2023 IEEE 6th International Conference on Multimedia Information Processing and Retrieval (MIPR) | 979-8-3503-0781-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/MIPR59079.2023.00023

Transformer Based Image-Text Consistency Analysis for Infographic Articles

Yuwei Chen Ming-Ching Chang
University at Albany, State University of New York, NY 12222, USA
ychen69@albany.edu mchang2@albany.edu

Abstract

We present a multi-modal T5 Transformer-based method for image-text semantic consistency analysis that are targeted at infographic articles. Infographics provide graphical presentations of information and have wide applications in visualization and news media. Many media forensics methods are developed for consistency analysis among texts and images, however methods that work specifically for infographic semantic consistency are still scarce. We integrate the Natural Language Inference (NLI) of the T5 Transformer with scene text extraction for infographic consistency analysis. Experimental study is performed on the DARPA Semantic Forensics (SemaFor) Evaluation dataset for the task of news article inconsistency detection. The test set contains 248 infographic articles that are evenly split between manipulated and pristine samples. For evaluation, low False Positive Rate (FPR) is desired in media forensics. We use the metrics of True Positive Rate (TPR) at low 0.08 FPR, as well as the standard ROC AUC and Equal Error Rate (EER). Results show that our method outperforms mainstream methods from other SemaFor performer teams.

1 Introduction

Semantic consistency analysis is an important central pillar in multi-modal/cross-modal media forensics [6, 17] and fake news detection [2, 14]. From the image captioning task [7] to longer form media such as news articles, the ability to process and analyze the different aspects of media has increasing importance. With the rise of daily media consumption, the ability to authenticate the contents of the news or social media is crucial. The advancement of Deep Learning (DL) and Natural Language Processing (NLP) had enabled the development of multi-model methods to automatically screen articles for semantic consistency analysis among the texts, images, captions, and paragraphs for content verification. With the debut of the Large Language Models (LLM) [16] such as the GPT-4 that can handle both images and long-form texts, the multi-modal consistency analysis technology can evolve into an important building block for content analysis and misinformation detection.

There are many variables to consider when analyzing the contents of the media, which might include multiple modalities such as images, texts, charts, audios, videos, and metadata. In this paper, we specifically focus on articles with **infographics** and analyze the media consistency in the image and text domains; see Fig. 1 for an overview. Infographics provide rich graphical presentations of information and data, which have wide usage in visualization, news and social media. Infographics take various forms, with the two main groups being those including summary texts and those featuring graphs or charts. We aim to develop a multi-modal method that can handle articles with *long-form* paragraphs. In contrast, CLIP [7] based methods can only handle a single image-text pair with short text length. Our method is distinct from the very large scale foundation models such as GPT-4, that requires huge amount of training data and computation resources.

Traditional semantic consistency analysis [17] is often carried out on media with everyday images, and those methods does not work on inforgraphic images. Infographics often include scene texts that requires additional inference to extract semantic meaning; see Fig. 3 for examples. Extracting semantic meaning from values within bar and line graphs necessitates associating them with their corresponding topic, which is a non-trivial task. We hypothesize that scene text extraction from the infographics can be fed into a general NLP analysis module for consistency analysis. Our approach comes with an advantage of versatility, that (1) the NLP module can be trained on a large corpus that are purely text-based, (2) scene text detection and recognition on the infographics can be reliably trained using synthetic data augmentation, and (3) the infographic consistency dataset can also be easily produced and augmented using media containing infographics that are available online.

Our method takes an input article that may contain a title, one or more infographic charts (each might come with a caption), and a few body paragraphs. The method then determines a semantic consistency score among the components. For example, a wrongly inserted figure that is out of context of the article can be detected this way. Fig. 2 provides an overview of our proposed pipeline, which starts with scene texts detection and recognition from the infographic image, and performs natural language preprocessing on the text parts (title, image captions, body

2770-4319/23/$31.00 ©2023 IEEE
DOI 10.1109/MIPR59079.2023.00023

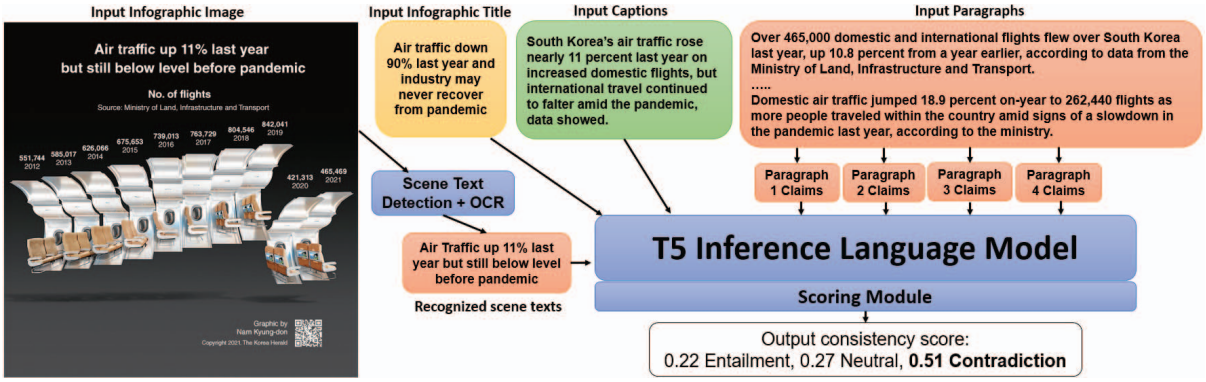

Figure 1: Overview of the proposed infographic image-text semantic consistency analysis pipeline.

paragraphs). All localized textual claims are fed into the Natural Language Inference (NLI) module of the T5 Transformer [9] to estimate a consistency score among the pairs of image-title, image-paragraph, *etc.* Finally, the semantic consistency scores are calculated after condensing and normalization.

This work is part of the efforts of the DARPA Semantic Forensics (SemaFor) program [11] [1] that aims to detect visual media manipulation and combat falsified information at large scale. The SemaFor media forensics process comprises three goals, namely the detection, attribution, and characterization of media forgery. To our knowledge, performers within the SemaFor program represent the State-of-The-Art (SoTA) research and development teams on the topic of multi-modal media consistency analysis. Due to the lack of suitable public datasets, we report evaluation and comparison of our method to other SemaFor performer teams on the task of *semantic consistency analysis of infographic news articles*.

We summarized our contributions in the following:

1. To our knowledge, this is the first multi-modal method that integrates infographic scene text from the image domain with NLP T5 Transformer for semantic consistency analysis. The proposed pipeline works well on practical, real-world news articles.
2. Experiments on the SemaFor evaluation dataset shows the our method outperforms significantly over the other mainstream methods within the DARPA program.
3. Additional ablation study shows the individual benefits of our design in the selected modules of DB [4] scene text, MORAN [5] OCR, and the effect of the T5[9] Transformer.

2 Related Works

Multi-modal image and text analysis has long been a standing challenge for the semantic consistency. Traditional approaches focus on shorter-form media consisting of a single image and a caption or brief description. Many new methods have shifted focus to longer-form media, which introduces additional challenges such as semantic consistency localization, converting diverse modalities to the same semantic context, and greater inference-based reasoning. These newer methods usually fall into two categories.

The first category involves exploring ways to combine the image and textual feature space to create a more complex multi-modal semantic space. Models such as CLIP [7] use two separate encoders to process both the image and text domains, then the two semantic spaces are combined for inference. Although powerful and data-driven, such end-to-end methods suffer from a lack of explainability regarding where and why the image and text are inconsistent.

The second category involves creating a semantic **knowledge graph** based on objects, attributes, and actions within the image. These methods often offer much higher explainability. Knowledge graph-based methods, such as Face-KEG [12], leverage large knowledge graphs to perform fact-checking across various domains. Similarly, InfoSurgeon [2] creates a dynamic knowledge graph based on the scene presented within the news article. This allows the model to compare semantic consistency between localized paragraphs while preserving high-level explainability. The trade off for knowledge graph based methods is that it often requires an external knowledge base to make inference decisions. The size and quality of the external knowledge base are essential for this category of methods, as the accuracy of the consistency score relies solely on the semantic relationships in the knowledge base. Hence, knowledge graphs like ClaimsKG [10] have complex claim models to ensure that the relationships are accurately and efficiently represented.

Both categories of methods established the foundation for general image-text consistency analysis, but there has been limited research on the specific challenge of analyzing articles with infographics. Infographic based media contain images with scene texts, often in the form of graphs or charts. Our findings suggest that existing SoTA methods, which do not specifically address infographic-specific media, face challenges in detecting semantic inconsistencies in news media. For instance, methods such as VinVL [15]

[1] https://www.darpa.mil/program/semantic-forensics

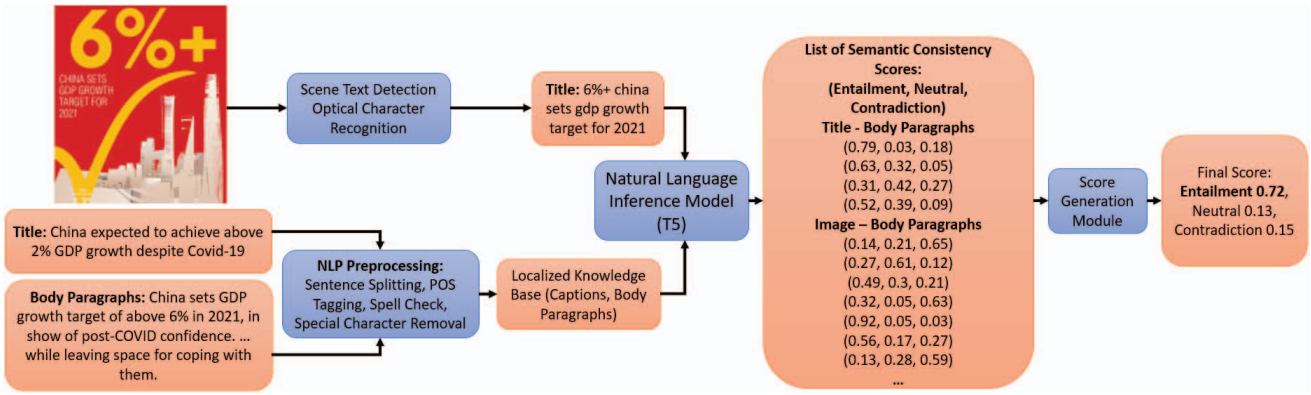

Figure 2: **An example showing the proposed infographic consistency analysis pipeline.** Infographic images are processed by the scene text extractor. Titles and body paragraphs are then pre-processed into a knowledge base. The T5 inference model is then used to verify all textual claim pairs based on the structure of the input according to Table 1. All inference scores are then normalized and condensed using the score generation module to create a final semantic consistency score for the whole article.

and CLIP [7] are designed to analyze general image and text media, but they struggle to detect manipulations in the scene text of the images since they do not explicitly focus on scene text. In particular, CLIP has faced difficulties in analyzing text-based images as it cannot disambiguate the intended meaning of the words.

3 Method

We introduce a two-stage, multi-modal method that utilizes scene text extraction and a T5 Transformer [9] based language model. Our method extracts all relevant scene text from the images in the news media and converts semantic claims into the textual domain. We then use the T5 Natural Language Inference (NLI) model to verify all textual claims. Our approach can effectively capture the semantic meaning of the scene text, thus addresses the limitations of existing main stream methods that do not work well on infographic articles.

An article comprising of infographics and body texts is first separated into distinct images and texts, such that consistency scores can be calculated among localized claims. The format and length of the infographic media being processed determine how we structure the premise/hypothesis statements for NLI. Scene texts are extracted from the infographic image (§ 3.1). The resulting text claims are then passed on to the T5 NLI module for semantic claim generation (§ 3.2). This list of semantic pair inference scores are then normalized and condensed into a single consistency score for the article (§ 3.3).

3.1 Scene text extraction from infographic images

We employ a fine-tuned version of Differentiable Binarization (DB) [4] for scene text detection. The resulting localized text regions are then forwarded to a fine-tuned MORAN [5] character recognition module, to segment and recognize possible texts in the area. Subsequently, we use a series of rule-based techniques to combine the recognized characters and generate meaningful keywords or sentences. The resulting list of keyword/sentence candidates is then filtered based on the length of each claim, with a chosen length filter of 5 or more words. The filtered sentence claims are sent to the textual language model for inference.

News Article Input	Premise	Hypothesis
Title Paragraphs	Paragraphs	Title
Title Images	Images	Title
Images Paragraphs	Paragraphs	Images
Title Images Paragraphs	Paragraphs	Title and Images

Table 1: The **Premise Hypothesis Structure** shows which claim within the news media should be verified for semantic consistency based on the given input. Based on what information is provided, the T5 language model is feed specific sections of the articles as the premise or hypothesis for consistency analysis.

3.2 Natural language inference score generation

In the text domain, we leverage a fine-tuned version of the T5 Transformer [9] for language feature extraction. We opted for the T5 Transformer due to its adaptability, performance, and versatility in performing multiple tasks using a single model. We start with a T5 model that is pre-trained on both the common crawl-based dataset of Colossal Clean Crawled Corpus (C4) [8] and the Multi-NLI [13] datasets. Thereafter, we fine-tuned the Transformer further by using the Stanford Natural Language Inference (SNLI) [1] dataset to enhance its NLI capabilities. The SNLI dataset contains 570K human written English sentence pairs that are manually annotated and balanced.

The input body text is initially segmented into distinct localized paragraphs and captions, which are further partitioned into individual textual claims. Upon receiving the extracted scene text claims from the image modules, the NLP modules authenticate the claims via the T5 NLI inference based on the textual claims from the paragraphs. Table 1 il-

SemaFor Performer ID	AUC	TPR @ 0.08 FPR	EER
1	0.72	**0.43**	0.36
2	0.74	0.42	0.34
3	0.79	0.35	0.24
4	**0.86**	0.3	**0.24**
5	0.6	0.26	0.44
6	0.27	0.12	0.71
7	0.54	0.1	0.47
8	0.59	0.1	0.43
Our Method	**0.84**	**0.61**	**0.24**

Table 2: (I) Title and image consistency results.

SemaFor Performer ID	AUC	TPR @ 0.08 FPR	EER
1	0.73	**0.43**	0.36
2	0.63	0.21	0.43
3	0.79	0.35	**0.24**
4	**0.84**	0.38	0.26
5	0.64	0.34	0.42
6	0.27	0.12	0.71
8	0.59	0.17	0.47
9	0.56	0.19	0.44
10	0.48	0.09	0.52
11	0.46	0.07	0.52
Our Method	**0.79**	**0.57**	**0.33**

Table 3: (II) Title and body paragraphs consistency results.

SemaFor Performer ID	AUC	TPR @ 0.08 FPR	EER
1	0.67	**0.36**	0.43
3	0.76	0.32	**0.27**
4	**0.81**	0.25	0.28
5	0.59	0.25	0.45
6	0.31	0.16	0.7
8	0.51	0.09	0.49
Our Method	**0.63**	**0.21**	**0.45**

Table 4: (III) Image and body paragraphs consistency.

lustrates how our T5 inference model processes entailment pairs by categorizing certain claims as either hypotheses or premises. Following verification of all premise-hypothesis pairs by T5, a set of entailment scores is generated for each localized sentence claim within the article.

3.3 Condensing and normalizing score

Let n_P denote the number of Premise sentences, and n_H denote the number of Hypothesis sentences. The NLI model estimates the semantic consistency scores for the total of $N = n_P \times n_H$ sentence pairs $\{P_i | i = 1, ..., N\}$ by nesting the number of Premise and Hypothesis sentences. This forms N Premise-Hypothesis pairs. What textual statements are considered premises or hypothesis are shown in Table 1. For each Premise-Hypothesis pair P, the T5 Transformer generates a tuple of three semantic consistency scores s reflecting the semantic *entailment* e, *neutrality* n, and *contradiction* c, namely, $s = (e, n, c)$. To estimate the semantic consistency score S_A for the entire article, we perform a weighted average of each type of score in s over all possible N sentence pairs from the article. The weight w_i for each sentence pair P_i indexed by i is calculated by performing a word-level comparison between the Premise and Hypothesis. This design aims to reward sentences with larger similarity in terms of entities, attributes, and events.

An overall average is taken over all weighted scores to normalize the scores over the whole article:

$$A_e = \frac{\sum_{i=1}^{N}(w_i \cdot e_i)}{N}, A_n = \frac{\sum_{i=1}^{N}(w_i \cdot n_i)}{N}, \quad (1)$$

$$A_c = \frac{\sum_{i=1}^{N}(w_i \cdot c_i)}{N} \quad (2)$$

The final article-level semantic consistency type is determined by finding the maximal score among the three types, *entailment* A_e, *neutrality* A_n, or *contradiction* A_c. The article-level semantic score $S_A = \max(A_e, A_n, A_c)$.

4 Experimental Evaluation

This work is part of the DARPA SemaFor program development, and evaluation is performed as a SemaFor evaluation done by an independent evaluator institution. Our method underwent evaluation in the form of the news article inconsistency detection task, which aimed to identify semantic inconsistencies present within the infographic specific news article. This task was further subdivided into four distinct evaluation sub-tasks, each focusing on different sections of the news media. The purpose behind these subtasks is to evaluate how well different methodologies adapt to being presented with limited sections of the news media. § 4.1 reports the SemaFor evaluation results. § 4.2 provide additional ablation study of our methods.

SemaFor Performer ID	AUC	TPR @ 0.08 FPR	EER
1	0.68	0.36	0.43
2	0.63	0.21	0.43
3	0.74	0.32	0.3
4	**0.85**	**0.4**	**0.23**
5	0.56	0.19	0.47
6	0.32	0.16	0.72
8	0.59	0.18	0.45
9	0.56	0.2	0.44
10	0.48	0.09	0.53
11	0.47	0.07	0.52
Our Method	**0.79**	**0.57**	**0.33**

Table 5: (IV) Title, image, and body paragraphs consistency results.

4.1 DARPA SemaFor evaluation results

Test dataset and methods for comparison. Each of the four SemaFor evaluation sub-tasks comprises a total of 248 samples, with 124 of them being manipulated (Contradiction) and the remaining 124 being authentic (Entailment) articles. Every news media sample includes a title, multiple images, captions, body paragraphs, and meta data. We note that, in accordance with program guidelines, the origins of other performer teams/methods cannot be disclosed unless they are publicly published. We refer to all participating methods with a consistent Performer ID across all evaluations. Participating teams are provided with a subset of these media assets to conduct forensic detection, with any possible falsifications guaranteed to be present within the given sub-task media assets.

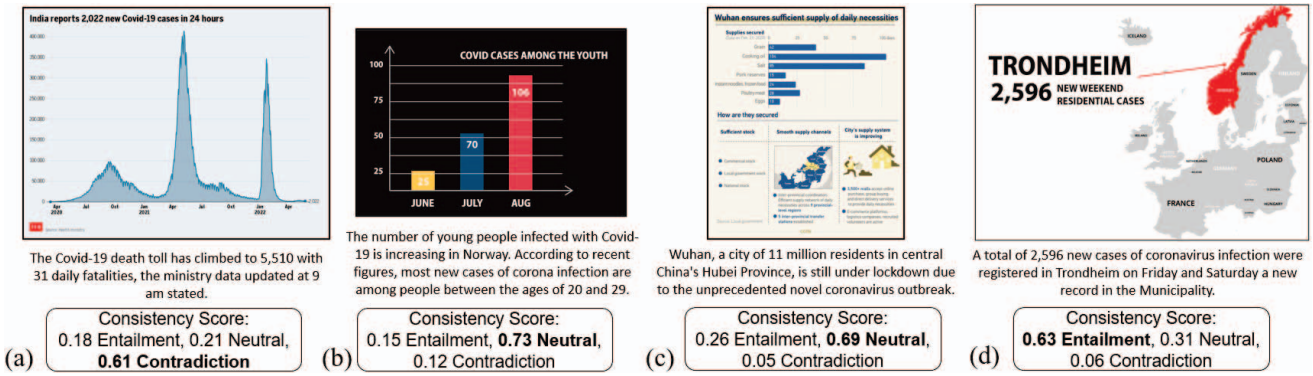

Figure 3: Examples of the infographic image caption of news articles for consistency analysis in the SemaFor evaluation test set: Additional images, captions, and body paragraphs are present in the news article along with meta data (not shown, which are used within the test set).

Evaluation metric. We employ various metrics of ROC analysis to assess the performance of our proposed method. In media forensic applications, the ability of a model to maintain high TPR while minimizing the FPR is crucial, as false positives can waste time, resources, and potentially mis-trust in the forensic model. To this end, we use the True Positive Rate (TPR) at the low 0.08 False Positive Rate (FPR) to measure the performance of methods in high liability situations. Lastly, we report the standard ROC-AUC and the Equal Error Rate (EER) to gain insight of how well each test method performs. We next report the evaluation results in the following **four** sub-tasks for the semantic consistency detection of infographic articles.

(I) Title and image consistency: Our method performed significantly better than most mainstream performers in the title image semantic consistency sub-task. As shown in Table 2, the only comparable method is Performer 4. However, our method's TPR at a low FPR is twice that of Performer 4. This is particularly significant for media forensics, where false alarms can have disastrous consequences in the fast-paced, click-bait-based internet ecosystem.

(II) Title and body paragraphs consistency: Similarly to the previous sub-task, Table 3 shows that our proposed method and Performer 4 outperform all other participating methods here. Our method achieves a significantly better TPR at low FPR compared to all other methods. However, the EER of our proposed method is higher than in the previous sub-task, at 33% compared to the 29% of Performer 4. This indicates that Performer 4's TPR and FPR are slightly more balanced than our method. We argue that our significantly better performance in low FPR outweigh the 4 % difference in EER in the context of media forensics.

(III) Image and body paragraphs consistency: It is evident in Table 4 that this task was the most challenging for all SemaFor performers. The performance of our method decreased significantly in this sub-task. However, our results are still comparable to other performers' methods. The decrease in performance is likely due to the larger number of premise-hypothesis claims that need to be verified based on the premise-hypothesis structure shown in Table 1.

(IV) Title, image, and body paragraphs consistency: Table 5 shows that our method perform significantly better at low FPR making it more reliable in high liability conditions. The remaining metrics such as AUC and EER is in second place slightly behind Performer 4.

Discussion. Our method overall outperforms all other participating methods in all three metrics across all four sub-tasks. Although Performer 4 has comparable results, it falls significantly behind in low FPR conditions. An insight gathered from these results is that detecting infographic text semantic consistency in news articles is very challenging. Most participating methods have near-random AUCs as seen in Fig. 4. We believe this is due to the language, number of moving parts, and infographic complexities within the articles. The models need to accurately extract the infographic and textual semantic features and correctly make inference. This is particularly challenging, as no manipulation infographic news article datasets are publicly available. We believe our high performance in low FPR is attributed to the integration of scene text extraction and language inference.

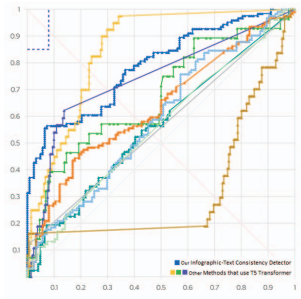

Figure 4: Title, image, and body paragraph consistency ROC

Qualitative Analysis. We next discuss some observed qualitative traits of our method. Regarding strengths, our model was able to detect the numeric difference in COVID cases between the title and caption in Fig. 3(a). In terms of weaknesses, our method struggled with chart-based recognition, shown in Fig. 3(b). We were not able to associate the increasing bar graph values to increasing COVID cases. Additionally, our model has the ability to reason in the neutral case. This is shown in Fig. 3(c). While the image scene text and caption is not directly consistent, our model is able

Sub-Modules	F1	AUC
CLIP [7]	0.56	0.51
OSCAR [3] and T5	0.62	0.59
Our method using DB [4], MORAN [5], and T5 [9]	**0.78**	**0.72**

Table 6: **Ablation Study** on the effectiveness of scene text recognition and the T5 language model toward the performance. CLIP was used for a baseline representing results without any of our sub-modules. OSCAR was used for image captioning to replace the infographic scene text OCR.

to recognize that the majority of the claims are neutral.

4.2 Ablation Study

We conducted an ablation study to gain a deeper understanding of the impact of the sub-modules on the overall performance of our method. We evaluate the replacement of our scene text extraction models with the OSCAR [3] and utilized CLIP [7] as a baseline image text consistency model. The purpose of this experiment is to assess the significance of considering scene text within the image, while analyzing semantic consistency in the infographic articles. We evaluated these new hybrid methods on a real-world dataset of 50 manipulated samples and 50 pristine samples of infographic news articles.

Table 6 presents a breakdown of scores for each hybrid method. Observe that there is a significant improvement in performance after integrating infographic scene text OCR into the pipeline. The results without OCR are close to random chance, which highlights the importance of capturing scene text for infographic semantic consistency analysis.

Limitations of our method. The quality of real-world media can vary significantly, with spelling mistakes, irrelevant information, personal bias, and grammatical errors. It is crucial for language models to be able to handle noisy inputs effectively. We have observed that our model struggles to handle noisy inputs. This limitation might stem from that our model being trained on the SNLI dataset, which contains smaller human-annotated research statements. We have opted to clean the input text before feeding it into our inference model. While this is not a perfect solution, it has helped to mitigate the above-mentioned issues.

5 Conclusion

We presented a two-stage multi-modal Transformer based method that detects semantic inconsistencies in infographic based articles. Our method significantly outperforms other mainstream methods within the four SemaFor evaluation sub-tasks. Our future work entails enhancing our model to recognize graphs and charts, refining our language model to better handle unstructured input, and augmenting the explainability of the model.

Acknowledgements. This work is supported by the U.S. DARPA Contract HR001120C0123. We thank William Corvey, Arslan Basharat, Kirill Trapeznikov, and Sam Blazek for providing guidance.

References

[1] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. 2015.

[2] Y. Fung, C. Thomas, R. G. Reddy, S. Polisetty, H. Ji, S.-F. Chang, K. McKeown, M. Bansal, and A. Sil. InfoSurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *ACL*, pages 1683–1698, 2021.

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

[4] M. Liao, Z. Wan, C. Yao, K. Chen, and X. Bai. Real-time scene text detection with differentiable binarization. In *Proc. AAAI*, 2020.

[5] C. Luo, L. Jin, and Z. Sun. MORAN: A multi-object rectified attention network for scene text recognition. *Pattern Recognition*, 90:109–118, 2019.

[6] S. McCrae, K. Wang, and A. Zakhor. Multi-modal semantic inconsistency detection in social media news posts. In *MMM*, 2022.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. *arXiv:2103.00020*, 2021.

[8] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv:1910.10683*, 2019.

[9] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(140):1–67, 2020.

[10] A. Tchechmedjiev, P. Fafalios, K. Boland, M. Gasquet, M. Zloch, B. Zapilko, S. Dietze, and K. Todorov. A knowledge graph of fact-checked claims. In *ISWC*, 2019.

[11] M. Turek and N. F. Johnson. Using AI to detect multimedia manipulation at scale. https://community.apan.org/cfs-file/__key/docpreview-s/00-00-17-41-39/Neil_5F00_Johnson_5F00_MediFor_2D00_SemaFor.pdf.

[12] N. Vedula and S. Parthasarathy. FACE-KEG: Fact checking explained using knowledge graphs. In *WSDM*, 2021.

[13] A. Williams, N. Nangia, and S. R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL*, pages 1112–1122, 2018.

[14] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi, and L. Wei. Detecting fake news by exploring the consistency of multimodal data. *Information Processing and Management*, 58:102610, 2021.

[15] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao. VinVL: Making visual representations matter in vision-language models. *CVPR*, 2021.

[16] W. X. Zhao, K. Zhou, J. Li, X. Wang, Y. Hou, and *et al.* A survey of large language models. *arXiv:2303.18223*, 2023.

[17] Z. Zhao, H. Zhu, Z. Xue, Z. Liu, J. Tian, M. C. H. Chua, and M. Liu. An image-text consistency driven multimodal sentiment analysis approach for social media. *IPM*, 2019.