UCM's Participation to the 2024 DIMEMEX Task: Automatic Detection of Inappropriate Memes in Mexico

Aisha Aman Parveen[1]

[1]*Universidad Complutense de Madrid (UCM), Av. Complutense, s/n, Moncloa - Aravaca, 28040 Madrid (Spain)*

Abstract

Social media has a huge impact in our world. It has transformed how we share and communicate. Nevertheless, detection of offensive and malicious messages is still an area that must be improved. The conference IberLef 2024 has prepared a competition aimed to detect abusive content. In this paper, we will describe the process of analysing and classifying different Facebook memes. We will describe our datasets and the process we have followed in order to pre-process and prepare the data to be used to train the Machine Learning (ML) model. Our main focus is on different techniques to process the data and looking for the model with the best combination of hyperparameter in order to classify properly each meme.

Keywords

Natural Language Processing, Hate Speech, Social Media.

1. Introduction

Social networks are online platforms that enable people, companies and governments to connect, communicate and share information. These networks promote virtual communities where users can establish personal and professional relationships, share information and collaborate on diverse activities. Different social media platforms are created for purposes such as marketing, business communication and education. Some of the most popular social media include Facebook, Instagram, Twitter and TikTok. Nevertheless, inappropriate content and hate speech are significant challenges in today's digital world. Inappropriate content includes offensive or harmful material unsuitable for certain audiences, while hate speech involves discriminatory communication against individuals or groups based on attributes like race, religion, or gender. These issues can harm mental health, foster division, and incite violence. Research to detecting that type of content has become a critical issue in order to prevent online harassment growth. However, despite the notable advances made, there are still challenges for a deeper understanding.

In this study, we aim to address these challenges by analysing Facebook memes as part of the DIMEMEX task [1] of the IberLef 2024 conference [2]. First, we will preprocess text and images. In order to prepare the text dataset, we will make some transformations to obtain a text embedding. Same process will be carried out with images but with a Transformer [3]. Once we have both text and image embeddings, we will merge them into a single embedding and train it with different models. Also, we will implement the starting kit provided by the organization and analyse the results. For this submission, we have participated on Task 1. The predictions labels are: hate speech, inappropriate and neither. Finally, we will analyse the predictions and compare them with the final results of the competition.

2. Dataset

Dataset is organized in 3 different blocks: train, validation and test. First, we will train our model with all the data available in train. Then, we will run our model with validation dataset and submit our predictions and receive feedback of our performance. Finally, after the adjustment that we consider, we will run our model with test dataset and submit final results.

IberLEF 2024, September 2024, Valladolid, Spain

aaman@ucm.es (A. A. Parveen)

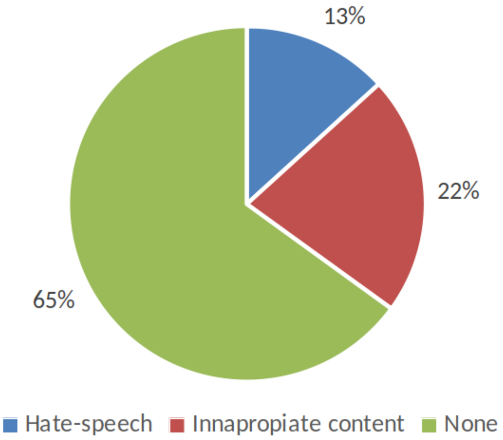

Figure 1: Distribution of the labels of the training set.

Table 1
Dataset description

	Data		Output (Y)
	Id	Meme	Label
Train: 2263	✓	✓	Hate-speech: 386 Inappropiate content: 472 None: 1405
Validation: 323	✓	✓	X
Test: 649	✓	✓	X

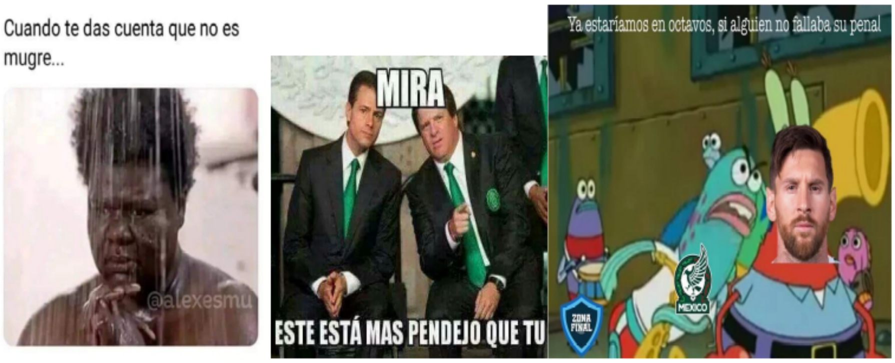
Cuando te das cuenta que no es mugre...
@alexesmu
MIRA
ESTE ESTÁ MAS PENDEJO QUE TU
Ya estaríamos en octavos, si alguien no fallaba su penal
ZONA FINAL
MEXICO

Figure 2: Classification of memes.

As we can observe in Table 1, train dataset is the only data which we know not only the date, but also the output, in other words, the label of each meme. From Figure 1, we can observe that 65% of train-dataset is None, 13% is hate-speech and 22% is inappropriate content. This information is revealing as our model will be better trained to predict memes with the label "none" than those with "inappropriate content". It is important to remark that predicting the category of hate speech will be much more challenging as we have less data to train our model.

Each post includes a meme and a sentence. A meme is a photography with a text that emphases the idea of the sentence written. Figure 2 presents an example of each post with their corresponding image. The image order from left to right matches the following labels: hate-speech, inappropriate content and none. The text associated with hate-speech is “Cuando te das cuenta que no es mugre @alexesmu”; for the inappropriate content is “[RA ESTE ESTÁ MAS PENDEJO QUE TU” and for the label none is “Ya

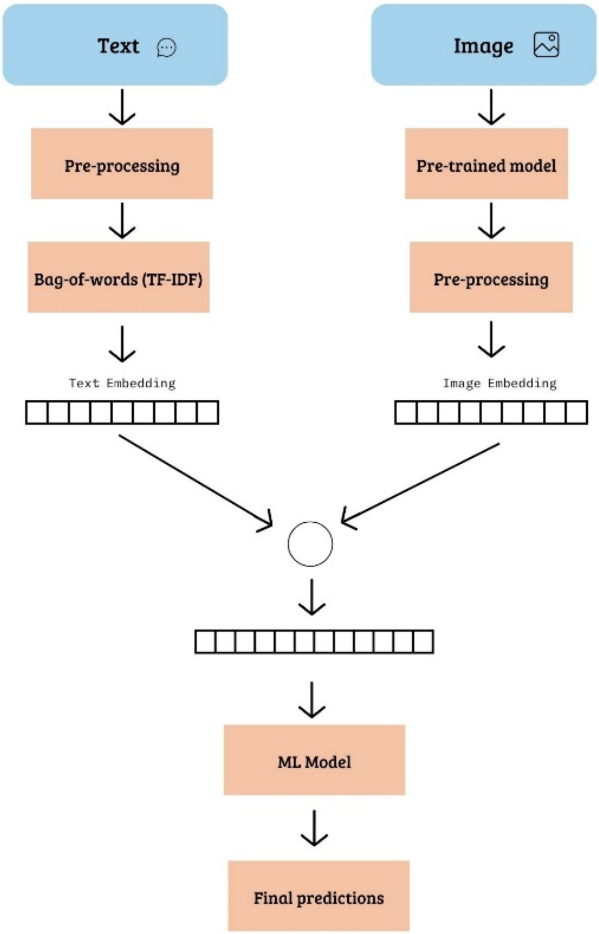

Figure 3: Description of the steps followed in the data analysis.

estaríamos en octavos; si alguien no fallaba su 00 MEXICO Zona peTexnal".

3. Text mining approach

We will now describe the models we implement to prepare and analyse out datasets. In the first approach, we have applied different classification models adjusting the hyperparameters associated to obtain the best combination. In the second approach, we use the starting kit provided by the shared task organizers [1], which is a combination of an image and text neural network.

3.1. Classification models

As our dataset combine text and image, we will process each one individually and then concatenate them into a single embedding to train our Machine Learning (ML) model and make predictions. Figure 3, describes schematically the process that has been followed. The processes followed in each case is different, for the images we have used a Transformer pretrained and for the text we have applied different techniques. In both cases, we will with details the process in the following section.

3.1.1. Text approach

In order to process data properly, we have made some adjustments on the dataset and then applied spaCy and Nltk libraries. The steps followed are:

1. We remove spaces and transform the text to lower case.

Table 2
Dataset description

Models	F1 result	Hyperparameters
Logistic Regression	0.42	C’: 0.1, ’class_weight’:’balanced’, ’penalty’: ’l2’,’solver’: ’liblinear’
Random Forest Classifier	0.36	class_weight’:’balanced’, ’max_depth’: 23, ’max_features’: 5, ’min_samples_leaf’: 10, ’n_estimators’: 100
MLP Classifier	0.40	alpha’: 1e-05’, batch_size’: 500, ’hidden_layer_sizes’: 100, ’learning_rate_init’: 0.05, ’random_state’: 1, ’solver’: ’lbfgs’

2. With the TweetTokenizer we split the document into words and replaced URLS, mentions and numbers by tokens. Then, we merge the words into a new document.
3. With the spaCy and Nltk libraries, we search for text lemmas, punctuation symbols and stopwords to remove them.
4. We merge the modified tokens into a new text.
5. We create 3 new columns where we will count the number of mentions, URLs and numbers in each post. Once the steps are finished, we will have the text dataset ready to join it with the embedding of the images.

3.1.2. Image approach

In this section, we extract features from each image with a pre-trained Hugging Face pipeline. This pipeline applies the principles of Transformers that have been so successful in NLP to computer vision. The pre-trained model is called “google/vit-base-patch16-384” [4]. No preprocessing has been applied apart from the one implemented in the “google/vit-base-patch16-384” pipeline. With the previous model we obtain the embedding of the images in order to merge them with the text.

3.1.3. Modelling

Once we have prepared the dataset, we proceed to transform the normalised text into a feature matrix using the TF-IDF technique. The parameters we set in the model are:

- We will convert the text to lowercase.
- We will use between 1 and 4 n-grams, i.e. we take into account unigrams, bigrams, trigrams and quadrigrams.
- We set the minimum number of times a term has to appear to be ignored 0.1%, which means that any term with less than 0.1% in the dataset is not going to be taken into account. Words that represent only 0.1% of the total amount of words, that is, words that have appeared very few times, will not be taken into account as they do not contribute significantly to the training and learning of the model.

Then, we split the train dataset into train and dev, with train been 80% of the dataset and dev 20%. During training, we will look for the model with highest performance. Once we know which model gives us the best result, we will train it with the whole dataset. In the next step, we download the libraries of LogisticRegression, RandomForestClassifier and MLPCLassifier models from Sklearn. We will set up a dictionary with a number of hyperparameters to test in each model in order to obtain the best combination for each model. To do so, we will use a 5-fold cross-validation scheme.

Table 2, shows the best combination of hyperparameters for each model. As the first model is the one with best performance, the Logistic Regression model will be trained with the whole data for the final submission.

Table 3
Confusion Matrix obtained with Logistic Regression Model.

35	16	46
26	31	61
64	82	205

Table 4
Classification results for each label with the Logistic Regression model. The support column refers to the number of samples used to evaluate the model for each class label.

	precision	recall	f1-score	support
0	0.28	0.36	0.32	97
1	0.24	0.26	0.25	118
2	0.66	0.58	0.62	351
accuracy	0.48			

Table 5
Results comparison between our submission and the competition baselines

User	f1	Precision	Recall
CLTL	0.58 (1)	0.61 (2)	0.56 (1)
michaelibrahim	0.56 (2)	0.63 (1)	0.53 (2)
Aaman (starting kit)	0.49 (3)	0.49 (4)	0.49 (4)
VickBat	0.48 (4)	0.48 (5)	0.47 (5)
fariha32	0.47 (5)	0.52 (3)	0.50 (3)
mashd3v	0.42 (6)	0.48 (6)	0.42 (6)
GarciaRodriguezMario	0.36 (7)	0.36 (7)	0.36 (7)
Baseline TXT	0.46	0.51	0.49
Baseline IMG	0.43	0.46	0.43
Baseline MM	0.49	0.50	0.48

3.2. Starting kit

The multimodal (MM) starting kit provided by the organization is a model that combines both image and text Transformers. For the image, the "google/vit-base-patch16-224-in21k" is used, whereas the "dccuchile/bert-base-spanish-wwm-uncased" is used for the text part. Both encodings are then merged together, followed by a linear layer. In this case, the weights of all models are updated using gradient descent. In contrast, in our previous approach, the weights of the image Transformer were frozen.

3.3. Results

As we can observe in Table 3, the model predicted correctly class 1 (hate-speech) for 35 examples, class 2 (inappropriate content) for 31 examples and class 3 (none) for 205 examples of our internal dev set. In other words, those numbers reflects that the class prediction corresponds to what it actually is. On the other hand, for those values outside the diagonal, the model predicted erroneously the type of class they are.

As Table 4 shows, the model has an accuracy of 48%. In this case the model was evaluated on 97 samples for class 0 (hate speech), 118 samples for class 1 (inappropriate content) and 351 samples for class 2 (none). The support samples reinforce the fact that there is less data available for class 0 and 2 than class 3. The f1-score of class 2 (none) is the one that has predicted the best, followed by class 1 (inappropriate content) and finally class 0 (hate-speech). One of the explanations for this, is that the model has more samples to train class 2 (1405 samples) and 1 (472 samples) than class 0 (386 samples). We observed this at an early stage when we analysed our initial dataset.

In contrast to this, the starting kit model obtained 0.59 accuracy and 0.47 f1 score. Although we have made a distinction between train and dev, we have always worked on the train dataset. The model with the best results was trained with the train dataset and the labels obtained associated to each image/text were uploaded to the conference platform, since we do not have the output (Y) of the test to check the accuracy of our model. We submitted both our proposed approach and the starting kit.

Once this process has been carried out, we observe in Table 5 that our accuracy, precision and recall is 49%, obtained with the starting kit. The Logistic Regression model obtained a significantly lower score, 27%. Among the participants, CLTL has the highest F1 score (0.58), precision (0.61) and recall (0.56). Garcia Rodriguez Mario has the lowest F1 score (0.36), which is lower than the competition Baseline TXT, IMG and MM. All the participants have a higher score in Precision, which means that they have classified properly the predicted class. Regarding the Baseline results, we can observe that same patter is repeated as in competitors, precision is higher than F1. Aaman and GarciaRodriguezMario and Baseline have obtained the same score in all metrics. Michaelibrahim, Vickbat, fariha32 and mashd3v has higher precision than recall.

During our internal experimentation, there was not a large difference. Nevertheless, if we compare with the results of the submissions, the Logistic Regression model is not competitive when compared with the starting kit.

4. Conclusions

We have presented our participation at the IberLef 2024 DIMEMEX Task 1, which consists of detecting hate speech or inappropriate content in social media. We have tested two multimodal approaches that process both text and image. On our proposed text mining approach, the accuracy was lower than the starting kit. The results are 49% and 27% accuracy, respectively.

References

[1] 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, in: Procesamiento del Lenguaje Natural, 2024.

[2] 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.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 5998–6008. URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net, 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.