A Multimodal Approach to Classify Prejudiced Speech in Brazilian Portuguese Memes

Hanna F. Menezes, Herman M. Gomes, Eanes T. Pereira
Federal University of Campina Grande
{hanna,herman,eanes}@copin.ufcg.edu.br

Sylvia Iasulaitis
Federal University of São Carlos
si@ufscar.br

Brett M. Drury
Liverpool Hope University
druryb@hope.ac.uk

Abstract

This paper proposes a method for classifying discriminatory speech in Brazilian Portuguese memes. Social media platforms host large volumes of multimodal content, particularly memes that combine text and images to convey ideas, critique and humour. Within this setting, the spread of discriminatory content has increased, yet most existing detection models are trained on English data.

The multimodal nature of memes poses challenges for automated analysis, as interpretation depends on the interaction between textual and visual elements, as well as the cultural context. To address this, we evaluate a multimodal model that integrates visual and linguistic features using late fusion. This is particularly relevant as prior work on Brazilian Portuguese has focused on unimodal approaches.

We constructed a dataset of 1,500 labelled memes for fine-tuning. The proposed method achieves an F1 score of 0.60, comparable to results reported for multilingual settings, and provides insights specific to Brazilian Portuguese.

1 Introduction

The word *meme* was first introduced by Dawkins (1976), which referred to a unit of human culture similar to a gene. The use of the term *meme* in the Internet context was first used by Mike Godwin in a text for Wired. In this paper, the word meme refers to a type of content composed of text and images, which are used in the semantic construction of the message. Currently, memes are notably used for communication on social media platforms. However, it has been observed that the dissemination of prejudiced speech with this type of artefact has generated a high impact on social relations, mainly in relations mediated by instant messaging applications and social networks.

This paper considers *prejudice* as the discourse which has an inclination against, or in favour of, a person or a group, related to their inherent or acquired characteristics, in a considered unjust manner (Mehrabi et al., 2021; Panch et al., 2019).

According to Roy et al. (2024), Waseem and Hovy (2016), and Ahmed and Lin (2022), due to the universal access to the Internet in the last decade, the web now has more than 4.5 billion active users, about 59% of the total global population, who are particularly connected to social media. In this context, we can observe the growing use of memes on social networks as a way of expressing ideas, beliefs, and humour.

Numerous studies, (Kiela et al., 2020; Fersini et al., 2022; Pramanick et al., 2021; Sharma et al., 2022), have been conducted to detect abuse/prejudice in memes. The majority of the publicly available datasets are in English (Das and Mukherjee, 2023). This impacts the study of memes in non-English languages, such as Brazilian Portuguese. Furthermore, understanding memes requires external knowledge related to social, cultural and local contexts, which are sometimes poorly represented in existing datasets and are often not taken into account by existing solutions.

Associated with the element 'idiom', the Internet meme has unique characteristics of composition and grammar. In the language aspect, when interpreting this type of content, it is necessary to consider that, in online interaction, content development occurs rapidly. The language is informal, with numerous abbreviations, fragmented syntax, hidden terms and semantics altered due to alterations to the original context (Gouveia, 2009).

Several studies in non-English contexts have addressed the detection of harmful multimodal content, typically comprising both text and images (Burbi et al., 2023; Pramanick et al., 2021; Lin et al., 2024). In contrast, prior research in Brazilian Portuguese has focused predominantly on unimodal data, particularly text-based analysis, for example (Silva and Freitas, 2022; Nascimento et al., 2019; Rosa et al., 2023). This article describes a controlled analyses of cross-lingual degradation in multimodal meme hate detection and proposes a multimodal approach employing late fusion for the binary classification of memes in Brazilian Portuguese.

The main contributions of this paper are (i) the construction of a dataset of memes in Brazilian Portuguese, named Prejudice Memes; (ii) the adaptation of a multimodal late fusion approach to the context of Brazilian Portuguese; and (iii) a critical discussion regarding the use of large multimodal models (predominantly developed from a vast English language dataset) showing lower performance in content analysis in other languages compared to English.

Thus, this research seeks to answer the following question: is it possible to build a prejudiced-meme detector for Brazilian Portuguese that achieves performance comparable to English-language detectors by using (i) knowledge transfer with fine-tuning and (ii) translation combined with a Portuguese pre-trained language model?

2 Related work

According to Rocha Junqueira et al. (2023), the availability of pre-trained language models for Portuguese remains limited. Notable examples include BioBERTpt, a BERT-based model adapted for clinical and biomedical texts that achieves competitive performance in Named Entity Recognition tasks (Schneider et al., 2020); BERTaú, a model trained from scratch on Portuguese-language corpora from the banking domain (Finardi et al., 2021); and Portuguese BERT base cased QA, which has been fine-tuned on the Stanford Question Answering Dataset (SQuAD) (Guillou, 2021). In addition, BERTimbau, developed by Neuralmind, is a model for Brazilian Portuguese that has achieved state-of-the-art results in Semantic Textual Similarity and Named Entity Recognition (Souza et al., 2020).

There are a wide range of pre-trained visual processing models with good results in the classification task. According to Liu et al. (2024), the growing use of transformer architecture in Computer Vision has presented competitive results in different tasks, such as detection (Carion et al., 2020), segmentation (Wang et al., 2021; Cheng et al., 2021), tracking (Chen et al., 2021), generation (Jiang et al., 2021), and enhancement (Chen et al., 2021).

According to Jabeen et al. (2023), Multimodal Deep Learning (MDL) has the purpose of creating models which can process and link information using different modalities. MDL to better comprehend and analyze when various senses are engaged in the information process. Despite extensive development through unimodal learning, MDL cannot yet cover every aspect of human learning.

Multimodal learning enables the development of robust, high-performing models, despite potential redundancy across modality-specific features. Certain tasks inherently benefit from a multimodal approach, as learning from multiple sources facilitates the capture of cross-modal correspondences and supports a deeper understanding of complex phenomena, including cognition, emotion, attention, and human interaction (Luo et al., 2024). Access to multiple modalities also allows for the integration of complementary information that may not be observable within a single modality (Jabeen et al., 2023). Moreover, in some applications, multimodal systems can maintain acceptable performance even when one modality is unavailable (Bayoudh et al., 2022).

There are several multimodal fusion methods, and the main difference between them lies in the stage at which the characteristics of different modalities are combined (Bayoudh et al., 2022): (i) early fusion, also known as feature-level fusion, combines raw features near the input data; (ii) late fusion, or decision-level fusion, processes each modality independently and then combines their outputs; and (iii) middle fusion, which fuses features from different modalities at an intermediate stage of the workflow.

According to Verma et al. (2022), approaches to social computing tasks frequently rely on

(a)

(b)

Figure 1: Examples of prejudice and no-prejudice memes: (a) Prejudiced Meme (Racism) ("Black folks aren't like they used to be"). (b) Non-prejudiced Meme (General Humor)("Who's my warrior?")

large language models. Such tasks include the detection of information related to humanitarian crises (Alam et al., 2018), fake news detection (Shu et al., 2017), and emotion recognition (Thang Duong et al., 2017). The authors highlight disparities in performance between models developed for English and those for other languages. They further emphasise the broader implications associated with the use of large language models, including environmental and financial costs (Strubell et al., 2019), as well as issues of linguistic bias (Bender, 2019) and gender bias (Costa-juss and Casas, 2019).

Among the main challenges in developing non-English language models are (i) large language models have architectures that require lots of computer resources to train the model, which is quite challenging; (ii) the need for a large dataset labelled in the target language, which, depending on the language and the task at hand, often requires building a dataset, as there are no publicly available datasets to use; and (iii) most NLP research is disproportionately concentrated on English and a few languages with more resources (e.g., Spanish and French) (Bender, 2019).

3 Methodology and Dataset

For this study, a multimodal dataset was developed to detect memes with prejudiced content in the Portuguese language. The development of the dataset was inspired by the processes performed by Kiela et al. (2020), and Gasparini et al. (2022).

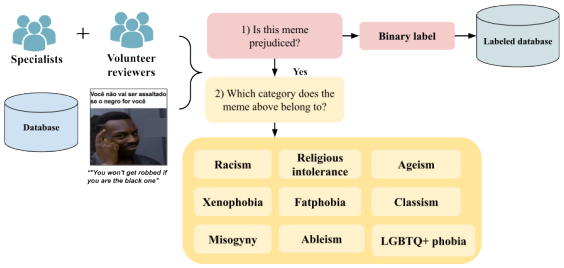


Figure 2: Meme Labelling Process. Source: Elaborated by the authors.

3.1 Data Description

The process of building the dataset was divided into three main stages: (i) collection of memes on popular social media platforms in Brazil, such as Facebook, X (formerly Twitter) and Instagram; (ii) annotating and labelling the memes with the participation of both experts and trained volunteers; and (iii) extracting the textual content from each meme using the Keras OCR software.

Regarding the prejudice categories, the sample sizes are uneven, as some types of prejudice appear more often than others in social media posts. Additionally, a single meme may include more than one category of prejudice. A majority-voting method was used to assign labels. The final label for each meme was determined by agreement of two out of three votes, where two social science experts each cast one vote, and the third vote came from the majority of three volunteers reviewing each meme, as shown in Figure 2.

The survey, executed through an online form, aimed to gather information on how individuals who access and share such content perceive the presence (or absence) of prejudice in these artefacts. The research sample consisted of 300 volunteers (recruited via institutional email, academic groups, and social media outreach) and two social science experts.

For the evaluation stage with volunteer reviewers, Google Forms was used to create the assessment instruments. Each form contained 15 memes, and each meme was evaluated independently by three different participants. Agreement among the evaluators was 50%: 750 memes received the same label from all three evaluators. To statistically measure how much the observed agreement exceeded what would be expected by chance, the Fleiss' kappa co-

efficient (Fleiss et al., 2013) was calculated. The resulting value was 0.35, which, according to Fleiss (Fleiss et al., 2013), indicates fair agreement, showing that classifying memes is inherently challenging.

The final dataset, Prejudiced Memes[1] , comprises 1,500 memes, categorised into two main groups: prejudiced and non-prejudiced. The prejudiced memes include examples of the following types of offensive content: racism, LGBTQ+ phobia, religious intolerance, xenophobia, fatphobia, misogyny, ableism, ageism, and classism.

In addition to the constructed dataset, the datasets of the Hateful Memes challenge formulated by Facebook (Kiela et al., 2020) and the Automatic Misogyny Identification dataset proposed by Fersini et al. (2022) for the SemEval-2022 challenges were also used. Both datasets have nearly ten thousand memes in the English language, labelled (binary classification) in two categories: hateful and not hateful and misogynist and not misogynist, for the hateful memes dataset and the misogynist dataset memes, respectively.

3.2 Model Description

The deep neural network architecture employed in this paper uses scalar product attention to learn the multimodal representation from text and images with late fusion. The architecture used was proposed by Sharma et al. (2022), which is composed of vision and language models previously trained as extractors of features from text and image to jointly learn the multimodal representation.

Visual and textual features were used as query parameters in attention. The context vectors obtained from the attention layer go through a one-dimensional convolution layer to learn the final multimodal features, and these features are finally flattened and fed into the final classification layers (Figure 3).

The attention mechanism uses Queries (Q), Keys (K) and Values (V) to calculate the context vectors. The Query (Q) and Key (K) are used to calculate attention weights, and the final features are a linear combination of attention weights over the values (V). Attention is calculated as the product between the matrices

[1]Dataset access request: https://forms.gle/m81ygbc4AqJomVYn9

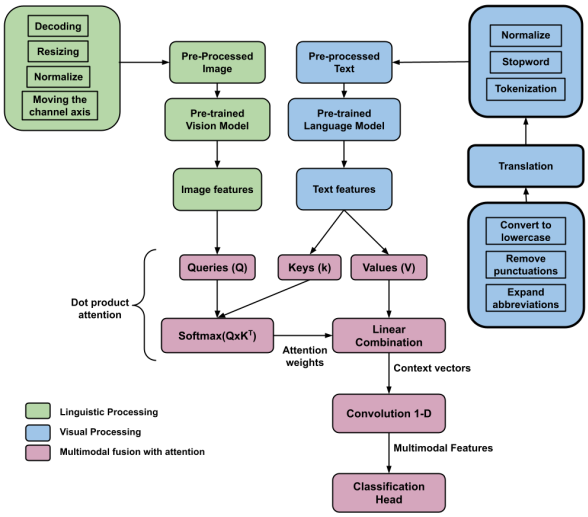


Figure 3: Proposed Model Architecture.

Q and K transposed, divided by the square root of the scale factor (dK) applied to the softmax function, and multiplied by V. The output from this stage of attention represents the vectors of context (Equation: (1)).

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

We adopted transfer learning to decrease the model training time and to compensate for the scarcity of training data. Pre-trained vision and language models were fine-tuned with data from a new target task. In addition to the developed dataset containing Brazilian Portuguese memes, the Hateful Memes and Misogyny Memes datasets were also used to fine-tune the models.

Two scenarios were investigated: (1) English, and (2) Portuguese. To conduct the experiments in scenario 1, the developed dataset was translated to English, and for the experiments from scenario 2, the datasets from the Hateful Memes and the Misogyny Memes were translated to Brazilian Portuguese. The translation model T5 (Lopes et al., 2020) was used in both scenarios.

As it may be observed in Figure 3, the architecture proposed by Sharma et al. (2022) was adapted. Considering the necessity of translation of datasets from English to Portuguese or vice versa, stages of pre-processing textual content (e.g. stopwords, expand abbreviations) were added before and after the translation

Dataset	Train	Val	Test
Misogyny Memes	6,000	2,000	2,000
Hateful Memes	6,000	2,000	2,000
Prejudice Memes	1,000	250	250

Table 1: Dataset split

process, with the purpose of changing the raw text in an appropriate format to be used by the model. Considering a large number of model parameters (more than one hundred and ninety million), the initial layers of the visual and language models were frozen.

3.3 Experimental Setup

The experiments were conducted with transformer models: TensorFlow, Keras, and Hugging Face, using A100 GPU and T4 on Google Colab Pro.

For the English text feature extraction, pretrained BERT (Devlin et al., 2019) model was used, and the BERTimbau (Souza et al., 2020) for Portuguese text. Image processing was performed using the ViT (Dosovitskiy et al., 2021) vision model.

To evaluate the proposed model, experiments were conducted with different scenarios to compare the results obtained with other datasets, both in English and Brazilian Portuguese, focused on verifying the model's efficiency in classifying prejudiced content in Brazilian Portuguese memes.

In both scenarios (English and Portuguese) the following datasets were used: (i) dataset of Misogyny Memes (10k); (ii) dataset of Hateful Memes (10k) and (iii) the dataset of Prejudice Memes (1.5k). Table 1 shows the division of data between training, testing and validation for the three datasets used.

Importantly, for the result of the Hateful Meme dataset, the weighted average F1-score was considered as the primary evaluation metric, taking class imbalance into account. For the other datasets, the macro average F1-score was considered, given that these datasets are balanced.

Other adjustments to improve the performance were made to the model proposed by Sharma et al. (2022). The adjustments were: batch size, number of epochs, early stop, addition of regularization (L1) and adaptive learning rate.

In addition to the configurations presented in Table 2, the data were preprocessed using standard techniques, including stopword removal, tokenisation, conversion to lowercase, punctuation removal, and abbreviation expansion.

Parameter	Value
Epochs	30
Max seq. length	80
Optimizer	Adam
Learning rate	0.001
Early stopping (pat.)	2
Batch size	32
Loss	Cross-entropy
L1 regularization (λ)	0.01

Table 2: Hyperparameter configuration

Model / Scenario	BERT + ViT		
	Prec.	Rec.	F1
Misogyny Memes (EN)	0.85	0.85	0.85
Hateful Memes (EN)	0.82	0.81	0.81
Prejudice Memes (EN trans.)	0.71	0.70	0.70

Table 3: Performance of BERT + ViT on meme classification tasks

4 Results and Discussion

The classifiers achieved their best performance on the English-language Misogyny Meme Dataset, attaining an F1-score of 0.85. This dataset is balanced and focuses on a single category of offensive content. The second-highest F1-score was obtained on the Hateful Memes Dataset, which is unbalanced and encompasses multiple, distinct categories of offensive content. Furthermore, the Hateful Memes Dataset includes benign confounding memes, that is, instances that are visually or textually similar to hateful memes but are in fact non-harmful, often referred to as contrastive or counterfactual examples.

Despite these characteristics, the Hateful Memes Dataset was constructed in a synthetic manner, with a standardised format in terms of text presentation, including uniform font and capitalisation, as well as consistent image size and quality. In contrast, for the experiments presented in Table 3, the result obtained on the Prejudice Meme Dataset was considered satisfactory, with an F1-score of 0.70.

When evaluated with the BerTimbau model, the three datasets produced lower results compared to those obtained with the English BERT model. However, as shown in Table 4, the best result came from the Prejudice Memes

Model / Scenario	BERTimbau + ViT		
	Prec.	Rec.	F1
Misogyny Memes (PT trans.)	0.61	0.60	0.59
Hateful Memes (PT trans.)	0.59	0.62	0.59
Prejudice Memes (BR-PT)	0.60	0.60	0.60

Table 4: Results of the 3 Portuguese language datasets.

dataset. Despite having fewer samples (1,500 memes) and containing a diverse range of categories (distinct types of prejudice content), this dataset consists of data in Brazilian Portuguese, which aligns with the language the model was trained on. The other two datasets, despite having specific features, achieved the same result (F1-score 0.59). For the experiments shown in Table 4, the texts from the Hateful Memes and Misogyny Memes datasets were translated into Portuguese using the T5 automatic translation model (Lopes et al., 2020).

As observed in the confusion matrix (Figure 3) for English evaluation and from the F1-scores in Table 3, the percentage of true positives and true negatives for the Misogyny Memes (Figure 3a) and Hateful Memes (Figure 3c) datasets follow a similar pattern, with a slightly higher percentage of true positives in both cases. In contrast, for the Prejudice Memes (Figure 3e) dataset, despite the factors mentioned above, the performance on the test set achieved competitive results compared to state-of-the-art approaches, with a higher-than-expected percentage for class 0 (no prejudice memes).

In the case of the Portuguese data, as shown in Table 4, the classifier performed worse on all three datasets compared to the English language experiments. This drop in performance can be attributed to the changes made, which included translating the Hateful and Misogyny Memes datasets into Portuguese and using the BerTimbau language model.

With regards to the evaluation of the test data from the Prejudice Memes dataset (Figure 3f), a more balanced situation is observed, with close percentages of true positives and true negatives, indicating that the model is managing to correctly classify both classes (prejudice memes and no-prejudice memes).

The translation process between English and Portuguese requires further refinement in order to more accurately preserve the intended semantic content.

Approach	Lang.	Dataset (Size)	Model	F1
(Moctezuma et al., 2021)	Span.	Memotion (8k)	RNN + Vis. attn.	0.21
(Li, 2021)	Tamil	Tamil Memes (2.9k)	BERT + ResNet152	0.55
Ours	**BR-PT**	Prejudice Memes (1.5k)	BERT-imbau + ViT	**0.60**
(Hossain et al., 2022)	Beng.	MUTE (4.15k)	VGG16 + B-BERT	0.67

Table 5: Results in other languages

Although the experiments conducted in Portuguese yielded lower performance than those carried out in English, the results remain competitive when compared with studies in other languages (Table 5).

Direct comparison between the evaluated approaches is not feasible due to the specific characteristics of each study, including differences in dataset composition and scale. Nevertheless, the studies presented in Table 5 highlight the respective strengths and limitations of multimodal approaches that analyse combined textual and visual content, and which aim to achieve state-of-the-art performance within their respective languages.

According to Table 5, the results indicate that it is challenging to develop solutions of high performance in the binary classification task of harmful memes. Another aspect to be observed is that all studies presented in Table 5 used a lower dataset in comparison with the dataset in the English language (over 10k samples), and in some cases they also used data originally in the English language translated to the target language (e.g., Bengali and Spanish), in other words, the amount of data is a bottleneck when it comes to languages with few resources.

According to Verma et al. (2022), the multimodality, for example, the textual modality with the visual modality (e.g. text and image), can bring considerable gains in the multimodal model performance since the visual modality transcends the language and can be extremely informative. When analyzing classification errors presented with the constructed dataset, it was observed that when harmful content is

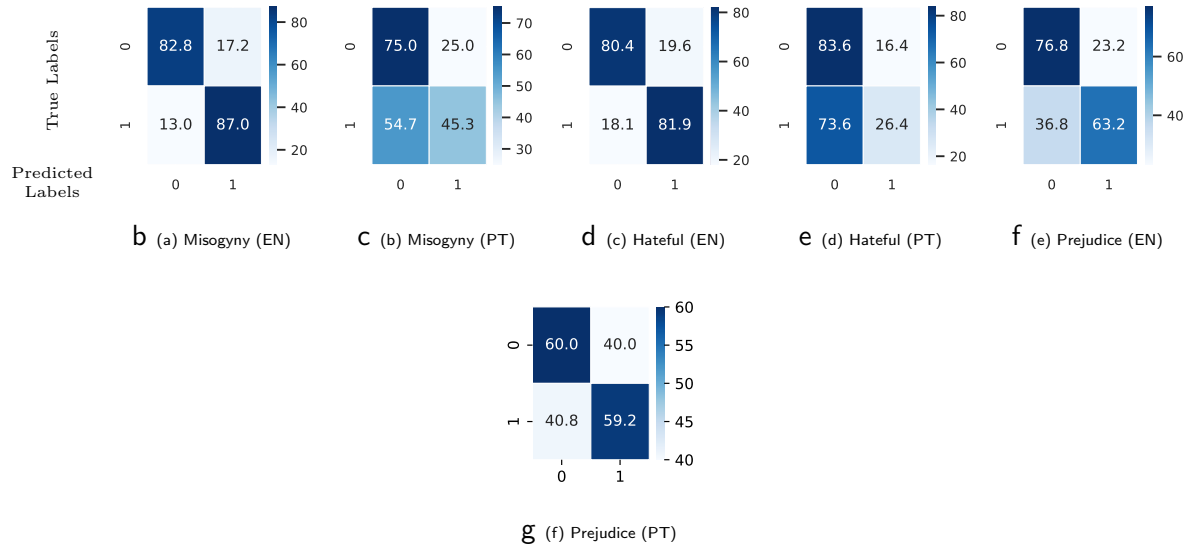


Figure 4: Confusion matrices of test data in English (EN) and Portuguese (PT) for the three datasets: Misogyny Memes, Hateful Memes and Prejudice Memes.

present in the text and also in the image, the error rates were considerably lower in comparison to error rates of memes when the harmful content is present only in the text or only in the image.

In contrast, as presented in the experiment results, such as in the results for other language studies for the same task (multimodal binary classification) (Table 5), Verma et al. (2022) also emphasize that the vision models depend mostly of the informative content from the images to the classification task.

Despite multilingual models, projected to more than a hundred other languages besides English, it's a crucial factor that these model performances are directly related to the amount of data (of each language) with which the model was pre trained, once that generate a dependency related to the corpus of the linguistic model. In other words, a superior representation in the pre training corpus is related to better performance in subsequent tasks (Verma et al., 2022).

4.1 Error Analysis

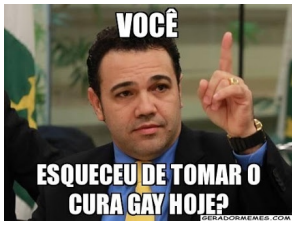
To obtain a wider view of the predictive capacity of the model and understand the recognition and classification of memes with prejudiced content, an error analysis was conducted, in both scenarios (English and Brazilian Portuguese).

In scenario 1, the Misogyny Memes and the Hateful Memes datasets presented a balanced amount of false positives and false negatives. The Prejudice Memes dataset, the situation is a litte different, the number of false positives is higher in comparison to the false negatives, indicating that for that dataset, the model faced more difficulty in classifying correctly the memes of class 1 (prejudice memes).

When investigating which categories presented the highest number of false negatives in scenario 1, it was found that for the dataset of Misogyny Memes, the stereotype category presented the highest number of false negatives (36) followed by the objectification category (32). Regarding the dataset of Hateful Memes, the religion category presented the highest number of false negatives (31) followed by the sexism and racism category (29 in both). The data on Prejudice Memes, the racism category was the one that presented the most errors (28) followed by the homophobia category (20).

It is important to notice that a specific meme may belong to more than one category (e.g., racism and misogyny), in other words, multilabel. Such a condition occurs in the three datasets used. According to Kassim et al. (2024), on literature, there are two main approaches to handling the multilabel classification: (i) problem transformation, in which the original problem of multilabel classification is turned into one or more problems of unique label; and (ii) algorithm adaptation, which modifies an existent algorithm to attend the

(a)



(b)

(c)

Figure 5: Examples of memes incorrectly classified: (a) Harmful content generally is present only in the text, (b) references to the news, celebrities, and cultural practices, and (c) sarcastic content.

restriction of multilabel data.

In scenario 2, the linguistic model used was the Brazilian Portuguese (BerTimbau), to the Hateful and Misogyny Memes datasets the number of false positives was higher than false negatives, indicating that, to this dataset, the model has found more difficult to classify correctly the memes of class 1 (hateful and misogyny memes). While, for the Prejudice Meme dataset, the model had the most difficulty correctly classifying the memes of class 0 (not prejudice memes).

When analyzed by category, it was observed that the most number of false positives in scenario 2, to the Misogyny Memes dataset was the same categories as from scenario 1, stereotype (279) and the category objectification (243). Related to the Hateful Memes dataset, the category of religion presented the highest number of false positives (158) followed by the category of racism (147). For the prejudice memes dataset, the categories fatphobia, ageism, and classism obtained the highest number of false negatives (3 in each category).

A detailed analysis of the memes revealed the following observations: (i) in many cases, the harmful content is predominantly conveyed through the textual component alone (Figure 5a); (ii) several memes include references to news events, celebrities, and cultural practices (Figure 5b); and (iii) some memes contain sarcastic content, which is often only fully understood when the image and text are interpreted jointly, together with external knowledge related to the cultural or contextual references embedded in the meme (Figure 5c).

5 Conclusion

Considering the speed of production of content on the web, harmful content can not be combated merely with a simple manual inspection by humans. In this sense, automatic learning and artificial intelligence can play a very important role in the mitigation of social problems, such as the predominance of prejudiced speech.

This paper presents a multimodal approach to the binary classification of prejudiced speech in Brazilian Portuguese memes, having yielded results compatible with the state of the art. The discussion of the results suggests that the learning transfer technique could be further improved by using a larger dataset.

Although a lower performance was obtained than models trained and evaluated on English-language memes, when the results obtained are indirectly compared with those of models developed for non-English languages, it is observed that such results are consistent with the recent work (F1-score of 0.60).

Therefore, the translation process requires further refinement to effectively incorporate the intended message and expand the dataset in the target language, ultimately leading to improved results. Additionally, it was observed that the presence of numerous categories, along with their limited representativeness (i.e., few examples per category), can hinder the model's learning process by restricting its generalization ability.

The main threats to the validity of this study comprised the high demand for computational resources required to train the model, the scarcity of available data in the target language, and the noise potentially introduced by the machine translation process.

In future work, we plan to expand the meme dataset and focus on one or two specific types of prejudice. This is essential, as effective pattern recognition in multimodal models requires a substantial amount of labelled data for each targeted category of prejudice. In addition, we intend to evaluate alternative model architectures in order to improve classification performance in the target language. We also plan to analyse the visual and textual modalities separately to gain a clearer understanding of the contribution of each modality to meme classification.

References

Usman Ahmed and Jerry Chun-Wei Lin. 2022. Deep Explainable Hate Speech Active Learning on Social-Media Data. *IEEE Trans. on Computational Social Systems*.

Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. CrisisMMD: Multimodal Twitter Datasets from Natural Disasters. In *Proc. of the Int. AAAI Conf. on Web and Social Media*, volume 12.

Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. 2022. A Survey on Deep Multimodal Learning for Computer Vision: Advances, Trends, Applications, and Datasets. *The Visual Computer*, 38(8):2939–2970.

Emily M. Bender. 2019. The #BenderRule: On Naming the Languages We Study and Why It Matters. *The Gradient*.

Giovanni Burbi, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. 2023. Mapping Memes to Words for Multimodal Hateful Meme Classification. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2824–2828. IEEE.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-End Object Detection with Transformers. In *Proc. of ECCV*, pages 213–229. Springer.

Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. 2021. Transformer Tracking. In *Proc. of CVPR*, pages 8126–8135.

Bowen Cheng, Alex Schwing, and Alexander Kirillov. 2021. Per-Pixel Classification Is Not All You Need for Semantic Segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875.

Christine Basta Marta R Costa-juss and Noe Casas. 2019. Evaluating the Underlying Gender Bias in Contextualized Word Embeddings. *GeBNLP*, page 33.

Mithun Das and Animesh Mukherjee. 2023. Transfer Learning for Multilingual Abusive Meme Detection. In *Proc. of the 15th ACM Web Science Conference*, pages 245–250.

Richard Dawkins. 1976. *The Selfish Gene*. Oxford University Press, Oxford.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification. In *Proc. of SemEval*, pages 533–549.

Paulo Finardi, Jos'e Di'e Viegas, Gustavo T. Ferreira, Alex F. Mansano, and Vinicius F. Carid'a. 2021. BERTaú: Itaú BERT for Digital Customer Service. *arXiv preprint arXiv:2101.12015*.

Joseph L. Fleiss, Bruce Levin, and Myunghee Cho Paik. 2013. *Statistical Methods for Rates and Proportions*. John Wiley & Sons.

Francesca Gasparini, Giulia Rizzi, Aurora Saibene, and Elisabetta Fersini. 2022. Benchmark Dataset of Memes with Text Transcriptions for Automatic Detection of Multi-Modal Misogynistic Content. *Data in Brief*, 44:108526.

Carlos AM Gouveia. 2009. Texto e Gramática: Uma Introdução à Linguística Sistémico-Funcional. *Matraga-Rev. PPGL UERJ*, 16(24).

Pierre Guillou. 2021. Portuguese BERT Base Cased QA (Question Answering), Finetuned on SQUAD v1.1. Website. Https://huggingface.co/pierreguillou/bert-base-cased-squad-v1.1-portuguese.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. MUTE: A Multimodal Dataset for Detecting Hateful Memes.

In *Proc. of the 2nd Conf. Asia-Pacific Chapter of the Assoc. for Comp. Linguistics and the 12th Int. Joint Conf. on Natural Language Processing: Student Research Workshop*, pages 32–39.

Summaira Jabeen, Xi Li, Muhammad Shoib Amin, Omar Bourahla, Songyuan Li, and Abdul Jabbar. 2023. A Review on Methods and Applications in Multimodal Deep Learning. *ACM Trans. on Multimedia Computing, Communications and Applications*, 19(2s):1–41.

Yifan Jiang, Shiyu Chang, and Zhangyang Wang. 2021. TransGAN: Two Pure Transformers Can Make One Strong GAN, and That Can Scale Up. *Advances in Neural Information Processing Systems*, 34:14745–14758.

Mohammed Awal Kassim, Herna Viktor, and Wojtek Michalowski. 2024. Multi-Label Lifelong Machine Learning: A Scoping Review of Algorithms, Techniques, and Applications. *IEEE Access*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. *Advances in Neural Information Processing Systems*, 33:2611–2624.

Zichao Li. 2021. CodeWithZichao@DravidianLangTech-EACL2021: Exploring Multimodal Transformers for Meme Classification in Tamil Language. In *Proc. of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 352–356.

Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards Explainable Harmful Meme Detection through Multimodal Debate between Large Language Models. In *Proc. of the ACM on Web Conf. 2024*, pages 2359–2370.

Yang Liu, Yao Zhang, Yixin Wang, Feng Hou, Jin Yuan, Jiang Tian, Yang Zhang, Zhongchao Shi, Jianping Fan, and Zhiqiang He. 2024. A Survey of Visual Transformers. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):7478–7498.

Alexandre Lopes, Rodrigo Nogueira, Roberto Lotufo, and Helio Pedrini. 2020. Lite Training Strategies for Portuguese-English and English-Portuguese Translation. In *Proc. of the Fifth Conference on Machine Translation*, pages 833–840. Association for Computational Linguistics.

Xueming Luo, Nan Jia, Erya Ouyang, and Zheng Fang. 2024. Introducing Machine-Learning-Based Data Fusion Methods for Analyzing Multimodal Data: An Application of Measuring Trustworthiness of Microenterprises. *Strategic Management Journal*.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6):1–35.

Daniela Moctezuma, Víctor Muníz, and Jorge García. 2021. Multimodal Data Evaluation for Classification Problems. *Computer Science & Information Technology*, 11(21).

Gabriel Nascimento, Flavio Carvalho, Alexandre Martins da Cunha, Carlos Roberto Viana, and Gustavo Paiva Guedes. 2019. Hate Speech Detection Using Brazilian Imageboards. In *Proc. of the 25th Brazillian Symposium on Multimedia and the Web*, pages 325–328.

Trishan Panch, Heather Mattie, and Rifat Atun. 2019. Artificial Intelligence and Algorithmic Bias: Implications for Health Systems. *J. of Global Health*, 9(2).

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. MOMENTA: A Multimodal Framework for Detecting Harmful Memes and Their Targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455, Punta Cana, Dominican Republic. Association for Computational Linguistics.

J'ulia Rocha Junqueira, Felix da Silva, Wesley Costa, Rodrigo Carvalho, Alexandre Bender, Ulisses Correa, and Larissa Freitas. 2023. BERTimbau in Action: An Investigation of Its Abilities in Sentiment Analysis, Aspect Extraction, Hate Speech Detection, and Irony Detection. In *The Int. FLAIRS Conf. Proc.*, volume 36.

C'assia CS Rosa, F'abio V Martinez, and Renato Ishii. 2023. Natural Language Processing Techniques for Hate Speech Evaluation for Brazilian Portuguese. In *Int. Conf. on Computational Science and Its Applications*, pages 104–117. Springer.

Sanjiban Sekhar Roy, Akash Roy, Pijush Samui, Mostafa Gandomi, and Amir H. Gandomi. 2024. Hateful Sentiment Detection in Real-Time Tweets: An LSTM-Based Comparative Approach. *IEEE Trans. on Computational Social Systems*.

Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra. 2020. BioBERTpt: A Portuguese Neural Language Model for Clinical Named Entity Recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72.

Mayukh Sharma, Ilanthenral Kandasamy, and WB Vasantha. 2022. R2D2 at SemEval-2022 Task 5: Attention Is Only as Good as Its Values! A Multimodal System for Identifying Misogynist Memes. In *Proc. of SemEval*, pages 761–770.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.

F'elix Silva and Larissa Freitas. 2022. Brazilian Portuguese Hate Speech Classification Using BERTimbau. In *The Int. FLAIRS Conf. Proc.*, volume 35.

F'abio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Proc. BRACIS, Part I 9*, pages 403–417. Springer.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Chi Thang Duong, Remi Lebret, and Karl Aberer. 2017. Multimodal Classification for Analysing Social Media. *arXiv e-prints*, pages arXiv–1708.

Gaurav Verma, Rohit Mujumdar, Zijie J. Wang, Munmun De Choudhury, and Srijan Kumar. 2022. Overcoming Language Disparity in Online Content Classification with Multimodal Learning. In *Proc. of the Int. AAAI Conf. on Web and Social Media*, volume 16, pages 1040–1051.

Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. 2021. Mask2Former: End-to-End Panoptic Segmentation with Mask Transformers. In *Proc. CVPR*, pages 5463–5474.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proc. of the NAACL Student Research Workshop*, pages 88–93.