HINDI ENGLISH - Meme Classification

Teammembers :
SRI KARTHIK VALA – 20MIA1032
SHREYAS V – 20MIA1063
NIKITHA AR – 20MIA1025

Abstract. The influence of internet culture has given rise to a diverse variety of memes which revolve around humor and satire. This project delves into the complex task of classifying memes into 4 different categories which are humor, sarcastic, offensive and motivational. The primary aim is to develop an efficient and accurate classification system capable of automatically labeling memes with their appropriate categories, and then further telling the intensity of that category. To accomplish this objective, we used the memotion 3 dataset which comprises 10,000 Hindi-English memes. With help of natural language processing (NLP) and computer vision techniques, we extract salient features from textual and visual elements within memes, enabling us to build a robust classification model. Beyond classification, the outcomes of this research hold immense importance for content moderation and meme-driven cultural insights. It paves the way for more responsible content sharing on social media platforms and empowers users with a better understanding of the intent and impact of memes.

Keywords: Meme classification, Hindi-English memes, Natural language processing (NLP), Computer vision, Content moderation

1 Introduction

The internet has become an inexhaustible source of humor, and at the heart of this digital comedy renaissance lies the meme—an emblem of online culture that transcends linguistic and cultural borders. Memes have evolved into a form of expression that encapsulates a wide spectrum of emotions, opinions, and social commentary. In a world of multilingual internet users, the proliferation of memes in various languages, including Hindi and English, has given rise to a vibrant ecosystem of online humor and satire.

This project delves into the fascinating world of Hindi-English memes with a specific focus on the classification of these digital artifacts into distinct categories. While memes often serve as vehicles of humor and satire, they also encompass diverse genres, including humorous, offensive, sarcastic, and more. Understanding and categorizing memes according to their intent and tone is not only a compelling challenge but also holds significant practical implications for content management, online discourse analysis, and cultural exploration in the digital age

. The primary objective of this research is to develop a robust and accurate classification system that can automatically identify and categorize Hindi-English memes into relevant genres. To achieve this, we leverage advanced techniques in natural language processing (NLP) and computer vision, harnessing the power of linguistic analysis and image recognition. By dissecting the textual and visual elements within memes, our aim is to shed light on the intricate nuances of online humor and enable users to navigate this vast digital landscape more effectively. This project confronts several intricate challenges, including humor comprehension across languages, sarcasm detection, and the identification of offensive content, all within the context of memes.

The outcomes of this research promise to enrich our understanding of internet humor while offering practical tools for content moderation and personalized content recommendation on social media platforms.

2 Related Works

Web Mining and Minimization Framework Design on Sentimental Analysis for Social Tweets Using Machine Learning - A sentimental tweets segregation and classification based on content is proposed under the objective of web minimization for optimized search results. The methodology includes K Nearest Neighboring (KNN) approach for extraction of similar tweets strength using the pre-learnt logs threshold values.

Text mining with sentiment analysis on seafarers' medical documents - Adoption of both lexicon and Naïve Bayes' algorithms was done to perform sentimental analysis and experiments were conducted over R statistical tool. Visualization of symptomatic information was done through word clouds and 96% of the correlation between medical problems and diagnosis outcome has been achieved.

A novel approach to stance detection in social media tweets by fusing ranked lists and sentiments - Stance detection is a relatively new concept in data mining that aims to assign a stance label (favor, against, or none) to a social media post towards a specific predetermined target. In this paper, a three-phase process is used for stance detection: (a) tweets preprocessing; clean and normalize tweets (e.g., remove stop-words) to generate words and stems lists, (b) features generation; c) classification; all the instances of the features are classified based on the list of targets.

Meme Sentiment Analysis Enhanced with Multimodal Spatial Encoding and Face Embedding - It shows performance gains from incorporating the spatial position of visual objects, faces, and text clusters extracted from memes. In addition, it also presents facial embedding as an impactful enhancement to image representation in a multimodal meme classifier.

Sentiment analysis of extremism in social media from textual information - This paper focuses on the sentimental analysis of social media multilingual textual data to discover the intensity of the sentiments of extremism. It classifies the incorporated textual views into any of four categories, including high extreme, low extreme, moderate, and neutral, based on their level of extremism.

Sentiment analysis on the impact of coronavirus in social life using the BERT model - Sentiment analysis is done using the BERT model on tweets. It is performed on two data sets; one data set is collected by tweets made by people from all over the world, and the other data set contains the tweets made by people of India.

Sentimental analysis of COVID-19 tweets using deep learning models - Data analysis was conducted by Bidirectional Encoder Representations from Transformers (BERT) model, which is a new deep-learning model for text analysis and performance and was compared with three other models such as logistic regression (LR), support vector machines (SVM), and long-short term memory (LSTM).

Sentimental analysis of Indian regional languages on social media - This paper provides information about whether the tweets posted by the customer are positive, negative, or neutral. For this the proposed model first scrape the tweets from Twitter by using Twitter APIs, then later by using text blob, the customer reviews are given different sentiment scores and classify them as positive, negative, or neutral by using text classification model.

Overview of Memotion 3: Sentiment and Emotion Analysis of Codemixed Hinglish Memes - Over 50 teams registered for the shared task and 5 made final submissions to the test set of the Memotion 3 dataset. CLIP, BERT modifications, ViT etc. were the most popular models among the participants along with approaches such as Student-Teacher model, Fusion, and Ensembling.

3. Data

In this section we describe the data collection, annotation and data analysis.

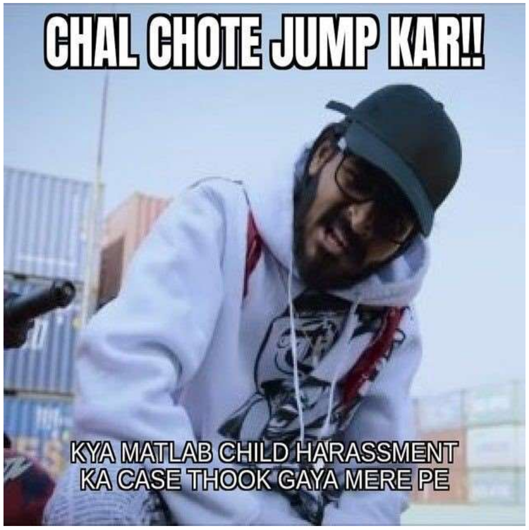


Figure 1: Example for Task A. People found this meme to have a negative sentiment.

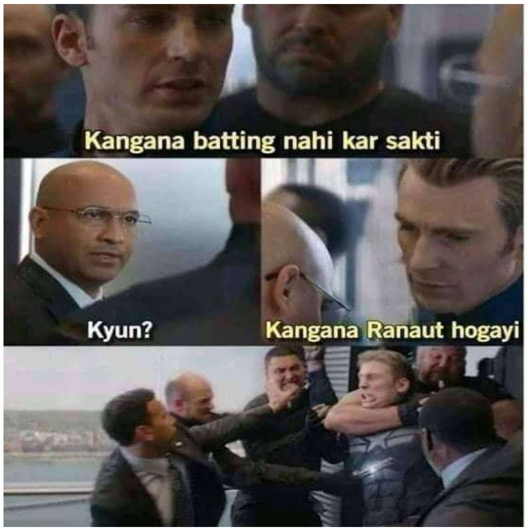
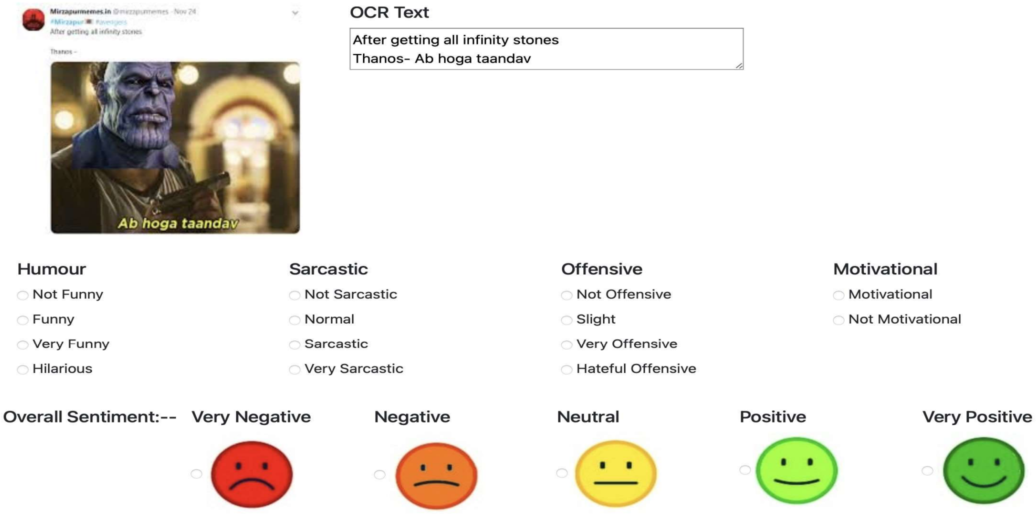


Figure 2: Example for Task B and C. Majority of annotators found this meme's humor intensity as very funny, sarcasm as twisted meaning, offensive as not offensive and motivational as not motivational. The corresponding labels for Task B will be funny, sarcastic, not offensive and not motivational.

4. **Data Collection**

We downloaded the memes after on topics of interest, such as politics, sports etc. We also collected memes using a Selenium-based web crawler. All memes are gathered from public websites Reddit and Google images. We cleaned the data to remove redundancies and performed random manual quality check. The memes are release along with the source URLs and OCR text. For OCR, we utilized the Google Vision API[1].


Mirzapurmemes.in @mirzapurmemes · Nov 24
#Mirzapur #avengers
After getting all infinity stones
Thanos –
Ab hoga taandav

OCR Text

After getting all infinity stones
Thanos- Ab hoga taandav

Humour	Sarcastic	Offensive	Motivational
○ Not Funny	○ Not Sarcastic	○ Not Offensive	○ Motivational
○ Funny	○ Normal	○ Slight	○ Not Motivational
○ Very Funny	○ Sarcastic	○ Very Offensive	
○ Hilarious	○ Very Sarcastic	○ Hateful Offensive	

Overall Sentiment:-- ○ Very Negative ○ Negative ○ Neutral ○ Positive ○ Very Positive

Figure 3: Annotator Interface. The annotators see a meme and have to mark the sentiment and emotion intensities of the meme. They also have to correct the OCR extracted using the Google Vision API, if there are any discrepancies.

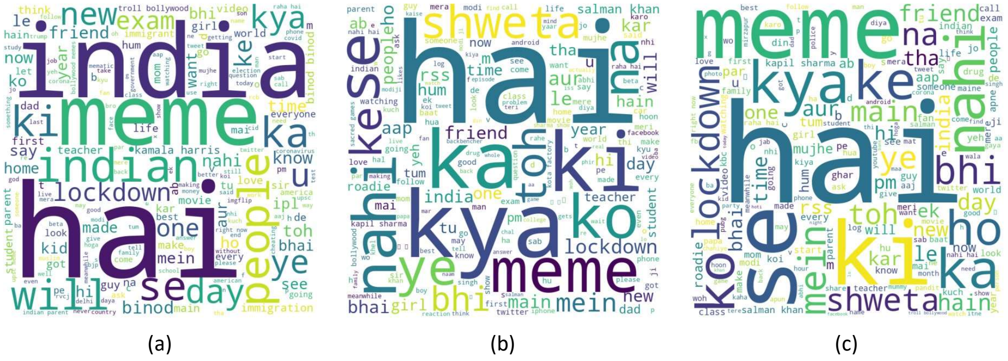


Figure 4: Word clouds indicating top words used for the (a) train, (b) validation and (c) test sets.

5. **Data Annotation**

We recruited Undergraduate student proficient in English, Hindi and meme knowledge. For annotation, they use an interface built by us, as shown in Figure 3. The annotators were asked to assess whether the meme's creator intended it to be positive, negative, or neutral in Task A.

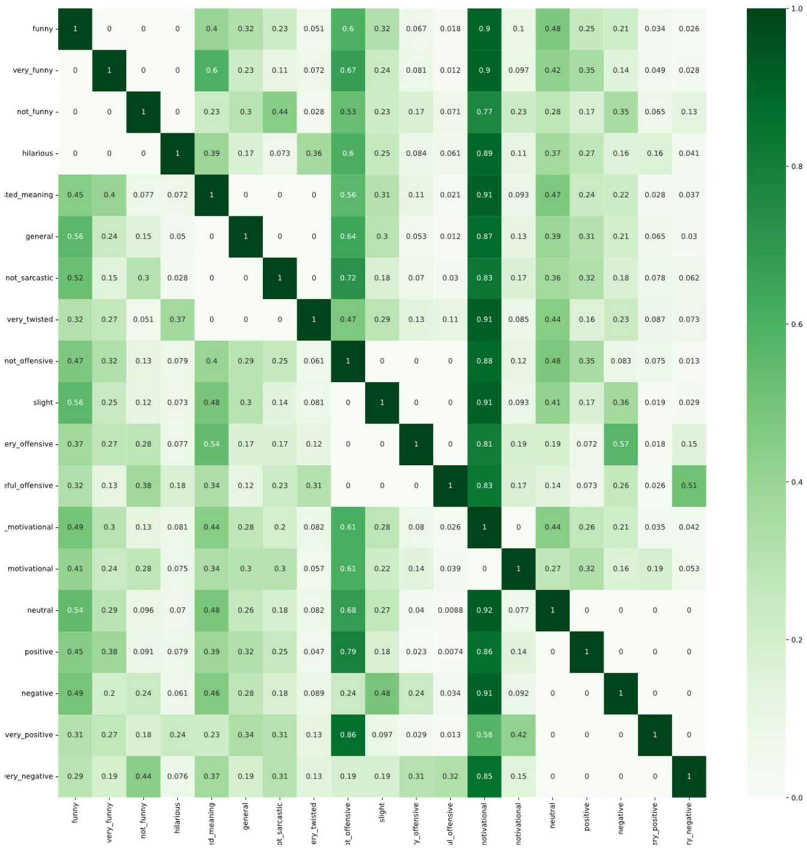


Figure 5: Overall distribution of the dataset showing overlap between all 20 labels

The annotators were asked to provide their thoughts on the emotion of the meme for Tasks B and C. The perception of a meme and societal elements may vary from person to person. Each meme is annotated by three separate annotators. The decision of the final annotations is made using a majority voting system.

6. Methodology

A series of feature extraction methods and a multimodal network, incorporating both text and image modalities, were employed to discern the sentiment within the meme collection. Initially, both text and image data underwent preprocessing before feature extraction. Image features were derived using a pre-trained CNN network, leveraging a comprehensive image dataset. In parallel, text features were extracted by converting token sequences into numerical representations, serving as inputs to the multimodal network.

For image feature extraction, a set of preprocessing techniques was employed:

- Resizing: This essential step ensures uniformity in image dimensions, a prerequisite for effective processing with the CNN model.
- Normalization: Pixel values of the images were normalized to fall within the range of [0, 1], enhancing the performance of the CNN model by standardizing the input data.

- Padding: This step maintains consistent dimensions across all images.

Following these preprocessing steps, the images were fed into the CNN model for feature extraction. The CNN model had been pre-trained on a comprehensive dataset containing images and corresponding labels. This prior training enabled the CNN to autonomously identify and extract pertinent features beneficial for various tasks. The extracted image features underwent a subsequent normalization process to ensure uniformity in scale and distribution. Ultimately, the processed image features were saved to a CSV file, serving as input to the multimodal network.

Text feature extraction comprised several sequential steps designed to convert textual data into a numerical format for the multimodal network. The initial phase focused on preprocessing the textual data to eliminate noise and rectify inconsistencies. This encompassed:

- Lowercasing: Standardizing the representation by converting all words to lowercase.
- Special Character Removal: Stripping punctuation marks and other special characters from the text to refine the data.
- Stemming or Lemmatization: Normalizing words to their root forms, reducing vocabulary size, and capturing core semantics.
- Stop Word Removal: Eliminating commonly occurring, non-discriminative words, known as stop words, to reduce noise and enhance focus on content-rich words.

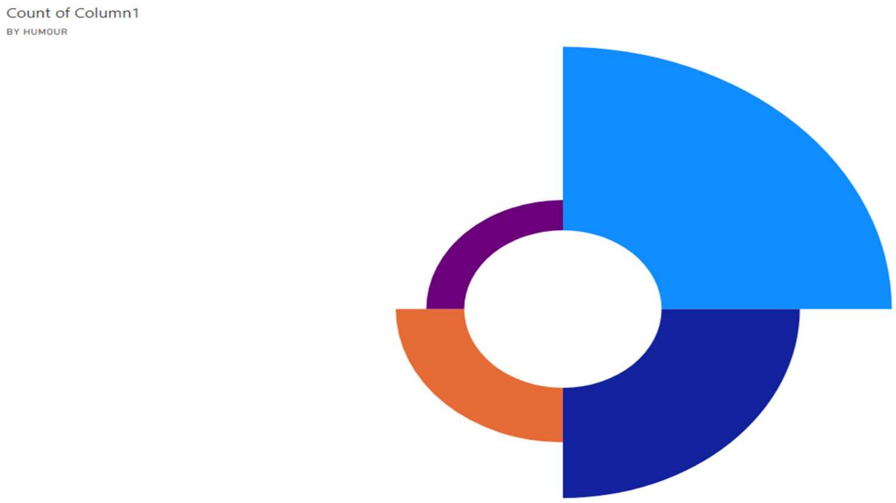
The processed text was then tokenized, segmenting it into a sequence of tokens, each representing a word, punctuation mark, or other meaningful symbol. Next, a text embedding model was applied to create a numerical representation for each token. These models were pre-trained on extensive datasets containing text and code, enabling them to learn representations capturing semantic information. Finally, the processed text features were saved to a CSV file, serving as input to the multimodal network.

The training data, following preprocessing and extraction, was labeled for the multimodal network to establish a consistent standard for the specific prediction task. These labels provided crucial context for the model to effectively learn and predict the correct class. The multimodal architecture was meticulously designed to accommodate both text and image inputs. It featured two dedicated input layers for processing textual and image-based information. The text input underwent transformation through a fully connected layer, yielding a single output value. Simultaneously, the image input underwent a parallel operation through a fully connected layer, producing a corresponding single output value. These outputs were concatenated, amalgamating the extracted features from both modalities. This concatenated output then passed through another fully connected layer, culminating in a singular output prediction.

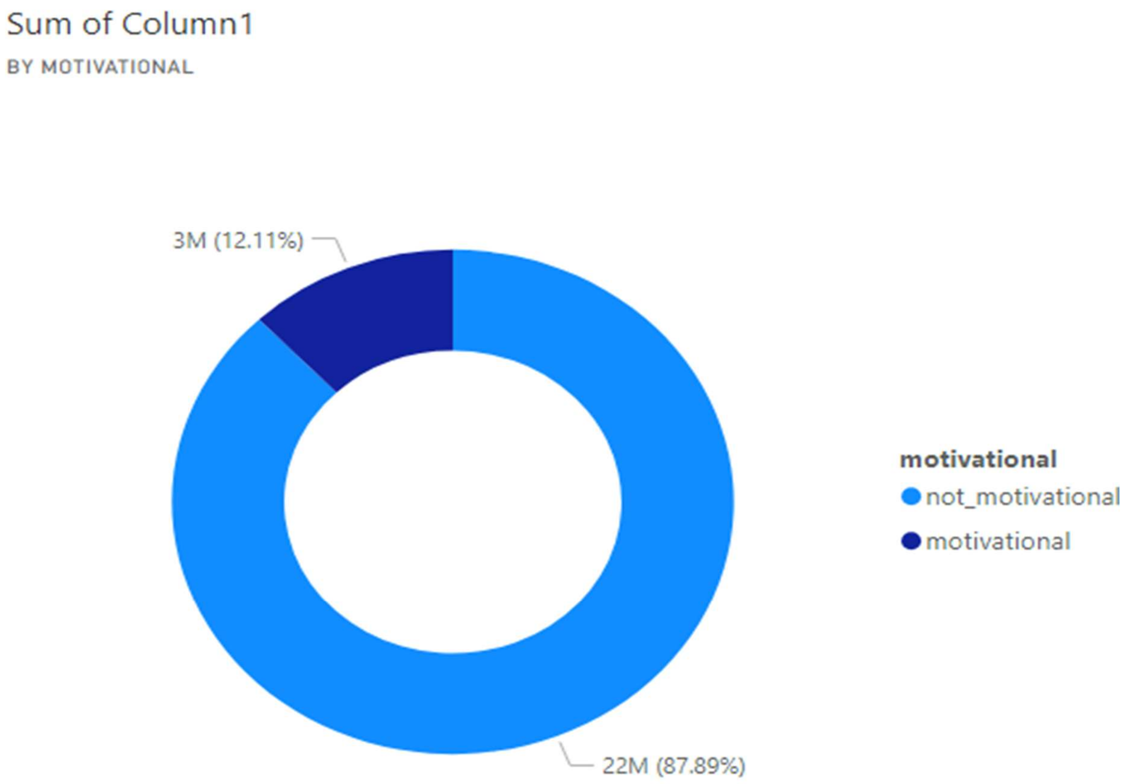
TASK	SUB-TASK	CONTENT
TASK A	TASK A	Classify a meme as positive, negative, or neutral.
TASK B	TASK B1	Classify a meme as humorous or not.
	TASK B2	Classify a meme as sarcastic or not.
	TASK B3	Classify a meme as offensive or not.
	TASK B4	Classify a meme as motivational or not.
TASK C	TASK C1	Quantify a meme as not funny, funny, very funny, or hilarious.
	TASK C2	Quantify a meme as not sarcastic, general, twisted meaning, or very twisted.
	TASK C3	Quantify a meme as not offensive, slight, very offensive, or hateful offensive
	TASK C4	Quantify a meme as motivational or not.

Additionally, the model incorporated a custom loss function sparse categorical entropy prior for training. This function, an adapted variant of the standard categorical cross-entropy loss, integrated a prior distribution over the output classes to enhance the model's accuracy in predicting less frequently occurring classes. For optimization, the model was compiled using the Adam optimizer, a widely favored algorithm for training deep-learning models. The choice of metric for evaluation was the sparse categorical accuracy, quantifying the model's precision in assigning the correct output class. This versatile model could be applied to a diverse range of tasks, including text and image classification, as well as question answering. The labels played a pivotal role in providing better context for the model to learn and predict the class effectively.

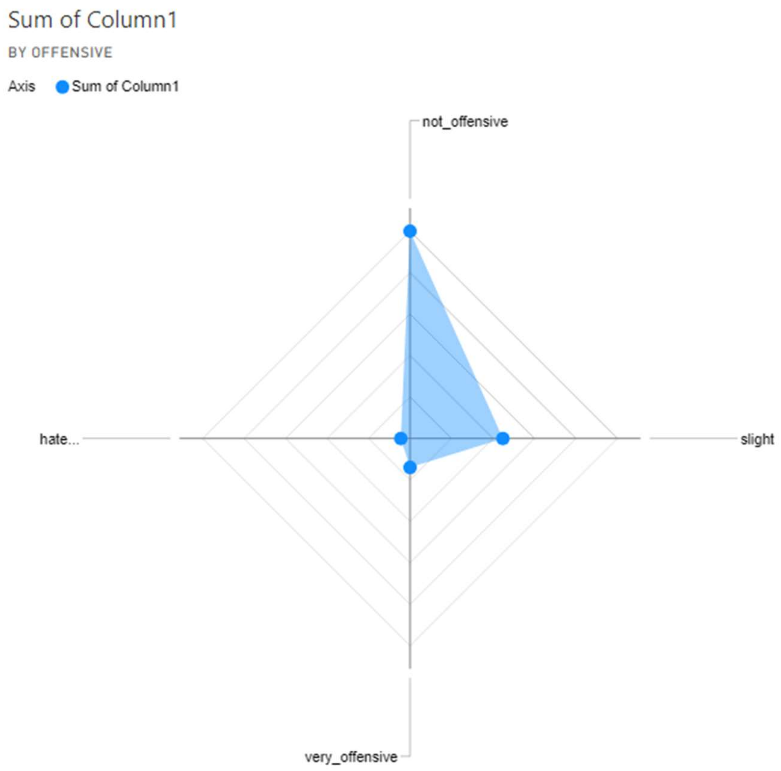
7. VISUALIZATION



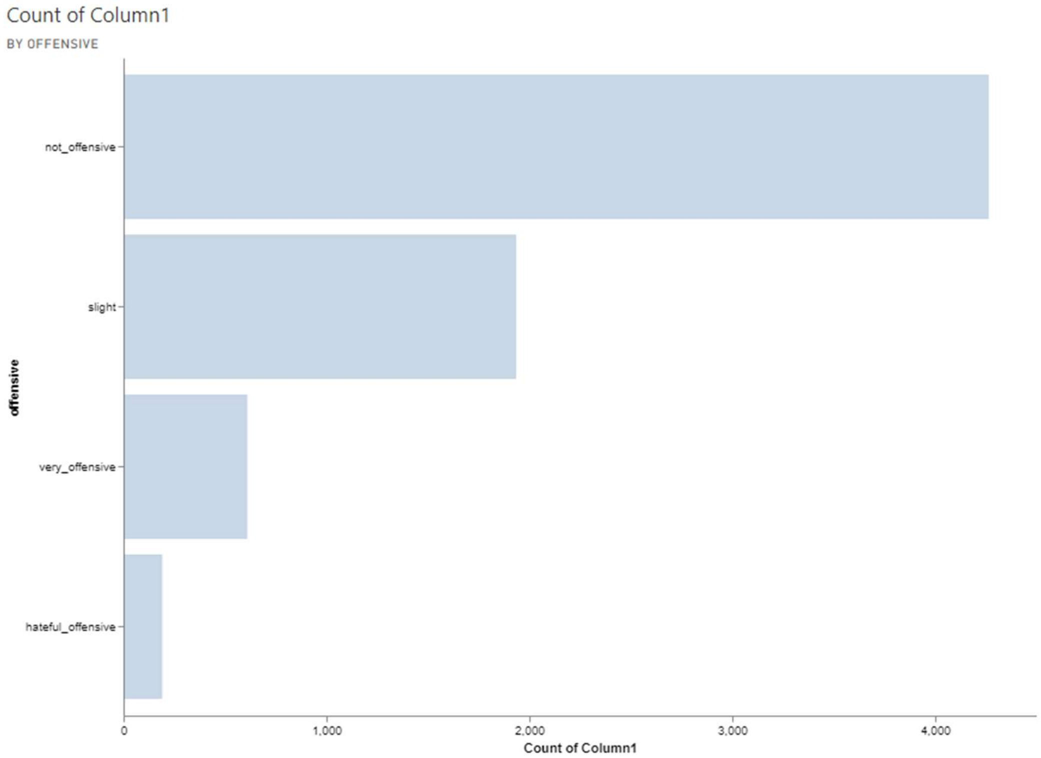
Open in Power BI
report
Data as of 4/23/24, 12:35 AM



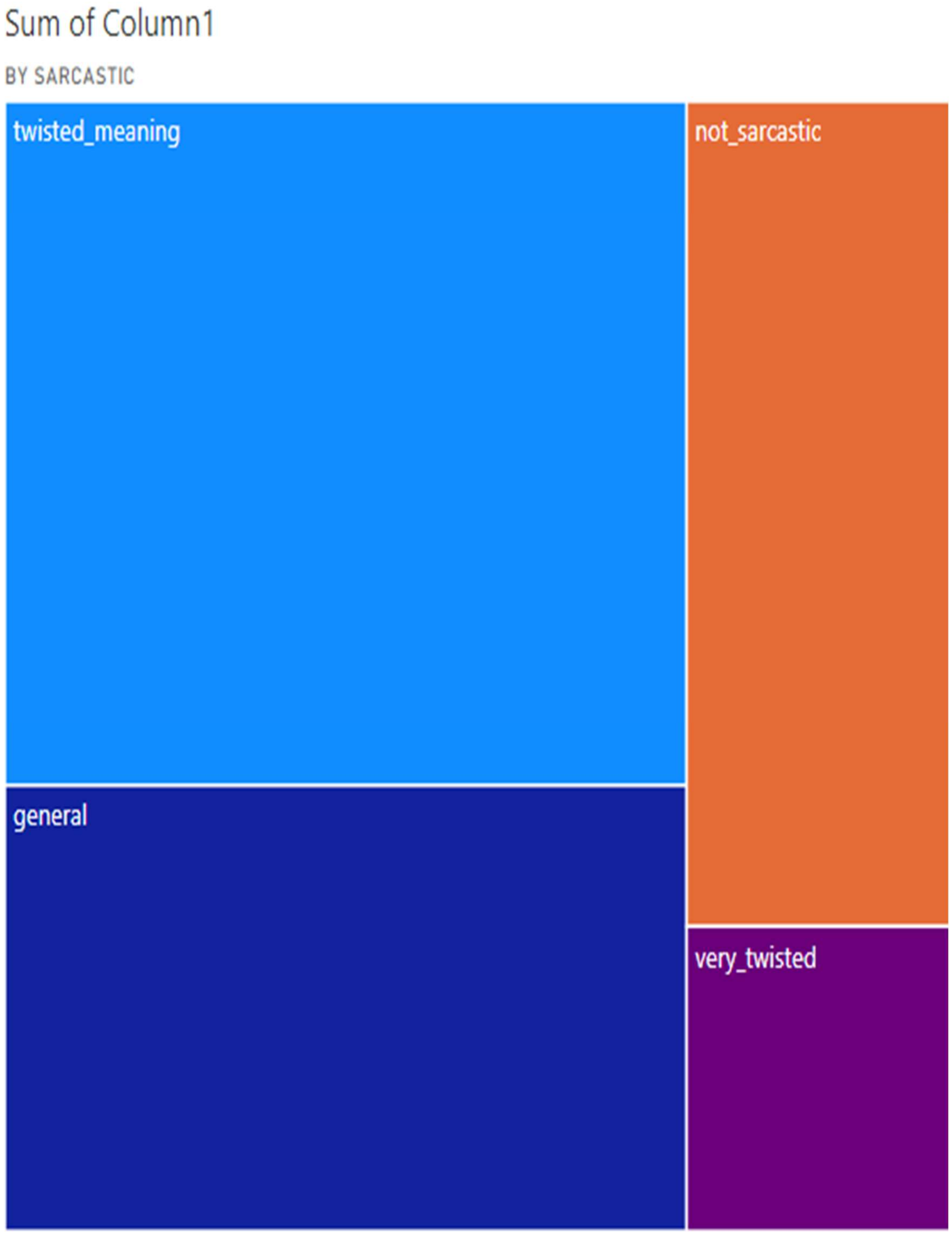
Open in Power BI
report
Data as of 4/23/24, 12:35 AM



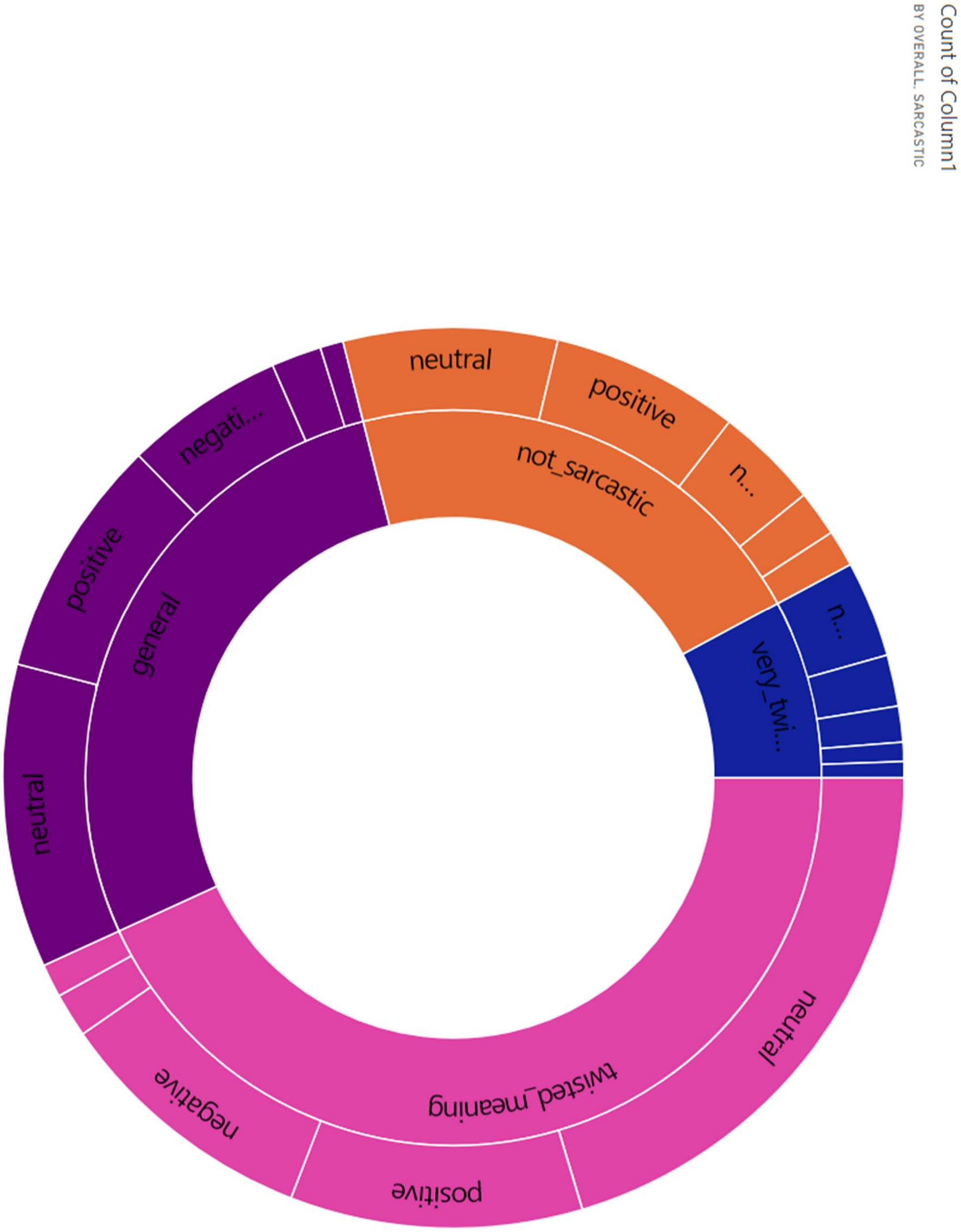
Open in Power BI
report
Data as of 4/23/24, 12:35 AM



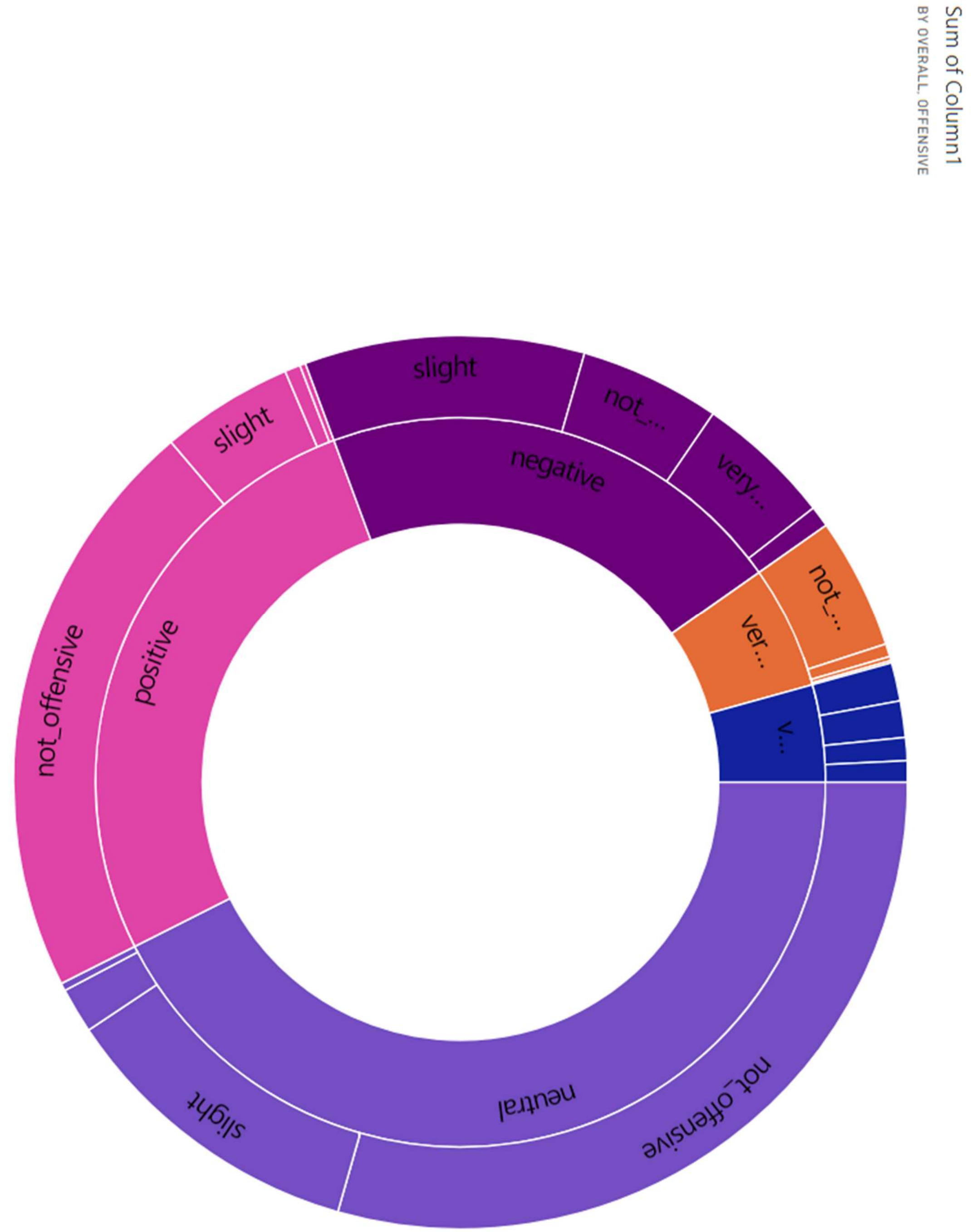
Open in Power BI
report
Data as of 4/23/24, 12:35 AM



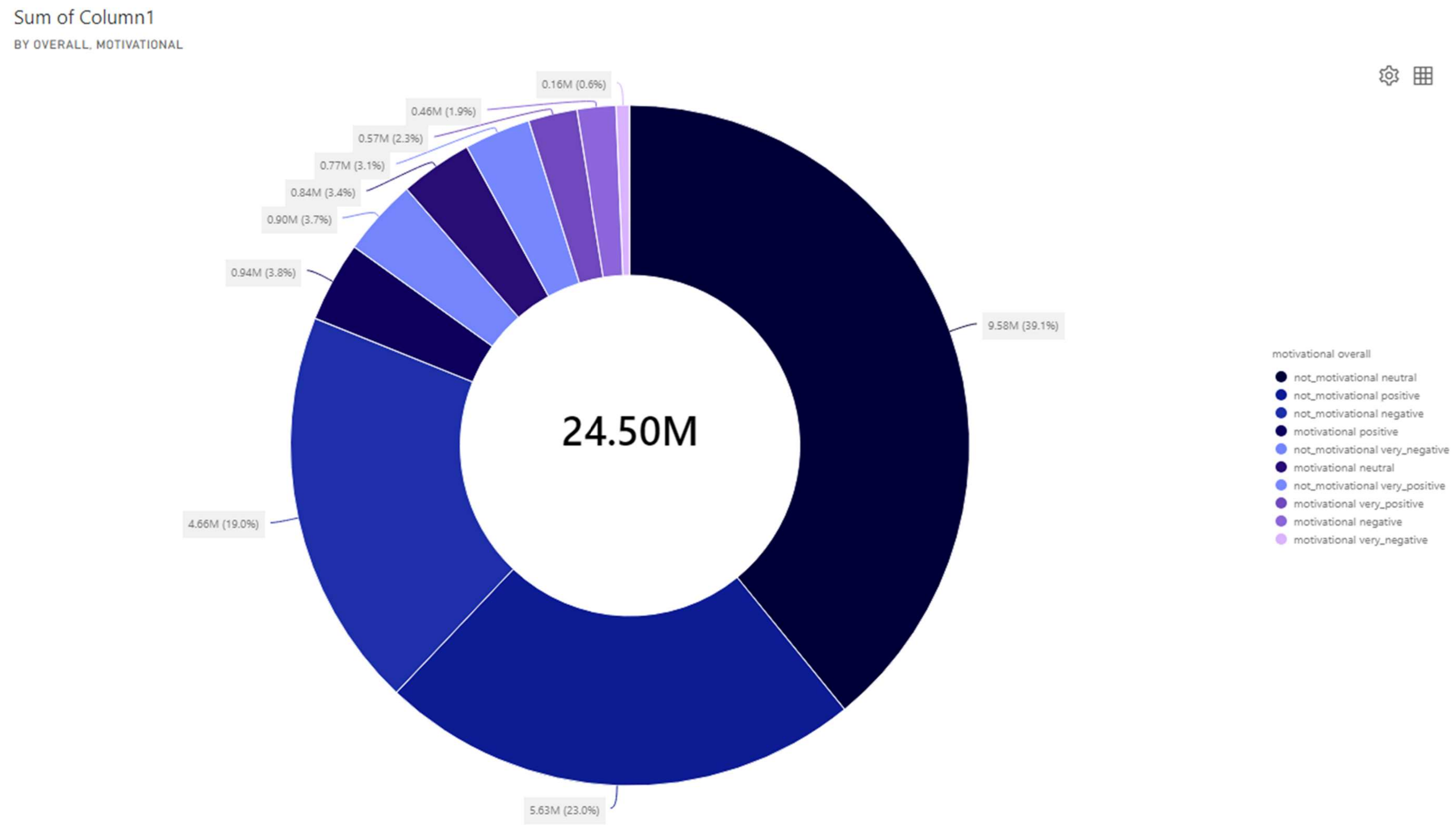
Open in Power BI
report
Data as of 4/23/24, 12:35 AM



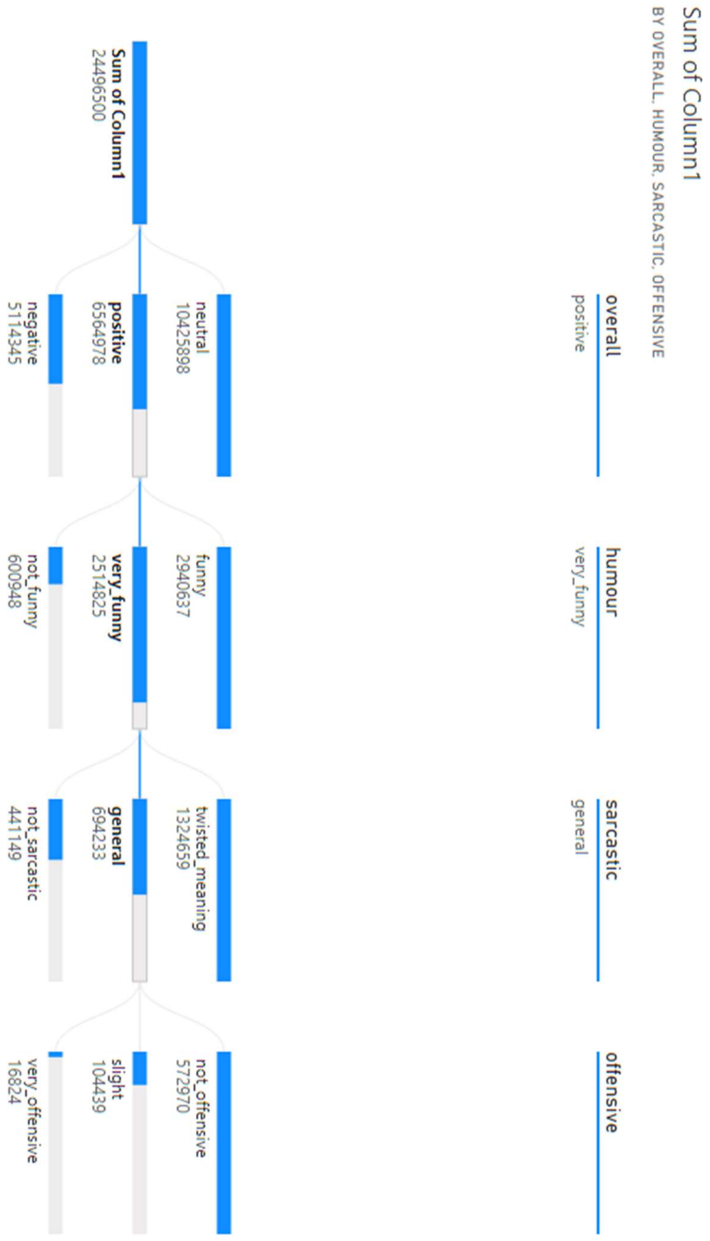
Open in Power BI
report
Data as of 4/23/24, 12:35 AM



[Open in Power BI](#)
report
Data as of 4/23/24, 12:35 AM



Open in Power BI
report
Data as of 4/23/24, 12:35 AM



[Open in Power BI]

report

Data as of 4/23/24, 12:35 AM

8. Experiments and Results

8.1 Dataset

The dataset used for our experiments was released by the organizers of the Memotion 3 task. Each entry in the dataset contains the following fields: image, text, and label. The field of label varies for the different tasks. The dataset contains a total of 10,000 samples, including 7,000 for training, 1,500 for validation, and 1,500 for test. For the experimentation, we rely on the training, validation, and test data as split by the organizers. Table shows the distribution of labels of different tasks across training, validation, and test sets.

	TRAIN	**VALIDATION**	**TEST**	**SUM**
POSITIVE	2275(33%)	341(23%)	586(39%)	3202(32%)
NEUTRAL	2970(45%)	579(39%)	533(36%)	4082(40%)
NEGATIVE	1755(25%)	580(39%)	381(25%)	2716(27%)
SUM	7000	1500	1500	10000

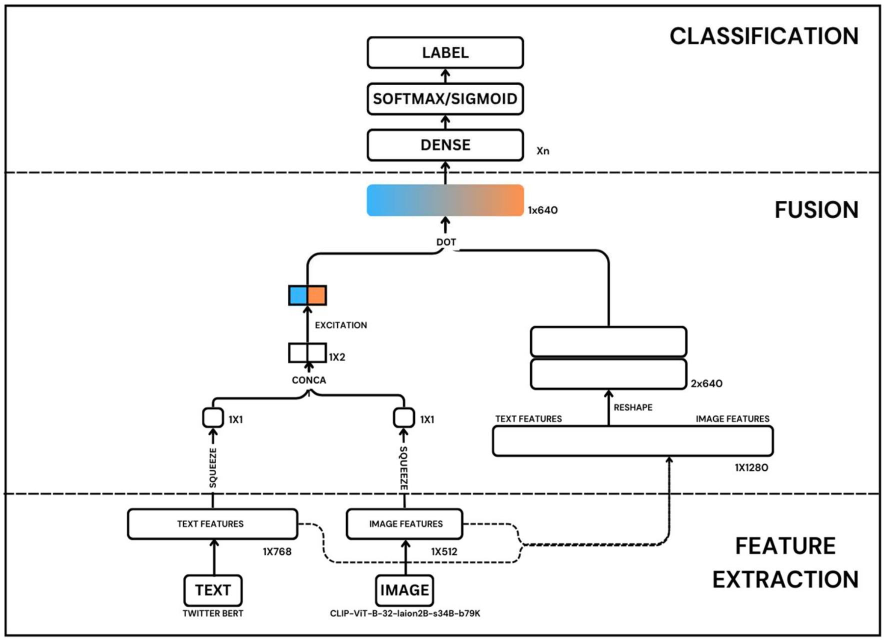
8.2 Implementation

Model named twitter-roberta-base was used for text extraction. RoBERTa-base model is trained on ~124M tweets. This model was used as it is better at classifying text based data and gave good results than the rest of Bert Models.

The model used for image extraction was CLIP-ViT-B-32-laion2B-s34B-b79K. Here, contrastive Language-Image Pre-training, works on text encoder and vision based model to create relationship between them and classify. It is trained on 2 Billion sample English subset of LAION-5. This model learns the relationship between both text and images, providing better classification on memes dataset. It gave 66.6 zero-shot top-1 accuracy.

Multimodal neural network(original), multimodal neural network(test - 1) and multimodal neural network(test - 2) are used for both image and text extraction. These models are able to process and learn from multiple types of data, such as text, images, and audio. In this case, we have trained with images and texts. Multimodal neural network(original) was used as it is a customized model based on the CLiP with defined categorical and sparse loss functions. The F1 score is 47.06. Multimodal neural network (test - 1) was used as it has altering model layers and establishing a new parameter function for the above model, more dense layers were added. (best.h5). The F1 score was 33.73. Multimodal neural network (test - 2) was used because there was changing dense layers in the model just after concatenation(best2.h5). Here the F1 score was 29.78

8.3 Model



8.4 Evaluation Metrics

We employ the weighted-F1 evaluation metric, the official measure proposed by the competition organizers. The weighted-F1 score is computed by averaging per-class F1 scores, with consideration for the support of each class. The formula for F1 score is given by:

F1 = 2 * Precision * Recall
Precision + Recall

Where 'Precision' and 'Recall' represent precision and recall, respectively, and 'C' denotes the total number of classes.

The weighted-F1 score is defined as:

Weighted-F1 = Σ(Fl) / C
l=1

Where 'C' is the total number of classes, and 'Fl' represents the F1 score for each class (l).

For task A, we directly use the weighted-F1 for evaluation. However, for tasks B and C, we calculate the weighted-F1 score for each sub-task and then compute the average of these scores to obtain an average weighted-F1 score.

9. Comparative Study

Text Feature Extraction po	Bert-base-multilingual-cased	A multilingual BERT model trained on 104 languages. Achieves state-of-the-art results on a variety of natural language processing tasks, including text classification, question answering, and machine translation.
	RoBerTa-base	A robustly optimized BERT model with improved performance on downstream tasks. Achieves state-of-the-art results on a variety of natural language processing tasks, including text classification, question answering, and natural language inference.
	twitter-roberta	A RoBERTa model trained on a massive dataset of tweets. Achieves state-of-the-art results on a variety of natural language processing tasks related to Twitter, such as sentiment analysis, topic modeling, and spam detection. Performed better currently on the given dataset.
Image Feature Extraction	CLIP-ViT-B-32-laion2B-s34B-b79K	A Vision Transformer model with 32 billion parameters trained on a massive dataset of text and images. Achieves state-of-the-art results on a variety of image classification and retrieval tasks. A RoBERTa model trained on a massive dataset of tweets. Achieves state-of-the-art results on a variety of natural language processing tasks related to Twitter, such as sentiment analysis, topic modeling, and spam detection. Performed better currently on the given dataset.
	openai/clip-vit-base-patch32	A Vision Transformer model with 1 billion parameters trained on a subset of the LaMDA dataset. Achieves state-of-the-art results on a variety of image classification and retrieval tasks.

In our experiments, we found that the combination of Twitter-RoBERTa and

CLIP-ViT-B-32-laion2B-s34B-b79K achieved the highest accuracy on the meme classification task.

Twitter-RoBERTa is a transformer-based language model that was trained on a massive dataset of tweets. It is well-suited for meme classification tasks because it can capture the nuances of human language, including slang and humor.

CLIP-ViT-B-32-laion2B-s34B-b79K is a vision transformer model that was trained on a massive dataset of text and images. It is well-suited for meme classification tasks because it can learn to represent the visual and textual features of memes in a unified way.

This is likely because the two models complement each other well. Twitter-RoBERTa can capture the nuances of the textual content of memes, while CLIP-ViT-B-32-laion2B-s34B-b79K can capture the visual content of memes.

Furthermore, we found that the combined model was able to generalize well to different categories of memes. This is important because it suggests that the model can be used to classify memes in real-world applications.

TASK	CATEGORY	F1-Scores(Weighted)	F1-Scores(Weighted)
		(Previous Best Model)	
TASK - A	Sentiment	33.28%	34.29%
TASK - B	Humor	84.55%	85.29%
	Sarcasm	74.82%	64.10%
	Offensive	48.84%	52.55%
	Motivation	90.78%	86.11%
	Average	74.74%	72.01%
TASK – C	Humor	43.03%	52.37%
	Sarcasm	32.89%	52.10%
	Offensive	42.40%	37.26%
	Motivation	90.78%	86.11%
	Average	52.27%	56.96%

10.Conclusion

In conclusion, our study introduces ModuFusion, an innovative multi-modal fusion technique designed to harmonize text and image features for emotion classification in internet memes. Our ModuFusion approach achieved the highest ranking in task A and secured the second position in task C during the Memotion 3 challenge.

ModuFusion employs straightforward yet powerful operations, encompassing compress and intensify techniques, to combine the features extracted from memes. These operations involve fully connected layers with suitable activations, reshaping, and matrix multiplication. In the spirit of the Squeeze-and Excitation Block, ModuFusion offers adaptability and can be applied to merge diverse sets of features derived from different models. Furthermore, ModuFusion can seamlessly integrate more than two types of features, provided that the dimensions are appropriately adjusted.

However, while our work has made significant strides, it is not without its limitations and offers avenues for future research. Specifically, our model learned weight vectors for each modality, but these weights were not directly applied to their respective modalities due to the mixing of text and image features during reshaping. We intend to explore the unification of feature dimensions from different modalities before executing the ModuFusion process. Additionally, the field of internet meme emotion analysis is still in its early stages. Despite our model's top performance in task A, there is room for improvement, as it only marginally outperformed the baseline model. This underscores the need for further research, ideally in conjunction with related tasks such as sentiment detection and identification of hateful content in memes.

11.References

1. Web Mining and Minimization Framework Design on Sentimental Analysis for Social Tweets Using Machine Learning by TS Raghavendra and TG Mohan
2. Text mining with sentiment analysis on seafarers’ medical documents by Nalini Chintalapudi, Gopi Battineni, Marzio Di Canio, Getu Gamo Sagaro, Francesco Amenta
3. A novel approach to stance detection in social media tweets by fusing ranked lists and sentiments by Abdulrahman I. Al-Ghadir, Aqil M. Azmi , Amir Hussain
4. Meme Sentiment Analysis Enhanced with Multimodal Spatial Encoding and Face Embedding by Muzhaffar Hazman, Susan McKeever and Josephine Griffith
5. Sentiment analysis of extremism in social media from textual information by Muhammad Asif, Atiab Ishtiaq, Haseeb Ahmad , Hanan Aljuaid , Jalal Shah
6. Sentiment analysis on the impact of coronavirus in social life using the BERT model by Mrityunjay Singh, Amit Kumar Jakhar and Shivam Pandey
7. Sentimental analysis of COVID-19 tweets using deep learning model by Nalini Chintalapudi, Gopi Battineni,, Francesco Amenta
8. Sentimental analysis of Indian regional languages on social media by Kakuthota Rakshitha, Ramalingam H M, M Pavithra, Advi H D, Maithri Hegde
9. Overview of Memotion 3: Sentiment and Emotion Analysis of Codemixed Hinglish Memes by Shreyash Mishra, S Suryavardan, Megha Chakraborty, Parth Patwa, Anku Rani, Aman Chadha,Aishwarya Reganti, Amitava Das, Amit Sheth, Manoj Chinnakotla, Asif Ekbal and Srijan Kuma.