DIMEMEX 2025: Solution based on Open-Clip of the INFOTEC+CentroGeo team

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

[1] *Centro de Investigación en Ciencias de Información Geoespacial (CentroGeo), Aguascalientes, Ags., 20213, Mexico*
[2] *SECIHTI, Secretaria de Ciencia, Humanidades, Tecnologia e Innovacion, Benito Juárez, Mexico City, 03940, Mexico*
[3] *INFOTEC Centro de Investigación en Tecnologías de la Información y Comunicación, 112 Circuito Tecnopolo Sur, Parque Industrial Tecnopolo 2, Aguascalientes, 20326, México.*

Abstract

This manuscript presents the INFOTEC+CentroGeo team's solution for task 1 of the DIMEMEX@IberLEF2025, which asks for solutions to a three-way classification problem using a meme and its text as data. The categories included are hate speech, inappropriate content, and harmless content. Our approaches include the use of CLIP models, our EvoMSA framework, and traditional machine learning models with different encodings for text and image data. Our result in the internal evaluation was a f1-score of 0.75 yet our result in the gold standard was relatively low, achieving a f1-score of 0.42. However, we draw a set of useful recipes and conclusions to help people select models for multimodal tasks.

Keywords

DIMEMEX, Meme classification, Text and image representation, CLIP models

1. Introduction

Memes are a well-known way to communicate ideas, especially on social media platforms. A meme is an image that includes a small text; the sense of a meme is usually humorous, funny, satirical, or ironic. Using these multimodal data (image and text) to tackle the problem of hate speech, inappropriate, and harmless content has become more frequent in the natural language processing research community. Nevertheless, this kind of problems are also addressed through text; unfortunately, hate speech and harmless content are very common phenomena in social networks, particularly Twitter (now X), Facebook, Youtube, and also in more professional networks such as LinkedIn.

Chetty and Alathur [1] define hate speech as "any speech that attacks an individual or a group with the intention of hurting or disrespecting based on identity." Paz et al. [2] present a systematic review surveying different proposals dealing with the recognition of hate speech using the Web of Science as a source of information. This task has also been deeply studied with Social Media data, for instance, Chetty and Alathur [1] presented a review of several methodologies applied to data from Twitter and Facebook. The main conclusions of this review are that laws on hate speech are not the same in all countries, and there are also a variety of communities dealing with this kind of discrimination on social networks.

In the harmless content detection task, the community has made some efforts. Arora et al. [3] survey the existing methods and suggests future research lines for this topic. Kirk et al. [4] developed an analytical scheme to categorize harms in text in three aspects: (1) type of harm (misinformation, racial stereotypes), (2) if harm is sought or intentionally addressed, and (3) who the affected people are.

IberLEF 2025, September 2025, Zaragoza, Spain

[*] Corresponding author.

[†] These authors contributed equally.

dmoctezuma@centrogeo.edu.mx (D. Moctezuma); tramirez@centrogeo.edu.mx (T. Ramirez-delreal); eric.tellez@infotec.mx (E. Tellez); mario.graff@infotec.edu.mx (M. Graff); luis.ruiz@infotec.mx (G. Ruiz)
0000-0003-3038-4642 (D. Moctezuma); 0000-0002-1638-5086 (T. Ramirez-delreal); 0000-0001-5804-9868 (E. Tellez); 0000-0001-6573-4142 (M. Graff); 0000-0001-7422-7011 (G. Ruiz)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2025-12-10

Also, some research has classified memes containing text and image into hateful or not-hateful, depending on whether the message is funny. Some works use feature extraction; in some cases it could be with attention mechanism methods or classic methods of machine learning [5]. Another approach dealing with multimodal data is presented in [6], where it compares and fine-tunes VisualBERT pretrained on several datasets, such as the Conceptual Caption dataset. For images was used the ResNeXT-152 Aggregated Residual Transformations–based Masked Regions with Convolutional Neural Networks was used, and for the text representation, the well-known uncased BERT (Bidirectional Encoder Representations from Transformers) model.

In addition, Burbi et al. [7] present a method called ISSUES, which is based on a pretrained CLIP model. Outstanding results were reached with two datasets, the Hateful Memes Challenge and HarMeme.[1] After reviewing different approaches, one can conclude that the image and text are not homogeneous and that the Transformer-enhanced ensemble models perform better, see [8]. Nevertheless, in our case, we tested both a more traditional approach (EvoMSA) and CLIP to represent the combination of par-associated image and text, trying to explore which one achieved better results in this particular problem.

Furthermore, this is a very complex task because sometimes the same image with different text is about two different and opposite classes; this is not the case of the dataset provided by the DIMEMEX 2025 organizers, since each meme has a unique text and classification.

This manuscript is organized as follows. Section 2 describes the target task of our solution. Section 3 describes the dataset used and provided by the competition organizers. Section 4 shows our efforts to solve Task, while Section 5 is dedicated to present and discuss our results. Finally, Section 6 concludes with some general ideas and future directions.

2. Task description

In general, DIMEMEX 2025[9] involves three different tasks, especially our solution is only for task 1, Detection of Hate Speech, Inappropriate, and Harmless Memes. It is important to say DIMEMEX is part of a set of text classification contests hosted by IberLEF 2025 [10].

This task has a three-way classification in which each meme should belong exclusively to one of the following classes: hate speech, inappropriate content, and harmless. These classes are defined as follows.

Hate Speech: The meme presents "Any kind of communication in speech, writing, or behavior that attacks or uses pejorative or discriminatory language with reference to a person or a group based on who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender, or other identity factors" (United Nations, n.d.).

Inappropriate Content: The meme presents any kind of manifestation of offensive, vulgar (profane, obscene, sexually charged), and/or morbid humor content.

Harmless: The meme does not present any kind of manifestation that presents neither hate speech nor inappropriate content.

3. Dataset

The data set provided for DIMEMEX 2025 consists of approximately 3,000 memes, extracted from public Facebook groups based in Mexico, and manually annotated for the presence of hate speech, inappropriate content, and harmful content.

Figure 1 shows the distribution of the three classes of task 1. Here, it can be seen that the most frequent class is Hate Speech, having a higher number of 1's in comparison with inappropriate and harmful content, being the last two more or less balanced between them.

[1] https://github.com/miccunifi/ISSUES

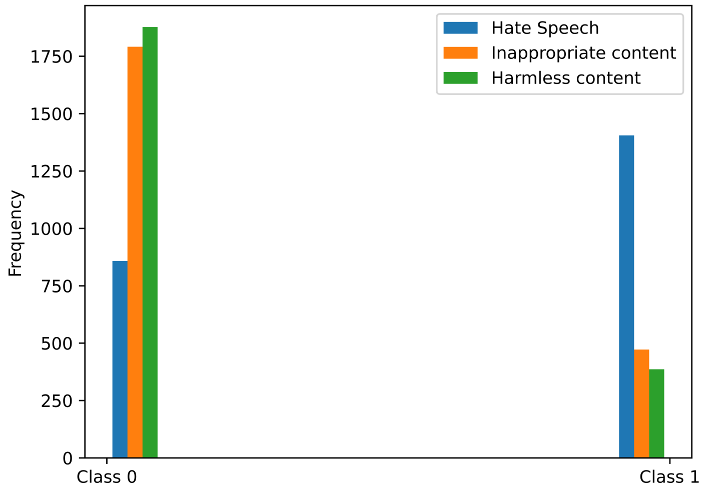

Figure 1: Distribution of the classes in task 1

4. Methodology

For internal purposes, we split the train data into training and testing (70% training and 30% for testing), the second where all our proposals were evaluated, and the highest was submitted to the Codalab platform. In this section, all our efforts will be explained.

Figure 2 shows the general methodology that consists of obtaining the features contained in the images of the dataset, depending on the model it can be text, image or both. Then, a machine learning model is used to train with these features, and finally, the classification is obtained. In our case, there are two general ways to extract features with text or image, or both. As for the machine learning approaches, two were tested, Support Vector Machine and Random Forest.

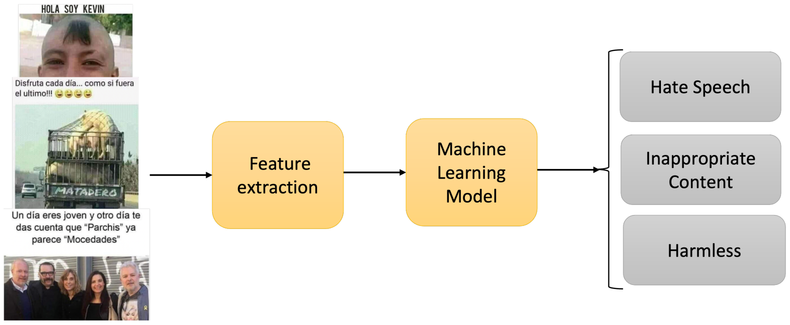
HOLA SOY KEVIN
Disfruta cada día... como si fuera el ultimo!!! 😂😂😂😂
MATADERO
Un día eres joven y otro día te das cuenta que "Parchis" ya parece "Mocedades"
Feature extraction
Machine Learning Model
Hate Speech
Inappropriate Content
Harmless

Figure 2: Approach methodology

4.1. Text and image representation with CLIP

We used CLIP for the representation of text and images. Specifically, we used the OpenAI version available [2]. All technical details can also be found in [11]. We tested several CLIP models, but with *ViT-B/32* the best result was reached, so our final submission was with it. In this case, three ways were tested, the first one was using only the embeddings provided by CLIP of the images, to finally classify using an SVM classifier with default parameters (Radial Basis Function, RBF, with gamma selected

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

automatically). The second one is the same but using only text data (in this case, we only use the meme inside text), and finally, the last way is using both embeddings, text, and image.

All of the approaches used the same configuration of the SVM algorithm. The vector size for both image and text was 512, so in the case of using both types of data, the vector size was 1024.

4.2. EvoMSA for text representation

The other kind of representation used was based on EvoMSA [12], which is based on a kind of Bag-of-words scheme. This scheme combined traditional supervised learning estimators as well as their decision functions.

EvoMSA has been tested in a variety of international text classification competitions, achieving outstanding results compared to Deep Learning solutions in terms of computational resources, dataset sizes, and explicability. As mentioned before, this representation was only tested with the text, both the text inside the meme and the text description provided by the organizers; just in the case of the other representations, the text inside the meme was only used. Similarly to CLIP representations, EvoMSA vectors were classified using an SVM with the same configuration (Radial Basis Function, RBF, with gamma selected automatically).

4.3. Count-Vectorizer for text representation

This approach uses the count vectorizer [13, 14] from scikit-learn. We then obtain the normalization of the meme text by transforming the textual data into a numerical representation that can be used by machine learning algorithms.

The count vectorizer converts each text into a vector whose dimension corresponds to the total vocabulary, and each value represents the frequency of a specific word in the text. Since the simplicity of this representation, in contrast to other text representations tested, here we used a set of elements that are filtered out if necessary (e.g., punctuation marks, very common or very rare words). Each unique word is assigned a numerical index in the vocabulary. This process is carried out for both the training and test data, ensuring that both sets share the same representational structure.

The classification approach used is a probabilistic model based on the frequency of the text features. The model is trained with transforms into vectors and their corresponding labels, the machine learning method is the random forest classifier (RF) with n_estimators=100 and criterion="gini" [15].

5. Results and analysis

As mentioned above, we use the training set to test our solutions internally. The training set consists of 2263 memes with their associated text as well as a general description of the image provided by the competition organizers. For our internal evaluation, we split this dataset into training and test, 70% for training (1584), and 30% for testing (679).

Table 1 shows our internal results with this dataset split. Here, it can be seen that the best internal result was achieved by CLIP using only the image embeddings; nevertheless, very similar results were achieved using text and both text and image. Bigger differences were reached with Count-Vectorizer with text (F1-score of 0.38) and with EvoMSA (F1-score of 0.63).

Our official results with the test data were: f1-score of 0.422, precision of 0.422, and recall of 0.423. Table 2 shows the results and the team name shown on the Codalab platform [3]. In this table, it can be seen that the best achieved F1-score was by *Ryuan* team with a score of 0.58, and in terms of precision, the highest was achieved by the same team with a value of 0.58 as well as the same value for recall. Despite achieving an equal value for the three metrics, our score remains low compared to the winner team, which places us at the seventh position, in general, tied with *VeronicaNeriMendoza.*

[3] https://codalab.lisn.upsaclay.fr/competitions/22012#results

Table 1
Results with proposal approaches

Approach	F1-score	Precision	Recall
CLIP-IMG+TEXT+SVM	0.74	0.89	0.67
CLIP-IMG+SVM	**0.75**	**0.92**	**0.65**
CLIP-TEXT+SVM	0.74	0.90	0.66
Count-Vectorizer+TEXT+RF	0.38	0.63	0.38
EvoMSA (TEXT)+SVM	0.63	0.62	0.64

Table 2
Final official results

Team	F1-score	Precision	Recall
Ryuan	0.58	0.58	0.58
Onarion	0.57	0.58	0.56
ymlopez	0.55	0.57	0.55
VickBat	0.52	0.54	0.51
michaelibrahim	0.44	0.46	0.43
John94	0.43	0.45	0.42
VeronicaNeriMendoza	0.42	0.42	0.42
Infotec+Centrogeo	**0.42**	**0.42**	**0.42**
csuazob	0.34	0.42	0.36
AngelBaron	0.33	0.35	0.34

6. Conclusions

The results obtained in the DIMEMEX 2025 challenge show the superiority of models based on the combination of visual and textual features, especially those using the CLIP model in conjunction with SVM. The CLIP-IMG+SVM model achieved the best F1-score, indicating a solid balance between precision and recall. Although the CLIP-IMG+TEXT+SVM model showed a slight decrease in this metric, it achieved better recall, suggesting that including text can help improve classification. However, the CLIP-TEXT+SVM model also maintained a competitive F1-score, demonstrating that even using only the textual part of the CLIP model, quite high performance can be achieved.

In contrast, more traditional approaches such as Count-Vectorizer+TEXT+RF showed inferior performance. This may reflect the limitations of classical methods in the multimodal and semantic aspects.

The EvoMSA (TEXT)+SVM model, while outperforming the Count Vectorizer-based approach, still lagged behind the CLIP models, indicating a lack of ability to match the multimodal understanding offered by CLIP.

In summary, the results reinforce the importance of integrating both visual and textual features to effectively classify memes; nevertheless, more efforts must be made to achieve better results, which in our case were low.

Generative AI Declaration

The authors declare they did not use any kind of Generative model to write this manuscript.

References

[1] N. Chetty, S. Alathur, Hate speech review in the context of online social networks, Aggression and violent behavior 40 (2018) 108–118.

[2] M. A. Paz, J. Montero-Díaz, A. Moreno-Delgado, Hate speech: A systematized review, Sage Open 10 (2020) 2158244020973022.

[3] A. Arora, P. Nakov, M. Hardalov, S. M. Sarwar, V. Nayak, Y. Dinkov, D. Zlatkova, K. Dent, A. Bhatawdekar, G. Bouchard, I. Augenstein, Detecting harmful content on online platforms: What platforms need vs. where research efforts go, ACM Comput. Surv. 56 (2023). URL: https://doi.org/10.1145/3603399. doi:10.1145/3603399.

[4] H. R. Kirk, A. Birhane, B. Vidgen, L. Derczynski, Handling and presenting harmful text in nlp research, arXiv preprint arXiv:2204.14256 (2022).

[5] P. C. d. Q. Hermida, E. M. d. Santos, Detecting hate speech in memes: a review, Artificial Intelligence Review 56 (2023) 12833–12851.

[6] A. Hamza, A. R. Javed, F. Iqbal, A. Yasin, G. Srivastava, D. Połap, T. R. Gadekallu, Z. Jalil, Multimodal religiously hateful social media memes classification based on textual and image data, ACM Trans. Asian Low-Resour. Lang. Inf. Process. 23 (2024). URL: https://doi.org/10.1145/3623396. doi:10.1145/3623396.

[7] G. Burbi, A. Baldrati, L. Agnolucci, M. Bertini, A. Del Bimbo, Mapping memes to words for multimodal hateful meme classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 2832–2836.

[8] F. Yu, J. Tang, W. Yin, Y. Sun, H. Tian, H. Wu, H. Wang, Ernie-vil: Knowledge enhanced vision-language representations through scene graphs, in: Proceedings of the AAAI conference on artificial intelligence, volume 35, 2021, pp. 3208–3216.

[9] H. Jarquín-Vásquez, I. Tlelo-Coyotecatl, D. I. Hernández-Farías, H. J. Escalante, L. Villaseñor-Pineda, M. Montes-y Gómez, Overview of DIMEMEX at IberLEF2025: Detection of Inappropriate Memes from Mexico, Procesamiento del Lenguaje Natural 75 (2025).

[10] J. Á. González-Barba, L. Chiruzzo, S. M. Jiménez-Zafra, Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org, 2025.

[11] 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.

[12] M. Graff, D. Moctezuma, E. S. Téllez, Bag-of-word approach is not dead: A performance analysis on a myriad of text classification challenges, Natural Language Processing Journal (2025) 100154.

[13] J. Brownlee, Deep learning for natural language processing, Machine Learning Mystery, Vermont, Australia 322 (2017).

[14] S. S. Vel, Pre-processing techniques of text mining using computational linguistics and python libraries, in: 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), IEEE, 2021, pp. 879–884.

[15] S. K. Madhav, P. J. HimanshuRawat, A brief study on random forest using python (2021).