Deep Learning and Machine Learning Approaches Including Ensemble Techniques for Multimodal Sentiment Analysis: Memotion Classification

Aldi Halili[1], Chunxue Liu[2], Valeriya Herrlein[3]

[1]Aldi.Halili@campus.lmu.de,[2]Chunxue.Liu@campus.lmu.de,[3]V.Shevchenko@campus.lmu.de

Center for Information and Language Processing, LMU Munich

July 29, 2024

Abstract

This study focuses on sentiment analysis of memes, employing ResNet-50 and BERT to extract visual and textual features, respectively. We applied both deep learning and machine learning techniques, specifically three methods: a Fusion Model based on Self-Attention (MMFA) [1], a Fusion Model without a Self-Attention Block (MMF), and Machine Learning with scikit-learn. For the deep learning models, we utilized five different classifiers: Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Multi-Layer Perceptron (MLP), and Feedforward Neural Network (FFNN). We constructed an ensemble model using majority voting based on these classifiers and evaluated the model performance using the Macro F1 score. Experimental results show that our individual models and the ensemble model outperform the baseline and other participating models in SEMEVAL-2020 TASK 8: MEMOTION ANALYSIS-Task A. [2]

1 Keywords

Ensemble model; Multimodal sentiment analysis; Multimodal fusion models; Self-Attention mechanism; Textual nuances capture; Visual features capture

2 Introduction

In the digital age, the rapid proliferation of social media and online platforms has led to an unprecedented richness of multimodal content, integrating text, visual, and audio data. This fusion of modalities presents both opportunities and challenges for effectively interpreting the human emotions embedded within these diverse forms of communication. Multimodal Sentiment Analysis (MSA) [3] aims to address these challenges by combining multiple data types to provide a more accurate and comprehensive portrayal of sentiments, surpassing the limitations inherent in text-only analysis.

Memes, which combine text and visual elements, have become a significant component of social media expression. While memes enhance the expressive capabilities of users on these platforms, their often-ambiguous meanings can result in misunderstandings and communication barriers. Thus, analysing the sentiment of memes presents a unique challenge that necessitates advanced approaches in multimodal sentiment analysis.

For example, Figure 1 is a classic ”Distracted Boyfriend” meme [4], and by analysing it separately from the perspectives of image and text, the importance of multimodal content in emotional expression can be demonstrated.

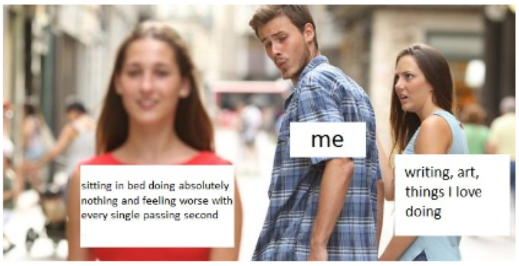

me
sitting in bed doing absolutely nothing and feeling worse with every single passing second
writing, art, things I love doing

Figure 1: Meme: 'Distracted Boyfriend' (image_6260 in 'memotion_dataset-7k').

Image Analysis From the perspective of image analysis alone, the expressions and postures of the characters in the picture already convey a scenario of attention shift. The man in the middle turns his head to look at the girl on the left, showing obvious interest, while the girl on the right looks at the man with a dissatisfied and jealous expression. This scenario can be understood as a typical attention shift even without any text. From an image-only perspective, this image may be marked as negative.

Text Analysis The text labels the three characters and objects in the picture: the girl on the left is labelled "sitting in bed doing absolutely nothing and feeling worse with every single passing second," the man in the middle is labelled "me," and the girl on the right is labelled "writing, art, things I love doing." When analysing each piece of text individually, "sitting in bed doing absolutely nothing and feeling worse with every single passing second" can be marked as negative due to the clear negative expression "feeling worse," while the other two texts, "me" and "writing, art, things I love doing," may be marked as neutral or positive. Combining these three texts makes it difficult to derive a clear emotional label, and this may conflict with the label derived from image analysis alone.

Combined Image and Text Analysis

Combining the image and text, the text provides specific context for the picture, making it easier for the audience to relate this scenario to their own experiences. The audience can immediately understand the specific message the meme is trying to convey: lying in bed doing nothing leads to a shift in attention and neglects activities like writing and art, which are things one loves doing. This emotional conflict and guilt are vividly expressed through the combination of image and text, allowing the audience to empathise.

The combined analysis of image and text, through the information provided by the image's "attention shift scenario," clarifies the overall "negative" emotional expression. The scene depicted in the image also enhances the emotional expression. The image provides a direct visual impact, showing the scenario of attention shift through the characters' expressions and postures, while the text reveals the underlying emotions and struggles, expressing regret and negative feelings. When image and text are combined, it not only enhances the effectiveness of information transmission but also deepens the audience's understanding and resonance with the emotions. The audience can not only see the attention shift scenario but also deeply feel the emotional conflict within it. Therefore, through multimodal analysis of image and text, the emotional expression is more comprehensively and powerfully conveyed.

In this study, we focus on sentiment analysis of memes for SEMEVAL-2020 TASK 8: MEMOTION ANALYSIS-Task A. We have implemented three main approaches: two deep learning (DL) approaches and one machine learning (ML) approach. We employ BERT and ResNet models to extract text and image features, respectively. The features from both modalities are concatenated using a fusion model, with or without self-attention (MMF, MMFA). These combined and refined features are then used to train models equipped with Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Multi-Layer Perceptron (MLP), and Feedforward Neural Network with Softmax (FFNN Softmax) classifiers. For our ML approach, we apply the following classifiers using the same method as above: Decision Tree Classifier, Logistic Regression, Multilayer Perceptron (MLP), Adaptive Boosting, and Support Vector Machine (SVM). Ultimately, we applied an ensemble approach to both the deep learning models and one ML approach. During training for our ensemble approach, we apply 5-fold cross-validation to allocate data for each model and select the best prediction from the five predictions obtained as input data for the ensemble model. For the ensemble model training, we use logits and the Majority Voting mechanism. We evaluate the model performance using the Macro F1 score.

Experimental results demonstrate that our individual models and ensemble model consistently outperform the baseline and other competing models in SEMEVAL-2020 TASK 8: MEMOTION ANALYSIS-Task A, and show strong performance in terms of macro F1 score and accuracy on the 7K dataset.

3 Related Work

This section describes related work on sentiment analysis of text, images, and multimodal content, including studies on fusion methods based on self-attention mechanisms.

In **text analysis**, Mozetič et al. (2016) adopted a hybrid approach combining dictionary-based and machine learning methods for text sentiment analysis. Their method predicts sentiment by identifying a set of emotion-laden words in the text using a dictionary [5]. Ghiassi and Lee (2018) developed a set of domain-transferable Twitter dictionaries to extract information from tweets for sentiment analysis tasks[6]. Li et al. (2019) proposed a sentiment feature-enhanced deep neural network (SDNN) that detects text sentiment by combining sentiment linguistic knowledge and attention mechanisms [7].

In **image analysis**, Kumar and Jaiswal (2017) proposed a model using Convolutional Neural Networks (CNN) to detect image sentiment. They trained their model on the Flickr image dataset and tested it using the Twitter image dataset [8]. Akhtar et al. (2020) introduced a stacked ensemble model for predicting the intensity of sentiments and emotions. They utilised a multilayer perceptron network to combine the outputs of feature-based models and deep learning models [9].

In **multimodal analysis**, The team led by S.

Singh (2015) employed deep convolutional neural networks (CNNs) combined with domain-specific fine-tuning methods, adjusting based on image type or description for image sentiment analysis[10]. The team led by N. Mittal (2018) tested the effectiveness of several deep learning methods for image sentiment analysis, including convolutional neural networks (CNNs), region-based convolutional neural networks (RCNNs), and three-dimensional convolutional neural networks (3D CNNs) [11]. Poria et al. (2018) explored different deep learning architectures for multimodal sentiment classification, utilising deep convolutional neural networks (CNNs) to extract features from both visual and textual modalities [12]. Jiang et al. (2020) proposed a fusion extraction network model for multimodal sentiment analysis. Their model employs an interactive information fusion mechanism to learn two types of representations: visual-specific textual representations and text-specific visual representations [13]. The team led by P. Behera (2020) proposed a multimodal framework combining textual and visual features of memes to enhance sentiment classification using attention mechanisms. Addressing the imbalance in the SemEval-2020 Task 8 dataset, they expanded it with new annotations and applied sampling strategies for balance. For text processing, they used pre-trained GloVe embeddings to convert text to vectors and extracted features with a convolutional neural network (CNN), employing attention mechanisms to capture relevant features. For visual processing, they used a pre-trained VGG-19 model to extract image features and applied attention mechanisms to identify key regions. Finally, the extracted text and image features were fused via a fully connected layer to create a combined feature vector for classification [14]. Nayan Varma Alluri's team (2021) proposed three models: the IMGTXT model, which combines Vision Transformers (ViT) and RoBERTa models, using a fusion method to integrate their embeddings; the IMGSEN model, which combines ViT and SBERT-RoBERTa embeddings, also using a fusion method to integrate their embeddings; and the CAPSEN model, which converts images into textual descriptions and processes the text using the SBERT model, then integrates the image and text embeddings using a fusion method[15]. A.-M. Bucur's team (2022) proposed two classification methods: a text classification method, which fine-tunes the BERT model and focuses on using only textual features for meme sentiment classification; and a Multi-Modal-Multi-Task transformer network, which combines image and text features by processing images with EfficientNetV4 and CLIP, and processing text features with a Sentence Transformer, culminating in a final classification using a multi-modal transformer network. They also adopted the CORAL (Cumulative Ordinal Regression) method to perform ordinal regression for emotion intensity, improving the prediction of emotion strength. Furthermore, they utilised advanced pre-trained models, such as EfficientNetV4 and CLIP, to enhance the performance of the multimodal model [16].

In **Fusion Method Based on Self-Attention Mechanism**, Hu Zhu et al. (2020) proposed a multimodal fusion method that combines self-attention mechanisms with low-rank tensor representations. Their approach addresses the high computational complexity of traditional tensor-based methods by using low-rank weight tensors with an attention mechanism, thereby improving efficiency. Evaluations on datasets such as CMU-MOSI, IEMOCAP, and POM show that this method excels in capturing both global and local connections, outperforming other state-of-the-art models [17].

4 Dataset

To rigorously assess our multimodal sentiment analysis framework, we utilised the 'memotion-dataset-7k' obtained from Kaggle [18]. This dataset is from the semeval challenge called "Memotion Analysis" in 2020 [19], it comprises a substantial collection of memes. For our project, we specifically employed its image folder containing internet memes and the labeled dataset in a CSV file.

4.1 Memotion Dataset-7K

The memotion-dataset-7k facilitates several classification tasks within the Memotion Analysis challenge. Although our project exclusively focuses on Task A, here is a brief overview of each task:

Task A - Sentiment Analysis: This task requires participants to assess an internet meme and determine whether its underlying sentiment is positive, negative, or neutral, providing a fundamental analysis of its emotional tone.

Task B - Humor Detection: In this task, the system must discern the specific humor type portrayed in a meme. The categories defined are sarcastic, humorous, and offensive. Memes that do not align with these predefined categories are labeled as 'other'. Notably, a single meme might be categorized under multiple humor types if it exhibits a blend of these elements.

Task C - Scales of Semantic Classes: This task requires quantifying the extent to which a particular sentiment or humor is expressed in a meme. Appropriate annotated data will be provided to support detailed assessments.

Image Dataset - This dataset contains image data for 6,992 internet memes. Each image represents a meme that typically conveys humor, irony, or emotional content through a combination of image and text. The dimensions and resolutions of the images vary based on their content, and they are commonly found in formats such as JPEG or PNG. See Fig. 2 below.

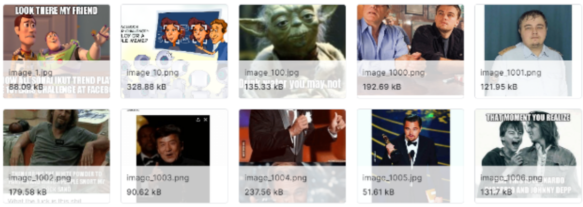
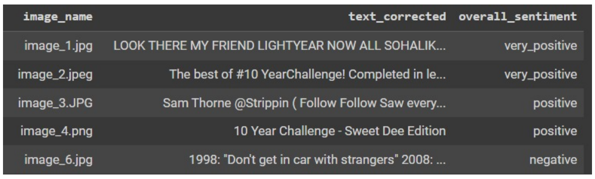


Figure 2: Examples of Image Section.

Original Text Dataset - The original labeled dataset, provided as a CSV file, contains label information for each meme. Each row in the dataset corresponds to a meme and includes the following fields:
- Image_Name
- text_ocr
- text_corrected
- Overall_Sentiment
- humour
- sarcasm
- offensive
- motivational

These labels are derived from manual annotations, assigning appropriate classifications to each meme, which aids in training models for a broad range of analytical tasks. The detailed annotations not only enable sentiment analysis by classifying a meme's emotional tone as positive, negative, or neutral, but also support other tasks such as Humor Classification (Task B) and Scales of Semantic Classes (Task C).

Processed Dataset - For Task A, we refined the original dataset to focus specifically on sentiment classification. From the entire CSV file, we retained only the following columns essential for our analysis:
- Image_Name: The filename or identifier of the meme image, used to match with the corresponding image file.
- text_corrected: The corrected textual content extracted from the memes.
- Overall_Sentiment: This column represents the overall sentiment classification of the meme, which can be categorized as very positive, positive, very negative, negative, or neutral. The dataset includes 1,033 instances of very positive sentiment, 3,127 instances of positive sentiment, 151 instances of very negative sentiment, 480 instances of negative sentiment, and 2,201 instances of neutral sentiment. Each row in the dataset corresponds to one of the 6,992 memes, ensuring a comprehensive representation of sentiment across the collection. (Refer to Fig. 3 for details). Figure 4 shows the labels and distributions of the 5-class labels in Task A - Sentiment Classification.



image_name	text_corrected	overall_sentiment
image_1.jpg	LOOK THERE MY FRIEND LIGHTYEAR NOW ALL SOHALIK...	very_positive
image_2.jpeg	The best of #10 YearChallenge! Completed in le...	very_positive
image_3.JPG	Sam Thorne @Strippin (Follow Follow Saw every...	positive
image_4.png	10 Year Challenge - Sweet Dee Edition	positive
image_6.jpg	1998: "Don't get in car with strangers" 2008: ...	negative

Figure 3: Examples of labels.csv for Task A - Sentiment Classification.

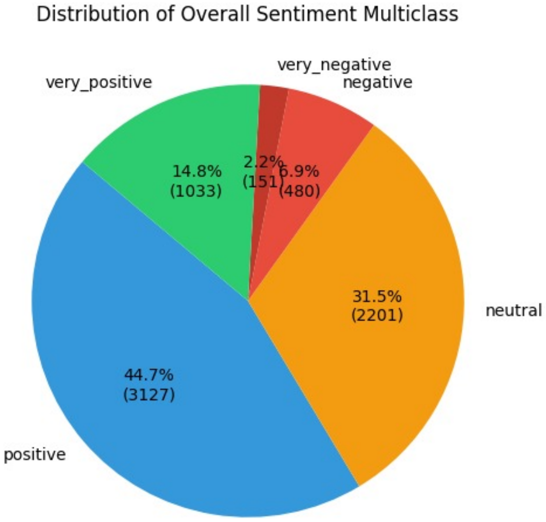


Figure 4: Distributions of the 5-class labels in Task A - Sentiment.

Filtered Dataset for Binary Classification - In our task, unlike the original Task A's categorization (very positive/positive/very negative/negative/neutral), we adopted a binary classification approach (positive/negative). We removed the neutral labels from the original dataset and merged the very positive and positive labels into one class, labeled as "positive" (1). Similarly, we merged the very negative and negative labels into one class, labeled as "negative" (0). Based on this principle, we filtered 4,791 memes from the original dataset, with 4,160 labelled as positive and 631 labelled as negative. Figure 5 shows the filtered version of the class distributions of the dataset specifically for the binary Sentiment Classification approach.

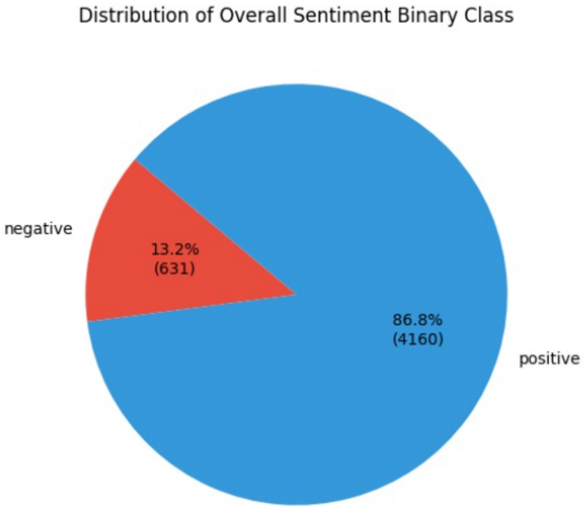
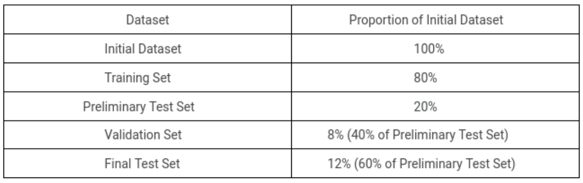


Figure 5: Class distributions of dataset for binary Classification.

4.2 Split of Dataset

We allocated the data for the five models using the same distribution method. First, we split 20% of the initial dataset into a preliminary test set, with the remaining 80% allocated to the training set. Next, we further divided 60% of the preliminary test set into the final test set, leaving the remaining 40% for the validation set. This splitting method ensures a reasonable proportion distribution among the datasets, providing a reliable foundation for model training and evaluation.The specific distribution is shown in the table below:



Dataset	Proportion of Initial Dataset
Initial Dataset	100%
Training Set	80%
Preliminary Test Set	20%
Validation Set	8% (40% of Preliminary Test Set)
Final Test Set	12% (60% of Preliminary Test Set)

Figure 6: Split Data.

5 Models

In our project we use the concept of multimodal fusion, which involves combining multiple modalities such as text and image. Goal of this model is to leverage the complementary information present in multimodal data. One of the challenges in multimodal fusion is to extend the fusion process to multimodal data while keeping the model and calculation complexity reasonable, another challenge is to preserve the correlations and dependencies between modalities and mode interaction. [20] To address these challenges, attention - or self-attention in our case - mechanisms are used. The attention mechanism allows the model to focus on relevant information by dynamically attending to different parts of each modality based on the current context or query. It involves three main components: the query, keys, and values, and output are all vectors. The attention mechanism measures the correlation between the query and the keys to determine the importance of each key - the output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key. The output - combination of the corresponding values - results then in a fused representation that combines information from multiple modalities. [21]

The self-attention mechanism enables the model to focus on different parts of the input sequence and determine their relevance to each other. It calculates attention weights for each position in the sequence based on the relationships between all positions. This allows the model to capture long-range dependencies and contextual information more effectively, leading to a better understanding of the overall context and more accurate predictions or meaningful output.

In fusion models, the features of different modalities are combined by concatenation, and a self-attention block is used to capture important features. The combined and refined features from the attention mechanism are then fed into the classification module for making predictions.

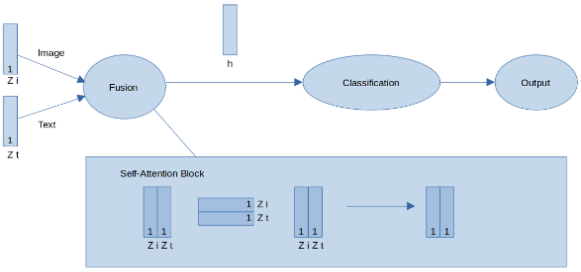


Figure 7: Multimodal Fusion Model with Self-Attention Mechanism. Based on Research Artikle, Multimodal Fusion Method based on Self-Attention.

For our experiment we use the following Models - Fusion Model with Self-Attention, based on pre-trained BERT and ResNet models for text and image representations respectively; this Model is trained with five different classificators - RNN, LSTM, CNN, FFNN with SoftMax and MLP. The same Fusion Model, but without Self-Attention mechanism. At the end we also experiment with the Scikit-learn framework for our task.

5.1 BERT and ResNet

BERT, short for Bidirectional Encoder Representations from Transformers, deployed for processing textual data, is a state-of-the-art transformer model for text classification introduced by Devlin et al. in 2018. BERT's key innovation is its bidirectional attention mechanism, allowing it to capture the context of a word based on all surrounding words, unlike previous models that processed text in a single direction.

BERT's architecture is a stack of transformer encoders with multiple self-attention heads, which weigh the importance of different words in a sentence, enabling deep understanding of context. BERT is pre-trained on two unsupervised tasks: Masked Language Modeling (MLM), where it predicts randomly masked input tokens, and Next Sentence Prediction (NSP), where it predicts if two input sentences are adjacent.

Post pre-training, BERT can be fine-tuned for specific tasks with minimal adjustments. Fine-tuning involves training the pre-trained model on a smaller, task-specific dataset with an additional task-specific layer on top. The entire model, including BERT and the task-specific layer, is fine-tuned jointly on the task-specific dataset, allowing BERT to adapt its representations to the nuances of the specific task. [22]

ResNet (Residual Network) is utilized for visual feature extraction. This deep learning model, trained using the ImageNet dataset, is primarily used for object classification. ResNet's key innovation is the residual network structure, incorporating shortcuts and batch normalization layers after each convolutional layer. This approach addresses issues with deeper networks, such as overfitting and vanishing gradients, by enabling efficient parameter passing to deeper layers.

Before ResNet, models like VGGNet and GoogLeNet had up to 19 and 22 layers, respectively. While deeper networks initially showed better performance, excessive depth led to degraded performance due to issues like overfitting and gradient problems.

Residual learning's main idea is adding shortcut paths to convolutional layers, ensuring that even if new layers learn nothing new, the network's performance doesn't degrade. ResNet structures include the basic block (ResNet-18/34) and the bottleneck block (ResNet-50/101/152), which use combinations of 3×3 and 1×1 convolution kernels to enhance training speed and accuracy.

ResNet50 begins with a large 7×7 convolution kernel, reducing a 224×224 input image to 56×56, lowering storage costs while maintaining performance. [23]

We also experimented with GloVe and CNN for image and text feature extraction in the preprocessing step, but the results were not superior. GloVe has only 100 dimensions compared to BERT, which has 768 dimensions, and CNN has only 512 dimensions compared to BERT's 1000 dimensions. For these reasons, we decided to use BERT and ResNet, as they provide a deeper understanding and have delivered better results. For the ensemble process using majority voting, we wanted to use different classifiers for our first-stage models, which focus on different aspects of classification.

5.2 Classifiers for Deep Learning Models

Classifiers such as RNN, LSTM, CNN, FFNN with SoftMax, and MLP were utilized. These deep learning models were carefully selected for their specific strengths in handling various types of data and complexities in pattern recognition. Each classifier focuses on different aspects of classification, making them ideal for approaches 1 and 2 in our project.

A **Recurrent Neural Network** (RNN) classifier is well-suited for sequence data, making it ideal for processing textual information in memotion classification tasks. RNNs maintain an internal memory that helps capture temporal dependencies and contextual information across time steps. This capability allows RNNs to sequentially process each word in a text, updating their hidden state to understand the overall sentiment and emotion conveyed in a meme.

The key advantages of RNN classifiers include their ability to handle variable-length input sequences and their effectiveness in modeling sequential data patterns. However, traditional RNNs can suffer from issues like vanishing gradients, which can hinder their ability to capture long-term dependencies. [24]

Long-Short-Term Memory (LSTM) addresses these issues by incorporating mechanisms to retain and forget information selectively. LSTMs use memory cells with input, output, and forget gates to control information flow, making them particularly effective for understanding complex patterns and relationships in text. This is crucial for tasks where emotions and sentiments are influenced by context spanning several words or sentences. LSTMs process each word in a sentence sequentially, updating their hidden state based on current and previous inputs. This enables the model to understand complex patterns and relationships within the text. [25]

A **Convolutional Neural Network** (CNN) is another classifier we used for our Model. In the context of multimodal memotion classification, CNNs can process text by applying convolutional filters to word embeddings, capturing local patterns and n-gram features. This approach is efficient for detecting specific emotional cues and sentiments within the text portion of memes. CNN classifiers are able to detect local patterns irrespective of their position in the sequence, are efficient in parallel processing, and have a relatively simple architecture compared to recurrent models. [26]

A **Feedforward Neural Network** (FFNN) with a Softmax classifier is another one we used for our Models. FFNNs process input data in a single direction, from input layers through hidden layers to output layers, without feedback loops. The Softmax classifier, typically used in the output layer, converts raw scores into probabilities across multiple classes, facilitating multi-class classification.

FFNNs with Softmax classifiers are employed to predict sentiment labels or emotional categories based on input text features. In multimodal memotion classification, FFNNs extend their applicability by integrating features from diverse modalities such as text and images. This holistic approach allows the FFNN to capture nuanced emotional expressions across different types of media, offering a comprehensive understanding of multimodal content. [27]

A **Multilayer Perceptron** (MLP) classifier we used for our purposes consists of multiple layers of neurons organized in a feedforward manner, where each neuron connects to every neuron in the subsequent layer. This architecture enables MLP to learn complex non-linear relationships between input features and target outputs.

MLP classifiers process text features by transforming them through multiple hidden layers with non-linear activation functions, ReLU (Rectified Linear Unit) in our case. This allows them to capture intricate patterns in textual data, making them suitable for tasks like sentiment analysis and text classification. In multimodal memotion classification, MLPs integrate features from different modalities (e.g., text and images) using concatenated or parallel input layers. This fusion approach enhances the model's capability to analyze and classify emotions expressed in multimodal content, combining textual sentiment with visual cues. [27]

5.3 Machine Learning (ML) Classifiers

As part of our experimental setup, we implemented an ML model using the scikit-learn framework [28], employing the same dataset used in our other models. This implementation was based on a fusion model from scikit-learn but excluded the self-attention mechanism, instead utilizing alternative algorithms to assess performance improvements. Our objective was to evaluate each model's performance before integrating them into an ensemble approach. We experimented with several classifiers including the Decision Tree Classifier, Logistic Regression, Multilayer Perceptron (MLP), Adaptive Boosting, and Support Vector Machine (SVM). Below are concise descriptions of these classifiers:

Decision Tree Classifier is a non-parametric supervised learning method that partitions the data into subsets based on feature values, making predictions by traversing the tree structure.

Logistic Regression is a linear model used for binary classification. It models the probability of the default class and is widely favored for its simplicity and interpretability.

Multilayer Perceptron (MLP) - as previously described, is a type of artificial neural network with multiple layers of perceptrons, adept at learning complex patterns and relationships within data.

Adaptive Boosting (AdaBoost) is an ensemble method that sequentially trains models, adjusting weights for misclassified instances to prioritize difficult cases, thereby creating a strong combined model.

Support Vector Machine (SVM) - a robust supervised learning algorithm for classification tasks, finding an optimal hyperplane in high-dimensional space to separate classes with maximum margin.

Through this approach, our goal was to harness the complementary strengths of diverse classifiers within an ensemble framework, thereby enhancing predictive accuracy and robustness for complex multimodal sentiment analysis tasks. Scikit-learn is an up-to-date, reliable and stable framework, that offers an accessible way to apply modern machine learning techniques without unnecessary complexity.

6 Ensemble

Ensemble learning is a powerful technique used in machine learning to improve the accuracy and robustness of predictive models by combining the predictions of multiple individual models. The core idea behind ensemble methods is that a group of diverse models, when aggregated together, often performs better than any single constituent model. This approach leverages the wisdom of crowds, where different models may excel in capturing different aspects of the data or exhibit varying strengths and weaknesses. [29]

Ensemble methods typically involve training multiple base models, which can vary in their architecture, training data subsets, or even the learning algorithms used. This diversity ensures that each model brings a unique perspective to the prediction task.

In our project, we applied an ensemble model using majority voting for our three approaches. The first two approaches involved deep learning models, and the last approach utilized a machine learning model. We employed previously tested models and five distinct classifiers for each approach, thus constituting five separate models. This structured approach allows us to maximize the strength of each model while minimizing any inherent weaknesses through a technique known as majority voting. [30]

6.1 Majority Voting

For each classifier in the Ensemble Model, we divided the training data into five using k-fold cross-validation, resulting in five outputs from each classifier. This means that under each architecture, we should have 5 models with the same architecture but different training/development sets. Then we have about [CNN, LSTM, RNN, FFNN, MLP] * 5 models, which allows us to conduct an ensemble. First, we extract the output logits on the test set (not output labels) before applying argmax. Then we average these logits and input them into argmax to get the labels.

We have conducted experiments with models within the same architecture, and then we applied majority voting for results from different architectures.

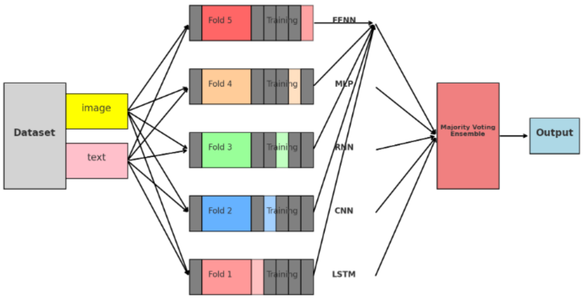


Figure 8: Majority Voting Ensemble Approach.

6.2 Implementation steps

1. Model Selection - We chose a variety of models that we had previously tested, including a fusion model with self-attention mechanisms and five different classifiers. Each of these models was trained independently.

2. Data Preparation - The training data for each classifier was divided into five subsets using k-fold cross-validation. This division enabled us to train each model multiple times on different subsets of the data, enhancing the models' ability to capture diverse patterns and nuances present in the dataset, and return best Models for Ensemble.

3. Model Training - Each classifier was trained separately on its respective subset of training data. This process involved optimizing the model parameters to minimize prediction errors and improve accuracy.

4. Prediction Collection - After training, each model generated predictions for the test data based on its learned patterns. From each classifier, we collected multiple sets of predictions corresponding to the five different subsets of training data.

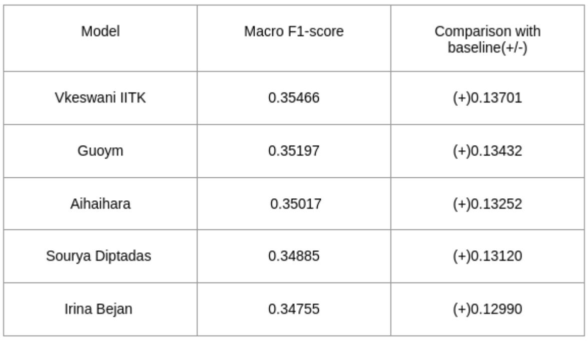
5. Selection of Best Predictions - From the multiple sets of predictions generated by each classifier, we selected the best predictions. This selection was based on criteria such as accuracy or performance metrics specific to our task.

6. Ensemble Aggregation - The final step involved aggregating these selected predictions from all classifiers using a method called majority voting. This means that for each prediction, the final output was determined based on the majority decision of the models.

7. Evaluation - Finally, the performance of the ensemble model was evaluated using metrics like macro F1 score. This evaluation helped assess how well the ensemble method improved prediction accuracy compared to individual models.

7 Results

Our models demonstrated significantly better performance compared to the baseline model in the experiments. The baseline model had a Macro-F1 score of 0.21765, whereas our best models across different methods achieved excellent results. Specifically, in the MMFA (Multimodal Fusion with Self-Attention) approach, the best-performing model, the Multilayer Perceptron (MLP), reached a Macro-F1 score of 0.5056, showcasing outstanding performance. Additionally, compared to the scores of the top participants, which ranged from 0.34600 to 0.35466, our models demonstrated competitiveness, with several models substantially exceeding these scores.

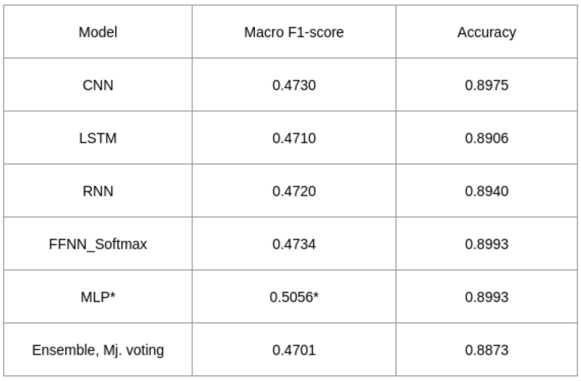


Model	Macro F1-score	Comparison with baseline(+/-)
Vkeswani IITK	0.35466	(+)0.13701
Guoym	0.35197	(+)0.13432
Aihaihara	0.35017	(+)0.13252
Sourya Diptadas	0.34885	(+)0.13120
Irina Bejan	0.34755	(+)0.12990

Figure 9: Macro-F1 scores of the models from the top five participating teams and the competition baseline model

7.1 Deep Learning Approach

In the Deep Learning Approach 1 (MMFA), our best model, MLP, achieved an F1 score of 0.5056. Figure 10 shows the Macro-F1 scores and Accuracy of the multimodal fusion models based on a Self-Attention Mechanism (MMFA) for 5 Models and Ensemble approach.

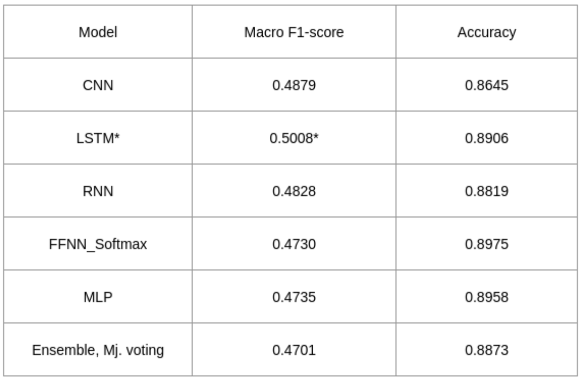


Model	Macro F1-score	Accuracy
CNN	0.4730	0.8975
LSTM	0.4710	0.8906
RNN	0.4720	0.8940
FFNN_Softmax	0.4734	0.8993
MLP*	0.5056*	0.8993
Ensemble, Mj. voting	0.4701	0.8873

Figure 10: Evaluation results of multimodal fusion model based on a Self-Attention Mechanism (MMFA) for 5 Models and Ensemble approach

7.2 Deep Learning Approach 2

In the Deep Learning Approach 2 (MMF), our top-performing LSTM model attained an F1 score of 0.5008. Figure 11 shows the Macro-F1 scores and Accuracy of the multimodal fusion models without Self-Attention Mechanism (MMFA) for 5 Models and Ensemble approach.

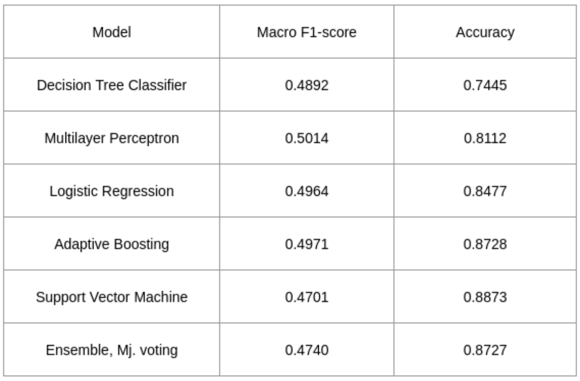


Model	Macro F1-score	Accuracy
CNN	0.4879	0.8645
LSTM*	0.5008*	0.8906
RNN	0.4828	0.8819
FFNN_Softmax	0.4730	0.8975
MLP	0.4735	0.8958
Ensemble, Mj. voting	0.4701	0.8873

Figure 11: Evaluation results of multimodal fusion model (MMF) for 5 Models and Ensemble approach

7.3 Machine Learning Approach

In the Machine Learning Approach using Scikit-Learn, our best-performing independent model, the Multilayer Perceptron (MLP), reached an F1 score of 0.5014. Figure 12 shows the Macro-F1 scores and Accuracy of the models with Machine Learning Approach using Scikit-Learn.



Model	Macro F1-score	Accuracy
Decision Tree Classifier	0.4892	0.7445
Multilayer Perceptron	0.5014	0.8112
Logistic Regression	0.4964	0.8477
Adaptive Boosting	0.4971	0.8728
Support Vector Machine	0.4701	0.8873
Ensemble, Mj. voting	0.4740	0.8727

Figure 12: Evaluation results of Machine Learning Approach using Scikit-Learn

These results indicate that our models excel in the multimodal sentiment analysis task, demonstrating not only advantages over the baseline but also competitive and superior performance compared to the top participants' models.

8 Conclusion

Our models have demonstrated a significant improvement over the baseline, which had a Macro-F1 score of 0.21765. The Multilayer Perceptron (MLP) model within the MMFA approach achieved a Macro-F1 score of 0.5056, indicating a substantial performance enhancement. Furthermore, when compared to the top-performing participants, whose scores ranged from 0.34600 to 0.35466, our models not only performed competitively but also exceeded these scores by a notable margin.

9 Limitations, Future Work

Despite the excellent performance of our models, there are still some shortcomings, primarily related to the quality of the dataset.

Firstly, the total amount of data is relatively small, and the classifiers we employed require a large amount of training data to achieve optimal performance, which has led to suboptimal model results.

Secondly, as illustrated by the confusion matrices for the MMF_LSTM and MMFA_LSTM models in Figures 13 and 14, the label distribution in the dataset is highly imbalanced, with a significant shortage of negative labels (631 out of 4791 samples, or 13.2%), making it challenging for the model to accurately identify negative cases.

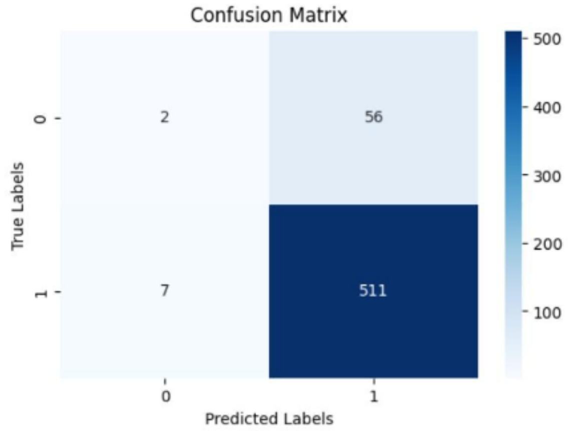


Figure 13: Confusion Matrix: MMF_LSTM

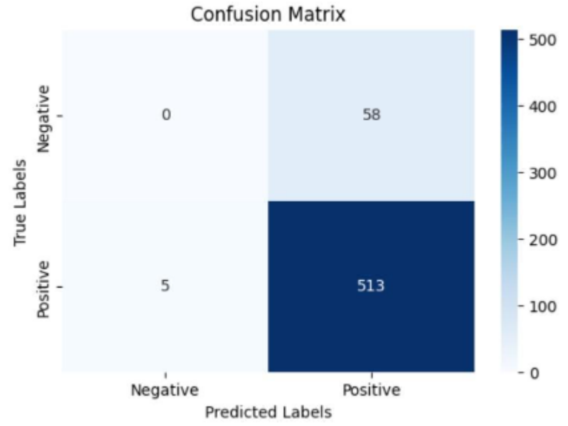


Figure 14: Confusion Matrix: MMFA_LSTM

To further enhance the performance and robustness of our models, future work will focus on the following areas: Expanding the Dataset Size: By increasing the size of the dataset and ensuring a more balanced distribution of positive and negative samples, we aim to improve the overall quality of the dataset. This will help the model learn various patterns in the data more effectively during training. Training on Larger Datasets: We plan to train our models on larger datasets to boost their performance and achieve more reliable results. This approach will help the models perform more robustly across different practical application scenarios. Exploring Advanced Methods: In addition to the existing BERT and CNN methods, we will explore other advanced feature extraction techniques. This will aid in further improving the model's ability to capture complex features and enhance prediction accuracy. By implementing these improvements, we expect to significantly enhance the overall performance of our models and provide stronger support for practical applications.

10 Group Composition

- Aldi Halili is responsible for the main idea of our project. He is also mainly responsible for model architectures;
- Chunxue Liu is mainly responsible for models' evaluation;
- Valeriya Herrlein is mainly responsible for data preprocessing, feature extraction, Overleaf formatting;

All team members wrote the paper with input from all authors - we researched and tried out such topics as Bert, ResNet, GloVe, CNN, Ensemble, Multimodal Fusion, Self-Attention, Deep learning classifiers, skikit-learn tools, sentiment analysis, memotion classification and a lot more. In the end we discussed the results and contributed to the final manuscript.

11 References

[1] Zhu, Hu, Wang, Ze, Shi, Yu, Hua, Yingying, Xu, Guoxia, Deng, Lizhen, Multimodal Fusion Method Based on Self-Attention Mechanism, Wireless Communications and Mobile Computing, 2020, 8843186, 8 pages, 2020. https://doi.org/10.1155/2020/8843186

[2] Sharma, chhavi, et al. "SemEval-2020 Task 8: Memotion Analysis–The Visual-Lingual Metaphor!." arXiv preprint arXiv:2008.03781 (2020).

[3] Lai, S., Hu, X., Xu, H., Ren, Z., & Liu, Z. (2023). Multimodal Sentiment Analysis: A Survey. arXiv. https://arxiv.org/abs/2305.07611

[4] image_6260 in 'memotion_dataset-7k' from Kaggle (https://www.kaggle.com/datasets/williamscott701 /memotion-dataset-7k?resource=download&select =memotion_dataset_7k)

[5] Igor Mozeti˘c, Miha Gr˘car, and Jasmina Smailovi´c.2016. Multilingual twitter sentiment classification: The role of human annotators. PloS one, 11(5):e0155036.

[6] Manoochehr Ghiassi and S Lee. 2018. A domain transferable lexicon set for twitter sentiment analysis using a supervised machine learning approach. Expert Systems with Applications, 106:197–216.

[7] Wenkuan Li, Peiyu Liu, Qiuyue Zhang, and Wenfeng Liu. 2019. An improved approach for text sentiment classification based on a deep neural network via a sentiment attention mechanism. Future Internet, 11(4):96.

[8] Akshi Kumar and Arunima Jaiswal. 2017. Image sentiment analysis using convolutional neural network. In International Conference on Intelligent Systems Design and Applications, pages 464–473. Springer.

[9] Md Shad Akhtar, Asif Ekbal, and Erik Cambria. 2020.How intense are you? predicting intensities of emotions and sentiments using stacked ensemble. IEEE Computational Intelligence Magazine, 15(1):64–75.

[10] International Conference on Information Processing (ICIP), S. Singh et al., Eds., 2015. [Online]. Available: https://ieeexplore.ieee.org/document/7489424

[11] N. Mittal, D. Sharma, and M. L. Joshi, "Image Sentiment Analysis Using Deep Learning," ACM International Conference on Web Intelligence (WI), pp. 684–687, 2018.

[12] Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Erik Cambria, Alexander Gelbukh, and Amir Hussain. 2018. Multimodal sentiment analysis: Addressing key issues and setting up the baselines.IEEE Intelligent Systems, 33(6):17–25.

[13] Tao Jiang, Jiahai Wang, Zhiyue Liu, and Yingbiao Ling. 2020. Fusion-extraction network for multimodal sentiment analysis. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pages 785–797. Springer.

[14] Pranati Behera, Mamta, and Asif Ekbal. 2020. Only text? only image? or both? Predicting sentiment of internet memes. In Proceedings of the 17th International Conference on Natural Language Processing (ICON), pages 444–452, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

[15] N. V. Alluri and N. Dheeraj Krishna, "Multi Modal Analysis of memes for Sentiment extraction," 2021 Sixth International Conference on Image Information Processing (ICIIP), Shimla, India, 2021, pp. 213-217, doi: 10.1109/ICIIP53038.2021.9702696.

[16] Bucur, A., Cosma, A., & Iordache, I. (2022). BLUE at Memotion 2.0 2022: You have my Image, my Text and my Transformer. ArXiv, abs/2202.07543.

[17] Hu, Z., Wang, Z., Shi, Y., Hua, Y., Xu, G., & Deng, L. (2020). Multimodal Fusion Method Based on Self-Attention Mechanism. Wireless Communications and Mobile Computing, 2020, Article ID 8843186. https://doi.org/10.1155/2020/8843186

[18]'memotion_dataset-7k' from Kaggle(https://www.kaggle.com/datasets/williamscott701/ memotion-dataset-7k?resource=download&select =memotion_dataset_7k)

[19] Memotion analysis" in 2020 (https://competitions.codalab.org/competitions/20629)

[20] Shivam Sharma, Ramaneswaran S, et al. "Emotion-Aware Multimodal Fusion for Meme Emotion Detection" 2024.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. 'Attention Is All You Need', 2017.

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', 2019.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun 'Deep Residual Learning for Image Recognition', 2015.

[24] Alex Sherstinsky 'Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network', 2020.

[25] Ralf C. Staudemeyer, Eric Rothstein Morris 'Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks', 2019.

[26] Farhana Sultana, Abu Sufian, Paramartha Dutta 'Advancements in Image Classification using Convolutional Neural Network', 2018.

[27] Yann LeCun, Yoshua Bengio, Geoffrey Hinton. 2015. Deep learning.

[28] Jamie J. R. Bennett, Yan Chak Li, Gaurav Pandey 'An Open-Source Python Package for Multi-modal Data Integration using Heterogeneous Ensembles', 2024.

[29] Alireza Ghorbanali, Mohammad Karim Sohrabi, Farzin Yaghmaee "Ensemble transfer learning-based multimodal sentiment analysis using weighted convolutional neural networks." 2022.

[30] Alican Dogan; Derya Birant 'A Weighted Majority Voting Ensemble Approach for Classification', 2019