Multimodal Hate Speech Detection in Memes from Mexico using BLIP

Fariha Maqbool[1], Elisabetta Fersini[1]

[1]*Dipartimento di informatica, sistemistica e comunicazione*
University of Milano-Bicocca
Viale Sarca 336, 20126 Milan, Italy

Abstract

The proliferation of online platforms has introduced a novel challenge in identifying inappropriate and hateful content in digital discourse. This paper describes our approach to detect such content on social media platforms, for Task 1 of DIMEMEX challenge in IberLEF 2024 [1]. We employed vision-language based pre-trained model BLIP to extract the combined image text embeddings. Subsequently, a Gradient Boosting Classifier was employed for sample classification. Our findings highlight the potential for further enhancements in multi-modal analysis and classification frameworks.

Keywords

Hate Speech, Inappropriate Content, BLIP

1. Introduction

The emergence of online social media platforms has transformed communication by enabling people to instantly connect with each other worldwide. Despite all of their advantages, these platforms also bring with them new challenges. They have evolved into major channels for the spread of fake news, hate speech, cyberbullying and harassment, playing a crucial role in the recent rise of cyber-hate crimes [2]. The instantaneous and viral nature of content dissemination on social media enables this harmful content to reach vast audiences rapidly. Consequently, it becomes extremely difficult to monitor this content effectively due to the sheer volume of information available on these platforms.

Hate speech has been persistently a social problem, and its forms have evolved significantly over time. It encompasses any kind of expression that targets individuals or groups based on their gender, sexual orientation, race, religion, ethnicity, or nationality [3]. This type of speech can incite violence, promote prejudice, and cause various other harmful effects on individuals and communities. In addition to hate speech, social media platforms also encourage the spread of other types of inappropriate content, such as profane, obscene, offensive, and macabre humor. These all types of content on social media spread through various means, such as text, images, multimedia, and other forms of digital communication. Despite the negative nature of this content, it unfortunately possesses certain qualities that contribute to its rapid dissemination.

Memes are ubiquitous form of multimedia that are created by overlaying text onto images. These humorous or satirical messages have gained immense popularity as a means of communication, spreading rapidly among individuals. Although the majority of internet memes are harmless and amusing, some of them are the source of spreading inappropriate or hateful content. It is extremely challenging to manually identify and stop the propagation of such harmful memes due to the enormous amount of data. Furthermore, automated detection methods face additional challenges due to the complex and multimodal nature of the problem, which necessitates a thorough understanding of the image, text, and context of both modalities. While humans possess an inherent capability to comprehend the meaning conveyed by the fusion of text and images in memes, machines struggle to perform this type of complex task. Detection of Inappropriate Memes from Mexico (DIMEMEX) [4] proposed shared tasks in IberLEF

IberLEF 2024, September 2024, Valladolid, Spain
f.maqbool@campus.unimib.it (F. Maqbool); elisabetta.fersini@unimib.it (E. Fersini)
0009-0008-2587-9417 (F. Maqbool); 0000-0002-8987-100X (E. Fersini)

2024 [1] to detect hate speech in memes written in the Spanish language. The DIMEMEX shared tasks aim to foster advancements in the field of meme analysis and contribute to the creation of safer and more inclusive online spaces.

In this paper, we describe the overview of the approach we adopted to detect hateful content in memes. We utilized a pre-trained BLIP model for this task to predict if each meme was hateful, inappropriate, or harmless. The paper is structured as follows: Section 2 reviews the literature on hateful memes and multilingual tasks. In Section 3, we describe the task and dataset utilized. Section 4 details the proposed approach, and Section 5 presents the results obtained from proposed method.

2. Related Work

Hate speech detection in memes is a challenging task that has garnered significant attention in research and academia. Numerous studies have explored various approaches to identify hate speech in memes. One of the most notable efforts in this area was the challenge proposed by Facebook AI, which focused on multimodal classification to identify hate speech in memes [5]. A preliminary study related to misogyny, a type of hate against women, was conducted by E. Fersini et al. [6] using unimodal (Visual or text) and multimodal-based approaches (text-visual) on a dataset consisting of misogynous content.

The majority of published methodologies and resources for detecting offensive language and hate speech were designed for the English language [7]. Therefore, researchers tried to generate resources for cross-lingual and cross-cultural perspectives. E. Hossain et al. [8] introduced a novel Memes dataset in Bengali language consisting of 7,148 memes. The researchers proposed a multimodal deep neural network called DORA (Dual cO-attention fRAmework) to combat the challenge of detecting hate speech in memes. They performed experiments for both binary classification to identify hate speech and to identify the targeted social entities within the memes. To address the issue of multilingual resources, the authors in [9] expanded Spanish resources with a new dataset of 9834 tweets. They also developed a comparative framework for evaluating models and organized a repository to make it easier to access multilingual datasets.

Other notable efforts to promote research in the Spanish language related to hateful content are DA-VINCIS [10] and HOMO-MEX [11] tasks from the shared evaluation campaign of Natural Language Processing systems in Spanish and other Iberian languages (IberLEF 2023) [12]. These tasks primarily focused on detecting harmful content using textual data. This year DIMEMEX [4]challenge introduced tasks based on memes in Spanish-language, aiming to categorize memes as either hateful, inappropriate or harmless. These initiatives demonstrate the ongoing advancements and the critical need for improved methods in detecting and moderating harmful content across different languages and modalities.

3. Task Description and Dataset

Our team participated in the first task of DIMEMEX challenge which consists of a classification of memes into three categories: hateful, inappropriate, or harmless. Memes that display a clear bias or prejudice against a particular group of individuals are labeled as hateful. On the other hand, memes that do not promote hate but contain vulgar, obscene, or morbid humor are considered inappropriate. Finally, memes without any hateful or inappropriate content are deemed harmless.

Table 1
Dataset description for Task 1 of DIMEMEX

Data Type	Total Size	hateful	inappropriate	harmless
Train	2263	386	472	1405
Test	648	–	–	–

The dataset for the first shared task consists of 2263 memes for training set and 648 for testing. All the text included in the memes is written in Spanish, and each meme is assigned one of three labels:

harmful, inappropriate, or harmless. For development purpose, we split the training set into ratio of 80, 10 and 10 for training, validation and test sets. Table-1 shows the details of the dataset for Task 1 of this challenge.

4. Proposed Approach

We proposed a BLIP model based approach to perform multiclass classification of memes. We utilized BLIP model to extract embeddings and then used a classifier to detect the class of each meme. The flow of the approach is shown in Figure 1.

4.1. Model

We utilized a vision-language based pre-trained model named BLIP [13] for this task. BLIP exploits noisy web data by generating synthetic captions, filtering out noisy ones and pre-training a multimodal mixture of encoder-decoder model. It integrates both language and image modalities into a unified model, aiming to enhance the performance of tasks that require multimodal reasoning. We adopted BLIP for its main capabilities to encode vision and text. Since it has been designed to perform vision-language tasks such as Visual Question Answering (VQA), 0-shot retrieval, and our goal was to predict hateful, inappropriate and harmless memes, we included straightforward fine-tuning.

BLIP is a multimodel Mixture of Encoder and Decoder that consists of a Text Encoder, Image-grounded Text Encoder, and a Decoder. The Unimodal encoder employs Image-Text Contrastive Loss (ITC) to favor positive image-text pairs with similar representations as opposed to negative pairs. Using Image-Text Matching Loss (ITM), the Image-grounded text encoder seeks to capture the finely grounded alignment between language and vision. In the ITM, the model predicts the match positive and unmatch negative pair in a binary classification task. The last part of the model is an image-grounded text decoder that employs causal self-attention layers rather than bi-directional self-attention. It is equipped with Language Modeling Loss (LM) to facilitate the generation of textual descriptions based on an input image. The purpose of this loss is to train the decoder in an autoregressive manner, maximizing the likelihood of the generated text. The architecture and training strategy of BLIP model enables remarkable performance across a wide range of vision-language tasks, demonstrating the effectiveness of its integrated framework.

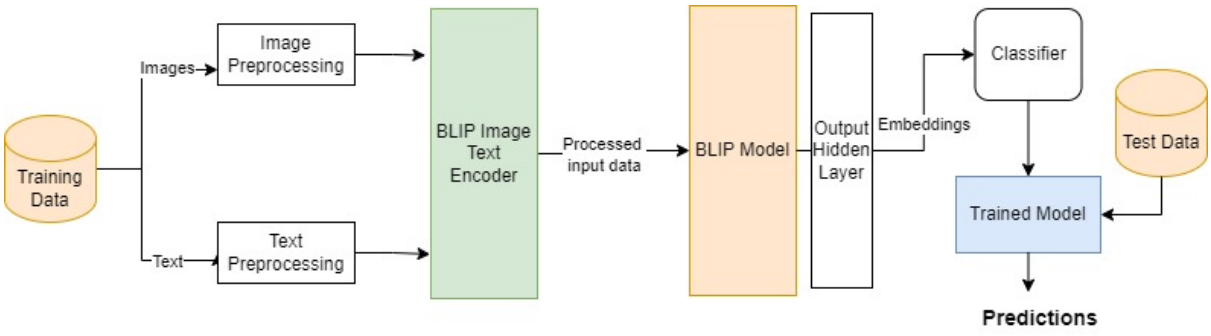

Figure 1: Workflow of the Proposed Approach

4.2. Preprocessing

In the preprocessing step of our model, we utilized the transformers library to handle the initial processing of our data. The associated text is translated to English using the GoogleTranslator API to ensure uniform language processing. Both the image and translated text are processed using a pre-defined processor that tokenizes the text and resizes the image as needed. The encoders of BLIP leverages pre-trained image encoders and frozen large language models (LLMs) to train a lightweight,

12-layer Transformer encoder. This processor converts the data into tensors, applying padding and truncation to ensure consistent input lengths, with a maximum of 128 tokens. The processed inputs are stored in a dictionary and squeezed to remove any unnecessary dimensions, ensuring compatibility with the input requirements of the model.

4.3. Feature Extraction

The dataset is divided into training and validation sets, and DataLoader objects are created for each set. To obtain feature embeddings, the preprocessed input is forward passed through the model in batches. The model processes the batch inputs to generate the last hidden state, which contains the combined embeddings. Subsequently, these embeddings are appended to a list for later concatenation. After processing all batches, the collected embeddings and labels are concatenated into single tensors. These tensors contain the feature representations and corresponding labels for the entire dataset, ready for further analysis or training downstream models.

4.4. Classification

Following the preprocessing and feature extraction stages, we proceeded with the classification of the extracted embeddings. To reduce the dimensionality of the embeddings and create a more manageable feature set, we employed mean pooling across the token dimension. This step computes the mean of the embeddings for each instance. The labels for the test and training datasets were converted from one-hot encoded format to their corresponding class indices. This step is essential for compatibility with the classifier, which requires labels in integer format. We utilized the Gradient Boosting Classifier from scikit-learn, a powerful ensemble method that builds an additive model in a forward stage-wise manner. Each iteration fits a new base-learner to the residual errors made by the previous model. Once the classifier was trained, we used it to predict the labels of the test embeddings.

5. Experimentation and Results

We utilized PyTorch library in Python in our implementation. After data preprocessing, the combined image-text embeddings are extracted using BLIP model using batch size of 16. The Gradient Boosting Classifier is trained on these embeddings using 100 estimators. This classifier is then used to predict the labels for unseen test dataset.

The challenge uses the Macro-average of Precision, recall and f1-score as evaluation measures. Table 2 shows the results of our approach based on labels produced by our model. The model was able to achieve the macro average F1-score of 0.47 on the test dataset.

Table 2
Official results of the proposed approach

Evaluation Metric	Best Scores	Proposed Model	
		Scores	Rank
Precision	0.63	0.52	3
Recall	0.56	0.50	3
F1-score	0.58	0.47	5

6. Conclusion

In this paper, we present our approach for the Task-1 of DIMEMEX challenge. The task is a multi-classification problem to categorize memes as hateful, inappropriate, or harmless. We used the visual language model BLIP to extract the combined image and text embeddings. These embeddings were

then used to train a Gradient Booster Classifier with 100 estimators. The performance of the model was evaluated on the test data provided by the task organizers, using precision, recall, and macro F1 score metrics. Our model achieved precision, recall, and macro F1 scores of 0.52, 0.50, and 0.47, respectively. In conclusion, our approach demonstrates the potential of visual language models and ensemble learning techniques in addressing complex multi-modal classification tasks. By using BLIP to extract image-text embeddings, the complex relationships between visual and textual content in memes can be captured. In the future, classification performance can be improved by using an ensemble of multiple models, such as combining BLIP with other vision-language models or classifiers.

7. Acknowledgments

We acknowledge the support of the PNRR ICSC National Research Centre for High Performance Computing, Big Data and Quantum Computing (CN00000013), under the NRRP MUR program funded by the NextGenerationEU.

References

[1] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org, 2024.

[2] N. S. Mullah, W. M. N. W. Zainon, Advances in machine learning algorithms for hate speech detection in social media: A review, IEEE Access 9 (2021) 88364–88376. URL: https://doi.org/10.1109/ACCESS.2021.3089515. doi:10.1109/ACCESS.2021.3089515.

[3] A. Rawat, S. Kumar, S. S. Samant, Hate speech detection in social media: Techniques, recent trends, and future challenges, WIREs Computational Statistics 16 (2024). doi:https://doi.org/10.1002/wics.1648.

[4] H. J. Vásquez, I. Tlelo-Coyotecatl, I. H. Farías, M. Casavantes, H. J. Escalante, L. Villaseñor-Pineda, M. M. y Gómez, Overview of DIMEMEX at IberLEF 2024: Detection of Inappropriate Memes from Mexico, Procesamiento del Lenguaje Natural (2024).

[5] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.

[6] E. Fersini, G. Rizzi, A. Saibene, F. Gasparini, Misogynous MEME recognition: A preliminary study, in: S. Bandini, F. Gasparini, V. Mascardi, M. Palmonari, G. Vizzari (Eds.), AIxIA 2021 - Advances in Artificial Intelligence - 20th International Conference of the Italian Association for Artificial Intelligence, Virtual Event, December 1-3, 2021, Revised Selected Papers, volume 13196 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 279–293. doi:10.1007/978-3-031-08421-8_19.

[7] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, Semeval-2022 task 5: Multimedia automatic misogyny identification, in: G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, S. Ratan (Eds.), Proceedings of the 16th International Workshop on Semantic Evaluation, SemEval@NAACL 2022, Seattle, Washington, United States, July 14-15, 2022, Association for Computational Linguistics, 2022, pp. 533–549. doi:10.18653/V1/2022.SEMEVAL-1.74.

[8] E. Hossain, O. Sharif, M. M. Hoque, S. M. Preum, Deciphering hate: Identifying hateful memes and their targets, CoRR abs/2403.10829 (2024). doi:10.48550/ARXIV.2403.10829. arXiv:2403.10829.

[9] A. A. Monnar, J. Perez, B. Poblete, M. Saldaña, V. Proust, Resources for multilingual hate speech detection, in: K. Narang, A. M. Davani, L. Mathias, B. Vidgen, Z. Talat (Eds.), Proceedings of the

Sixth Workshop on Online Abuse and Harms (WOAH), Association for Computational Linguistics, Seattle, Washington (Hybrid), 2022, pp. 122–130. doi:10.18653/v1/2022.woah-1.12.

[10] H. J. Jarquín-Vásquez, D. I. H. Farías, L. J. Arellano, H. J. Escalante, L. V. Pineda, M. Montes-y-Gómez, F. Sánchez-Vega, Overview of DA-VINCIS at iberlef 2023: Detection of aggressive and violent incidents from social media in spanish, Proces. del Leng. Natural 71 (2023) 351–360.

[11] G. Bel-Enguix, H. Gómez-Adorno, G. Sierra, J. Vásquez, S. T. Andersen, S. Ojeda-Trueba, Overview of HOMO-MEX at iberlef 2023: Hate speech detection in online messages directed towards the mexican spanish speaking LGBTQ+ population, Proces. del Leng. Natural 71 (2023) 361–370.

[12] M.-y.-G. Jiménez-Zafra, Francisco Rangel, Overview of IberLEF 2023: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), co-located with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), CEUR-WS.org, 2023.

[13] J. Li, D. Li, C. Xiong, S. C. H. Hoi, BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, S. Sabato (Eds.), International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, volume 162 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 12888–12900.