MIMEME ATTRIBUTE CLASSIFICATION USING LDV ENSEMBLE MULTIMODEL LEARNING

Dhivya. K[1], Akoramurthy. B[2], T. Sivakumar[3], M. Sathya[3]

[1]M. Tech Scholar, Department of Computer Science & Engineering,
Pondicherry University Puducherry, India.
[2]Assistant Professor, Department of Computing & IT, REVA University,
Bangalore, Karnataka, India.
[3]Assistant Professor, Department of Computer Science & Engineering,
Pondicherry University, Puducherry, India.

ABSTRACT

One of the most common types of social networking interaction is memes. Memes are innately multimodal, so studying and processing them is a hot issue currently. This study's analysis of the DV dataset comprises classifying memes according to their irony, humour, motive, and overarching mood. The effectiveness of three different creative transformer-based strategies has been carefully examined. The DV Dataset used here is created by own meme data for this implementation analysis of hateful memes. Out of all of our strategies, the proposed ensemble model LDV obtained a macro F1 score of 0.737 for humour classification, 0.775 for motivation classification, 0.69 for sarcasm classification, and 0.756 for overall sentiment of the meme.

KEYWORDS

Mimeme, multimodal classification, LDV, Social-Media, hateful meme.

1. INTRODUCTION

Social media is now widely used and accepted, which has led to the emergence of many new communication methods. Memes are one of the many unusual forms of communication. These are versatile, context-aware communication techniques. Digital memes are frequently inspired by the entertainment industry, politicians for vote banking, television programs, religion, and popular culture. Creative replication and intertextuality are two separate traits of memes. In contrast to the word "intertextuality," which refers to the mixing of many customs, "creative replication" alludes to the merging of multiple scenes. The study of multimodal issues has grown in popularity recently, with applications ranging from visual question answering (VQA) [1] to image captioning [2] and ahead. Since many real-world issues involve multimodal by design, such as the way people view and comprehend their surroundings and the way they handle data that is gathered "in the public" on the Web, multimodal tasks are interesting. Numerous practitioners think that multimodality holds the solution to a variety of issues, including embodied AI, computer vision evaluation [3], and language processing translation [4]

Despite with their achievement, it is rarely clear how much sincerely multimodal understanding and reasoning are needed to solve the several daily duties and datasets. For instance, it has been noted that language can unintentionally impose powerful priors that produce achievement that appears to be impressive, even when the image data of the model structure is not understood. Similar issues were discovered in multimodal translation software, in which it was discovered

that the image mattered comparatively little, where basic baselines lacking sophisticated multimodal comprehension performed amazingly well. In this work, we present a task set with clear evaluation metrics and an immediate application in the actual world, intended to assess sincerely multimodal overview and justification.

Facebook posts have developed into a significant medium for articulation and are an important factor of the cultural context on social media platforms. Memes are widely used on current social media websites since they are simple to absorb and contain a humorous component. Examples of these widely disseminated memes can be found in Figure 1. Additionally, social networks has make things exceedingly simple to post memes; everything it needs is 2 hits to ensure that all ofyour friends and followers are showing the same meme on their feeds. They can also help the spreading of falsehood and hateful speech, similar to other kind of online information. Eventually, individual memes contribute significantly to the construction of our faiths and serve as a secondary source of data for us.

There is an immediate necessity to quick and flexible computerized systems because of how much of this information is created, making it hard for humans to weed out the offensive memes. The results of Facebook AI's Hateful Memes Challenge [5] show how difficult this challenge is to solve, as comprehending a meme necessitates a lot of cultural, political, and historical knowledge as well as multimodal awareness. Some other crucial point to make is that all of the human observers struggled to categorise individual memes, and also the annotating process required an average of 27 minutes per meme. And finally, despite the fact that findings of earlier studies may have seemed encouraging, considering the size of the internet, this still translates to a number of ugly memes becoming posted in publicly.

The motivation is to predict whether a meme is hateful or non-hateful. This is essentially a binary classification problem with multimodal input data consisting of the meme image itself (the image mode) and a string representing the text in the meme image (the text mode).

Figure 1: These two memes are well understood, but it's visible that one encourages hateful and negativity of the Hindu temple while the other does the opposite to that it is not hateful with peace.

2. LITERATURE SURVEY

During the initial stage of the project, we began examining pertinent books and articles on the categorization of hateful memes. Numerous studies that make an effort to answer the issue were discovered throughout our literature review. We mainly took into account articles that suggest a multimodal strategy although this challenge necessitates knowing both aspects at once, whereas

unimodal approaches exclusively concentrate on one modality, i.e., either image or text. The below is a list of publications with brief summaries:

2.1. Hateful Speech Recognition in Multimodal Memes [21]

In this study, it was found that simply by identifying the meme's appearance, the most of data sets in the Facebook Meme Dataset that have been initially nasty were transformed into hate less. This would be performed well because, in just this content, word or visual-only unimodal models might fall short and also only multimodal models will also be capable of producing correct reasoning. Many multimodal models prioritise hateful speech a little more in the different languages. Consequently, while concentrate on the visual modality, images labelling and object detection are being used. The real text of the meme is retrieved using object detection and images captioning, as well as the characteristics from the real text are then blended with multimodal modelling to produce the final estimate. In parallel to the methods described previously, multimodal characteristics using pre-trained models are employed as well as, this is together as unimodal emotional characteristics for both the text and images. To convey the content and connection between the 2 modalities, this is achieved.

2.2. Prize-winning response to the Hateful Memes Challenge: Detecting Hateful Speech in Memes Using Multimodal Deep Learning Approaches [20]

In this study, multimodal representations is obtained using VisualBERT [10]. The method is divided into four stages: training, dataset extension, image encoding, and ensemble learning. Several memes out from Emotion Dataset are chosen for the dataset extension phase. The choice is made depending on how closely the memes resemble those in the dataset of hateful memes. That used a ResNeXT-152 driven Mask-RCNN model, 2048-dimensional region-based picture characteristics are extracted during the Image Encoding phase. The text embedding space is therefore constructed with all these properties. In order to obtain the multimodal images, a pre-trained visualBert model is used. The training process is used to adjust the VisualBert model. The identity from the transformers model is utilised to sync the verbal paradigm and visual paradigm. The speech and image portions were joined and serve as the transformer's source. The result of a linear transformation (Wx + b) on the result of the transformer is classified via SoftMax. Several models are utilised for includes all kinds, and the eventual data point is predicted using a clear majority technique.

2.3. Multimodal Learning for the Detection of Hateful Memes [7]

The suggested model used in this study consists of a classifier, an object detector, a triplet-relation network, and a photo captioner. They take into account the picture caption, meme text, and visual elements as 3 distinct types of information gleaned out of each meme using object detection. The proposed triplet-relation network adopts the cross-attention model to understand the most discriminant information using cross-modal embeddings, modelling the triplet interactions among some of the caption, objects, and OCR words. In order to create sentence embeddings, picture features are first extracted who used an image encoder, and then visual data are converted into sentences using a sentence decoder. OCR text extracted features and image caption semantic similarity are combined in the end. Faster R-CNNs are used to create image encoding after they have been trained. The cross-modality interactions among both textual features and image features are modelled using the triplet-relation network, which is effectively a transformers network. To determine the predictions probability, a full-connected layer is given, then a SoftMax layer, along with a combined representations of the textual and visual information that was generated from the transformers model.

The suggested model used in this study consists of a classifier, an object detector, a triplet-relation network, and a photo captioner. They take into account the picture caption, meme text, and visual elements as 3 distinct types of information gleaned out of each meme usingobject detection. The proposed triplet-relation network adopts the cross-attention model to understand the most discriminant information using cross-modal embeddings, modelling the triplet interactions among some of the caption, objects, and OCR words. In order to create sentence embeddings, picture features are first extracted who used an image encoder, and then visual data are converted into sentences using a sentence decoder. OCR text extracted features and image caption semantic similarity are combined in the end. Faster R-CNNs are used to create image encoding after they have been trained. The cross-modality interactions among both textual features and image features are modelled using the triplet-relation network, which is effectively a transformers network. To determine the predictions probability, a full-connected layer is given, then a SoftMax layer, along with a combined representations of the textual and visual information that was generated from the transformers model.

The proposed model in this paper involves an ensemble approach. To use the detection 2 architecture, picture characteristics are gathered in the initial stage. The ensemble models are then feed the meme content in addition to the characteristics. The five models in the ensemble are VisualBERT, OSCAR, UNITER, and ERNIE-ViL (small, large).

2.4. Detecting Offensive Memes for Automatic Moderation in "Hate Speech in Pixels" [9]

A tiny dataset of 5020 memes is introduced in this work, although the categorization wasn't performed in an exact and regulated way. This model is fairly straightforward; its extracts text from images with OCR and afterwards analyzes this using BERT to achieve a characteristic embedding again for phrase. Images are treated using a pre-trained VGG net to produce a visual feature.

2.5. Multimodal Learning for Detecting Hateful Memes [10]

The suggested model used in this study comprises of a classifier, an object detector, a triplet-relation network, as well as an image captioner. Those who take into account that image caption, meme text, and visual elements as 3 distinct types of information gleaned out of each meme using object detection. This suggested triplet-relation network adopts the cross-attention model to extract ever more discriminant information using cross-modal embeddings, modelling the set of twins among caption, objects, and OCR words. In order to create phrase embeddings, picture characteristics are first extracted using an image encoder, and afterwards visual data are converted into sentences to use a word decoder. OCR text embeddings with image caption embeddings were combined in the end. Faster R-CNNs are used to create image embeddings after they have been learned.

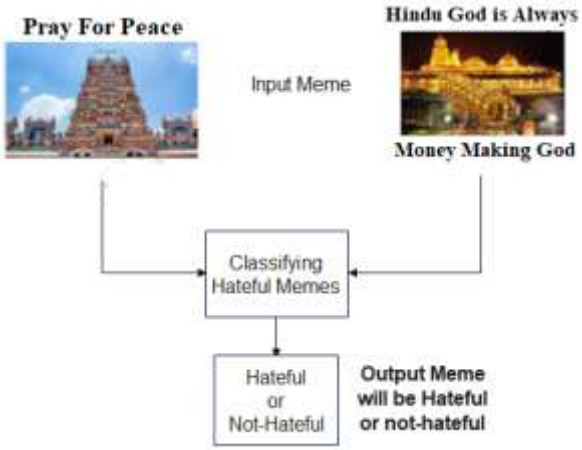

Figure 2: Expected outcome for Hateful meme classification

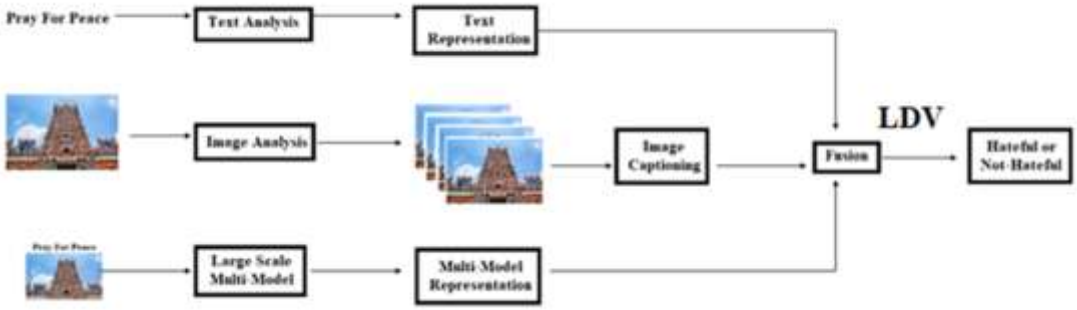

Figure 3: Process Model for LDV Multi-Model

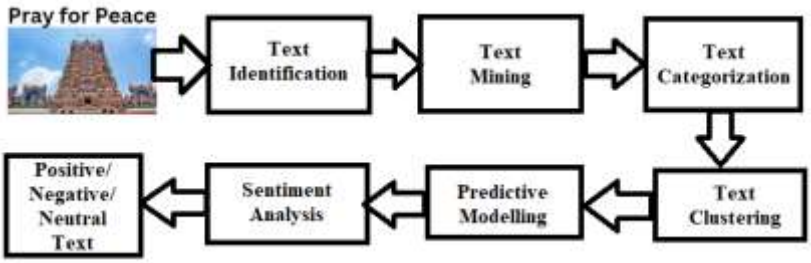

Figure 4: Text Analysis Process for Multimodel

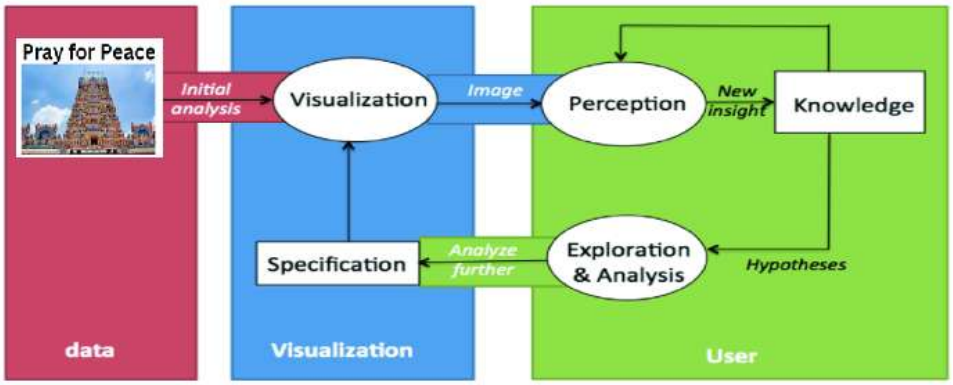

Figure 5: Visual / Image Analysis Process for Multimodel

3. LDV ENSEMBLE PROCESS MODEL

3.1. Expected Outcome for Hateful Memes Classification

This approach recognizes the meme's images and text as inputs and produces a tag that indicates whether the memes is hateful or not. We will use the process and design described in Section 4 for implementation and analysis. We expect that model will perform well. Figure 2 depicts the expected result of our project based on an input meme.

3.2. Process Model for Proposed Multimodel Approach

The Proposed LDV multimodel approach (Figure 3) is used for analysis the memes with Text Analysis Process (Figure 4) with text representation and fusion to classify it and Image Analysis Process (Figure 5) for Multimodel process with image analysis, image capturing and fusion to classify both to identify the Divi dataset is hateful or not hateful is shown in the figure.

4. DATA & METHODOLOGY

4.1. Data

There are 1809 photos in the DV dataset. Each image has an equal amount of Comical, Mockery, Inspiration, and basical sentiment. Each humorous example is split into two groups: comical and non-comical. The mockery and non-mockery subcategories of each sarcastic sample are separated. Both motivational and non-motivational samples are classified as inspirational/motivational. Positive, negative, and neutral sentiment are the three categories used to categorise basical sentiment. Table 1 shows how unbalanced the data set is. The fraction is therefore resampled in order to resemble the population using a method known as "synthetic minority oversampling" [16]. After rebalancing the dataset using the SMOTE approach, we utilised the pareto principle, i.e., 80% of the train data and 20% of the text data for every mental state.

Table 1. Data Diffusion

Diffused Data	Positive	Neutral	Negative
Comical	**236**	**1001**	**572**
Mockery	**1232**	**2**	**575**
Inspirational /Motivational	**980**	**0**	**829**
Basical sentiment	**786**	**0**	**1023**

4.2. Methodology

A particular form of multi-media messaging is a meme. It contains some words and a picture etched into it. Typically, the picture is created by drawing inspiration from well-known sources like cinema, politics, or animations. The wording is distinct for each user and environment.

Each of these is capable of expressing the perspective and significance of the specific context in terms of emotion. In this work, we only utilised transformer models. The integration of word embeddings and picture embeddings results in a powerful depiction of the mimeme. We'll discuss further details in the subsections that follow.

A.Pictorial Illustration: Image Transformation: Image Transformers is the most comprehensive visual representation currently available on the market. They use a mild and periodic focus structure that makes it easier to recognise the key components of the image. The 14-by-14-inch blocks that make up each picture are separated, and before being delivered to the image transformer network, each tile is positionally encoded. To guarantee that the areas with the greatest effect are prioritized, we use a range of multi-head attention and equalization strategies within the transformer.

B.Photo descriptions: It's conceivable that the picture won't be able to capture the context in a meaningful way. Graphical aspects are frequently believed to be less effective in expressing contextual characteristics than verbal ones. In order to effectively provide context, we developed a transformer-based image tagging model that could describe pictures. Then, a conventional image transformer model is used to translate these descriptions into embeddings.

C.Linguistic illustration: BERT, often referred to as [20] bidirectional encoder representation from transformers, is a more appropriate way to express words than glove embeddings or quick text. Compared to LSTMs or gated recurrent units, BERT has the benefit of being faster and better able to describe the relationship between words. They have a certain directionality about them. Their methodology is based on self-attention. The context of the statement may be faithfully represented using BERT embeddings. The Masked Language Model (MLM) and Next Sentence Prediction phases of BERT training are sequential. Akin to the Bag of Words approach, the two phases run concurrently and support one another's learning. The entire statement is encoded in this way.

D.Multimodal Representation-Fusion Method: This technique is used to create multivariate data by merely concatenating the picture and text values along the first axis. It is good to provide this multivariate input to the LSTM model.Facebook AI Research offered pre-trained models, as well as the coding and environments required to duplicate the findings, under the mmf framework. We initially examined the models' efficiency on the pre-trained values, then trained all 11 models to see whether their performance is improved. The hyper - parameters employed have been those specified in the challenge report. Independent of batch size, the were trained or fine-tuned for a max of 22000 updates.

Table 2 shows the outcomes of fine-tuning the values of baseline methods on the Divi dataset.

Types	Model	Accuracy	AuROC
Unimodal	Image-Grid	61.6%	0.56
	Image-Region	60.8%	0.57
	Text BERT	62.8%	0.60
Multimodal with Unimodal Pre-training dataset	Late Fusion	65.1%	0.64
	Concat BERT	65.7%	0.65
	MMBT-Grid	66.1%	0.66
	MMBT-Region	68.3%	0.72
	ViLBERT	71.4%	0.71
	Visual BERT	71.0%	0.73
Multimodel with Multimodel Pre-training dataset	ViLBERT CC	70.7%	0.72
	Visual-BERT COCO	69.1%	0.74

The accuracy & AuROC scores during finetuning are shown in Table 2. A few of our procedure observes that are congruent with those listed in the contest article are included below.

The improvement in accurate and AuROC values during fine-tuning has been insufficient, and in a few cases were negative for a few of the baseline methods, showing that there needs to be a better approach to train the data with multimodal goals. The efficiency of unimodally pre-trained as well as multimodally pre-trained models differs little across multimodal designs, showing that multimodal model pre-training could be enhanced. Including sophisticated merging of the two different modalities, performance increases. ConcatBERT is the example for outperforms Late Fusion after a little fine-tuning. It would be supplanted by multimodal designs such as MMBT, ViLBERT, and VisualBERT. Human AuROC of 0.8465 is still much worse than the best baseline AuROC of 0.74. As a result, addressing such gap is becoming a driving factor in experimenting with other solutions in the work. The appearance of the optimal updating earlier in several models implies an overfitting condition. Providing the tiny amount of the datasets in comparison to the model designs chosen, a lower number of iterations should be adequate to achieve acceptable results.

4.3. Adding Additional Training Models(LDV)

We trained three more models, DeVLBERT [22] LXMERT [18] and VilBERT [19] using the hateful memes dataset using existing PyTorch[14] implements as well as introduced them towards the Vilio ensembles to try to increase the result. Both of these are multimodal transformer designs with double streams. A summary comparing various designs is provided below.

4.3.1. DeVLBERT [22]

DeVLBERT design seeks for de-confound visuo-linguistic representations through addressing the issue from out pre-training and compares their efficiency to Vil-BERT. Since of variables in the information being used pretraining, spurious associations were discovered in models. These suggest a large conditional distribution from one token supplied another in the lack of a substantial link among them; for example, the ViLBERT model demonstrates an extremely high conditional probability again for visual object shirt providing the word instrument. The design mitigates the issues of false correlates and increases generalisation capabilities including out pre-training data by incorporating the concept via covert modification via causal inference. They

suggest a great conditional distribution of that single token given by another in the lack of every powerful association among them; for example, the ViLBERT model displays an extremely great conditional distribution for the visual object shirt provided the word 'instrumental'. The design addresses the problem of correlation and enhances generalisation capabilities for the out of pre-training data by including the concept of backdoor modification from normal inference.

4.3.2. LXMERT[18]

LXMERT's design is divided into three separate transformer models: speech encoder for verbal modality, object related encoder for picture modality, & cross-modality encoder for combining two modalities. However that ViLBERT design, that has both the images and textual transformer streams entangled across the layers, the modalities were segregated for several layers in the this architecture, before the bridge encoder permits multimodal classification method. Information from 5 vision-and-language datasets are utilised for pre-training, using pictures from COCO[11] and otherwise Visual Genome datasets, as well as image feature caption datasets - VQA 2.0, GQA, and VG-QA.

4.3.3. ViLBERT

ViLBERT (Vision-and-Language BERT) is a paradigm for developing task-independent combined descriptions of picture content and natural language processing. They expand the well-known BERT framework to multi-modal models were compared which processes both visual as well as textual input data in distinct channels that communicate via co-attentional transformation layers. Its pre-train with us model on the massive think, instantly gathered Basic theoretical Descriptions dataset using 2 different proxy tasks, and afterwards transmit it to multiple formed vision-and-language tasks - graphical issue answering, graphical sensible logic, relating emotions, and caption-based image processing techniques - with only slight adjustments to the base architecture. We see considerable gains across activities as when compared to previous mission models, with all four levels obtaining state-of-the-art performance. These findings signal a movement away from studying groundwork among vision and language as parts of cognitive task and more towards seeing visual anchoring as a pre-trainable and transportable talent.

Table 3. AuROC evaluation of the Vilio model with only an expanded model which consisting of all three of these LXMERT, DeVLBERT and ViLBERT models.

Combination of modals	**Dataset Used**	**AuROC**
VisualBERT, UNITER, OSCAR, ERNIE-ViL	Divi Dataset	0.8319
	Divi-Test Dataset	0.8454
VisualBERT, UNITER, OSCAR, ERNIE-ViL, **LXMERT, DeVLBERT, VilBERT**	Divi Dataset	0.8323
	Divi-Test Dataset	0.921

The Table 3 illustrates that when two additional models are added to the initial Vilio modal, the achievement of the modals degrades (marginally). As a consequence, it chosen to stick with the core Vilio modal for additional testing.

4.4. Additional Enhancement – MoEDL

We chose to focus their future experiments purely on the Vilio design because to its far higher success than VisualBert. Here on estimates using Vilio models, they tested the Mixture of Experts in Deep Learning (MoEDL) assembling technique.

The MOEDL ensembles approach uses the result probability across numerous models to create the final result. Let's say there are N models. For each model, we already have outcome probability for an input data. There's a filtering network, that is essentially a Multi-Layer Perceptron Modal (MLPM), whose goal is to forecast N normalised primitives (ϕi). The last chance of both the input becoming hated is just a sum of the weights of the possibilities of each model, with the weighting (ϕi) calculated by using MLPM. To compute the loss, we utilise the ending probability (P). We utilised Binary Cross Entropy with Logit Loss as our loss function. We utilised a sigmoid activity over the ending probability (P) to generate the ending prediction as well as a threshold value is 0.5 to identify the input as hateful memes.

Let's, ending probability = P; Threshold value = 0.5, resulting probability = Ri;
R = (R1; R2; :::;RN)
ϕ = MLPM(R)
P = Σ(i = 1 to N) [ϕi * P^i]
Ri denotes the resulting probability from the ith model, and P is the weighed sum of each individual probability.

Table 4: MoEDL Ensemble Results show an improvement in Divi dataset Accuracy is shown above from the previous results (table 3).

Dataset Used	Accuracy Rate	AuROC
Divi Dataset	75.6%	-
Divi Test Dataset	81.5%	0.732

The findings in the table 4 show that the MoEDL ensemble does not significantly improve the efficiency. The efficiency improves with small number of margins (0.4%) for the Divi dataset, although it drops again for Test dataset for testing. Their investigations revealed that now the ideal value for i are comparable, implying that filtering system gives weighting to all existing models. While examining the individual performance of the models in the outfit, we discovered that their reliability and AUROC rating remained nearly identical. Since different models execute similarly, the filtering system assigns equal weight to each of those since no single outfit surpasses the others by a massive margin.

5. EXPERIMENTAL RESULTS AND RESULT ANALYSIS

The Performance evaluation of the techniques used to recognize offence and troll in social media memes. The experiments are carried out on a GPU-enabled platform called Google colaboratory. The pandas and numpy libraries are used for data processing as well as preparation. The Divi dataset is used to all models are built with Keras as well as TensorFlow. Scikit-learn packages are used for model assessment. Training, validation, and test the Divi dataset are used to build the models. For model development, train set examples are used, whereas the validation established is used for set of parameters fine tuning and selection. At last, using test dataset occurrences, the trained models are tested.

We performed a graphical study to determine which complex types of memes are rated as hateful meme or non-hateful meme by each of different classification systems:

1. COCO VisualBERT: That model was selected out of the baseline methods implemented since it was statistically stronger compared to all of the other 11 baseline methods from the issue article.
2. VisualBERT Group: It is one of the two SoTA baselines stated in section 3 that outperformed the baseline.
3. Vilio Group: The Vilio ensembles outperformed all other models. We compare the findings of the other two models to those of Vilio. Finally, we illustrate several instances of Vilio's failures.
4. LXMERT, DeVLBERT, VilBERT(LDV) Model: The LXMERT, DeVLBERT, VilBERT outperformed to all models which we compare the findings of all three models to those of LVD with several instances.

For all of the following analysis, we concentrated on the Divi dataset, which contained 1809 memes, 236 of which are hateful meme and 572 of which are not hateful meme.

Table 5. Accuracy percentage of Training

Diffused Data	**LXMERT**	**DeVLBERT**	**ViLBERT**
Comical	**72.33**	**83.83**	**78.23**
Mockery	**62.78**	**75.34**	**78.24**
Inspirational /Motivational	**64.22**	**74.41**	**75.09**
Basical sentiment	**57.60**	**77.78**	**76.15**

Table 5 explains about training accuracy of the DIVI dataset which is used to show in the table with percentage.

Table 6. Accuracy Percentage of Testing

Diffused Data	**LXMERT**	**DeVLBERT**	**ViLBERT**
Comical	**64.13**	**68.31**	**66.97**
Mockery	**59.47**	**61.44**	**66.94**
Inspirational /Motivational	**60.43**	**67.58**	**59.07**
Basical sentiment	**51.87**	**64.21**	**64.25**

Table 6 explains about F1 Score of the DIVI dataset which is used to show in the table with percentage.

5.1. Analysis Performance of LDV model

Among the factors behind LDV model significantly improved performance are as follows:

1. Use of a superior picture feature extraction methods
2. Variation in features of collections per multi-model types
3. Different model architectures
4. Pretraining of Tasks
5. Enhanced learning techniques and grouping

5.2. Advantages of Proposed Methods

● **Use of better feature extraction for images**

○ Use of bottom-up attention in detectron-2 framework

● **Diversity in feature set per architecture**

○ Use of 3-5 different feature sets for training models for each architecture
○ Simple-averaged predictions of the trained models are used for further ensembling

● **Diverse model architectures**

○ Single-stream architectures: OSCAR, UNITER, VisualBERT
○ Dual-stream architectures: ERNIE-Vil - Small & Large - using models pretrained on CC / VCR datasets

● **Task-adaptive pretraining**

○ OSCAR was pretrained on Image-text matching, MLM
○ VisualBERT was pretrained on MLM

● **Better training strategies and ensembling**

○ Stochastic Weighted Averaging for last 25% of training
○ Simplex Optimization for final ensembling

6. FUTURE WORK

This project may be expanded in many different ways. By fine-tuning the multimodal approaches on a Tamil dataset instead of an English language, one may examine the Tamil memes has greater detail. Or, attempts could be made to create classifiers, regardless of the dataset which are trained on, generalise to ideas of any social or verbal domain. Considering low-resource, badly supervised scenarios in several languages throughout the world, this might be quite helpful. Since we've seen that when models are trained on memes that have been converted into multiple languages, they don't perform with the same level of accuracy, and developing such annotated dataset for each and every language would also be a time-consuming, expensive effort. Additionally, it is possible to experiment with brand-new ensembling as well as data-augmentation methods that incorporate both textual and picture augmentation. The possibilities are unlimited because it is still a relatively young area of study for social cause.

7. CONCLUSIONS

Although there has been a substantial increase over the benchmark observed, there is still much room for development. Overall, the average assessment precision is 68.77%, and the mean F1 score is 66.045%. In this work, we presented and explored the value of using cutting-edge visual and verbal transformer models to improve accuracy and F1 score. We also discussed the barriers that prevent various models from doing effective meme analysis as well as potential future solutions.

ACKNOWLEDGEMENTS

I would like to thank Dr.S.Sathya, my course supervisor for doing this research work. I would like to express my deep gratitude to Dr.T.Sivakumar (my project coordinator), Dr.R.Subramanian (my research supervisors) and Dr.S.Sivasathya (my HOD) for their patient guidance of this research work.

I would also like to extend my thanks to Professor Mr.B.Akoramurthy for enthusiastic encouragement, great support and useful critiques in doing this research paper.Finally, I wish to thank my parents for their wonderful support and encouragement throughout my study.

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision, pages 2425–2433, 2015.

[2] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325, 2015.

[3] Donald Geman, Stuart Geman, Neil Hallonquist, and Laurent Younes. Visual turing test for computer vision systems. Proceedings of the National Academy of Sciences, 112(12):3618–3623,2015.

[4] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. In Proceedings of ICLR, 2017.

[5] Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. Probing the need for visual context in multimodal machine translation. arXiv preprint arXiv:1903.08678, 2019.

[6] Riza Velioglu and Jewgeni Rose. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. arXiv preprint arXiv:2012.12975, 2020.

[7] Yi Zhou and Zhenhao Chen. Multimodal learning for hateful memes detection. arXiv preprint arXiv:2011.12870, 2020.

[8] Niklas Muennighoff. Vilio: State-of-the-art visio-linguistic models applied to hateful memes. arXiv preprint arXiv:2012.07788, 2020.new https://arxiv.org/abs/2012.14891v1

[9] Y. Zhou, Z. Chen and H. Yang, "Multimodal Learning For Hateful Memes Detection," 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 2021, pp. 1-6, doi: 10.1109/ICMEW53276.2021.9455994.

[10] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557, 2019.

[11] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019.

[12] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. arXiv preprint arXiv:2005.04790, 2020.

[13] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016.

[14] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering, 2016.

[15] D Kevin W. Bowyer and Nitesh V. Chawla and Lawrence O. Hall and W. Philip Kegelmeyer, Synthetic Minority Over-sampling Technique, CoRR, abs/1106.1813, 2011.

[16] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images, 2016.

[17] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers, 2019.
[18] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019.
[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding," 2019.
[20] Abhishek Das, Japsimar Singh Wahi, and Siyao Li. Detecting hate speech in multi-modal memes. arXiv preprint arXiv:2012.14891, 2020.
[21] Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. Devlbert. Proceedings of the 28th ACM International Conference on Multimedia, Oct 2020.

AUTHORS

Mrs.Dhivya.K 1st, She has completed Bachelor of Engineering in Computer Science and Engineering from Anna University and pursuing Master of Technology in Computer Science and Engineering from Pondicherry University. Her motivation towards research sparked in industry research projects. Her research interests include Data Analytics, Nature Inspired Algorithm, Big Data, Network Security and Cyber Security.

Prof.Akoramurthy.B 2nd, He has received M.E (CSE) from Anna University and completed MBA (Banking Technology) from Pondicherry University. His teaching career started as a Lecturer and is currently working as Assistant Professor in the Department of Computing and IT, REVA University, Bangalore. His industrial experience has motivated him for taking up application-oriented researches. His Research interests include Cloud Computing, Big Data, Data Analytics, and Machine Learning, IOT, Theoretical foundations of Computer Science and Contemporary technologies.

Dr.T.SivaKumar 3rd, He received his M.Tech and Ph.D. from Pondicherry University, Puducherry, India. He is presently working as an Assistant Professor in Department of Computer Science, Pondicherry University, Puducherry. His field of interest includes Data Communication Networks, Network Security, MANET and VANET.

Dr.M.Sathya 4th**,** She is currently working as an Assistant Professor at the Department of Computer Science, Pondicherry University, Puducherry, India. She earned his Ph.D. and M.Tech. in Computer Science & Engineering from Pondicherry University and a B.E. in Computer Science & Engineering from Bharathidasan University.His research interests include Evolutionary Algorithms, Swarm Intelligence, Web Service Computing and Wireless Sensor Networks.