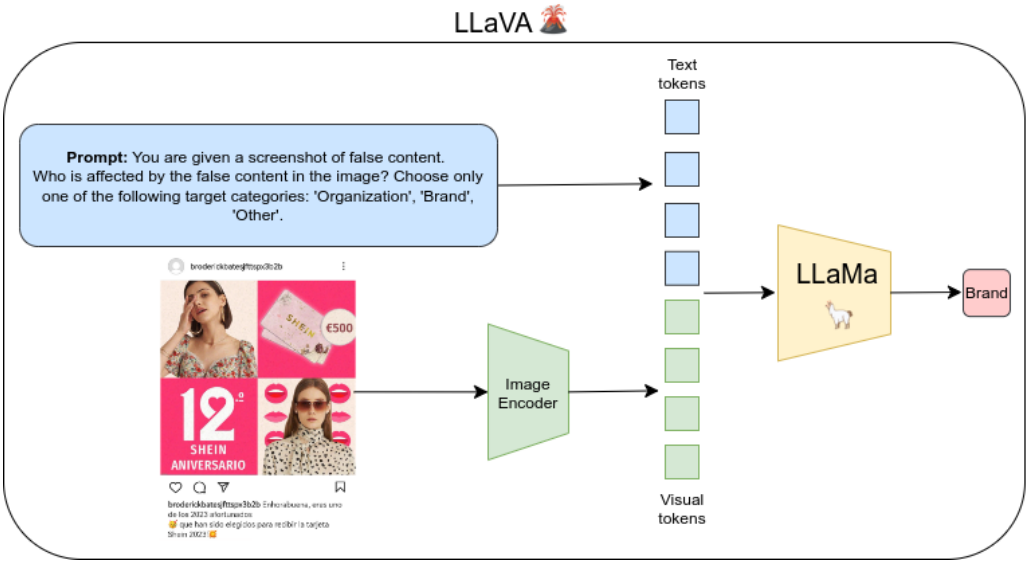

Figure 5: Zero-Shot Classification pipeline with LLaVA. LLaVa uses a language model (in our case, LLaMa) to process both visual information and language instructions, and generate an appropriate response. LLaVa leverages a pre-trained CLIP model to encode visual information from images. These embeddings are then projected into the same word embeddings space and fed into LLaMa. Finally, LLaMa generates a suitable language response.

that surpasses LLaMa-2 on all the tested benchmarks [25]. We chose these two models to evaluate their classification performance on our dataset based solely on the text found in the image and its caption. The text found in images is written in Spanish (as presented in Figure 1) and was extracted using Pytesseract[5]. The caption of the image was generated using BLIP-2 [29]. We conducted zero-shot and few-shot experiments using the aforementioned LLMs.

Although these LLMs are pre-trained on data that is mostly in English, LLaMa, for example, was pre-trained on 1.3B Spanish tokens (0.13% of the total corpus). This amount of pre-training tokens makes it capable of processing Spanish content, although the results may not be as accurate as for English data [30]. No information about the data used for pre-training Mistral models is available [25]. Because the text from the multimodal false content is in Spanish, we chose to include in our experiments a fine-tuned version of LLaMa-7B on Spanish instructions[6].

3.3. Multimodal Large Language Models

In our work, we also conduct experiments using the Multimodal LLM LLaVa [31], which is a general-purpose visual and language model (Figure 5). LLaVa uses a language model (in our case, LLaMa-2 [27]) to process both the visual information from the image and the text of the language instructions. LLaVa uses a pre-trained CLIP vision transformer to process visual input, which is then projected in the same embedding space as the text. The visual and text embeddings are then fed to LLaMa, which generates a suitable language response. In our experiments we use LLaVA-v1.5 [26] and LLaVA-v1.5 Q-Instruct [28]. We chose to use LLaVA-v1.5, as it is an improved version of the original LLaVA, and it achieves state-of-the-art results on various benchmarks related to visual question answering. LLaVA-v1.5 Q-Instruct improves over the aforementioned versions by demonstrating low-level visual perception [28].

[5] https://github.com/madmaze/pytesseract
[6] clibrain/Llama-2-7b-ft-instruct-es

4. Experimental Setup

As part of our experiments, we tested the zero-shot and few-shot (one-shot) capabilities of various models. Our test set is comprised of 519 samples, as 15 samples were kept to potentially be used for the few-shot settings. We used the open-source implementations for all the models. Due to computational limitations, we only experimented with 7B variants of LLMs and Multimodal LLMs. While generating the output, we use the default temperature of 0.7. Additionally, we post-processed the generated output to remove any punctuation, quotation marks, or explanations generated by the models. The prompts for LLaMa-2-7B and Mistral-7B were written in English. For LLaMa-2-7B-ES, given that it is a model fine-tuned for the Spanish language, we use prompts written in Spanish.

5. Results

			Target				Source		
Model	**Labels**	**Weighted-F_1**	**Brand**	**Org.**	**Other**	**Weighted-F_1**	**Adv.**	**Corp.**	**Other**
CLIP$_{base}$	Short	29.62	28.38	29.17	32.31	40.57	1.37	15.71	43.16
	Long	47.89	29.20	57.89	27.99	84.62	6.90	**48.78**	88.45
CLIP$_{large}$	Short	32.77	25.37	36.49	25.77	43.19	1.29	6.40	43.19
	Long	49.95	27.20	59.01	36.97	78.97	2.50	32.52	83.83
MetaCLIP$_{base}$	Short	20.80	31.81	15.10	31.80	48.73	1.22	12.32	52.49
	Long	50.53	30.38	60.03	33.71	70.51	3.28	34.21	74.35
MetaCLIP$_{large}$	Short	19.46	**45.69**	8.82	35.15	83.69	2.90	26.51	89.62
	Long	14.99	26.67	7.57	31.62	80.46	2.20	41.90	84.56
MetaCLIP$_{huge}$	Short	13.04	20.83	6.13	31.22	82.48	5.71	14.74	89.43
	Long	54.34	28.00	**66.37**	33.78	85.36	8.70	40.00	90.11
OpenCLIP$_{base}$	Short	10.10	25.11	0.58	31.42	82.73	3.70	0.00	91.14
	Long	36.64	31.10	38.10	36.18	63.83	1.93	23.66	68.02
OpenCLIP$_{large}$	Short	18.66	32.34	10.87	34.88	76.72	2.38	25.00	82.08
	Long	23.29	36.88	17.19	33.53	33.86	1.06	30.77	34.31
OpenCLIP$_{huge}$	Short	**55.05***	45.54	62.41	36.75	65.52	3.01	8.33	71.37
	Long	21.42	28.32	15.83	35.57	78.21	1.69	22.78	83.95
SigLIP$_{base}$	Short	21.82	31.40	16.40	33.02	82.90	**10.53**	7.23	90.60
	Long	29.54	29.43	28.02	35.10	17.14	0.96	29.17	16.03
SigLIP$_{large}$	Short	13.91	37.21	2.87	33.51	**86.18***	0.00	9.52	**94.03**
	Long	51.59	30.45	59.93	**39.81**	4.16	0.96	34.01	1.26

Table 1: Zero-shot classification using visual transformer models. We present the Weighted F_1-score, and the F_1-scores for each of the classes. We present the best results with **bold**, and with underline the second-best results. * denotes statistically significant differences between best and second-best models using the McNemar-Bowker Test ($p<0.05$).

We evaluate each model for the two tasks, either target or source classification, by computing F_1 scores for each class. We also measure the performance over each task using Weighted-F_1 score, given that the categories of our dataset are highly imbalanced. We present the results of the zero-shot classification using CLIP, MetaCLIP, OpenCLIP, and SigLIP in Table 1. For the majority of the models and variants, using longer descriptions of the class names improved the results of the classification. The best model for classifying the target of the false multimodal content was OpenCLIP$_{huge}$, obtaining a Weighted-F_1 score of 55.05%. Even if SigLIP$_{large}$ obtained an 86.18% Weighted-F_1 score for predicting the source of disinformation, it cannot accurately make predictions for all the categories.

In Table 2, we showcase the performance of the LLMs in zero-shot and few-shot settings. LLaMa-2-7B, Mistral-7B and LLaMa-2-7B-ES use only the text extracted from the image and its generated caption. By providing only one example in the prompt, the performance of LLaMa-2-7B improves by 28.15%. For Mistral-7B, there is a 10.49% improvement in Weighted-F_1 score for target classification, while, for LLaMa-2-7B-ES, the improvement is minimal between zero-shot and few-shot settings. However, the model fine-tuned on Spanish instructions, LLaMa-2-7B-ES, obtained the best Weighted F_1 score of 64.01% in the few-shot setting and second-best Weighted F_1 score of 62.31% in the zero-shot setting.

Model	Target				Source			
	Weighted-F_1	Brand	Org.	Other	Weighted-F_1	Adv.	Corp.	Other
LLaMa-2-7B (zero-shot)	14.33	0.00	12.90	31.85	80.71	0.00	0.00	88.94
LLaMa-2-7B (one-shot)	42.48	22.43	50.47	31.00	72.66	2.65	0.00	80.05
Mistral-7B (zero-shot)	49.89	23.53	59.51	38.04	**86.98**	0.00	4.26	**95.43**
Mistral-7B (one-shot)	60.38	**32.00**	74.89	32.62	86.35	0.00	0.00	95.15
LLaMa-2-7B-ES (zero-shot)	62.31	19.23	76.07	50.00	81.81	2.38	**41.24**	86.11
LLaMa-2-7B-ES (one-shot)	**64.01***	24.56	**76.41**	**53.42**	78.67	**2.96**	41.03	82.67

Table 2: Zero-shot and one-shot classification using LLMs. * LLaMa-2-7B-ES (one-shot) obtains statistically significant improvement over the best English counterpart Mistral-7B (one-shot) in Target prediction (McNemar-Bowker Test, $p<0.05$).

Model	Target				Source			
	Weighted-F_1	Brand	Org.	Other	Weighted-F_1	Adv.	Corp.	Other
LLaVA-v1.5-7B	**51.88***	21.37	**65.85**	27.89	61.68	1.89	8.60	67.12
LLaVA-v1.5-7B (Q-Instruct)	49.68	**24.84**	60.20	**33.22**	**68.72***	**2.65**	**15.93**	**74.16**

Table 3: Zero-shot classification using LLaVA. * denotes statistically significant differences between best and second-best models using the McNemar-Bowker Test ($p<0.05$).

Predicting the target of disinformation is easier, usually relying on specific cues, such as the presence of organizations' or brands' logos or names appearing in the picture or written in text. However, predicting the source of disinformation from multimodal content is a harder task, as in many instances, no information about it appears, and the source is unknown. For source classification, the LLMs sometimes only predict the *Other* class, failing to predict other categories. Using the LLaMa-2-7B-ES in one-shot setting with the text from the image and its caption as input was proven to be a suitable approach for target classification, surpassing all other visual models, such as CLIP, MetaCLIP, OpenCLIP and SigLIP. The limitations of general language models trained solely on English data are highlighted by the best performance of LLaMa-2-7B-ES, which was adapted to Spanish data. This further emphasizes the need to develop language-specialized LLMs.

In Table 3, we show the results of LLaVA-v1.5-7B for zero-shot classification. LLaVA-v1.5-7B obtains a better performance of 51.88% Weighted-F_1 score for target classification, while LLaVA-v1.5-7B (Q-Instruct) obtains a better performance for source classification (74.16% Weighted-F_1 score). In zero-shot settings, LLaVA-v1.5-7B outperforms the English-based language-only counterparts, LLaMa-2-7B and Mistral-7B, for target classification, obtaining a Weighted-F_1 score of 51.88%. However, it has a lower performance than LLaMa-2-7B-ES. According to our experiments, while general LLMs pre-trained on mostly English data can provide satisfactory results for identifying false content in our corporate multimodal disinformation dataset, models specifically adapted for a particular language perform better. This is because they can make use of the Spanish text present in the multimodal content, leading to enhanced performance.

6. Conclusion

In this paper, our aim was to create a valuable resource for characterizing corporate multimodal disinformation from digital media featuring both visual and textual elements in Spanish, annotated with details about the source and target of the false content. By publishing our dataset, we aim to encourage further research in this area and the development of more effective disinformation characterization technologies. Our comprehensive experiments have assessed the efficacy of state-of-the-art multimodal transformer models and LLMs in characterizing false content within images. Our findings reveal that predicting the target of the false content is easier than predicting the source, as the latter requires information that may not be easily represented in the multimodal data. In terms of zero-shot versus few-shot settings, providing one example for each class improved the performance for target classification by 28.15% for LLaMa-2-7B and 10.49% for Mistral-7B in terms of Weighted-F_1 score. LLaVA, the Multimodal

LLM that we had tested, obtained a Weighted-F_1 score of 51.88% in a zero-shot setting for target classification. The best result for target classification, of 64.01% Weighted-F_1 score, was obtained by LLaMa-2-7B-ES in one-shot setting, suggesting that LLMs specifically adapted for a particular language are needed when processing non-English data.

Our goal is to assist corporate entities in monitoring digital streams for fake news that could potentially harm their reputations. In our future work, we intend to expand our dataset and develop methods for identifying the specific brands and organizations targeted by false content. Moreover, we would like to expand our analysis to recently-released LLMs, such as LLama-3[7], LLaVA-NeXT[8], GPT-4V [32], Gemini Pro[9], InstructBLIP [33].

Limitations

One of the limitations of the current study is the small and imbalanced number of samples in each class from the collected dataset. Our approach relies on data that was already fact-checked, which is challenging to obtain. Due to the insufficient samples in some categories, our models struggle to accurately predict those classes. To address this limitation, our future work will focus on expanding the dataset. Specifically, we will target the collection of more samples for underrepresented classes, such as Brand for target classification and Corporate and Advertising for source classification.

Another limitation is the use of 7B variants of LLMs and Multimodal LLMs in our experiments due to computational limitations. Even if LLaMa-2-7B-ES and LLaVA-v1.5-7B have shown promising results of 64.01% and 51.88% Weighted-F_1 for source classification, using bigger variants of the models could lead to further improvements in the results [34].

Acknowledgments

The work of Paolo Rosso was in the framework of FAKE news and HATE speech (FAKEnHATE-PdC) funded by MCIN/AEI/10.13039/501100011033 and by European Union NextGenerationEU/PRTR (PDC2022-133118-I00), Iberian Digital Media Observatory (IBERIFIER Plus) funded by the EC (DIGITAL-2023-DEPLOY-04) under reference 101158511, and Malicious Actors Profiling and Detection in Online Social Networks Through Artificial Intelligence (MARTINI) funded by MCIN/AEI/ 10.13039/501100011033 and by European Union NextGenerationEU/PRTR (PCI2022-135008-2).

References

[1] C. Ireton, J. Posetti, Journalism, fake news & disinformation: handbook for journalism education and training, Unesco Publishing, 2018.

[2] P. Berthon, E. Treen, L. Pitt, How truthiness, fake news and post-fact endanger brands and what to do about it, NIM Marketing Intelligence Review 10 (2018) 18–23.

[3] S. A. Baker, Alt. health influencers: how wellness culture and web culture have been weaponised to promote conspiracy theories and far-right extremism during the covid-19 pandemic, European Journal of Cultural Studies 25 (2022) 3–24.

[4] M. De Veirman, V. Cauberghe, L. Hudders, Marketing through instagram influencers: the impact of number of followers and product divergence on brand attitude, International journal of advertising 36 (2017) 798–828.

[5] A. Christov, et al., Economic effects of the fake news on companies and the need of new pr strategies, Journal of Sustainable Development 8 (2018) 41–49.

[6] A. Reid, What's the damage?. measuring the impact of fake news on corporate reputation can act as a guide for companies to navigate a post-truth landscape, CommunicationDirector.com (2017).

[7] https://ai.meta.com/blog/meta-llama-3/

[8] https://llava-vl.github.io/blog/2024-01-30-llava-next/

[9] https://deepmind.google/technologies/gemini/pro/

[7] M. Peterson, A high-speed world with fake news: brand managers take warning, Journal of Product & Brand Management 29 (2020) 234–245.
[8] W. A. Galston, Is seeing still believing? the deepfake challenge to truth in politics, Brookings Institution (2020).
[9] S. Gomes-Gonçalves, Los deepfakes como una nueva forma de desinformación corporativa–una revisión de la literatura, IROCAMM: International Review of Communication and Marketing Mix, 5 (2), 22-38. (2022).
[10] M. Westerlund, The emergence of deepfake technology: A review, Technology innovation management review 9 (2019).
[11] M. Babakar, W. Moy, The state of automated factchecking, Full Fact 28 (2016).
[12] G. Ruffo, A. Semeraro, A. Giachanou, P. Rosso, Studying fake news spreading, polarisation dynamics, and manipulation by bots: A tale of networks and language, Computer science review 47 (2023) 100531.
[13] Y. Li, B. Jiang, K. Shu, H. Liu, Toward a multilingual and multimodal data repository for covid-19 disinformation, in: IEEE Big Data, IEEE, 2020, pp. 4325–4330.
[14] Q. Li, M. Gao, G. Zhang, W. Zhai, J. Chen, G. Jeon, Towards multimodal disinformation detection by vision-language knowledge interaction, Information Fusion 102 (2024) 102037.
[15] G. Zhang, A. Giachanou, P. Rosso, Scenefnd: Multimodal fake news detection by modelling scene context information, Journal of Information Science (2022).
[16] S. Tufchi, A. Yadav, T. Ahmed, A comprehensive survey of multimodal fake news detection techniques: advances, challenges, and opportunities, International Journal of Multimedia Information Retrieval 12 (2023) 28.
[17] A. Wilson, S. Wilkes, Y. Teramoto, S. Hale, Multimodal analysis of disinformation and misinformation, Royal Society Open Science 10 (2023) 230964.
[18] 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, CEUR Workshop Proceedings (2023).
[19] 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: Proceedings of ICML, 2021, pp. 8748–8763.
[20] H. Xu, S. Xie, X. E. Tan, P.-Y. Huang, R. Howes, V. Sharma, S.-W. Li, G. Ghosh, L. Zettlemoyer, C. Feichtenhofer, Demystifying clip data, in: Proceedings of ICLR, 2023.
[21] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, Sigmoid loss for language image pre-training, in: Proceedings of ICCV, 2023.
[22] Y. Peskine, D. Korenčić, I. Grubisic, P. Papotti, R. Troncy, P. Rosso, Definitions matter: Guiding gpt for multi-label classification, in: Findings of ACL: EMNLP 2023, 2023, pp. 4054–4063.
[23] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, L. Schmidt, Openclip, 2021.
[24] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al., Laion-5b: An open large-scale dataset for training next generation image-text models, in: Proceedings of NeurIPS, volume 35, 2022, pp. 25278–25294.
[25] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, arXiv preprint arXiv:2310.06825 (2023).
[26] H. Liu, C. Li, Y. Li, Y. J. Lee, Improved baselines with visual instruction tuning, in: Proceedings of ITIF Workshop, 2023.
[27] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).
[28] H. Wu, Z. Zhang, E. Zhang, C. Chen, L. Liao, A. Wang, K. Xu, C. Li, J. Hou, G. Zhai, et al., Q-instruct: Improving low-level visual abilities for multi-modality foundation models, arXiv preprint arXiv:2311.06783 (2023).
[29] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: bootstrapping language-image pre-training with frozen

image encoders and large language models, in: Proceedings of ICML, 2023.
[30] H. Choi, Y. Yoon, S. Yoon, K. Park, How does fake news use a thumbnail? clip-based multimodal detection on the unrepresentative news image, in: Proceedings of the CONSTRAINT Workshop, 2022, pp. 86–94.
[31] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, in: Proceedings of NeurIPS, 2024.
[32] OpenAI, Gpt-4v(ision) system card, preprint (2023).
[33] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, S. Hoi, Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. arXiv:2305.06500.
[34] J. Lucas, A. Uchendu, M. Yamashita, J. Lee, S. Rohatgi, D. Lee, Fighting fire with fire: The dual role of llms in crafting and detecting elusive disinformation, in: Proceedings of EMNLP, 2023, pp. 14279–14305.