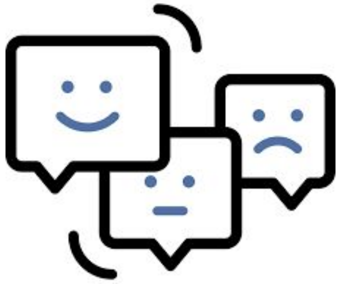
Project Presentation



Deep learning and Machine Learning Techniques
for Multimedia Sentiment Analysis

Group Members: Aldi Halili, Chunxue Liu, Valeriya Herrlein

Date: 17.07.2024

CONTENT

01 Introduction

02 Dataset

03 Features Extraction

04 Model Architectures

05 Models Evaluation

06 Conclusion

01

P

ART ONE

Introduction

Introduction

Background and Motivation

-The influence of social media and memes;

-Multimodal fusion - text and image data;

-The adoption of an ensemble model approach in order to see, whether the ensemble model will demonstrate enhanced performance and greater accuracy in sentiment analysis than any single fusion model approach

Research Question

-How effective are the developed Deep Learning and Machine Learning models compared to the baseline and the top participants of the Memotion competition in Task A of sentiment analysis?

02

PART TWO

Dataset

Dataset

Source

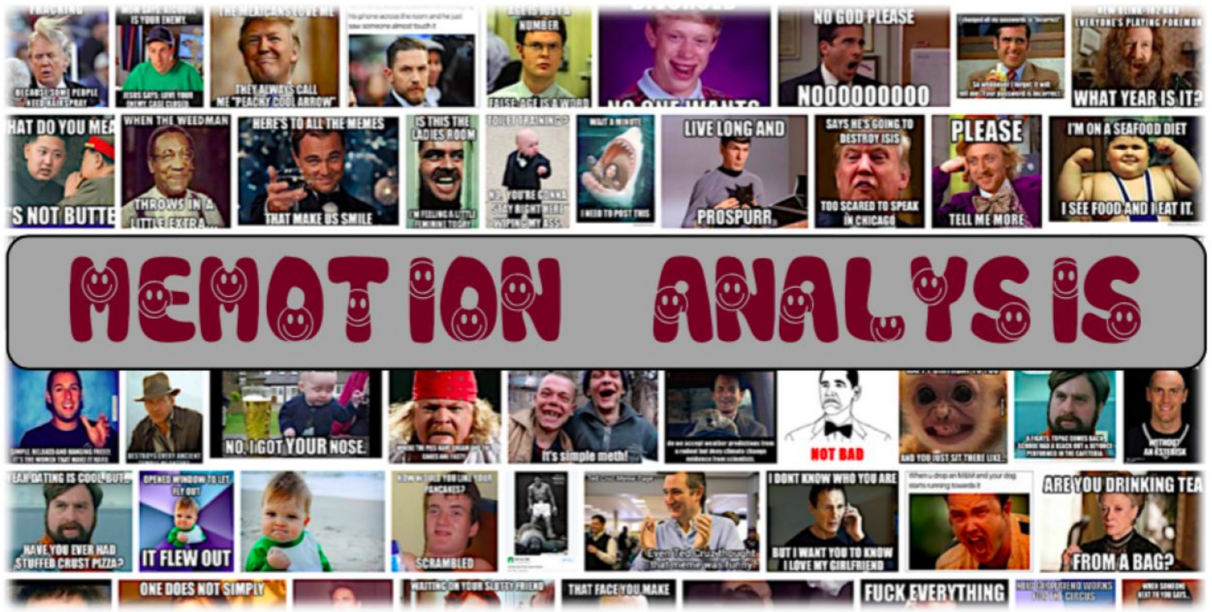
'memotion_dataset-7k' from Kaggle
Originally from Codalab competition (Memotion Analysis, Task A- Sentiment Classification)

Annotations

Human-annotated with "overall_sentiment" labels
(we used positive and negative labels for our task)

Volume & Usage

originally 6,992 memes, filtered to 4,791
Images (memes) and labels.csv file



Dataset

Image Data

image_1

image_2

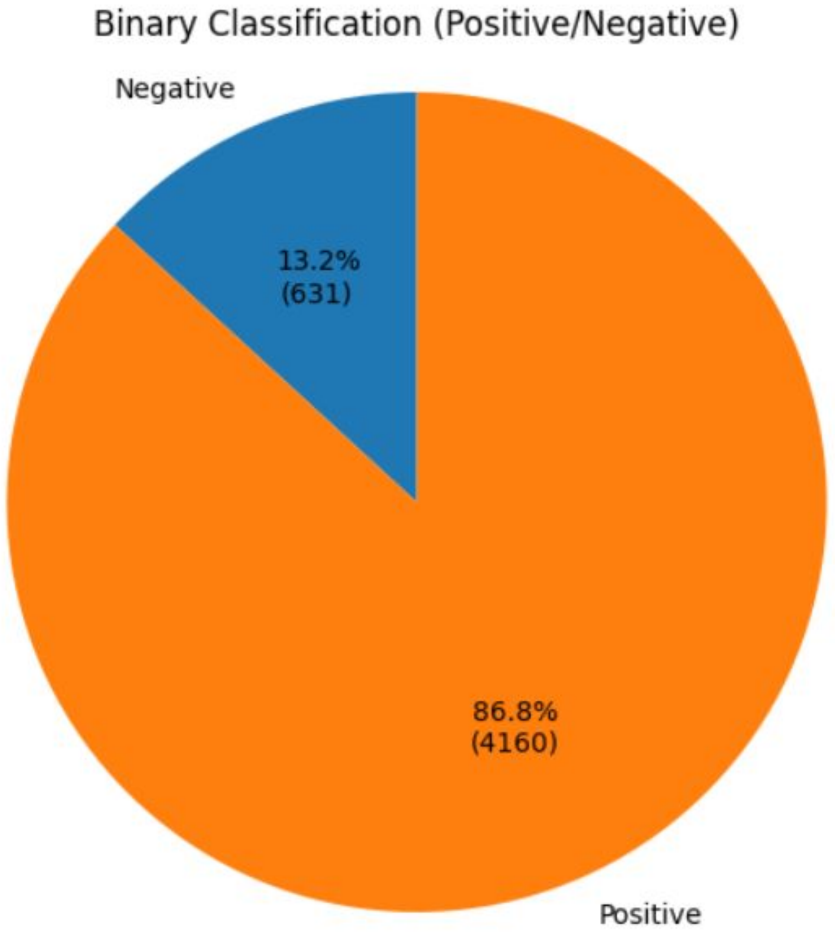
image_3

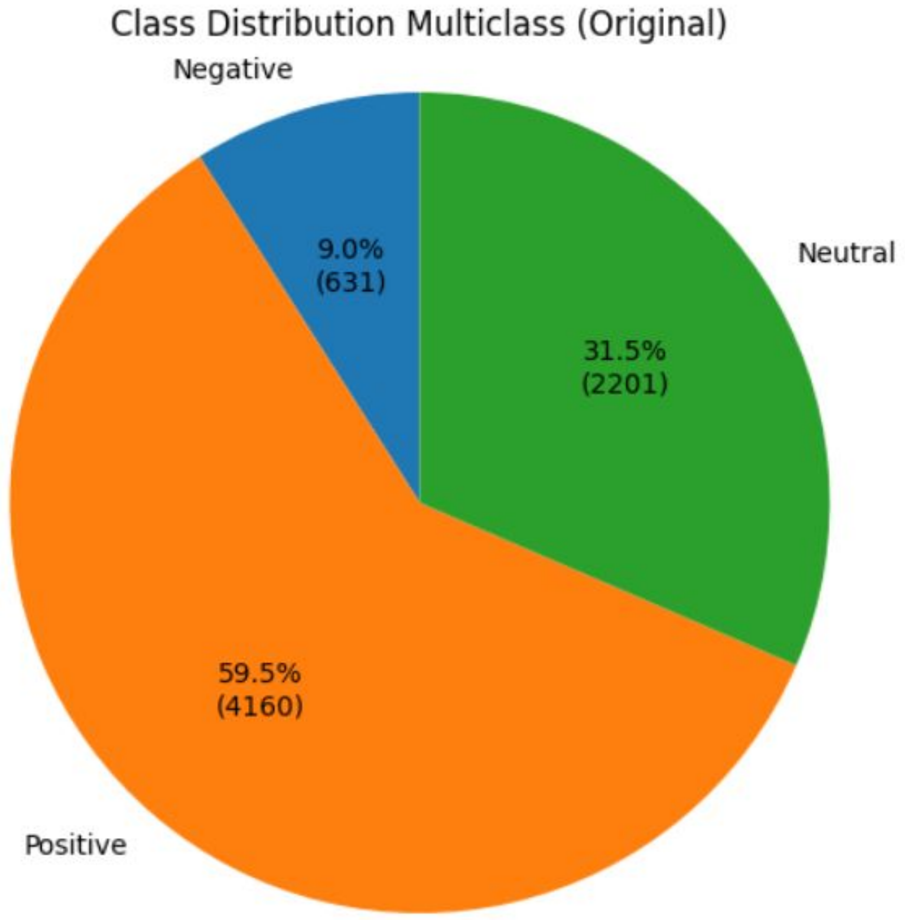
image_4

Labeled text data

image_name	text_ocr	text_corrected	humour	sarcasm	offensive	motivational	overall_sentiment
image_1.jpg	LOOK THERE MY FRIEND LIGHTYEAR NOW ALL SOHALIK...	LOOK THERE MY FRIEND LIGHTYEAR NOW ALL SOHALIK...	hilarious	general	not_offensive	not_motivational	very_positive
image_2.jpeg	The best of #10 YearChallenge! Completed in le...	The best of #10 YearChallenge! Completed in le...	not_funny	general	not_offensive	motivational	very_positive
image_3.JPG	Sam Thorne @Strippin (Follow Follow Saw every...	Sam Thorne @Strippin (Follow Follow Saw every...	very_funny	not_sarcastic	not_offensive	not_motivational	positive
image_4.png	10 Year Challenge - Sweet Dee Edition	10 Year Challenge - Sweet Dee Edition	very_funny	twisted_meaning	very_offensive	motivational	positive

Class distributions of dataset for task A





03

PART THREE

Features Extraction

Image Preprocessing

with pre-trained **ResNet-50**

- Resize images
- Crop images
- Convert images to tensors
- Normalize images
- Extract Image features

```
# Initializing the pre-trained ResNet-50 model
r_model = models.resnet50(pretrained=True)
re_model = r_model.to(device)
re_model.eval()
```

```
# define the transformation
image_trans = transforms.Compose([
    transforms.Resize(256),
    transforms.CenterCrop(224),
    transforms.ToTensor(),
    transforms.Normalize(
        mean=[0.485, 0.456, 0.406],
        std=[0.229, 0.224, 0.225]),
])
```

Text Preprocessing

with **BERT**

- Tokenization
- Text Embeddings extraction

```
# We use BERT Tokenizer and BERT Model for the preprocessing step of text and then extract embeddings

# Initialize the BERT tokenizer
tokenizer = BertTokenizer.from_pretrained('bert-base-uncased')

# Initialize the BERT model
model = BertModel.from_pretrained('bert-base-uncased')
```

04

PART FOUR

Model Architectures

Model Architectures

Binary Classification for Multimodal Sentiment Analysis Task

Three Approaches to Achieve a State-of-the-Art Model:

1. **First Approach: Deep Learning Model**

   Multimodal Fusion Model Based on Self-Attention (MMFA)

2. **Second Approach: Deep Learning Model**

   Multimodal Fusion Model (MMF) without Self-Attention

3. **Third Approach: Machine Learning Model**

   Instead of using deep learning, this approach use traditional Machine Learning techniques with the scikit-learn framework.

4. **Majority Voting Ensemble Approach**:

   Each approach (1, 2, 3) employs a majority voting technique.

First Approach: DL Model

Fusion Model based on Self-Attention (MMFA)

Extraction:

- Image: ResNet
- Text: BERT

Fusion:

- Combines features from all modalities.
- Use A Self-Attention Block to capture important features.

Classification:

- Produces the final output based on combined features and refined features from self-attention block
- Five Classifier: LSTM, CNN, RNN, MLP, FFNN

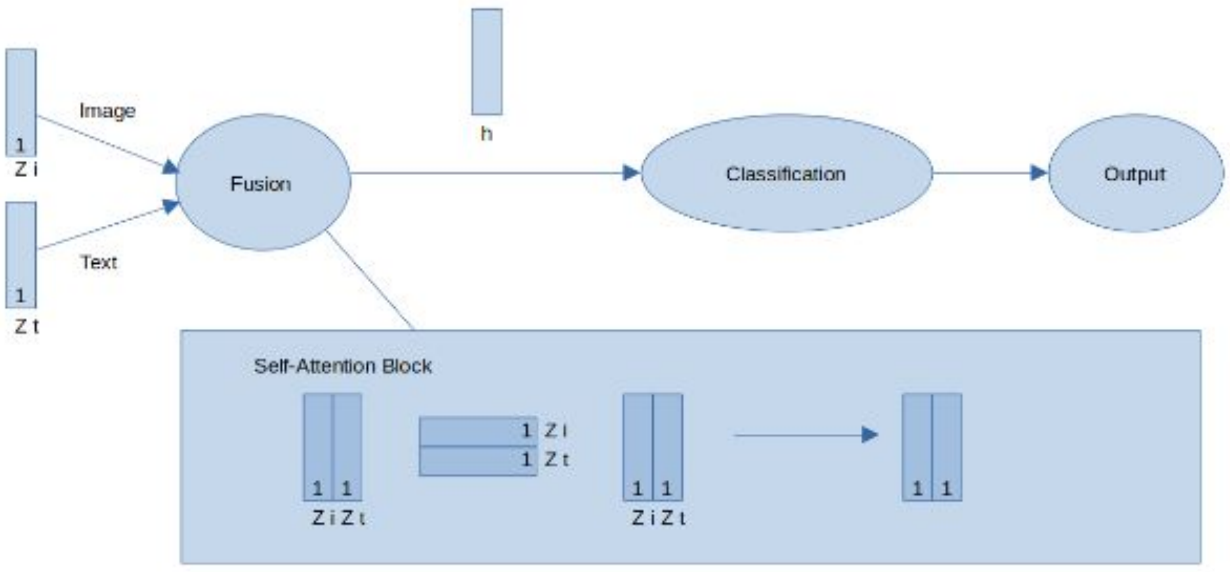


Figure3: Multimodal Fusion Model with Self-Attention Mechanism

(Zhu, Hu, et al., 2020)

Second Approach: DL Model

Fusion Model *without Self-Attention Block* (MMF)

Extraction:
- visual :ResNet
- Language: BERT

Fusion:
- Combines features from all modalities.

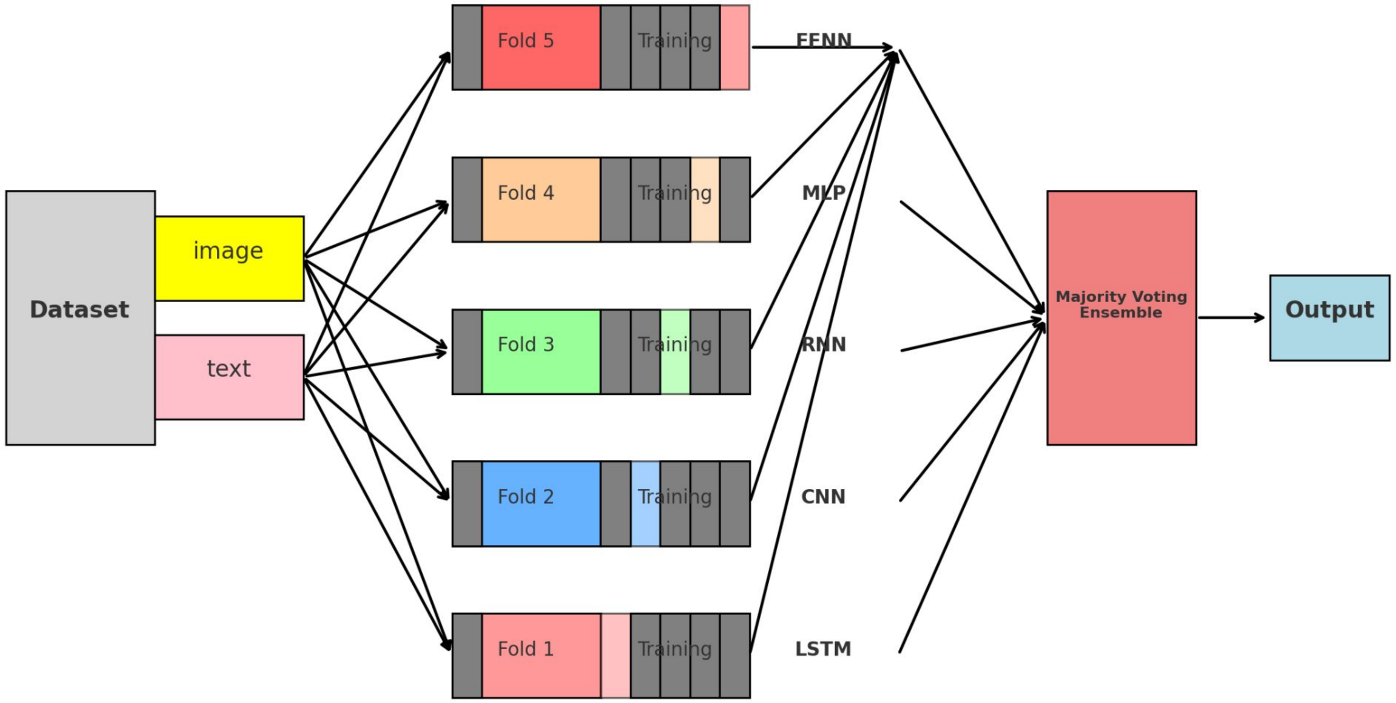
Classification:
- Produces the final output based on combined features
- 5 Classifier: LSTM, CNN, RNN, MLP, FFNN

Multimodal Fusion Model

```python
class MultimodalFusionClassifier(nn.Module):
    def __init__(self, text_dim, image_dim, hidden_dim, lstm_hidden_dim, num_classes):
        super(MultimodalFusionClassifier, self).__init__()
        self.text_model = nn.Linear(text_dim, hidden_dim)
        self.image_model = nn.Linear(image_dim, hidden_dim)

        # Self-attention module for refining text and image features by focusing on important elements.
        self.text_attention = SelfAttention(hidden_dim)
        self.image_attention = SelfAttention(hidden_dim)

        # LSTM Classifier
        self.classifier = LSTMClassifier(hidden_dim * 2, lstm_hidden_dim, num_layers=2, num_classes=num_classes)

    def forward(self, text_features, image_features):
        # Process text features and give attention
        text_features = self.text_model(text_features)
        text_features = self.text_attention(text_features)

        # Process image features and give attention
        image_features = self.image_model(image_features)
        image_features = self.image_attention(image_features)

        # Combine text and image features
        combined_features = torch.cat([text_features, image_features], dim=1)
        combined_features = combined_features.unsqueeze(1)

        # Classification
        #The combined features are passed through the LSTM classifier to produce the final output.
        output = self.classifier(combined_features)
        return output
```

Majority Voting Ensemble Approach: MMFA

- K-Fold Cross-Validation:
  - The dataset is split into 5 equal parts (folds).
  - We use these 5 pieces of data to obtain 5 * 5 models.
- Majority Voting
  - After training, predictions from each model are collected.
  - Majority Voting is employed where the final output label is determined by the most common prediction from all models.



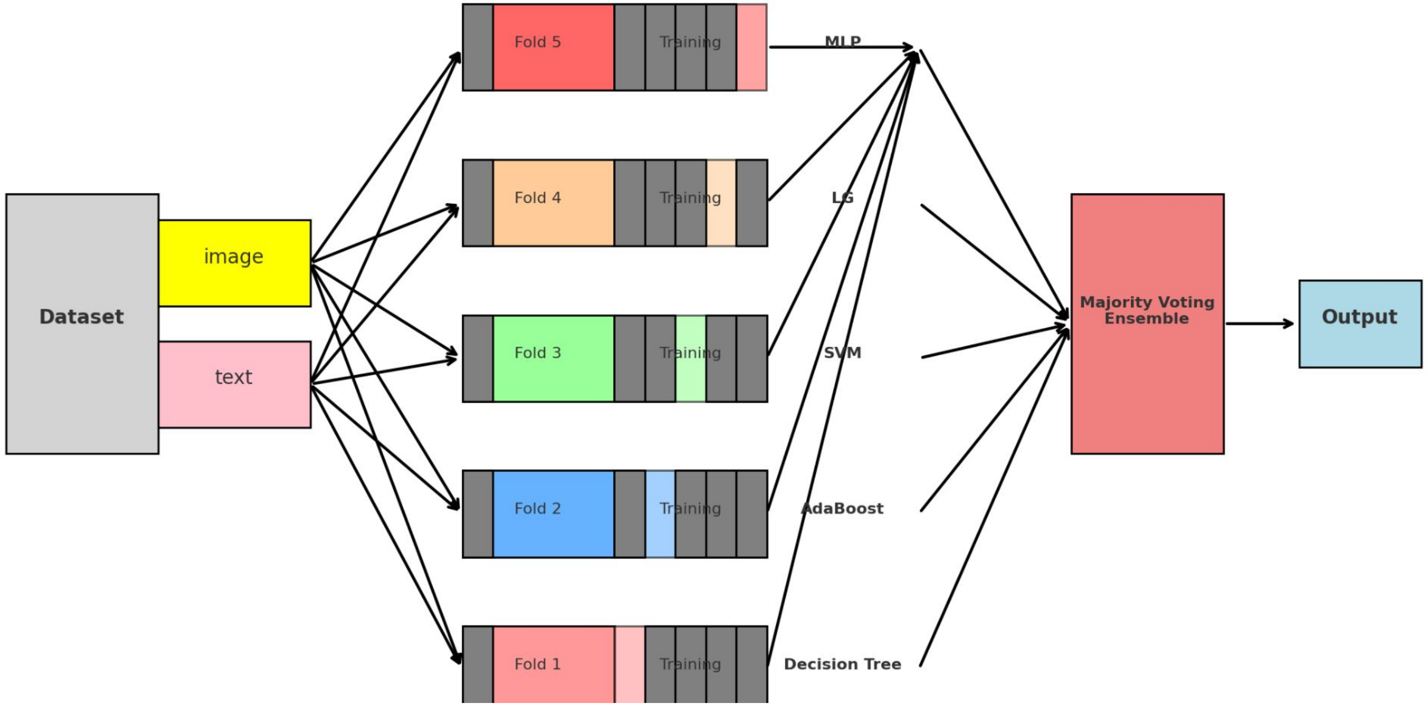
Third Approach：
Machine Learning with scikit-learn

Model Training:

- Decision Tree Classifier
- Logistic Regression
- Multilayer Perceptron
- Adaptive Boosting
- Support Vector Machine

K-Fold Cross-Validation

Majority Voting(Soft Voting)



04

PART FIVE

Models Evaluation

Results

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
SemEval-Baseline	0.2176	-

Results

Deep-Learning Approach 1: Multimodal Fusion with Self-Attention (MMFA)

- **Results:**

Model	Macro F1-score	Accuracy
CNN	0.4730	0.8975
LSTM	0.4710	0.8906
RNN	0.4720	0.8940
FFNN with Softmax	0.4734	0.8993
MLP ✦	0.5056 ✦	0.8993
Ensemble, Mj. Voting	0.4701	0.8873

Results

Deep-Learning Approach 2: Multimodal Fusion (MMF)

- **Results:**

Model	Macro F1-score	Accuracy
MLP	0.4735	0.8958
CNN	0.4879	0.8645
RNN	0.4828	0.8819
FFNN with Softmax	0.4730	0.8975
LSTM	0.5008	0.8906
Ensemble, Mj. Voting	0.4701	0.8873

Results

Machine Learning Approach Using Scikit-Learn

- **Results:**

Model	Macro F1-score	Accuracy
Decision Tree	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

Conclusion

Compare of the performance

our Models VS. SEMEVAL-Baseline Model

Models	Macro F1 score
MMFA-MLP	0.5056
MMF_LSTM	0.5008
Scikit_learn_ensemble	0.4741
SEMEVAL-Baseline	**0.21765**

Conclusion

1. Answer to the Research Question

- How effective are the developed Deep Learning and Machine Learning models compared to the baseline and the top participants of the Memotion competition in Task A of sentiment analysis?

- Compared to the baseline models and the performance of the participants', our models have a better performance.

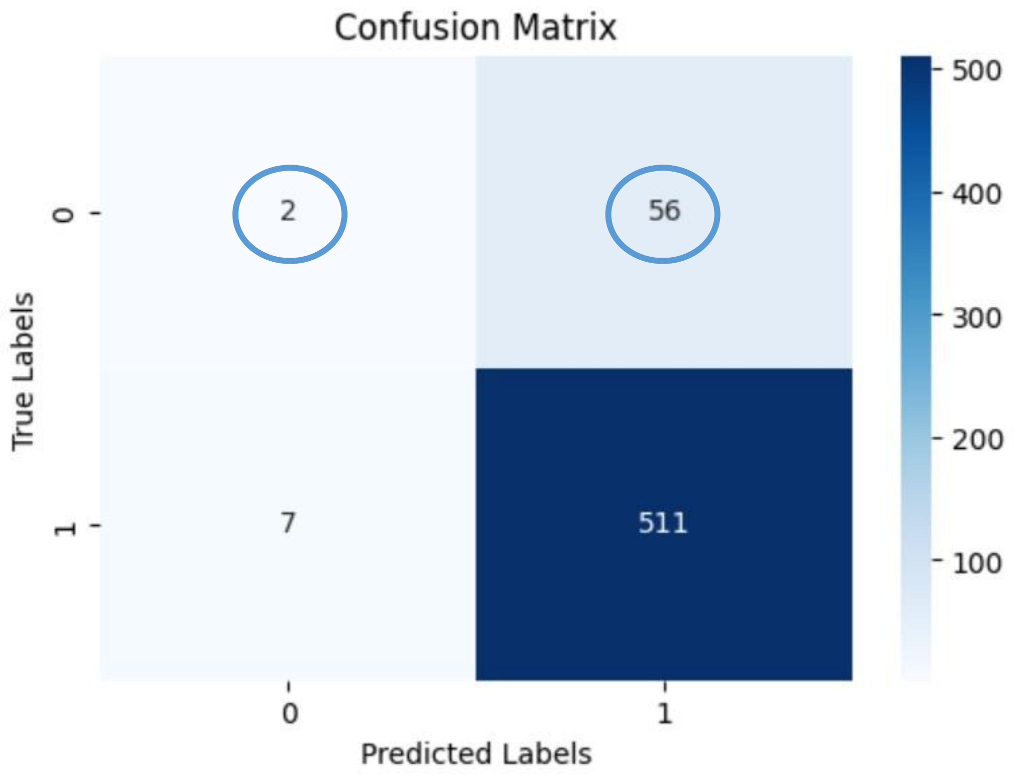
Limitation

Dataset:

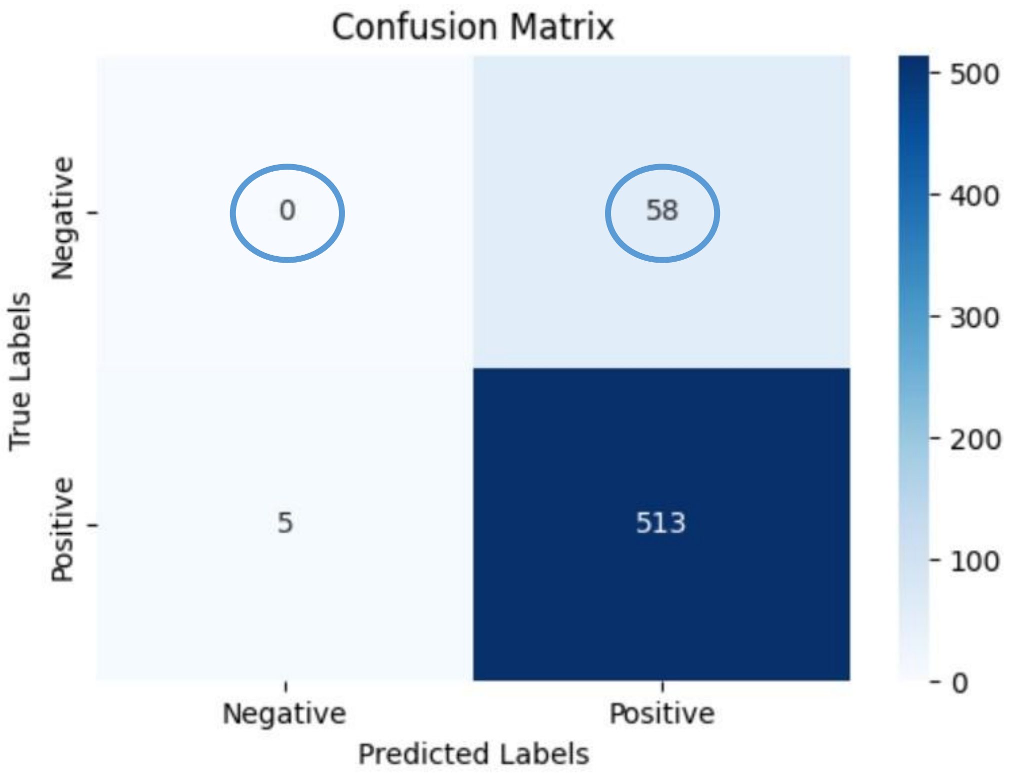
1. The total amount of data is relatively small, and the classifiers we employed require a large amount of training data to achieve optimal performance, which leads to suboptimal model results.

2. The label distribution is highly imbalanced, with too few negative labels in the dataset (631 out of 4791, 13.2%), making it difficult for the model to accurately identify negative cases.

Confusion Matrix



Confusion Matrix: MMF_LSTM



Confusion Matrix: MMFA_LSTM

Future Work

Dataset:

- Increase the dataset size and improve the quality of the dataset by ensuring a more balanced distribution of positive and negative examples.

- Train the models on larger datasets to improve their performance and ensure more robust results.

- Explore methods beyond BERT and CNN to improve feature extraction.

Literature

Dataset

https://www.kaggle.com/datasets/williamscott701/memotion-dataset-7k?resource=download

Restnet and transformation

https://medium.com/@engr.akhtar.awan/how-to-fine-tune-the-resnet-50-model-on-your-target-dataset-using-pytorch-187abdb9beeb
https://medium.com/@nitishkundu1993/exploring-resnet50-an-in-depth-look-at-th[Linktext](https://)e-model-architecture
-and-code-implementation-d8d8fa67e46f
https://datagen.tech/guides/computer-vision/resnet-50/

Models

blogs
https://dida.do/de/blog/ensembles-in-machine-learning
https://neptune.ai/blog/ensemble-learning-guide

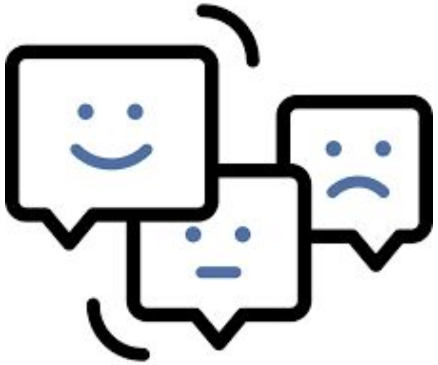
Paper
https://dl.acm.org/doi/pdf/10.1145/3589335.3651971
https://www.sciencedirect.com/science/article/abs/pii/S030645732200053X

Zhu, Hu, et al. "Multimodal fusion method based on self-attention mechanism." *Wireless Communications and Mobile Computing* 2020 (2020): 1-8.
https://downloads.hindawi.com/journals/wcmc/2020/8843186.pdf?_gl=1*3plcio*_ga*MTA5NTM2OTE1Ni4xNzE1MTcxODk1*_ga_NF5QFMJT5V*MTcxNTE3MTg5NC4xLjEuMTcxNTE3MzMxMi42MC4wLjA.&_ga=2.236610569.906314749.1715171895-1095369156.1715171895

Liu, Zhicheng, et al. "Ensemble Pretrained Models for Multimodal Sentiment Analysis using Textual and Video Data Fusion." *Companion Proceedings of the ACM on Web Conference 2024*. 2024.

Github
https://github.com/imadhou/multimodal-sentiment-analysis/blob/main/notebooks/multi_modal_model.ipynb
https://medium.com/@wangdk93/implement-self-attention-and-cross-attention-in-pytorch-1f1a366c9d4b#d075

Our implementation of the complete project for multimodal sentiment analysis: itsmeeeeeee/MML (github.com)



THANK YOU

Group Members: Aldi Halili, Chunxue Liu, Valeriya Herrlein