Infotec+CentroGEO at Touché: MCIP, CLIP and SBERT as Retrieval Score

Tania Ramirez-delreal[1,2,*,†], Daniela Moctezuma[2,†], Guillermo Ruiz[3], Mario Graff[1,3] and Eric Tellez[1,3]

[1]SECIHTI, Secretaría de Ciencia, Humanidades, Tecnología e Innovación, Benito Juárez, Ciudad de México, 03940, México
[2]Centro de Investigación en Ciencias de Información Geoespacial (CentroGeo), Aguascalientes, Ags., 20213, México
[3]Centro de Investigación e Innovación en Tecnologías de la Información y Comunicación (INFOTEC), Aguascalientes, Ags., 20213, México

Abstract

This manuscript presents the Infotec+CentroGEO solution for the Image Retrieval/Generation for Arguments challenge at Touché-2025. This shared task asks for systems that can retrieve an image from a given dataset to support a given argument; these arguments include only a single claim without supporting premises. The team's solutions included the usage of MCIP, CLIP, and SBERT to obtain a value for the representations of the images and claims.

Keywords

Image Retrieval, MCIP, CLIP, SBERT

1. Introduction

Visual information plays a crucial role in communication, often enhancing the impact and memorability of textual messages. This phenomenon, known as the *picture superiority effect* [1], suggests that images are processed and recalled more effectively than words due to their distinct perceptual and conceptual qualities. Consequently, images are powerful tools for constructing arguments and improving comprehension and persuasiveness.

Hung et al. [2] studied the relevance judgment of journalists when searching for photographs to support news stories; several key criteria for image search and selection were identified, highlighting the importance of associated textual information and personal feelings. On the other hand, Wang et al. [3] indicate how specialized applications have concentrated on examining the link between text messages and images in political campaigns circulated on social networks; studying the relationship between text, objects, and colors to comprehend the propaganda strategy. The work of Kiesel et al. [4] on image retrieval employs computational techniques that use argument mining, justify the use of images as evidence, and improve information retrieval and comprehension in argumentative contexts.

As described in [5, 6], the task of Touché 2025 involves selecting the best image to support an argument. During our data analysis and exploration, we saw that several images could equally support a single argument; nevertheless, the competition organizer did not share any ground truth data, so it was not easy to assess our solutions automatically.

So, we think the task proposed in the Touché 2025 competition is very complex because the dataset has many images to pair with each claim, some of which are very similar. A solution cannot apply supervised learning directly due to the lack of a training set to help learning models; therefore, tasks require heuristics that replicate human image search and selection strategies. The response solutions are

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
†These authors contributed equally.
tramirez@centrogeo.edu.mx (T. Ramirez-delreal); dmoctezuma@centrogeo.edu.mx (D. Moctezuma); luis.ruiz@infotec.mx (G. Ruiz); mario.graff@infotec.mx (M. Graff); eric.tellez@infotec.mx (E. Tellez)
0000-0002-1638-5086 (T. Ramirez-delreal); 0000-0003-3038-4642 (D. Moctezuma); 0000-0001-7422-7011 (G. Ruiz); 0000-0001-6573-4142 (M. Graff); 0000-0001-5804-9868 (E. Tellez)

manually evaluated through a human committee. So, in this manuscript, we describe all the approaches we used trying to select the best possible pair of image and argument to finally submit our best solutions to the human evaluation coordinated by the organizers of the Touché 2025 competition. In our case, only a solution based on image retrieval was submitted because although we tested some image generation models, we considered not submitting a generation image approach because of our limitations in computational resources and the time required.

This manuscript is organized as follows. Section 2 describes the task in which the team participated and the dataset the organizers provided. Section 3 shows the approaches proposed for retrieving images for arguments. In Section 4, the evaluation of the results is done with our tiny labeled dataset. Finally, Section 5 concludes the manuscript and discusses future directions.

2. Task description

The Image Retrieval/Generation for Arguments challenge corresponds to Task 3 in Touché 2025 [5]. Given a claim, the objective is to find relevant images from a provided image database and rank them according to how much these images can be used as an argument to support that claim.

Each image includes metadata automatically extracted from the picture using AI models. The metadata includes the text, labels, and keywords extracted from the image using Google Cloud Vision and a caption provided by LLaVA. Human experts will evaluate submissions according to defined relevance criteria. The results were uploaded to the TIRA platform [7].

The dataset is composed of images depicting scenes or visual concepts, as well as their associated texts, organized into reference captions representing the content of each image. It contains 32,339 images, with 128 claims derived from 27 different topics. Each claim is uniquely recorded in an XML file, although a topic may appear in multiple arguments. Each image includes its respective metadata, and in this study, text and captions were utilized to improve the representation of the images.

3. Methodology

We tried some state-of-the-art models such as CLIP [8], MCIP [9], and Sentence BERT [10]. In the following sections, each approach will be explained in detail.

3.1. A tiny labeled dataset for guiding our approach

We have to manually classify a small subset with 13 claims, from 3 topics, containing 995 images, to compare and determine whether our strategies improve our results quantitatively. We decided to build this ground truth data to have a guide to assess our efforts. The 13 claims (see Table 1) provided in the first stage were used to select among 995 images (also provided in the first stage of the competition, in a dataset called tiny version) the top three that most support each claim, so for each claim, we selected the intersection of the top three selected by two humans, to establish the final top three images per each argument. So, in this labeling produced at least two evaluations per image concerning claims, and then we selected the five best images for each argument.

After this manual annotation, different techniques for retrieving the images were evaluated. The architectures used are explained in the following subsections. The score was computed as if the top n images provided by the model were in the top three images manually selected; it was considered a success. The score has a higher probability when n is also higher. Although we did not measure the inter-annotator agreement, we consider it high enough because both persons have a high intersection between their selected images.

Table 1
Claims in the tiny dataset provided in the first stage

ID	Topic	Claim
1-1	Automation in the Workforce	Automation increases work efficiency
1-2	Automation in the Workforce	Automation allows workers to focus on complex tasks
1-3	Automation in the Workforce	Automation increases productivity in industries
1-4	Automation in the Workforce	Automation reduces human error in repetitive tasks
1-5	Automation in the Workforce	Automation leads to significant job displacement for workers
2-1	Renewable Energy	Renewable energy production is dependent on weather conditions
2-2	Renewable Energy	Fossil fuels contribute heavily to pollution
2-3	Renewable Energy	Renewable energy sources are virtually limitless
2-4	Renewable Energy	Renewable energy reduces greenhouse gases
3-1	Social Media's Role in Society	Social Media contributes to mental health issues
3-2	Social Media's Role in Society	Social media connects people across diverse cultures
3-3	Social Media's Role in Society	Social media connects people across countries
3-4	Social Media's Role in Society	Social media fosters cyberbullying

3.2. Multimodal encoding models

CLIP

Contrastive Language-Image Pretraining (CLIP) [8] is a multimodal model that is used to predict whether a pair, image and text is related or not. CLIP was trained on text paired with images on the internet, and one important limitation mentioned by its authors is the presence of social biases in the model because this pair of image-text was used unfiltered and uncurated; nevertheless, the results achieved by CLIP are impressive most of the time. Specifically, we used the CLIP version of OpenAI,[1] with the `ViT-L-14-336` model. We used the pair, image, and text, and obtained the similarity between them, so we sorted these similarities and chose the top five for the submission.

MCIP

Schall et al. [9] introduce the Multi-Caption-Image-Pairing (MCIP), while it is similar to CLIP, here the image encoder is trained with a different loss function and a collection of related captions per image. The text encoder is frozen to maintain alignment of the text and image embeddings.

Argumentative information is extracted from the claims, which constitutes a textual proposition. Each image could be a good illustration of the claim; the model used is the same as that used for the caption. This transforms each claim into a normalized semantic vector that can be compared with both the image and the reference captions.

Then, we derived textual propositions (arguments) from claims. Each image may effectively demonstrate the claim, using the same model employed for captions. This process converts each claim into a standardized semantic vector, enabling comparison with both the image and the associated captions; thus, we are able to evaluate semantic similarity by comparing the image vector with both the claim and the caption embeddings. This is done by applying cosine similarity, which evaluates how close the vectors are in the shared semantic space.

This methodology aims to determine the relevance of a claim relative to the visual content of an image by using reference captions as a context. A score is generated to measure the semantic compatibility of each claim concerning the image and its caption. Ultimately, results are organized that associate each image with a list of scores that reflect the similarity between the image and all claims. These scores can be ranked to identify the top five images with the highest score, which implies a stronger semantic association between the claim, the image, and its caption.

[1] https://github.com/openai/CLIP/

3.3. Text encoders

Sentence BERT (SBERT) [10] is a language model tailored for tasks involving similarity. Built on the foundation of a BERT model, SBERT is further optimized to transform text into vectors in such a way that semantically similar text points correspond within the target latent space. Training of SBERT involves semantically related pairs and triplets, utilizing a siamese neural architecture that encloses the base BERT model, along with latent-space pooling and a transformation to align vectors with similar text sentences.

A different approach we tried was to use text-based information only. Instead of using actual images, we concatenated the text within the image with the caption provided as input to the Sentence BERT model; in particular, we used the `multi-qa-mpnet-base-dot-v1` variant. So we obtained an embedding for each image that captures its semantic information. We then concatenated the topic with the claim and obtained the corresponding embeddings. We normalized all the embeddings from the images and claims. Finally, for each claim, we used cosine similarity to rank the image embeddings and retrieved the top five.

4. Results

As mentioned above, we did our own tagged dataset with a subset of the data provided by the competition organizers. This internal evaluation provided some information on the results of our efforts.

Table 2 shows the results of this evaluation. In the headings, we have all the different claims that are used to develop our models (which means the argument IDs). MCIP, CLIP, and SBERT refer to the corresponding models explained before and k is indicated, for which the top values k generated by the model were considered.

Table 2
Internal results using our labeled dataset

	1-1	1-2	1-3	1-4	1-5	2-1	2-2	2-3	2-4	3-1	3-2	3-3	3-4	Score
MCIP_k=5	0	0	1	0	1	1	0	0	0	1	0	1	0	0.3846
MCIP_k=3	0	0	1	0	1	1	0	0	0	0	0	1	0	0.3077
MCIP_k=1	0	0	1	0	0	0	0	0	0	0	0	0	0	0.0769
CLIP_k=5	0	0	1	0	1	0	1	0	0	1	0	1	0	0.3846
CLIP_k=3	0	0	1	0	1	0	1	0	0	1	0	1	0	0.3846
CLIP_k=1	0	0	1	0	0	0	0	0	0	0	0	1	0	0.1538
SBERT_k=5	1	1	1	1	1	1	0	0	0	1	1	0	0	**0.6154**
SBERT_k=3	1	1	0	1	1	0	0	0	0	1	1	0	0	**0.4615**
SBERT_k=1	1	0	0	1	1	0	0	0	0	0	0	0	0	**0.2308**

The score was calculated as follows; if one of the k best results is at least one of the three solutions manually established, it is considered a success (represented by 1), otherwise it is an error (represented by 0).

The k values are related to the number of images returned by the model, so the higher this k higher the probability of being (at least in one element) in our three manually chosen images. So, the score has the range value 0-1, 0 being the lowest possible result and 1 the contrary case. Then, the best results were obtained by SBERT $k = 5, 3, 1$.

Although SBERT reached the best result, we decided to send as the final solution the three outputs produced by CLIP, MCIP, and SBERT, respectively. The main reason we decided that is because MCIP and CLIP are very similar in their results and both considered image-text; on the other hand, SBERT is the unique solution text-based only and achieved our best results, so, as we had the opportunity to submit more than one solution, we took advantage of that.

According to the final results, our best performance was obtained by our CLIP solution, beating the results achieved by SBERT.

5. Conclusions

This work presented our proposals to tackle Touché in the CLEF2025 competition, a task asking language and vision models to find better images that support a specific claim (text).

Our methodology involved creating a tiny subset of arguments and images to guide our research efforts. In terms of models, we evaluated two multimodal models, a text-based language model and one text-based only model. Our most successful strategy used SBERT to evaluate and align images (their textual descriptions) with claims, allowing us to identify the top five images related to each claim. This undertaking is particularly complex due to the necessity for human assessment, which presents considerable obstacles in refining models to achieve specific objectives.

Our results suggest that the retrieval problem presents significant challenges and necessitates further research to address and meet individuals' informational needs effectively.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly and Writefull's model for grammar and spelling checks. After using these services, the authors reviewed and edited the content as needed and assumed full responsibility for the content of the publication.

References

[1] E. B. Thomas, Visual superiority effect in memory (pictures are worth a thousand words), in: Proceedings of the Human Factors Society Annual Meeting, volume 27, SAGE Publications Sage CA: Los Angeles, CA, 1983, pp. 714–714.

[2] T.-Y. Hung, C. Zoeller, S. Lyon, Relevance judgments for image retrieval in the field of journalism: A pilot study, in: Digital Libraries: Implementing Strategies and Sharing Experiences: 8th International Conference on Asian Digital Libraries, ICADL 2005, Bangkok, Thailand, December 12-15, 2005. Proceedings 8, Springer, 2005, pp. 72–80.

[3] M.-H. Wang, W.-Y. Chang, K.-H. Kuo, K.-Y. Tsai, Analyzing image-based political propaganda in referendum campaigns: from elements to strategies, EPJ Data Science 12 (2023) 29.

[4] J. Kiesel, N. Reichenbach, B. Stein, M. Potthast, Image retrieval for arguments using stance-aware query expansion, in: Proceedings of the 8th workshop on argument mining, 2021, pp. 36–45.

[5] J. Kiesel, Ç. Çöltekin, M. Gohsen, S. Heineking, M. Heinrich, M. Fröbe, T. Hagen, M. Aliannejadi, T. Erjavec, M. Hagen, M. Kopp, N. Ljubešić, K. Meden, N. Mirzakhmedova, V. Morkevičius, H. Scells, I. Zelch, M. Potthast, B. Stein, Overview of Touché 2025: Argumentation Systems, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 16th International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2025.

[6] B. Ionescu, H. Müller, D.-C. Stanciu, A.-G. Andrei, A. Radzhabov, Prokopchuk, Ştefan, Liviu-Daniel, M.-G. Constantin, M. Dogariu, V. Kovalev, H. Damm, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. G. Pakull, B. Bracke, O. Pelka, B. Eryilmaz, H. Becker, W.-W. Yim, N. Codella, R. A. Novoa, J. Malvehy, R. J. Dimitrov, Dimitar Das, Z. Xie, H. M. Shan, P. Nakov, I. Koychev, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, D. Fabre, C. Macaire, B. Lecouteux, D. Schwab, M. Potthast, M. Heinrich, J. Kiesel, M. Wolter, B. Stein, Overview of imageclef 2025: Multimedia retrieval in medical, social media and content recommendation applications, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 16th International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2025.

[7] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. doi:10.1007/978-3-031-28241-6_20.

[8] 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.

[9] K. Schall, K. U. Barthel, N. Hezel, K. Jung, Optimizing CLIP models for image retrieval with maintained joint-embedding alignment, in: E. Chávez, B. Kimia, J. Lokoc, M. Patella, J. Sedmidubsk (Eds.), Similarity Search and Applications - 17th International Conference, SISAP 2024, Providence, RI, USA, November 4-6, 2024, Proceedings, volume 15268 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 97–110. doi:10.1007/978-3-031-75823-2_9.

[10] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410.pdf.