HIJLI-JU-CLEF at MULTI-Fake-DetectiVE: Multimodal Fake News Detection Using Deep Learning Approach

Sandip Sarkar[1,*], Nripen Tudu[2] and Dipankar Das[3]

[1]Department of Computer Science and Application, Hijli College, Kharagpur
[2]Department of Computer Science and Engineering, Jadavpur University, Kolkata
[3]Department of Computer Science and Engineering, Jadavpur University, Kolkata

Abstract

This report presents the progress made in developing a system for participation in a shared task called MULTI-Fake-DetectiVE Task 1, which aims to detect and verify fake news in a multimodal environment comprising both textual and visual elements. The task primarily revolves around automatically identifying fake news by analyzing the combined use of text and images. Our system is structured into three distinct modules. The initial module is responsible for extracting textual information from images. The second module serves as a translation component, enabling the analysis of non-English text by converting it into English. Lastly, the classification module utilizes the outputs from the previous modules to predict the appropriate classes, allowing for accurate differentiation between various types of content. Our objective is to create a system that can predict these labels. To achieve this, we extract information from both the image and the text, specifically focusing on English language text and translating any textual data into English. Subsequently, both sets of data are utilized to train a classification-based model, which aims to predict the aforementioned labels. We were able to obtain a Weighted Average F1-Score of 0.393 by implementing a Multi-head attention mechanism.

Keywords

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

1. Introduction

Over the past few years, there has been a substantial increase in Internet and social media usage. Unfortunately, this growth has been accompanied by a significant rise in the dissemination of fake news and misinformation. As a result, the ability to share information has become more accessible, no longer limited to prominent news organizations. Visual content, such as images, holds greater prominence on social media platforms due to its intuitive nature. The intentional dissemination of fake news aims to harm the reputation of individuals or organizations. It can serve as a propagandistic tool targeting political parties or specific communities. Fake news proliferates effortlessly across various platforms such as social media, online news platforms, blogs, messaging applications, and group conversations. To enhance its credibility and facilitate its dissemination on social media and other online channels, manipulated images and videos are frequently employed within fake news content. Certain websites and blogs employ deceitful designs and deceptive domain names to mimic legitimate news sources. Within closed groups and messaging apps, fake news can rapidly circulate among like-minded individuals through forwarded messages and posts. These techniques facilitate the swift dissemination of misinformation, further perpetuating false narratives.

The combination of natural language processing (NLP) and computer vision is mutually beneficial when it comes to detecting fake news and analyzing textual and visual content. NLP focuses on examining the language used in news articles and social media posts to detect patterns and inconsistencies. Identify and highlight untrue assertions, evaluate the reliability of sources, and verify information using reputable sources. Computer vision analyzes images and videos, detecting manipulated images by examining content, metadata, and contextual information. Detect indications of manipulation, deepfakes, or deceptive depictions. Combined, these methods provide a comprehensive approach to counteracting fake news.

Our paper exhibits a clear and logical organization, ensuring that the content is presented in a cohesive manner. In Section 2, we give a summary of related work means what other researchers have found in this area. Then, in Section 3, we provide a detailed explanation of MULTI-Fake-DetectiVE Task 1, using ideas and methods from previous research. We also look at the dataset for MULTI-Fake-DetectiVE Task 1 and share some interesting numbers and facts in Section 4. In the next part, which is Section 5, we go into detail about our model. Then, in Section 6, we describe our results. Finally, we wrap up our paper in Section 7 with our conclusions.

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
*Corresponding author.
sandipsarkar.ju@gmail.com (S. Sarkar);
nripentudu010@gmail.com (N. Tudu);
Dipankar.dipnil2005@gmail.com (D. Das)

2. Related work

Social media serves as a double-edged sword for news consumption. While it offers easy access to information at a low cost, it also enables the spread of fake news with intentionally false information [1]. Because there has been an increase in the use of text combined with images on social media, numerous studies have been conducted to examine how the combination of visual content and text can be used to anticipate the spread of false information.

In this research, the focus is on examining how the fusion of textual and visual content can be used to forecast deceptive information and false news. The study presents a multi-modal method that examines both text and images to detect patterns and discrepancies that suggest the presence of fake news [2].

Nguyen and Kyumin gathered and examined a group of online users known as guardians, who play a role in rectifying misinformation and fake news within online discussions by referencing fact-checking URLs. They introduced an innovative model for recommending fact-checking URLs, aiming to motivate guardians to actively participate in fact-checking endeavors [3].

The suggested framework employs the explicit convolution neural network model for image processing and the sentence transformer for text analysis. The features extracted from both visual and textual inputs are then fed through dense layers and subsequently merged to predict the authenticity of images [4].

3. Task Description

The primary focus of "MULTI-Fake-Detective Task 1 - Multimodal Fake News Detection and Verification" is to automatically identify fake news within environments that involve both text and images [1]. This task is framed as a multi-category classification problem within a multimedia context [5, 6].

The problem is defined as follows: when provided with a content element, denoted as c = ⟨t, v⟩, comprising a textual component t and a visual component v (such as an image), the task is to classify it into one of the following labels: "Certainly Fake," "Probably Fake," "Probably Real," or "Certainly Real."

These labels represent the overall information conveyed by a piece of content rather than its individual components. For instance, even if fake news includes authentic photographs presented in a deceptive context, it can still be considered misleading or entirely false. The labels can be interpreted as follows:

- **Certainly Fake:** news that is most certain to be fake, whatever the context.
- **Probably Fake:** news that is still 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 most certain to be real and incontestable, whatever the context.

4. Dataset Description

The dataset used for this project consists of a wide range of social media posts and news articles that include both text and images. These posts and articles are closely connected to real-world events that are often targeted for spreading false information. Specifically, the dataset focuses on the Ukrainian-Russian conflict that began in February 2022.

The provided download script was used to obtain the actual data. Table 1 shows the description of the dataset. This script creates a TSV (Tab-Separated Values) file and a directory named "Media" in the current working directory. The "Media" directory contains the downloaded images. The TSV includes the following:

I. **ID**: A unique identifier for the data point.
II. **URL**: The web address (URL) of the data point.
III. **Date**: The date when the data point was created. Please note that newspaper articles may not have this information available.
IV. **Type**: Indicates whether the data point is a news article or a tweet.
V. **Text**: The complete text of the data point.
VI. **Media**: The names of image files associated with the data point, if any.
VII. **Label**: A numerical label that represents one of the four possible labels (i.e. Certainly Fake: 0, Probably Fake: 1, Probably Real: 2, Certainly Real: 3) specified in the task description.

The “Media” directory contains all the images specified in the Media column of the TSV file.

5. System Description

The system consists of three modules designed to analyze both textual and visual components of content in order to classify it as fake or real. The First module of the system is the Image to Text generation module, which employs a pre-trained model to generate textual descriptions for images. The subsequent module is the translation module, which handles the task of translating text between

[1] https://sites.google.com/unipi.it/multi-fake-detective/home?authuser=0

Table 1
Data Points

ID	URL	Type	Label
1498022438398877704	https://twitter.com/manuela_carloni/status/1498022438398877704	tweet	1

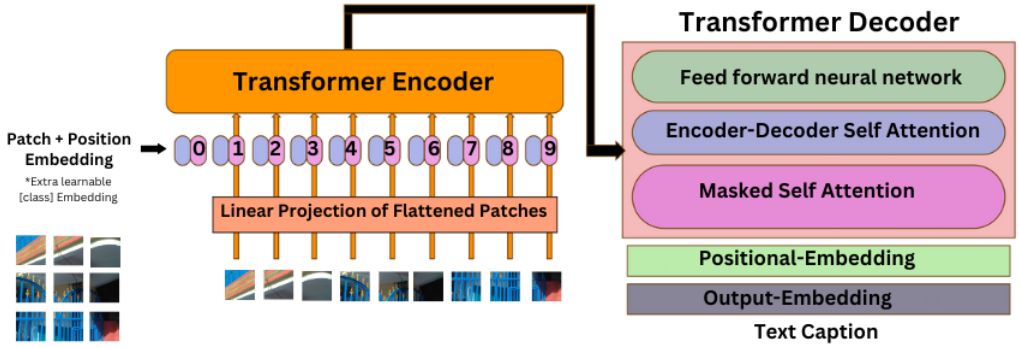

Figure 1: Image to Text Generation Model

different languages. The third module is dedicated to text classification and utilizes a combination of a BiLSTM (Bidirectional Long Short-Term Memory) model and a Multi-head attention model to classify text into different categories.

5.1. Image to Text generation Module

Image captioning is the process of generating a caption i.e., a description from the input image. It requires both Natural language processing as well as computer vision to generate the caption. Image captioning typically involves a deep learning approach, where a neural network model is trained on a large dataset of paired images and their corresponding captions.

This model learns to extract visual features from the images and then generates captions based on these visual features. The architecture of the Image to Text Generation Model is illustrated in Figure 1.

Image captioning is an example, in which the encoder model is used to encode the image, after which an autoregressive language model i.e., the decoder model generates the caption. The model we use to generate text from images is developed and trained by Ankur [2]. It is a Vision Encoder-Decoder Model, which is a type of neural network model used for tasks that involve processing both visual and textual information. It combines an encoder network, which processes the visual input, with a decoder network, which generates textual output. [3]

The encoder part of the model is responsible for extracting features from the visual input, such as an image. It typically consists of convolutional neural network (CNN) layers that can capture spatial and visual information from the input image. The encoder encodes the image into a fixed-length vector representation, often referred to as "image embedding" or "visual features."

The decoder part of the model takes the image embedding as input and generates textual output, such as captions or descriptions, related to the visual content. The decoder is typically implemented using recurrent neural networks (RNNs), such as long short-term memory (LSTM) or gated recurrent units (GRUs), which can process sequential data and generate coherent textual output.

The model was initialized with an image-to-text model with a pre-trained Transformer-based vision which involves dividing the input image into patches and treating them as tokens. These patches undergo self-attention and feed-forward networks within the Transformer encoder. Self-attention captures relationships between patches, while the feed-forward networks refine the representations. Positional embeddings capture spatial information, and a classification head predicts the final output. and a pre-trained language model as the decoder GPT2 (Generative Pre-trained Transformer 2) model operates through a mechanism called the Transformer architecture. It consists of multiple layers of self-attention and feed-forward neural networks. The model takes a sequence of tokens as input and processes them iteratively, attending to relevant tokens and generating context-aware representations. The self-attention mechanism captures relationships between different tokens in the sequence, while the feed-forward networks apply non-linear transforma-

[2]https://huggingface.co/nlpconnect/vit-gpt2-image-captioning
[3]https://ankur3107.github.io/assets/images/vision-encoder-decoder.png

tions.

The Vision Encoder-Decoder Model was trained using a dataset Common Objects in Context (COCO), a collection of more than 120 thousand images with descriptions [7]. The model is optimized to minimize the discrepancy between the predicted captions and the ground truth captions in the training data.

5.2. Link module

The Link module establishes a connection between Module 1 and Module 2. Since the data given by the organizer is in Italian, we need to convert it into a universal language, such as English. Subsequently, in the next module, we employ various classification methods.

5.3. Text Classification Module

The text classification module is a vital component of the multi-modal fake news detection system as it plays a pivotal role in analyzing the textual content and accurately categorizing it into one of the predefined labels: "Certainly Fake," "Probably Fake," "Probably Real," or "Certainly Real".

To perform text classification, we have employed two distinct models: 1) Bi-LSTM and 2) Multi-head attention model. Further information about these models can be found in the subsequent sections.

5.3.1. BiLSTM

BiLSTM, which stands for Bidirectional Long Short-Term Memory, is a recurrent neural network (RNN) model that takes into account information from both previous and future contexts when analyzing sequential data. Unlike traditional LSTM, which processes sequences in only one direction, BiLSTM processes the sequence in both forward and backward directions concurrently. It finds applications across various domains in the field of NLP, such as text simplification, machine translation, text similarity, and numerous other areas [8].

In the BiLSTM model, the input sequence is passed through two distinct LSTM layers: one layer handles the sequence in a forward direction, while the other layer handles it in a backward direction. This enables the model to grasp relationships and contextual information from both preceding and succeeding elements within the sequence.

Through the amalgamation of representations from both directions, BiLSTM adeptly captures a more holistic comprehension of the sequential data. This architecture is widely employed in various tasks, including natural language processing, speech recognition, and sentiment analysis, where incorporating context from both preceding and succeeding elements is crucial for precise predictions or classifications.

5.3.2. Multi-head Attention Mechanism

The Multi-head Attention Mechanism model is a component used in transformer-based neural networks that allows the model to attend to different parts of the input simultaneously and capture diverse relationships. It enhances the model's ability to process and extract information from the input sequence effectively.

In this mechanism, the input sequence is transformed into multiple query, key, and value representations through linear projections. These projections are then used to compute attention scores between different positions in the sequence. The attention scores determine the importance or relevance of each position with respect to others.

Multiple attention heads are employed in parallel, each attending to a different set of positions in the sequence. This enables the model to capture various patterns and dependencies at different levels of granularity. The output of each attention head is combined to form the final output, which incorporates information from multiple perspectives.

By utilizing the multi-head attention mechanism, the model can capture both local and global dependencies in the input sequence. It enhances the model's ability to model long-range dependencies, improve performance on complex tasks such as machine translation, text generation, and image captioning, and allows for efficient parallel computation during training and inference.

6. Results

Table 2 presents the rankings and performance of different teams in MULTI-Fake-DetectiVE Task 1, based on their weighted average F1 scores. The higher the F1 score, the better the performance of the team in that task. The "Weighted Avg. F1-Score" column displays the corresponding F1-score achieved by each team. The F1-score is a measure of a model's accuracy, combining precision and recall, and the weighted average takes into account the relative importance of each class in the evaluation.

7. Conclusion

In conclusion, the field of multi-modal fake news detection has witnessed significant research efforts aimed at detecting and verifying fake news in environments that involve both textual and visual elements. The studies mentioned above highlight the importance of considering multiple modalities, such as text and images, to enhance

Table 2
Result of MULTI-Fake-Detective Task 1 - Multimodal Fake News Detection

Rank	Team-Run	Weighted Avg. 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
11	**HIJLI-JU-CLEF-Bi-LSTM**	**0.314**

the accuracy and effectiveness of fake news detection systems.

One key aspect addressed by these studies is the combination of textual and visual content analysis to identify patterns, inconsistencies, and misleading information indicative of fake news. By analyzing the language used in news articles, social media posts, and the visual content accompanying them, researchers aim to detect and characterize fake news more comprehensively.

In conclusion, our system leveraged the combination of text and image analysis, employing state-of-the-art techniques in NLP and computer vision to detect and verify fake news in a multimodal environment. While we acknowledge the challenges associated with this task, our system demonstrates promising results and serves as a foundation for further advancements in combating fake news in multimedia contexts.

References

[1] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective, ACM SIGKDD Explorations Newsletter 19 (2017) 22–36.

[2] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, Y. Choi, Truth of varying shades: Analyzing language in fake news and political fact-checking, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2931–2937. URL: https://aclanthology.org/D17-1317. doi:10.18653/v1/D17-1317.

[3] N. Vo, K. Lee, The rise of guardians: Fact-checking url recommendation to combat fake news, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 275–284. URL: https://doi.org/10.1145/3209978.3210037. doi:10.1145/3209978.3210037.

[4] B. Singh, D. K. Sharma, Predicting image credibility in fake news over social media using multimodal approach, Neural Computing & Applications 34 (2022) 21503–21517. URL: https://doi.org/10.1007/s00521-021-06086-4. doi:10.1007/s00521-021-06086-4.

[5] 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.

[6] 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.

[7] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft COCO: common objects in context, CoRR abs/1405.0312 (2014). URL: http://arxiv.org/abs/1405.0312. arXiv:1405.0312.

[8] S. Sarkar, D. Das, P. Pakray, D. Pinto, A hybrid sequential model for text simplification, in: N. Priyadarshi, S. Padmanaban, R. K. Ghadai, A. R. Panda, R. Patel (Eds.), Advances in Power Systems and Energy Management, Springer Nature Singapore, Singapore, 2021, pp. 33–42.