Master Thesis

Visual Impact on Sentiment: Climate Change Tweets Analysis

Pranav Tyagi
(matriculation number 1937303)

February 14, 2025

Submitted to
Data and Web Science Group
Prof. Dr. Margret Keuper
University of Mannheim

Abstract

Summary of research goals, methodology, results, and contributions.

Contents

Abstract **ii**

1. Introduction **1**

1.1. Introduction . . . 1
1.2. Background and Motivation . . . 1
1.3. Problem Statement . . . 3
1.4. Research Objectives and Approach . . . 3
1.4.1. Primary Research Question . . . 3
1.5. Thesis Structure . . . 4

2. Literature Review **5**

2.1. Text-Based Sentiment and Emotion Analysis . . . 5
2.2. Vision-Based Sentiment and Emotion Analysis . . . 8
2.3. Multimodal Sentiment and Emotion Analysis . . . 12
2.3.1. Multimodal Models . . . 12
2.3.2. Fusion strategies . . . 14
2.4. Weakly Supervised Learning . . . 16
2.4.1. Weakly Supervised Learning in Text . . . 16
2.4.2. Weakly Supervised Learning in Vision . . . 20
2.5. Summary . . . 21

3. Research Methodology **22**

3.1. Overview . . . 22
3.2. Data Description and Preprocessing . . . 22
3.2.1. Dataset Source . . . 22
3.2.2. Data Filtering . . . 23
3.2.3. Preprocessing Pipeline . . . 23
3.2.4. Final Dataset Statistics . . . 23
3.3. Evaluation Metrics . . . 24
3.3.1. Zero-Shot Evaluation Metrics . . . 24
3.3.2. Fine-Tuning Evaluation Metrics . . . 24
3.4. Zero-Shot Experimentation . . . 25
3.4.1. Model Selection and Rationale . . . 25
3.4.2. Experimental Setup . . . 26
3.5. Weakly Supervised Learning . . . 27
3.5.1. Zero-shot-classification-boost-with-self-training . . . 27
3.5.2. Loss Reweighting . . . 28

3.6. Fine-Tuning with Soft Labels 29
3.6.1. Label Mapping to Ekman's Six Emotions 30
3.6.2. Architectural Overview 30
3.7. Experimental Setup 34

4. Experimental Results 35
4.1. Zero-Shot Experimentation Results 35
4.1.1. Model Performance Overview 35
4.1.2. Confidence Filtering 36
4.2. Weakly Supervised Learning Results 36
4.2.1. Zero-shot Classification Boost with Self-training 36
4.2.2. Loss Re-weighting 37
4.3. Text-Based Results 38
4.4. Image-Based Results 39
4.5. Multimodal Results 40
4.6. Sensitivity to Datasets and Random Initialisation 40
4.6.1. Comparison Across Datasets 40
4.6.2. Effect of Random Initialisation 42
4.7. Comparative Analysis: Performance Metrics, Model Configurations, and Hyperparameter Impact 42
4.7.1. Overall Performance Across Modalities 43
4.7.2. Hyperparameter Impact: Text-Only Models 43
4.7.3. Hyperparameter Impact: Image-Only Models 44
4.7.4. Hyperparameter Impact: Multimodal Models 45
4.7.5. Summary of Comparative Findings 45
4.8. Error Analysis 46
4.8.1. Cross-Modal Error Patterns 47
4.8.2. Class-Level Analysis 48
4.9. Qualitative Analysis 50
4.9.1. Key Findings 55
4.10. Conclusion 56

5. Discussion 57
5.1. Analysis of Findings 57
5.1.1. Zero-shot performance 57
5.1.2. Experimental Results: Distribution vs. Instance-Level Accuracy . 58
5.2. Future Directions 60

6. Conclusion 62
6.1. Summary of the Thesis 62

Bibliography 63

A. Program Code / Resources 69

B. Complete Experimental Results 70

C. Loss Curves 87

Ehrenwörtliche Erklärung 88

List of Figures

2.1. Contrastive Pre-training 13
2.2. Creation of dataset classifier and final prediction 13
2.3. Approach to training CLIP and Inference (Radford et al. 2021) 13

3.1. Single Modality Architecture 31
3.2. Multimodal Fusion Architecture 31
3.3. Finetuning setup with Base MLP 31
3.4. Multimodal Residual fusion architecture 32

4.1. Performance comparison of zero-shot models based on key evaluation metrics. 35
4.2. Emotion-wise comparison of Mean Squared Error (MSE) across models and baselines. Emotion order - Anger, Fear, Surprise, Disgust, Sadness, Joy. Lower is better. 46
4.3. Confusion matrices for text models. Darker blues indicate higher prediction frequency. 47
4.4. Confusion matrices for image-only models. 48
4.5. Multimodal confusion matrices. 49
4.6. Examples of high-error predictions (Case A on left, Case B on right). . . . 50
4.7. Examples of well-aligned predictions (Case C on left, Case D on right). . . 53

C.1. Train and Validation KL Divergence loss curves 87

List of Tables

3.1. Dataset Statistics for February and August 2019. 23
3.2. Distribution of Manual Emotion Labels 27
3.3. Mapping of Emotions into Ekman Categories 30
3.4. Architectural Variations 33

4.1. Performance of the zero-shot models on the annotated subset ($n = 99$). . 36
4.2. Performance of the zero-shot models on the annotated subset ($n = 99$) with confidence score > 0.9. 36
4.3. Zero-shot classification accuracy of entailment models. For each zero-shot entailment model and dataset, The test accuracy of the off-the-shelf model to its accuracy after 2 iterations of self-training. RoBERTa, DeBERTa, and BART correspond to the following models from Hugging Face Hub: roberta-large-mnli, deberta-large-mnli-zero-cls, and bart-large-mnli. 37
4.4. Per-class accuracy comparison. 38
4.5. Classification report comparison before and after training. 38
4.6. Performance comparison between Baseline and Our Model. The Baseline is the **CardiffNLP RoBERTa-Large**, which is used in a zero-shot setting. Higher values are better for ↑, and lower values are better for ↓. Bold values indicate combined scores. 39
4.7. Performance comparison between Baseline and Our Model for image-based experiments. The Baseline model is **CLIP ViT-L/14**, applied without fine-tuning. Higher values are better for ↑, and lower values are better for ↓. Bold values indicate combined scores. 39
4.8. Performance comparison between Baseline and Our Model for multimodal experiments. The Baseline predictions are obtained by averaging the outputs from the two unimodal Baseline models. Higher values are better for ↑, and lower values are better for ↓. Bold values indicate combined scores. 40
4.9. Comparison of mean performance on August vs. February data. For each month, the results are averaged across all experiments, including both random seeds (42 and 7), to provide a comprehensive performance summary for each modality (text-only, image-only, and multimodal). 41
4.10. Comparison of mean performance with seeds 42 vs. 7. For each seed, the results are averaged across all experiments, including data from both months, to provide a comprehensive performance summary for each modality (text-only, image-only, and multimodal). 42
4.11. Top and bottom performers across text, image, and multimodal settings. . 43

4.12. Impact of learning rate, epoch count, optimizer, and freezing on text-only performance. The best configuration is in bold. 44
4.13. Impact of learning rate, epoch count, optimizer, and freezing on image-only performance. The best configurations are in bold. 44
4.14. Impact of learning rate, epochs, dropout, optimizer, fusion strategy, and freezing on multimodal performance. The best configurations are in bold. 45

B.1. Complete multimodal experimental results 70
B.2. Complete image experimental results 84
B.3. Complete text experimental results 86

1. Introduction

1.1. Introduction

Climate change is one of the most urgent global challenges, affecting environmental, economic, and social dimensions worldwide. In the digital era, social media platforms such as X (formerly Twitter), Facebook, and Instagram serve as critical arenas for climate change discussions, where information—and misinformation—spreads rapidly. Given the multimodal nature of these platforms, where text blends with images, memes, and infographics, it is essential to understand not only the emotions these elements convey but also how they trigger emotional responses in viewers.

However, analyzing these emotional dynamics in social media posts poses several challenges. Social media content is typically noisy and unlabelled, complicating both training and evaluation of machine learning models. Moreover, the multimodal aspect means that images can reinforce, contradict, or add nuance to accompanying text, creating emotional impacts not captured by text-only approaches. Consequently, the goal shifts from identifying an inherent or "built-in" emotion in the content to understanding how specific text–image combinations *make people feel*.

This thesis aims to address these challenges by harnessing state-of-the-art (SOTA) text and image models. Specifically, it explores zero-shot classification and weakly supervised learning, including fine-tuning with soft labels, to infer how climate change–related tweets (which include images) may elicit emotional responses in viewers—even in the absence of explicit emotional labels. By examining how visual content shapes these emotional reactions, this work seeks to enhance our understanding of viewers' emotional responses to climate change messages on social media.

1.2. Background and Motivation

This research builds upon the datasets introduced in *Towards Understanding Climate Change Perceptions: A Social Media Dataset* by Prasse et al. (2023), which provide valuable resources for exploring climate discourse. The authors present two key datasets:

1. **ClimateTV**: Comprising over 700,000 climate-related images from Twitter, collected between January 1, 2019 and December 31, 2019. The images carry labels derived from associated hashtags, providing a broad visual overview of climate discourse.

2. **ClimateCT**: A curated set of 1,000 climate-related Twitter images (January 1, 2019 – December 31, 2022), each manually annotated across five dimensions: (i) Animals, (ii) Climate Action, (iii) Consequences, (iv) Setting, and (v) Type. These annotations offer a more detailed look at the visual narratives in climate discussions.

While ClimateCT includes extensive annotations, neither dataset was designed to capture how these images might influence viewers' emotional states. This gap provides a fertile ground for experimenting with zero-shot and weakly supervised approaches, enabling an investigation into the emotional impact of climate-related posts without the need for manually curated emotion labels.

In the realm of text-based analysis, transformer models such as BERT (Devlin et al. 2018) have shown strong performance in encoding textual data for classification. More recent developments, including RoBERTa-based models (e.g., roberta-large-mnli) (Liu et al. 2019), demonstrate robust zero-shot capabilities, expanding the potential for emotion-focused analysis across diverse contexts and domains.

Conversely, analyzing emotional cues in images remains relatively underexplored. Convolutional neural network (CNN) architectures like ResNet (He et al. 2015) and VGGNet (Simonyan and Zisserman 2015) are commonly used for extracting image embeddings, which are then classified. More recently, vision transformers (ViTs) (Dosovitskiy et al. 2021) have sought to replicate the success of transformers in text-based tasks for image processing. CLIP (Radford et al. 2021), trained on large sets of image–text pairs, pushes this further by enabling classification based on natural language descriptions—making it particularly suited for zero-shot prediction of viewer emotion.

In multimodal emotion classification, Zhu et al. (2023) introduce the MUlti-Level SEmantic Reasoning network (MULSER), which performs fine-grained image–text emotion analysis. This aligns closely with our goal of exploring how combined textual and visual elements affect viewers' emotional responses on social media. While MULSER emphasises fine-grained emotion classification, our research extends this approach by employing zero-shot and weakly supervised methods to estimate emotional responses to climate change-related tweets containing images. By integrating visual and textual cues, we aim to deepen our understanding of how social media audiences react emotionally to climate-related content.

Because the ClimateTV dataset provides rich textual (tweets and replies) and visual data but does not include explicit emotional labels, it presents an opportunity to test zero-shot or weakly supervised methodologies. By leveraging CLIP and similar architectures—alongside zero-shot text models for generating soft labels—we can investigate how likely people are to respond emotionally to climate-related tweets, thereby illuminating the role that visual media plays in shaping reactions to climate change messaging on social media.

1.3. Problem Statement

Despite increasing interest in how social media users perceive climate change, most existing approaches rely on labelled data and focus predominantly on text-based, polarity-oriented sentiment (e.g., positive vs. negative). Multimodal approaches, though promising, often assume access to large-scale labelled data for both text and images—an assumption that is impractical in many real-world scenarios.

The ClimateTV dataset highlights these challenges. Although it presents a wealth of text, replies, and images reflecting diverse perspectives, it does not include labels that capture the emotional effect on viewers. Understanding how such content influences emotional responses is critical for assessing the public's engagement with climate change.

Additionally, current multimodal emotion analysis techniques often fail to account for the specific contexts and semantic richness of climate imagery, especially when combined with conversation threads (replies). This thesis aims to bridge that gap by addressing the following key questions:

1. How can SOTA text and image models generate meaningful emotional insights for a climate change dataset with no existing emotion annotations—specifically, how does the content make viewers feel?
2. How can we refine these models to better capture multimodal cues and climate-specific contexts—especially in the absence of large-scale manual annotations—so that we can more accurately reflect how people emotionally respond to such content?

To answer these questions, we explore zero-shot and weakly supervised learning techniques, followed by fine-tuning experiments using soft labels generated by the best-performing models.

1.4. Research Objectives and Approach

1.4.1. Primary Research Question

How can state-of-the-art text and image models enhance our understanding of the emotional impact *(on viewers) of climate-related content on social media?*

From this central question, we define the following objectives:

- Evaluate and compare SOTA zero-shot models that can infer emotional responses in a dataset without explicit emotion labels, focusing on how the content makes viewers feel.
- Implement and assess weakly supervised approaches for emotion detection that account for both textual and visual cues.

- Fine-tune text, image, and multimodal models using soft labels generated by SOTA zero-shot methods to improve classification performance in terms of viewers' emotional response.
- Conduct error analysis and propose refinements to better capture how specific climate-related visuals affect public emotional reactions.

By systematically addressing these objectives, this thesis aims to demonstrate how SOTA models—combined with innovative inference and fine-tuning strategies—can provide deeper insights into the visual and textual cues shaping the emotional reception of climate change–related content on social media.

1.5. Thesis Structure

The thesis is structured as follows:

- **Chapter 2 – Literature Review:** Provides an overview of prior research in text-based and image-based emotion analysis, highlighting gaps in zero-shot and weakly supervised methods.
- **Chapter 3 – Research Methodology:** Details the dataset, preprocessing pipeline, and experimental protocols.
- **Chapter 4 – Experimental Results:** Presents quantitative and qualitative findings, contrasting model performance under various conditions and includes detailed error analysis.
- **Chapter 5 – Discussion:** Examines failure modes and proposes enhancements for improved emotion detection.
- **Chapter 6 – Conclusion:** Summarises key takeaways.

By re-centering the analysis on emotional responses rather than traditional positive, negative, or neutral sentiment this thesis seeks to provide nuanced insights into how climate change-related posts make people feel, thereby offering a richer understanding of public engagement with climate discourse on social media.

2. Literature Review

Sentiment and emotion analysis are two closely related yet distinct fields within natural language processing (NLP) and computational linguistics that aim to understand and interpret human affective states from textual data. The distinction between sentiment and emotion lies in their granularity and scope. While sentiment is broader typically focusing on identifying and categorizing the polarity of expressed opinions, such as positive, negative, or neutral sentiments, emotions are more intricate and multifaceted, reflecting the complexity of human psychological experiences. A foundational study by Ekman (1992) identified six basic emotions—joy, anger, sadness, fear, surprise, and disgust—that are universally recognized across cultures. These emotions serve as a cornerstone for emotion analysis, providing a framework for understanding how affective states are expressed and perceived in text.

This chapter provides an overview of the existing research on sentiment and emotion analysis, covering approaches that leverage textual, visual, and multimodal data. Section 2.1 focuses on text-based sentiment and emotion analysis, exploring classical machine learning and deep learning models used for this task. Section 2.2 shifts the focus to vision-based emotion analysis, discussing techniques for extracting emotions from images. Section 2.3 delves into multimodal sentiment and emotion analysis, outlining the models that integrate text and vision and the fusion strategies used to combine modalities effectively. Given the challenges of obtaining high-quality labeled data, Section 2.4 discusses weakly supervised learning approaches in both text and vision domains. Finally, Section 2.5 summarizes key findings from the literature and highlights the gaps that motivate the research in this thesis.

2.1. Text-Based Sentiment and Emotion Analysis

Text-based sentiment and emotion analysis has been a cornerstone of natural language processing (NLP) research for several decades, driven by the need to understand and quantify human emotions and opinions expressed in textual data. This field has evolved significantly, with methodologies ranging from simple lexicon-based approaches to sophisticated deep learning architectures. The ultimate goal of these techniques is to automatically classify the sentiment (positive, negative, or neutral) or emotion (e.g., joy, anger, sadness) conveyed in a given text, which has applications in areas such as customer feedback analysis, social media monitoring, and mental health assessment.

Lexicon-Based Methods

One of the earliest and most straightforward approaches to sentiment and emotion analysis is the lexicon-based method. This technique relies on predefined dictionaries or lexicons, where each word is assigned a sentiment or emotion score. For example, the word *happy* might be associated with a positive sentiment score, while *sad* might carry a negative score. The overall sentiment or emotion of a sentence or document is then computed by aggregating the scores of the individual words, often using summation or weighted combination techniques.

Popular resources for lexicon-based sentiment analysis include **SentiWordNet** (Baccianella et al. 2010), which extends the WordNet lexical database by assigning sentiment scores to synsets (sets of synonyms), and domain-specific dictionaries tailored to particular industries or contexts. While these methods are computationally efficient and interpretable, they often require additional preprocessing steps to account for linguistic nuances such as **negation** (e.g., *not happy*), **intensifiers** (e.g., *very happy*), and **modifiers** (e.g., *slightly happy*). These factors can significantly influence the final sentiment or emotion classification, making lexicon-based approaches somewhat limited in their ability to capture complex contextual information.

Classical Machine Learning Approaches

With the advent of machine learning, researchers began to explore more data-driven methods for sentiment and emotion analysis. Classical machine learning algorithms such as **Naïve Bayes**, **Support Vector Machines (SVMs)**, and **Logistic Regression** became popular choices for this task. These models are typically trained on labelled datasets, where each text sample is annotated with its corresponding sentiment or emotion label.

A critical aspect of these approaches is **feature engineering**, where handcrafted features are extracted from the text to represent it in a way that is suitable for machine learning. Common features include **word n-grams** (sequences of n words), **part-of-speech (POS) tags**, and **syntactic or semantic cues** such as dependency relations or sentiment-bearing phrases. The performance of these models heavily depends on the quality and relevance of the engineered features, as they do not inherently learn contextual representations of words. Despite their limitations, these methods laid the groundwork for more advanced techniques by demonstrating the potential of data-driven approaches in NLP.

Deep Learning and Neural Networks

The introduction of **deep learning** marked a significant shift in sentiment and emotion analysis, moving away from manual feature engineering toward **automatically learned representations** of text. Early deep learning architectures for this task often employed **Recurrent Neural Networks (RNNs)** (Mikolov et al. 2010), which are designed to process sequential data by maintaining a hidden state that captures information from previous time steps. Variants such as **Long Short-Term Memory (LSTM)** networks

(Hochreiter and Schmidhuber 1997) and **Gated Recurrent Units (GRUs)** (Cho et al. 2014) addressed the **vanishing and exploding gradient problems** commonly encountered in standard RNNs. These architectures are particularly effective for sentiment and emotion analysis because they can model long-range dependencies in text, capturing the influence of earlier words on the sentiment or emotion expressed later in a sentence.

Bidirectional RNNs further enhanced this capability by processing text in both forward and backward directions, allowing the model to incorporate context from both preceding and succeeding words (Graves and Schmidhuber 2005). This bidirectional approach proved especially useful for tasks requiring a comprehensive understanding of sentence structure and meaning.

Convolutional Neural Networks (CNNs) for Text

In parallel with RNNs, **Convolutional Neural Networks (CNNs)** were also adapted for text-based sentiment and emotion analysis (Krizhevsky et al. 2012). Originally designed for image processing, CNNs were repurposed for NLP tasks by applying one-dimensional filters to **word embeddings** (dense vector representations of words). These filters extract local n-gram features, capturing patterns and relationships within small windows of text. The most salient features are then aggregated through **pooling layers** before being passed to a classification layer.

While CNNs have demonstrated strong performance in certain NLP tasks, they are less effective at modeling **long-range dependencies** in text. This limitation arises because CNNs are designed to capture local spatial relationships, making them less suitable for tasks that require a global understanding of context or relationships between distant parts of a sentence. As a result, CNNs have gradually been overshadowed by more advanced architectures that better address these challenges.

Attention Mechanisms and Transformers

A major breakthrough in sentiment and emotion analysis came with the introduction of **attention mechanisms** (Bahdanau et al. 2014), which enable models to selectively focus on the most relevant parts of a sentence rather than processing it sequentially. Attention-based architectures improved the interpretability of results by highlighting which words or phrases contributed most to the predicted sentiment or emotion. This innovation paved the way for **Transformers** (Vaswani et al. 2023), which revolutionized NLP by eliminating the need for recurrent processing altogether.

Transformers rely on a **self-attention mechanism** that allows the model to attend to different positions in a sequence, enabling it to learn **contextualized embeddings** efficiently. This architecture has become the foundation for many state-of-the-art models in NLP, including **BERT** (Devlin et al. 2018) and **GPT** (Radford and Narasimhan 2018). These **pretrained language models** leverage vast amounts of unlabelled text to learn general-purpose representations of language, which can then be fine-tuned for specific tasks such as sentiment or emotion classification.

Pretrained Language Models and Fine-Tuning

Pretrained models like BERT and GPT have dominated the field of sentiment and emotion analysis due to their ability to capture rich contextual information. These models are typically fine-tuned on task-specific datasets by appending a classification layer on top of their final hidden representations. **RoBERTa**, introduced in the paper titled *RoBERTa: A Robustly Optimized BERT Pretraining Approach* by (Liu et al. 2019), improves upon BERT by optimizing several key aspects of the pretraining process. It trains longer with larger batches and more data, addressing BERT's undertraining. The next sentence prediction (NSP) objective is removed, as it was found unnecessary, and dynamic masking is introduced, generating new masks for each sequence instead of using static ones. RoBERTa also trains on longer sequences to capture broader dependencies and uses a larger, more diverse dataset. These changes collectively enhance performance, enabling RoBERTa to achieve state-of-the-art results on several benchmarks.

(Nguyen et al. 2020) introduce BERTweet, the first large-scale pre-trained language model specifically designed for English Tweets, using the same architecture as BERT-base and trained on an 80GB corpus of 850 million Tweets following the RoBERTa pre-training procedure. BERTweet outperforms strong baselines like RoBERTa-base and XLM-R-base on three downstream Tweet NLP tasks—Part-of-speech tagging, Named-entity recognition, and text classification (sentiment analysis and irony detection)—achieving new state-of-the-art results, particularly in novel and emerging entity recognition and text classification.

Moreover variants such as **DistilBERT** (Sanh et al. 2020) and **T5** (Raffel et al. 2023) have further optimized these architectures by modifying training objectives and data settings, resulting in more efficient and effective models. The widespread adoption of transformer-based models has led to significant improvements in performance across diverse datasets and languages. However, challenges remain, particularly in capturing domain-specific nuances and handling multimodal data (e.g., combining text with images or audio). Ongoing research continues to explore advanced techniques such as **domain adaptation**, **transfer learning**, and **multimodal sentiment analysis** to address these limitations and further enhance the capabilities of text-based sentiment and emotion analysis systems.

2.2. Vision-Based Sentiment and Emotion Analysis

While text-based methods have historically dominated the field of sentiment and emotion analysis, vision-based approaches have emerged as a powerful alternative, particularly in applications where facial expressions, body language, and other visual cues are central to understanding emotional states. Vision-based sentiment analysis leverages images or video data to infer the emotional states of individuals, offering a complementary perspective to text-based methods. This approach is especially valuable in scenarios where textual data is sparse, ambiguous, or unavailable, such as in video surveillance, or human-computer interaction.

The ability to analyze visual data for sentiment and emotion has broad implications across multiple domains, including psychology, marketing, healthcare, and entertainment. For instance, in healthcare, vision-based emotion analysis can be used to monitor patients' emotional well-being (Sariyanidi et al. 2015). Despite its potential, vision-based sentiment analysis presents unique challenges, such as the need to account for variations in lighting, pose, occlusion, and cultural differences in emotional expression (Zhang et al. 2018).

Traditional Feature Extraction

Early approaches to vision-based sentiment analysis relied heavily on handcrafted feature extraction techniques, which aimed to capture low-level visual patterns indicative of emotional states. Two of the most widely used methods were the Scale-Invariant Feature Transform (SIFT) (Lowe 2004) and the Histogram of Oriented Gradients (HOG) (Dalal and Triggs 2005). These techniques extract features such as edges, corners, and gradients from images, which are then used as input to classical machine learning models like SVMs or KNNs.

While these methods are computationally efficient and interpretable, they have significant limitations. Handcrafted features often fail to capture high-level semantic information, such as the context or subtle nuances of emotional expressions (Sariyanidi et al. 2015). For instance, they may struggle to distinguish between a genuine smile and a forced one or to recognize complex emotions like confusion or ambivalence.

Convolutional Neural Networks (CNNs)

The advent of Convolutional Neural Networks (CNNs) marked a paradigm shift in vision-based sentiment analysis (LeCun et al. 2015). Unlike traditional methods, which rely on handcrafted features, CNNs automatically learn hierarchical representations of visual data through multiple layers of convolutional filters. This capability allows CNNs to capture both low-level features (e.g., edges and textures) and high-level semantic information (e.g., facial expressions and emotional context).

The breakthrough moment for CNNs in computer vision came with the introduction of **AlexNet** in 2012 (Krizhevsky et al. 2012), which demonstrated the power of deep convolutional layers in learning discriminative features from images. AlexNet's success was followed by a series of increasingly sophisticated architectures, such as **VGGNet** (Simonyan and Zisserman 2015), **Inception** ((Szegedy et al. 2015), and **ResNet** (He et al. 2015). ResNet, in particular, introduced residual connections to address the vanishing gradient problem, enabling the training of very deep networks with hundreds of layers. These advancements significantly improved the ability of CNNs to recognize emotions from facial expressions and other visual cues, achieving state-of-the-art performance on benchmark datasets.

One of the key advantages of CNNs is their ability to leverage pretrained models, such as those trained on large-scale datasets like ImageNet (Deng et al. 2009). By fine-tuning these models for sentiment analysis tasks, researchers can achieve high accuracy and generalization even with relatively small amounts of labeled data (Yosinski et al. 2014). For example, a CNN pretrained on ImageNet can be adapted to recognize facial expressions by retraining its final layers on a dataset of annotated facial images. This transfer learning approach has become a standard practice in vision-based sentiment analysis, significantly reducing the need for large annotated datasets.

Recent studies have explored modifications to CNN architectures to improve their performance in emotion recognition. For instance, Limami et al. (2024) advance contextual emotion detection in images by proposing two deep learning models—a Deep Convolutional Neural Network (DCNN) and a VGG19-based model—that integrate 26 discrete emotion categories with three continuous emotional dimensions (valence, arousal, dominance) to improve emotion recognition accuracy. By combining body features and contextual features, the models achieved a mean Average Precision (mAP) of 78.39% and 79.60%, respectively, outperforming previous methods and demonstrating the importance of context in interpreting emotions.

Al-Halah et al. (2019) introduce SmileyNet, a novel approach for visual sentiment analysis that leverages emoji-based image embeddings to address the limitations of small-scale sentiment datasets. By constructing a large-scale dataset of 4 million images from Twitter, annotated with associated emojis, the authors train a deep neural network to predict emojis, creating a compact and sentiment-aligned embedding. This embedding outperforms traditional object-based representations (e.g., ImageNet) in visual sentiment analysis and fine-grained emotion classification, achieving state-of-the-art results on benchmark datasets. Additionally, hybrid models combining Multi-Scale Dynamic 1D CNN and Gated Transformer have shown high accuracy in EEG-based emotion recognition, highlighting the effectiveness of combining spatial-spectral features with global dependencies (Cheng et al. 2024)

Despite their success, CNNs are not without limitations. They require substantial computational resources for training and inference, particularly for deep architectures like ResNet (He et al. 2015). Additionally, CNNs are primarily designed to process local spatial information, which can make it challenging to capture global relationships within an image. This limitation has spurred the development of alternative approaches, such as Vision Transformers (ViTs), which aim to address these shortcomings.

Vision Transformers (ViTs)

The latest advancements in vision-based sentiment analysis have been driven by **Vision Transformers (ViTs)** (Dosovitskiy et al. 2021) and their variants, such as **Swin Transformers** (Liu et al. 2021). Unlike CNNs, which process images through convolutional filters, ViTs treat images as sequences of patches and apply self-attention mechanisms to capture global relationships within the visual data. This approach enables a more comprehensive understanding of complex emotional expressions, particularly in scenarios where subtle cues are critical.

ViTs were initially inspired by the success of transformers in natural language processing (NLP), where self-attention mechanisms have proven highly effective at modeling long-range dependencies in text (Vaswani et al. 2023). By adapting this architecture to visual data, ViTs can capture both local and global features, making them particularly well-suited for tasks like emotion recognition, where context and fine-grained details are important (Dosovitskiy et al. 2021). For example, a ViT can analyze the relationship between different facial regions (e.g., eyes, mouth, and eyebrows) to infer emotions like surprise or disgust.

One of the most notable variants of ViTs is the Swin Transformer, which introduces a hierarchical feature extraction process. Swin Transformers divide an image into non-overlapping windows and apply self-attention within each window, reducing computational complexity while maintaining the ability to capture global relationships. This hybrid approach combines the strengths of ViTs and traditional CNNs, enabling state-of-the-art performance on various benchmarks.

ViTs have also paved way for multi-modal approaches such as **CLIP (Contrastive Language-Image Pretraining)** (Radford et al. 2021) have shown remarkable potential in tasks that involve vision and language integration. This capability is particularly relevant for emotion recognition tasks that rely on both visual cues and associated text, such as social media posts containing images and captions.

The success of ViTs and their variants, along with multi-modal models like CLIP, has opened up new possibilities for vision-based sentiment analysis. These models are particularly effective at handling complex scenarios, such as group emotion recognition or the analysis of subtle micro-expressions (Liu et al. 2021). Soni et al. (2024) apply a ViT-based model was applied to the FER-2013 dataset for emotion detection in human-computer interaction (HCI). The study emphasized meticulous preprocessing, data augmentation, and fine-tuning of the ViT model, achieving a testing accuracy of 70%. However, they also come with challenges, including high computational costs and the need for large amounts of training data. Future research is likely to focus on improving the efficiency and scalability of these models, as well as exploring their application in real-world settings.

2.3. Multimodal Sentiment and Emotion Analysis

Multimodal sentiment/emotion analysis is a rapidly evolving field focused on understanding human emotions by integrating data from multiple modalities, such as text, audio, and visual inputs. Because human communication naturally spans these different modalities, relying on a single source of information often results in incomplete or inaccurate emotion recognition. As a result, multimodal approaches are increasingly critical for achieving more robust and comprehensive emotion analysis.

2.3.1. Multimodal Models

Several multimodal models have been developed to address the complexities of sentiment and emotion analysis. This section highlights three notable examples: VisualBERT, MuAL, and CLIP.

VisualBERT. VisualBERT (Li et al. 2019) is a multimodal transformer model designed for vision-language tasks. It unifies visual and textual inputs within the same architecture, leveraging BERT's bidirectional attention mechanism to jointly process image regions and text tokens. VisualBERT has shown strong performance in tasks such as visual question answering and image captioning. However, it requires aligned image-text pairs during training, which can constrain its flexibility in handling unaligned multimodal data.

MuAL (Multimodal Sentiment Analysis with Cross-Modal Attention and Difference Loss). MuAL (Deng et al. 2024) is a recent approach that integrates cross-modal attention and a difference loss to enhance model robustness in multimodal sentiment analysis. By minimizing the gap between image and text representations, MuAL improves performance over traditional unimodal methods and demonstrates strong transfer learning capabilities. Notably, it outperforms baseline models even when its pre-trained parameters are frozen, making it particularly suitable for real-world applications where computational efficiency and generalization are crucial.

Multi-Level Semantic Reasoning Network (MULSER) for fine-grained multimodal emotion classification. This paper introduced by Zhu et al. (2023) addresses the limitation of traditional sentiment polarity analysis (e.g., positive/negative) by distinguishing nuanced emotions (e.g., happiness vs. love). MULSER leverages graph attention networks to model semantic relationships at multiple levels: for images, it constructs object-level, global-level, and joint regional-global graphs to capture interactions between local objects and contextual concepts; for text, it builds word-level graphs to enhance interdependencies between words. A cross-modal attention fusion module integrates the enriched visual and textual features, enabling complementary reasoning. Experiments demonstrate MULSER's superiority over state-of-the-art methods, achieving significant improvements in accuracy and F1 score. Ablation studies confirm the effectiveness of each component, emphasizing the importance of semantic reasoning and cross-modal interaction.

CLIP

CLIP (Contrastive Language–Image Pre-training) is a neural network model that aligns textual and visual modalities within a shared embedding space. This alignment enables various multimodal tasks without the need for task-specific labeled datasets. Departing from traditional supervised learning approaches that rely on domain-specific annotations, CLIP employs a zero-shot learning paradigm, allowing it to generalize across tasks by capturing the relationships between text and images without explicit fine-tuning (Radford et al. 2021).

CLIP is trained using a contrastive learning objective that maximizes the cosine similarity for matched image-text pairs while minimizing it for mismatched pairs. By doing so, CLIP learns representations that effectively capture semantic relationships across modalities. This capability allows the model to perform tasks such as image classification by comparing embedded textual labels (e.g., "a photo of a dog") with image embeddings in a shared latent space. Figure 2.3 illustrates the contrastive learning framework employed in CLIP.

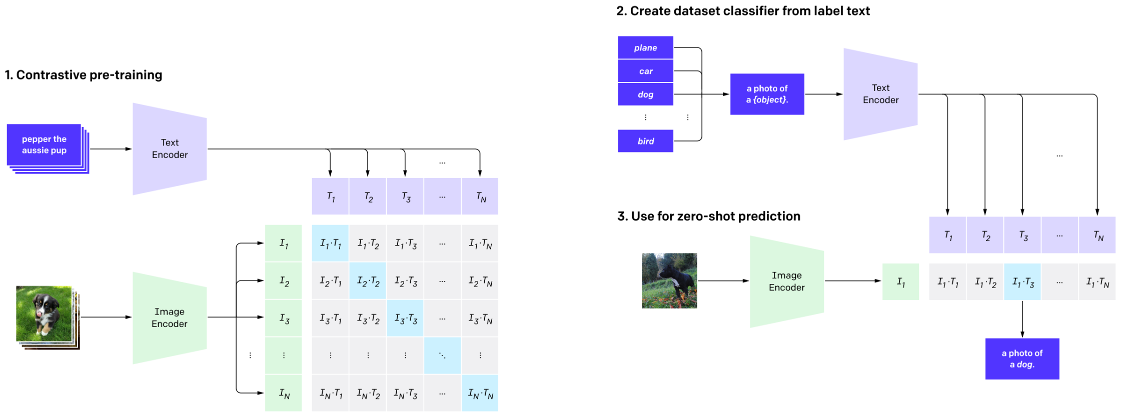

1. Contrastive pre-training
pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ... I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$
2. Create dataset classifier from label text
plane
car
dog
bird
a photo of a {object}.
Text Encoder
3. Use for zero-shot prediction
Image Encoder
T_1 T_2 T_3 ... T_N
I_1
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
a photo of a dog.

Figure 2.1.: Contrastive Pre-training

Figure 2.2.: Creation of dataset classifier and final prediction

Figure 2.3.: Approach to training CLIP and Inference (Radford et al. 2021)

CLIP was trained on approximately 400 million image-text pairs collected from the internet, leveraging the natural co-occurrence of images and their descriptions instead of manual annotations. This large and diverse dataset allows the model to learn broad visual-textual relationships, making it applicable to a variety of tasks. CLIP's architecture consists of two main components: an image encoder (commonly a Vision Transformer (Dosovitskiy et al. 2021)) and a text encoder (based on transformer architectures like GPT or BERT (Radford and Narasimhan 2018; Devlin et al. 2018)). Both encoders produce high-dimensional embeddings that are projected into a shared latent space, where contrastive learning aligns matched image-text pairs and separates mismatched ones.

To further improve generalization, CLIP leverages data augmentation techniques—such as resizing, cropping, and color adjustments—and large batch sizes that provide a diverse set of negative examples during contrastive learning. One of CLIP's most significant strengths is its zero-shot learning paradigm, which allows the model to interpret natural language as a "programming interface." In this setup, text prompts are embedded into the shared latent space and compared to image embeddings. Tasks like image classification, object detection, and text-based image retrieval can thus be performed without any task-specific fine-tuning (Figure 2.2).

CLIP's foundational principles have informed other multimodal models, such as LXMERT (Tan and Bansal 2019), which employs cross-attention mechanisms for visual question answering and image captioning. These advances extend CLIP's core ideas to a variety of multimodal applications, including affective computing and social media analysis.

2.3.2. Fusion strategies

Multimodal data fusion has emerged as a critical area of research in machine learning, driven by the increasing availability of diverse data sources such as text, images, audio, and video. The integration of these modalities offers opportunities to enhance model performance by leveraging complementary information. However, the challenges of aligning and fusing heterogeneous data types necessitate sophisticated techniques. This sub-section summarizes insights from two key papers: *Effective Techniques for Multimodal Data Fusion: A Comparative Analysis* by (Pawłowski et al. 2023) and *Multimodal Alignment and Fusion: A Survey* by (Li and Tang 2024). These works provide a comprehensive overview of fusion techniques, their applications, and the challenges associated with multimodal integration.

Fusion Techniques and Their Applications

Pawlowski et al. focus on three primary fusion techniques—**late fusion**, **early fusion**, and **sketch representation**—and evaluate their effectiveness in classification tasks. Their findings highlight the dominance of **late fusion** in scenarios where one modality is dominant or unimodal models already perform well. For instance, in the Amazon Reviews dataset, late fusion achieved the highest accuracy (0.969) by combining textual and visual modalities. This technique processes each modality independently and combines their outputs at the decision level, making it robust to modality-specific noise and variability.

In contrast, **early fusion** integrates modalities at the input level by concatenating their embeddings. While this approach is beneficial when modalities are interdependent, it often underperforms compared to late fusion, as seen in the MovieLens datasets. Early fusion's reliance on combined embeddings can lead to information loss or redundancy, particularly when modalities are not equally informative.

The **sketch representation** technique, which transforms modalities into a common space using hash functions, offers a memory-efficient alternative. Although it underperformed in classification tasks, its scalability and computational efficiency make it suitable for large-scale applications like recommendation systems. Pawłowski et al. emphasize that the choice of fusion technique should be guided by task requirements, modality impact, and memory constraints.

Li and Tang provide a comprehensive overview of multimodal fusion techniques by categorizing them into four main types: encoder-decoder fusion, kernel-based fusion, graphical fusion, and attention-based fusion. Each of these methods addresses the challenge of integrating information from different modalities, such as text, images, and audio, to improve performance in various machine learning tasks.

Encoder-Decoder Fusion: This approach involves using separate encoders to process different modalities and then combining their outputs through a decoder. The encoders transform raw input data (e.g., images, text) into a shared latent space, where the decoder generates a unified representation. This method is particularly useful in tasks like image captioning and video summarization, where the goal is to generate a coherent output from multiple input modalities Baltrušaitis et al. (2019).

Attention-Based Fusion: Attention mechanisms have gained significant attention in recent years due to their ability to dynamically weight the importance of features across modalities. Li and Tang emphasize the role of attention mechanisms in capturing long-range dependencies and inter-modal interactions. Advanced models like ALBEF (Align before Fuse) and BLIP (Bootstrapped Language-Image Pretraining) further refine this approach by aligning modalities before fusion and leveraging large-scale pretraining to improve performance (Li et al. 2021, 2022).

These attention-based methods excel in tasks like social media analysis and emotion recognition, where understanding the nuanced interactions between text, images, and other modalities is crucial. For example, in social media analysis, attention mechanisms can help identify the most relevant posts or images that contribute to a trending topic, while in emotion recognition, they can dynamically weight the importance of facial expressions, voice tone, and textual content to accurately infer emotional states (Poria et al. 2017).

Alignment Challenges and Conclusion

A critical aspect of multimodal fusion is alignment, which ensures that data from different modalities are synchronized and meaningfully combined. Li and Tang distinguish between explicit alignment and implicit alignment. Explicit methods, such as Dynamic Time Warping (DTW) and Canonical Correlation Analysis (CCA), directly measure inter-modal relationships and are particularly useful for tasks requiring precise temporal or spatial synchronization, such as video-audio alignment. In contrast, implicit alignment techniques learn a shared latent space through neural networks or graphical models, employing methods like attention mechanisms, Generative Adversarial Networks (GANs), and Variational Autoencoders (VAEs). Attention-based models, particularly transformers, have significantly enhanced multimodal learning by enabling the adaptive integration of modalities.

Despite these advancements, several challenges persist in multimodal fusion, including modal feature misalignment, computational inefficiency, and data quality issues. Pawłowski et al. highlight the importance of selecting meaningful modalities, citing the limited contribution of visual data in the MovieLens datasets, and stress the need for standardized benchmarks akin to GLUE in NLP. Li and Tang echo these concerns, advocating for scalable and adaptive frameworks to handle large-scale, heterogeneous datasets. They emphasize the potential of graphical fusion methods, such as Heterogeneous Graph-based Multimodal Fusion (HGMF), to model incomplete and noisy data effectively.

2.4. Weakly Supervised Learning

Weakly supervised learning (WSL) has emerged as a critical area of research in machine learning, addressing scenarios where labelled data is scarce, noisy, or incomplete. This section reviews recent advancements in WSL, focusing on its applications in both computer vision and natural language processing (NLP). The discussion is organized into two main subsections: *Weakly Supervised Learning in Text* and *Weakly Supervised Learning in Vision*, followed by a critical analysis of the limitations and future directions of WSL.

2.4.1. Weakly Supervised Learning in Text

Weakly supervised learning has also been extensively applied to NLP tasks, particularly in scenarios where annotated data is scarce or expensive to obtain. Recent work has focused on generating supervision signals from weak sources, such as language models or heuristic rules.

Song et al. (2022) in the paper *Learning from Noisy Labels with Deep Neural Networks: A Survey*, provide a comprehensive review of techniques for training deep neural networks (DNNs) in the presence of noisy labels. The main techniques discussed in the paper are categorized into five groups:

1. **Robust Architecture**
   - **Noise Adaptation Layer:** Adds a layer to model the noise transition matrix, which helps in learning the label transition behavior. This approach aims to mimic the label transition process by estimating the probability of label corruption.
   - **Dedicated Architecture:** Designs specialized architectures to handle more complex noise types, such as instance-dependent noise. These architectures often involve multiple networks or human-assisted constraints to improve robustness.

2. **Robust Regularization**
   - **Explicit Regularization:** Modifies the expected training loss to prevent overfitting. Techniques in this category often involve bilevel optimization, pre-training, or gradient clipping to control overfitting to noisy labels.
   - **Implicit Regularization:** Introduces stochasticity to improve generalization. Methods like adversarial training, label smoothing, and mixup are used to encourage the model to learn more robust representations.

3. **Robust Loss Function**
   - **Noise-Tolerant Loss Functions:** Modifies loss functions to be robust to label noise. These loss functions are designed to minimize the impact of noisy labels by ensuring that the loss remains stable even when labels are corrupted. Examples include mean absolute error (MAE) variants, generalized cross-entropy, and symmetric cross-entropy.

4. **Loss Adjustment**
   - **Loss Correction:** Adjusts the loss based on the estimated noise transition matrix. This involves correcting the loss values during forward or backward propagation to account for label noise.
   - **Loss Reweighting:** Assigns different weights to examples based on their likelihood of being correctly labelled. This approach reduces the influence of potentially noisy examples during training.
   - **Label Refurbishment:** Refurbishes noisy labels by combining them with model predictions. This technique dynamically updates the labels during training to reduce the impact of incorrect annotations.
   - **Meta Learning:** Automates the process of loss adjustment using meta-learning techniques. These methods learn to reweight examples or adjust labels based on a small clean validation set.

5. Sample Selection

- **Multi-network Learning:** Uses multiple networks to identify clean examples. By leveraging disagreements between networks or using a mentor-student approach, these methods filter out noisy examples.
- **Multi-round Learning:** Iteratively refines the set of clean examples over multiple training rounds. This approach gradually improves the quality of the selected examples by repeatedly training and filtering.
- **Hybrid Approach:** Combines sample selection with other techniques like semi-supervised learning. These methods treat selected examples as clean labelled data and the remaining examples as unlabelled, applying semi-supervised learning to improve robustness.

Additional Topics

- **Noise Rate Estimation:** Techniques for estimating the noise rate, including using the noise transition matrix, Gaussian Mixture Model (GMM), and cross-validation.
- **Experimental Design:** Discussion of publicly available datasets and evaluation metrics used to validate robust training methods.
- **Future Research Directions:** Identifies areas for future research, such as instance-dependent label noise, multi-label data with label noise, class imbalance data with label noise, robust and fair training, connection with input perturbation, and efficient learning pipelines.

The paper provides a detailed comparison of these methods based on six properties: flexibility, no pre-training, full exploration, no supervision, heavy noise, and complex noise. It also discusses the challenges and future directions in the field of learning from noisy labels.

Chen et al. (2022) propose a framework called *Multiple Weak Supervision (MWS)* for short text classification, addressing challenges such as insufficient labelled data, data sparsity, and imbalanced classification. MWS leverages multiple weak supervision sources, including keyword matching, regular expressions, and distant supervision clustering, to automatically label unlabelled data. The framework generates probabilistic labels through a conditional independent model, which helps mitigate class imbalance. Evaluated on public, synthetic, and real-world datasets, MWS demonstrates significant improvements in recall and F1-scores without compromising precision. This work highlights the potential of combining multiple weak supervision sources to address the challenges of short text classification.

Zeng et al. (2022) introduce a novel approach for weakly supervised text classification that leverages a masked language model (MLM) to generate supervision signals. By appending the sentence “This article is talking about [MASK]” to documents, the MLM’s predictions for the [MASK] token are used as weak supervision signals. These generated words are then used to train a latent variable model called *WDDC (Word Distribution and Document Classifier)*, which learns a word distribution over predefined categories and a document classifier without requiring annotated data. Evaluated on datasets like AGNews, 20Newsgroups, and UCINews, the method outperforms existing weakly supervised baselines by 2%, 4%, and 3%, respectively. This work demonstrates the potential of using MLMs to generate supervision signals for text classification tasks in low-resource settings.

Gera et al. (2022) in the paper *”Zero-Shot Text Classification with Self-Training”* address the challenge of improving zero-shot text classification performance using self-training. Zero-shot classification, where models classify text without task-specific labelled data, often underperforms compared to supervised models. The authors propose a self-training approach that fine-tunes zero-shot classifiers on their most confident predictions, leveraging only class names and an unlabelled dataset.

Key Contributions and Findings

- **Self-Training for Zero-Shot Models:** The authors adapt self-training, traditionally used in semi-supervised learning, to improve general-purpose zero-shot models. This involves generating pseudo-labels from the model’s confident predictions and iteratively fine-tuning the model.
- **Entailment-Based Classification:** The study focuses on Natural Language Inference (NLI)-based models, which map text classification tasks to textual entailment problems. The authors hypothesize that self-training helps these models better understand class name interactions and the specific entailment sub-types relevant to the target task.
- **Experimental Setup:** The method is evaluated on eight diverse text classification datasets. The authors use three off-the-shelf NLI models (RoBERTa, DeBERTa, and BART) and demonstrate significant performance improvements across all datasets after self-training.
- **Token Masking:** To enhance the informativeness of pseudo-labelled examples, the authors introduce a token masking heuristic that masks tokens most similar to the class name, forcing the model to rely on other contextual cues.
- **Cross-Task Effects:** The study explores the impact of self-training on one task for performance on another. Results show that self-training on related tasks (e.g., within the same domain) can be beneficial, while unrelated tasks (e.g., sentiment vs. emotion classification) may degrade performance.

- **Practical Implications:** The approach requires only a modest amount of unlabelled data (up to 10K examples) and does not need domain expertise, making it accessible for practitioners.

The paper concludes that self-training is a valuable tool for adapting general-purpose zero-shot models to specific tasks, offering significant performance gains with minimal effort. This work opens avenues for further research into combining self-training with other zero-shot learning paradigms and exploring its applicability to different types of NLP tasks.

2.4.2. Weakly Supervised Learning in Vision

Weakly supervised learning has been widely adopted in computer vision tasks, particularly in scenarios where obtaining large-scale, high-quality labelled datasets is prohibitively expensive. Several approaches have been proposed to leverage noisy or incomplete labels effectively.

Hu et al. (2019) propose a novel framework for weakly supervised image classification in the presence of noisy labels. Their method, which consists of a clean net and a residual net, leverages both clean and noisy labelled data to improve classification performance. The clean net learns a mapping from the feature space to the clean label space, while the residual net models the residual mapping between clean and noisy labels, acting as a regularization term to prevent overfitting. Evaluated on multi-label (OpenImage (Krasin et al. 2016), MS COCO (Lin et al. 2015)) and single-label (Clothing1M (Xiao et al. 2015)) datasets, the approach demonstrates significant improvements in mean average precision (mAP) and top-1 accuracy. This work highlights the importance of effectively utilizing noisy data in weakly supervised learning and provides a robust framework for handling label noise in practical applications.

Mahajan et al. (2018) explore the limits of weakly supervised pretraining by leveraging billions of Instagram images labelled with hashtags. Their study shows that models pretrained on such large-scale, weakly supervised datasets outperform those pretrained on traditional datasets like ImageNet in transfer learning tasks, including image classification and object detection. Key findings include the robustness of models to label noise, the importance of aligning source and target label spaces, and the potential of "hashtag engineering" to improve transfer learning results. This work underscores the value of leveraging naturally annotated, large-scale datasets for pretraining deep learning models.

Xie et al. (2020) introduce *Noisy Student Training*, a semi-supervised learning method that enhances model performance by leveraging unlabelled data. The approach involves training a teacher model on labelled data, generating pseudo-labels for unlabelled data, and then training a larger or equal-sized student model on both labelled and pseudo-labelled data while injecting noise (e.g., dropout, stochastic depth, and data augmentation). This method achieves state-of-the-art results on ImageNet (Deng et al. 2009), with an 88.4% top-1 accuracy, and demonstrates significant improvements in robustness on challenging datasets like ImageNet-A (Hendrycks et al. 2021), ImageNet-C, and ImageNet-P (Hendrycks and Dietterich 2019). The study highlights the effectiveness of combining weakly supervised learning with semi-supervised techniques to improve both accuracy and robustness.

Limitations

Despite its promise, weakly supervised learning (WSL) has notable limitations. Zhu et al. (2023) critically assess the effectiveness of WSL approaches, arguing that their benefits are often overestimated. Their experiments on eight NLP datasets reveal that fine-tuning models on even minimal clean validation data (e.g., five samples per class) often outperforms sophisticated WSL methods. Moreover, WSL fails to improve over weak labels without clean validation samples, and its advantages diminish when clean data is used for training instead of validation. The authors recommend that future research focus on fully leveraging available clean data and consider simple yet effective baselines, such as fine-tuning on weak labels followed by fine-tuning on clean samples ($FTw+CFT$).

2.5. Summary

This chapter reviewed existing work on sentiment and emotion analysis, emphasizing text-based, vision-based, and multimodal approaches. Text-based methods primarily rely on transformer-based architectures such as BERT and GPT, which have demonstrated state-of-the-art performance in sentiment and emotion classification tasks. Vision-based approaches leverage convolutional neural networks (CNNs) and vision transformers to extract emotion-relevant features from images, often using facial expressions or scene context.

Multimodal sentiment and emotion analysis integrates textual and visual data, improving predictive performance by capturing complementary cues from both modalities. Fusion strategies, such as early, late, and hybrid fusion, play a crucial role in effectively combining different modalities. Despite these advancements, one of the major challenges remains the scarcity of labelled data, which limits the generalization of supervised models. To address this, weakly supervised learning techniques have been explored, particularly in text and vision domains, leveraging noisy or incomplete labels to train models effectively.

The literature highlights the strengths and limitations of various approaches, underscoring the need for improved multimodal models and better weakly supervised learning strategies. These insights inform the research direction of this thesis, which aims to enhance emotion analysis through advanced multimodal techniques and more effective label-learning methods.

3. Research Methodology

3.1. Overview

Building on the dataset introduced in Chapter 1, this chapter details the methodology employed to investigate climate change perceptions through social media data. The study uses a multimodal dataset of tweets, their replies, and associated images, focusing on two critical phases: **zero-shot experimentation** and **fine-tuning with soft labels**. The methodology is structured as follows:

1. **Data Description and Preprocessing**: We further discuss how the dataset originally introduced in Chapter 1 has been filtered, cleaned, and prepared for analysis. This includes language detection, criteria for tweet selection, and other preprocessing steps to ensure high-quality input for downstream tasks.

2. **Metrics**: We define evaluation metrics tailored to both zero-shot and fine-tuning paradigms.

3. **Zero-Shot Experimentation**: We benchmark pre-trained language and vision models to select the best-performing architecture for zero-shot inference.

4. **Weakly Supervised Learning**: We analyse weakly supervised approaches for emotion classification.

5. **Fine-Tuning with Soft Labels**: The selected models are fine-tuned using soft labels generated using the best zero-shot text model.

6. **Experimental Setup**: A comprehensive evaluation of model performance across experimental setups is conducted, with ablation studies to validate design choices.

This structured approach ensures reproducibility while addressing challenges inherent to social media data, such as linguistic diversity, noise, and the subjective nature of climate change discourse.

3.2. Data Description and Preprocessing

3.2.1. Dataset Source

As noted in Chapter 1, the dataset is derived from the study *Towards Understanding Climate Change Perceptions: A Social Media Dataset* (Prasse et al. 2023), comprising tweets, replies, and images posted on Twitter (X) in 2019. To capture temporal variations in climate change discourse, this work focuses on the months of **February and August 2019**.

3.2.2. Data Filtering

1. **Temporal Filtering**: Tweets outside the months of February and August 2019 were excluded.

2. **Language Filtering**:

   - **Primary Tweets**: The `papluca/xlm-roberta-base-language-detection` model which is a fine-tuned variant of the XLM-RoBERTa (Conneau et al. 2019) classified tweet languages. Only English tweets were retained.

   - **Replies**: Replies were pre-translated in the dataset to English in the original dataset, ensuring linguistic consistency.

3.2.3. Preprocessing Pipeline

1. **Text Cleaning**:

   - Removed URLs, hashtags, and user mentions using regex patterns.

   - Retained emojis and punctuation to preserve sentiment cues.

2. **Multimodal Alignment**:

   - Paired tweets with their corresponding replies and images using tweet IDs.

   - Removed orphaned entries (e.g., tweets without replies or images).

3.2.4. Final Dataset Statistics

Metric	February 2019	August 2019
Total Tweets	1774	7,769
English Tweets	1,673	7,345
Replies	12,616	51,147
Valid Images	1,673	7,345

Table 3.1.: Dataset Statistics for February and August 2019.

This preprocessing ensures a focused, high-quality corpus for analyzing climate change perceptions while addressing linguistic and temporal complexities.

3.3. Evaluation Metrics

This research uses two sets of metrics: those that gauge ranking performance in zero-shot scenarios and those that measure distributional accuracy when models are fine-tuned with soft labels.

3.3.1. Zero-Shot Evaluation Metrics

In the zero-shot setting, the model predicts labels without task-specific training. The following metrics emphasize ranking quality and semantic relevance:

Exact Match (EM) Accuracy Measures the percentage of predictions for which the top-ranked label matches the ground truth. This provides a strict indicator of precision, especially for unambiguous classes.

Top-3 Accuracy Calculates the fraction of instances where the correct label appears among the top three predictions. This reflects real-world recommendation scenarios where multiple plausible labels can be acceptable.

Ranked Score (RS) Assigns a weighted score to each prediction, granting higher credit to correct labels placed earlier in the ranking. A correct prediction at rank r earns $\frac{1}{\log_2(r+1)}$, rewarding models that correctly prioritize relevant classes.

Normalized Discounted Cumulative Gain (NDCG@3) Compares the predicted ranking with the ideal ranking, truncated to the top three positions. This standard information retrieval metric highlights the importance of correctly ordering the most relevant labels.

$$\text{NDCG@3} = \frac{\sum_{i=1}^{3} \frac{\text{rel}_i}{\log_2(i+1)}}{\sum_{i=1}^{3} \frac{\text{rel}_i^{\text{ideal}}}{\log_2(i+1)}} \tag{3.1}$$

where rel_i represents the relevance score of the item at rank i, and $\text{rel}_i^{\text{ideal}}$ denotes the relevance score in the ideal ranking (i.e., sorted in decreasing order of relevance).

3.3.2. Fine-Tuning Evaluation Metrics

For the fine-tuning phase, we obtain soft labels (probabilistic annotations) from weak supervision. Accordingly, these metrics assess how well predicted distributions align with the target distributions:

Cosine Similarity Measures the cosine of the angle between predicted and target probability vectors. This captures directional alignment and is particularly useful for high-dimensional or multi-label tasks. It is computed as:

$$\text{Cosine Similarity} = \frac{\sum_i P(i)Q(i)}{\sqrt{\sum_i P(i)^2}\sqrt{\sum_i Q(i)^2}} \tag{3.2}$$

Kullback–Leibler Divergence (KLDiv) Quantifies how one probability distribution Q diverges from a true distribution P:

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{3.3}$$

It penalizes overconfident misclassifications, which is critical when working with noisy social media labels.

Mean Squared Error (MSE) Computes the average squared difference between predicted probabilities $Q(i)$ and target probabilities $P(i)$:

$$\mathrm{MSE} = \frac{1}{N} \sum_{i=1}^{N} \big(P(i) - Q(i)\big)^2 \tag{3.4}$$

MSE is sensitive to large deviations and helps maintain proper probability calibration, reducing the risk of assigning very high probabilities to rare classes.

Ranking-score Finally, to identify the best-performing experiments, we aggregated the mean values of metrics described above per experiment across both datasets and seeds. To standardize the evaluation, we normalized the metrics by preserving Cosine Similarity and inverting MSE and KL Divergence, ensuring higher values indicate better performance. A final ranking score was computed as the sum of these normalized metrics, and experiments were ranked accordingly. This approach provides a robust comparison, enabling the selection of the most effective configurations for further analysis.

$$\text{Ranking-score} = \text{Cosine Similarity} + (1 - \text{MSE}) + (1 - \text{KLDiv}) \tag{3.5}$$

3.4. Zero-Shot Experimentation

This section outlines our approach to selecting and evaluating zero-shot models for emotion classification in climate-change discourse. Our goal was to identify a pre-trained language model capable of handling informal, often noisy social media text when classifying emotions.

3.4.1. Model Selection and Rationale

We benchmarked five pre-trained models chosen for either their focus on social media content or their established zero-shot capabilities:

1. `cardiffnlp/twitter-roberta-base-emotion-latest` (Antypas et al. 2023)
    - Built on a **RoBERTa-base** architecture, fine-tuned for emotion detection on Twitter.

   - Part of the **SuperTweetEval** benchmark, addressing multilabel scenarios (tweets can have multiple emotions).
   - Trained on 154M tweets (up to December 2022) and fine-tuned on the **TweetEmotion** dataset.

2. `cardiffnlp/twitter-roberta-large-emotion-latest` (Antypas et al. 2023)
   - A **RoBERTa-large** variant of the above model, offering greater capacity for contextual reasoning.
3. `facebook/bart-large-mnli` (Lewis et al. 2019)
   - A BART model trained for Natural Language Inference (NLI), employed here for zero-shot classification.
   - Uses an entailment-based approach (Yin et al. 2019), framing candidate labels (e.g., *joy*) as hypotheses (e.g., "This text expresses joy") and inferring probabilities from entailment scores.
4. `MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7` (Laurer et al. 2022)
   - A multilingual DeBERTa model trained on cross-lingual NLI tasks, enabling zero-shot usage across different languages.
5. `MoritzLaurer/deberta-v3-large-zeroshot-v2.0` (Laurer et al. 2023)
   - A high-performing DeBERTa model fine-tuned to provide robust zero-shot classification.

Why These Models?

- The CardiffNLP models were pre-trained for **social media emotion detection** on Twitter which makes then highly suitable for our task.
- BART and DeBERTa-based models provide **general-purpose zero-shot capabilities**, making them strong baselines for cross-domain generalization.

3.4.2. Experimental Setup

Evaluation Dataset

We sampled **99 English replies** from our corpus, manually assigning a single ground-truth emotion label to each. The labels followed Ekman's six basic emotions—*anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise*—and their distribution is shown in Table 3.2.

Procedure

We evaluated each zero-shot model in two ways:

Manual Label	Count
Anger	33
Joy	16
Disgust	15
Fear	13
Sadness	12
Surprise	10

Table 3.2.: Distribution of Manual Emotion Labels

1) Primary Evaluation (Standard Metrics). All 99 samples were fed to each model, which returned a single emotion label per sample. We then computed the metrics described in Section 3.3.1.

2) Secondary Evaluation (Confidence Filtering). We additionally filtered out low-confidence predictions by retaining only those with a confidence score above 0.9. This second analysis measured (a) the proportion of samples meeting the threshold and (b) the accuracy of high-confidence predictions.

3.5. Weakly Supervised Learning

3.5.1. Zero-shot-classification-boost-with-self-training

As outlined in Section 2.4, (Gera et al. 2022) proposed a self-training framework to enhance zero-shot text classification by iteratively fine-tuning models on their own high-confidence predictions. Their approach addresses the scarcity of labelled data by leveraging unlabeled corpora and class names alone, aligning with our goal of ”classifying tweets into distinct emotion categories without task-specific labels”.

Reproduction of Gera et al.'s Methodology

To validate the reproducibility of the original study, we reimplemented their workflow as follows:

Model and Dataset Selection

- We used the same off-the-shelf NLI models (RoBERTa-large, DeBERTa-v3) and datasets *(AG's news (Zhang et al. 2015) and ISEAR (Shao et al. 2015))* as described in the original work.
- Weak supervision signals were derived solely from class names and unlabeled text, mirroring the zero-shot setup.

Self-Training Protocol

- **Pseudo-Label Generation:** For each dataset, we generated pseudo-labels by selecting the model's most confident predictions (threshold: $\tau = 0.9$, as in the original study).
- **Token Masking:** We implemented the token masking heuristic to mask tokens with high semantic similarity to class names (using cosine similarity over SBERT embeddings).
- **Fine-Tuning:** Models were iteratively fine-tuned on pseudo-labeled batches (batch size: 8) for $K = 2$ iterations, retaining the original learning rate (2×10^{-5}) and optimizer (AdamW).

We adapted the framework to our task of *classifying tweets into distinct emotion categories.*

Adaptations to the Self-Training Pipeline

- **Class Name Prompt Engineering:** To improve pseudo-label quality, we reformulated class names into natural language hypotheses (e.g., "The emotion in this text is joy" instead of "joy").
- **Cross-Task Sampling:** Since our task lacks related labelled datasets, we limited self-training to in-domain pseudo-labels, avoiding the cross-task degradation noted in (Gera et al. 2022).

3.5.2. Loss Reweighting

To mitigate label noise from weak supervision, we implemented a confidence-based loss reweighting strategy described in sub-section 2.4.1. This approach adjusts the influence of each training example based on confidence scores derived from the initial label generation process.

Implementation

This method consists of three key components:

- **Confidence-Weighted Loss Function**: The standard cross-entropy loss was modified to incorporate confidence scores as multiplicative weights. This ensured that higher-confidence samples contributed more to the learning process, while lower-confidence samples had a reduced impact.
- **Data Integration Pipeline**: The training dataset combined weak labels, raw text, and confidence scores, represented as:

$$\mathcal{D} = \{(\mathbf{x}_i, \tilde{y}_i, c_i)\}_{i=1}^{N} \tag{3.6}$$

where $\tilde{y}_i$ denotes weak labels, and $c_i \in [0, 1]$ represents the assigned confidence score.

- **Adaptive Training Protocol**: We fine-tuned BERT-base using a structured approach:
  - Applied class-balanced sampling when enabled
  - Used the AdamW optimizer with a learning rate of 2×10^{-5}
  - Implemented early stopping based on validation performance

Theoretical Basis

This method aligns with a noise-robust learning objective:

$$\min_{\theta} \sum_{i=1}^{N} c_i \cdot \ell(f_\theta(\mathbf{x}_i), \tilde{y}_i) \tag{3.7}$$

where ℓ is the cross-entropy loss, and c_i acts as a weighting factor. This formulation prioritizes high-confidence samples while minimizing the influence of potentially incorrect labels.

This strategy aimed to improve learning stability by focusing on reliable samples and reducing the effect of label noise in weakly supervised training.

3.6. Fine-Tuning with Soft Labels

This section presents our methodology for fine-tuning unimodal (text-only, image-only) and multimodal (text & image) models using soft labels. The **CardiffNLP RoBERTa-Large** model exhibited the best performance in generating reliable emotion probability distributions under zero-shot conditions. We leveraged this optimal model to produce inference distributions across all replies linked to each parent tweet. For each original tweet, we then aggregated these reply-level probability distributions into a single soft-label signal by summation and normalization.

The resultant aggregated distributions serve as target supervisory signals for fine-tuning, where model parameters are updated by minimizing the Kullback–Leibler (KL) divergence between model outputs and these soft label distributions. This approach facilitates knowledge transfer from the robust zero-shot model while preserving the nuanced emotional gradients captured in the crowd-sourced replies. Figure 3.3 illustrates the overall framework, including encoder components for each modality and the projection heads that map representations to the soft-label space. Training details regarding objective functions and implementation can be found in Section 3.7.

3.6.1. Label Mapping to Ekman's Six Emotions

Since the CardiffNLP model originally yielded predictions for 11 emotion categories (e.g., *love*, *trust*, *optimism*, *anticipation*), we consolidated these into Ekman's six basic emotions (*anger, disgust, fear, joy, sadness, surprise*) to maintain consistency and interpretability. Table 3.3 summarizes the mapping scheme, whereby categories such as *love*, *trust*, and *optimism* are merged into *joy*, and *anticipation* is subsumed under *surprise*. After merging, confidence scores for combined classes are normalized so that each distribution sums to 1.

Primary Emotion (Ekman)	**Mapped Emotions (Original-11)**
Anger	Anger
Disgust	Disgust
Fear	Fear
Joy	Joy, Love, Optimism, Trust
Sadness	Sadness
Surprise	Anticipation, Surprise

Table 3.3.: Mapping of Emotions into Ekman Categories

Rationale:

1. **Theoretical Robustness:** Ekman's framework is widely recognized and reduces ambiguity in noisy, multicultural social media contexts.

2. **Practical Efficiency:** Merging labels helps mitigate class imbalance (e.g., combining *love*, *optimism*, and *trust* under *joy*) and simplifies downstream training.

3. **Domain Relevance:** In climate discourse, broad labels like *anger* (e.g., policy opposition) or *surprise* (e.g., reactions to disasters) provide actionable insights more readily than finer-grained categories.

4. **Methodological Consistency:** Using a single taxonomy across different models enables uniform evaluation and aligns with prior work that employs Ekman's classification.

By consolidating all predictions under these six universal emotion categories, we ensure that both unimodal and multimodal models are trained on a well-defined and consistent label space.

3.6.2. Architectural Overview

In both cases i.e. Unimodal (Text or Image) and Multimodal, we attach a standard MLP as described below to the encoder output(s) for classification, unless specified otherwise.

Projection Head (MLP)

- **Depth:** 2 layers.
- **Hidden Dimension:** 512.
- **Activation:** ReLU.
- **Output Layer:** Produces 6 class probabilities (soft labels).

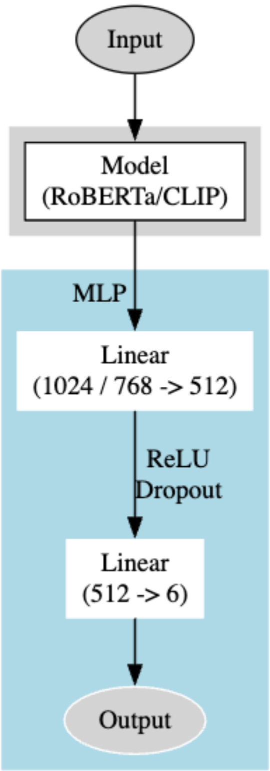


Figure 3.1.: Single Modality Architecture

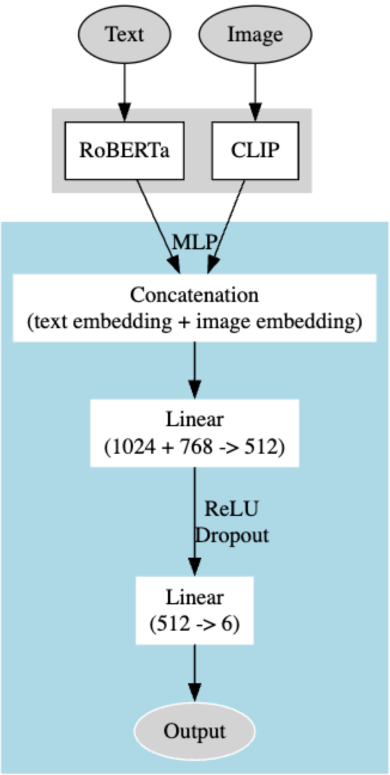


Figure 3.2.: Multimodal Fusion Architecture

Figure 3.3.: Finetuning setup with Base MLP

Unimodal-Text (Figure 3.1)

Uses a **Cardiffnlp RoBERTa-Large** encoder to produce text embeddings (1024-dim), followed by the standard projection head.

Unimodal Image (Figure 3.1)

Uses a **CLIP ViT-L/14** encoder to extract image embeddings (768-dim), again followed by the standard projection head.

Multimodal Fusion (Figure 3.2)

Combines text and image encoders from both unimodal settings in parallel. Text embeddings and image embeddings are concatenated. This fused representation (1792-dim) is passed to the standard projection head. For multimodal experiments, we tested many fine-tuned variations. The details are:

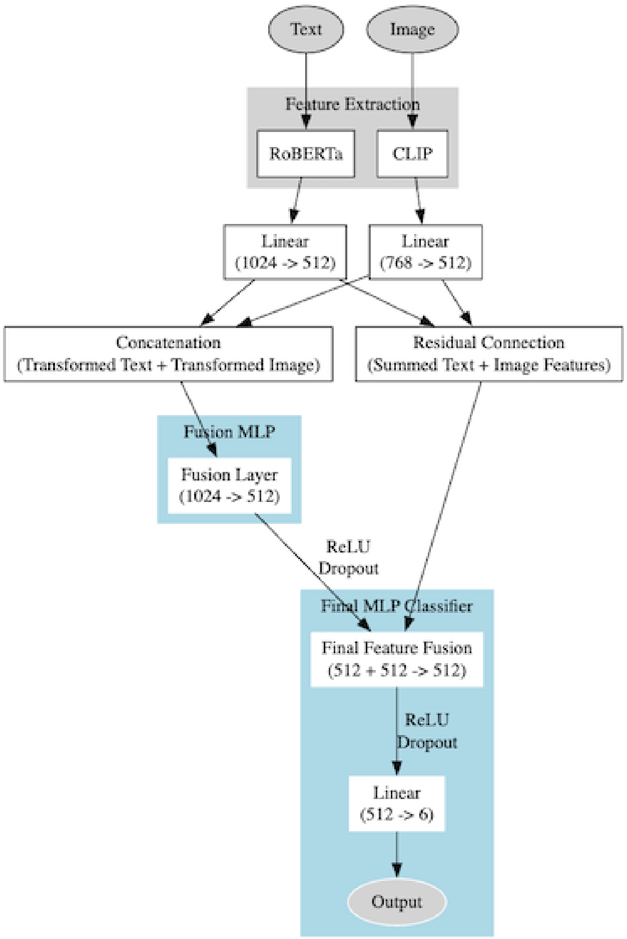

Text
Image
Feature Extraction
RoBERTa
CLIP
Linear (1024 -> 512)
Linear (768 -> 512)
Concatenation (Transformed Text + Transformed Image)
Residual Connection (Summed Text + Image Features)
Fusion MLP
Fusion Layer (1024 -> 512)
ReLU Dropout
Final MLP Classifier
Final Feature Fusion (512 + 512 -> 512)
ReLU Dropout
Linear (512 -> 6)
Output

Figure 3.4.: Multimodal Residual fusion architecture

- **Encoder Tuning**
  - **Frozen text encoder:** Only freeze text encoder weights, fine-tune MLP and image encoder

  - **Frozen image encoder:** Only freeze image encoder weights, fine-tune MLP and text encoder
  - **Both encoders Frozen:** Freeze both encoder weights, only fine-tune MLP.
  - **Full fine-tune:** Fine-tune MLP and both encoders.
  - **Staggered-unfreezing** Unfreeze Image encoder weights after 2 epochs and unfreeze text encoder weights after 4 epochs.

- **MLP variations**
  - Standard Projection Head (Figure 3.2)
  - Deeper (3-layer) MLP with ReLU Activation (hidden sizes 1024 and 512)
  - Deeper (3-layer) MLP with GELU Activation (hidden sizes 1024 and 512)

- **Fusion strategies**
  - Late Fusion - Concatenation of final layer embeddings of text and image models (Figure 3.2)
  - Residual Fusion (Concatenation + Residual Addition) (Figure 3.4)

All experiments were conducted on data from two distinct time intervals (February and August) for temporal robustness. These choices were combined systematically resulting in 320 unique experiment configurations (1280 in total for 2 datasets and 2 seeds)

A summary of the main architectural variations is given in Table 3.4.

Component	**Text-Based**	**Image-Based**	**Multimodal**
Model	RoBERTa-large	CLIP ViT-L/14	RoBERTa (Large/Base) + CLIP ViT-L/14
Input	Max 512 tokens	224 × 224 pixels	Text: 512 tokens, Image: 224 × 224
Embedding Dim	1024	768	1792
MLP Depth	2 layers	2 layers	2 or 3 layers
Hidden Sizes	512	512	1024, 512 (3-layer MLP)
Activation	ReLU	ReLU	ReLU or GELU
Tuning Mode	Frozen or Fine-tuned	Frozen or Fine-tuned	Frozen, Fine-tuned or Staggered,
Fusion Strategy	N/A	N/A	Concatenation / Residual Fusion

Table 3.4.: Architectural Variations

Our approach contrasts with baseline models which in our study are pre-trained zero-shot models, which are evaluated without any fine-tuning. Specifically:

- For the **text-based approach**, the Baseline model is **CardiffNLP RoBERTa-Large**, which is used in a zero-shot setting.

- For the **image-based approach**, the Baseline model is **CLIP ViT-L/14**, applied without fine-tuning.

- For the **multimodal approach**, the Baseline predictions are obtained by averaging the outputs from these two unimodal Baseline models.

3.7. Experimental Setup

This section details the computing environment, hyperparameter selection, reproducibility measures, and monitoring tools used for the experiments.

Hardware and Software Configuration

- **Hardware:** Two NVIDIA RTX A6000 GPUs with 48 GB VRAM each.
- **Programming Language:** Python 3.9
- **Key Libraries:**
  - PyTorch 2.5.0 with CUDA 11.8
  - HuggingFace Transformers 4.44.2
  - NumPy 1.26.4, pandas 2.2.2
  - scikit-learn 1.5.1 for evaluation metrics
  - TensorBoard 2.18.0 provided real-time loss and accuracy curves

Training Configuration and Hyperparameters

A systematic approach was employed to explore and validate hyperparameter choices:

- **Objective Function:** KL divergence to match soft-label distributions.
- **Optimizers:** Compared Adam vs. AdamW (decoupled weight decay $\lambda = 0.01$), with $\beta_1 = 0.9$, $\beta_2 = 0.999$.
- **Learning Rates:** $\{ 1 \times 10^{-5}, 5 \times 10^{-6} \}$.
- **Batch Size:** Typically 16 per GPU.
- **Epochs:** 2, 5 and 10
- **Regularization:**
  - Dropout $\{0.3, 0.5\}$ within MLP layers.
  - Weight decay (AdamW).

Reproducibility Protocol

- All experiments were conducted using two distinct random seeds (42 and 7) to ensure reproducibility.
- Randomness was controlled by setting the same seed for Python's random module, NumPy, PyTorch, and CUDA operations.
- Identical initialization and data splits were maintained across all runs.

4. Experimental Results

4.1. Zero-Shot Experimentation Results

This section presents the performance of five zero-shot models in predicting emotion labels for climate change discourse on social media. Each model's predictions were evaluated against an annotated subset of 99 English replies, using Exact Match (EM), Top-3 Accuracy, Ranked Score, and NDCG@3 as performance metrics.

4.1.1. Model Performance Overview

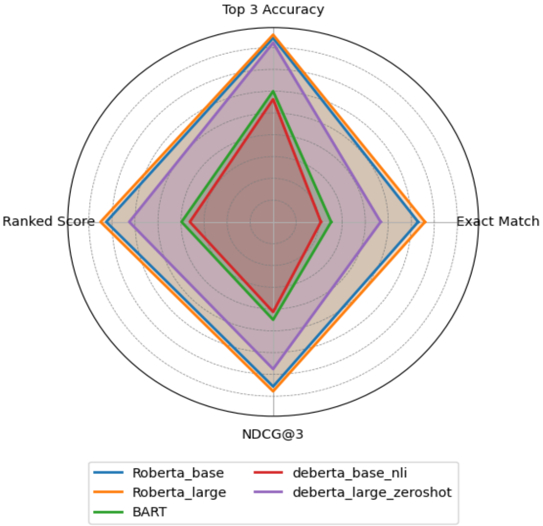


Figure 4.1.: Performance comparison of zero-shot models based on key evaluation metrics.

Table 4.1 and Figure 4.1 summarize the results. The CardiffNLP *twitter-roberta-large-emotion-latest* model achieved the highest scores across most metrics, with an Exact Match of 0.659, a Top-3 Accuracy of 0.859, a Ranked Score of 0.749, and an NDCG@3 of 0.778. In comparison, the other CardiffNLP model (*twitter-roberta-base-emotion-latest*) also performed well, outperforming the more general-purpose BART-large-MNLI and DeBERTa-based models in all metrics.

The pre-training of the CardiffNLP models on tweet datasets appears to be advantageous for emotion detection, as evidenced by their higher scores. The more general

Model	Exact Match	Top-3 Accuracy	Ranked Score	NDCG@3
RoBERTa-base (CardiffNLP)	0.630	0.844	0.725	0.755
RoBERTa-large (CardiffNLP)	**0.659**	**0.859**	**0.749**	**0.778**
BART-large-MNLI	0.252	0.600	0.398	0.449
mDeBERTa-base (multilingual)	0.207	0.563	0.362	0.413
DeBERTa-v3-large-Zeroshot	0.467	0.822	0.625	0.675

Table 4.1.: Performance of the zero-shot models on the annotated subset ($n = 99$).

BART and DeBERTa variants had lower performance, particularly in Exact Match and Top-3 Accuracy, suggesting less robust sensitivity to the nuances of climate change discourse.

4.1.2. Confidence Filtering

A confidence threshold of 0.9 was applied to refine the model outputs. Table 4.2 shows how many predictions remained after filtering and how often these filtered predictions were correct.

Model	Correct Predictions	Total Predictions	Percentage (%)
RoBERTa-base (CardiffNLP)	38	61	62.30
RoBERTa-large (CardiffNLP)	41	65	63.08
BART-large-MNLI	4	5	80.00
mDeBERTa-base (multilingual)	0	1	0.00
DeBERTa-v3-large-Zeroshot	24	33	72.73

Table 4.2.: Performance of the zero-shot models on the annotated subset ($n = 99$) with confidence score > 0.9.

Interestingly, the *twitter-roberta-large-emotion-latest* model produced the largest number of high-confidence predictions, although its overall precision at these high-confidence levels was slightly lower than that of BART-large-MNLI and DeBERTa-v3-large-Zeroshot. In contrast, DeBERTa-v3-large-Zeroshot had fewer total high-confidence predictions, indicating that it was more conservative in assigning high confidence but achieved a higher precision among those it did classify confidently.

4.2. Weakly Supervised Learning Results

4.2.1. Zero-shot Classification Boost with Self-training

We replicate the study conducted by (Gera et al. 2022), which investigates the impact of self-training on zero-shot entailment models. Table 4.3 presents the accuracy results on three datasets: **AG**, **ISEAR**, and **ClimateTV**. The first two datasets, AG and ISEAR, were originally used in the reference study, while we extend the evaluation to ClimateTV dataset.

Model	AG	ISEAR	ClimateTV
BART	66.2	56.0	**25.25**
+Self-training	74.2	65.3	**34.34**
DeBERTa	73.2	58.5	**46.46**
+Self-training	81.4	59.5	**46.46**
RoBERTa	62.4	52.0	**32.32**
+Self-training	76.5	56.7	**32.32**

Table 4.3.: Zero-shot classification accuracy of entailment models. For each zero-shot entailment model and dataset, The test accuracy of the off-the-shelf model to its accuracy after 2 iterations of self-training. RoBERTa, DeBERTa, and BART correspond to the following models from Hugging Face Hub: roberta-large-mnli, deberta-large-mnli-zero-cls, and bart-large-mnli.

For each model—**BART**, **DeBERTa**, and **RoBERTa**—we report the baseline accuracy of the off-the-shelf zero-shot model, followed by its performance after two iterations of self-training. Consistent with the findings of Gera et al. (2022), self-training substantially improves zero-shot classification performance across AG and ISEAR datasets, yielding gains of up to **12%**. However, the effect varies when applied to the ClimateTV dataset.

- **BART** shows a notable improvement on the ClimateTV dataset, increasing accuracy from **25.25% to 34.34%** after self-training.
- **DeBERTa**, which already performs significantly better than the other models on ClimateTV, **does not benefit from self-training**, maintaining an accuracy of **46.46%**.
- **RoBERTa**, similar to DeBERTa, shows **no performance gain** on the ClimateTV dataset, with self-training yielding identical results.

These observations suggest that while self-training generally enhances zero-shot performance, its effectiveness may depend on the dataset characteristics. In particular, the ClimateTV dataset appears to present challenges that self-training does not overcome for DeBERTa and RoBERTa.

4.2.2. Loss Re-weighting

To evaluate the impact of fine-tuning with loss re-weighting, we compare the model's performance on the manually annotated dataset before and after training. Table 4.4 provides per-class accuracy comparisons. Additionally, the classification reports in Table 4.5 present a detailed breakdown of precision, recall, and F1-score for each emotion class.

Emotion Class	Baseline Accuracy	After Training Accuracy
Anger	0.6667	0.7576
Disgust	0.4000	0.2667
Fear	0.3077	0.3077
Joy	0.8750	0.8750
Sadness	0.0833	0.0833
Surprise	0.3000	0.5000

Table 4.4.: Per-class accuracy comparison.

Class	Baseline			After Training		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Anger	0.76	0.67	0.71	0.71	0.76	0.74
Disgust	0.27	0.40	0.32	0.27	0.27	0.27
Fear	0.57	0.31	0.40	0.50	0.31	0.38
Joy	0.48	0.88	0.62	0.54	0.88	0.67
Sadness	0.50	0.08	0.14	0.50	0.08	0.14
Surprise	0.30	0.30	0.30	0.38	0.50	0.43
Accuracy	0.5051			0.5354		

Table 4.5.: Classification report comparison before and after training.

The results indicate a slight overall accuracy improvement of **3.03%** after training. The per-class accuracy shows that **"anger" and "surprise" saw the most significant improvements**, while **"disgust" slightly degraded**. The confidence scores on average decreased by **0.0552**, suggesting that while the model made some better predictions, it became slightly less confident overall. The classification report highlights improved F1-scores for **anger, joy, and surprise**, while other classes remained stable.

4.3. Text-Based Results

Table 4.6 summarizes the results for August and February, showing significant improvements across all metrics. Our model outperforms the Baseline by 16.3% in semantic alignment (Cosine Similarity), reduces distributional divergence by 58.5% (KL Divergence), and lowers error magnitude by 43.4% (MSE).

Our model achieves a combined Cosine Similarity score of 0.8028, a **16.3% improvement** over the Baseline ($\Delta = +0.1125$). Month-wise improvements are +17.0% (August) and +16.6% (February), showing stable performance over time.

For KL Divergence, our model reduces distributional mismatch by **58.5%** ($\Delta = 0.4898$), with the highest reduction in February (60.9%). This indicates better probability distribution calibration.

Metric	Our Model			Baseline		
	August	February	Combined	August	February	Combined
Cosine Similarity ↑	0.8094	0.7962	**0.8028**	0.6919	0.6830	**0.6903**
KL Divergence ↓	0.3473	0.3483	**0.3478**	0.8254	0.8909	**0.8376**
MSE ↓	0.0230	0.0232	**0.0232**	0.0408	0.0415	**0.0410**

Table 4.6.: Performance comparison between Baseline and Our Model. The Baseline is the **CardiffNLP RoBERTa-Large**, which is used in a zero-shot setting. Higher values are better for ↑, and lower values are better for ↓. Bold values indicate combined scores.

The MSE results show a **43.4% reduction** in error ($\Delta = 0.0178$), with consistent improvements of 43.6% (August) and 44.1% (February), confirming robust generalization.

4.4. Image-Based Results

Table 4.7 presents the results for image-based experiments. Our model improves semantic alignment by 56.4%, reduces distributional divergence by 72.2%, and lowers error magnitude by 64.2%.

Metric	Our Model			Baseline		
	August	February	Combined	August	February	Combined
Cosine Similarity ↑	0.8105	0.7887	**0.7996**	0.5017	0.5529	**0.5113**
KL Divergence ↓	0.3555	0.3611	**0.3583**	1.3696	0.9196	**1.2859**
MSE ↓	0.0236	0.0240	**0.0238**	0.0682	0.0583	**0.0664**

Table 4.7.: Performance comparison between Baseline and Our Model for image-based experiments. The Baseline model is **CLIP ViT-L/14**, applied without fine-tuning. Higher values are better for ↑, and lower values are better for ↓. Bold values indicate combined scores.

Our model achieves a combined Cosine Similarity of 0.7996, a **56.4% improvement** over the Baseline ($\Delta = +0.2883$), with month-wise improvements of 61.5% (August) and 42.6% (February).

For KL Divergence, our model reduces distributional mismatch by **72.2%** ($\Delta = 0.9276$), with a 74.0% reduction in August.

The MSE results show a **64.2% reduction** in error ($\Delta = 0.0426$), with reductions of 65.4% (August) and 58.8% (February). Fine-tuning the CLIP encoder was crucial, as aggressive tuning without regularization led to overfitting.

4.5. Multimodal Results

Table 4.8 summarizes multimodal results, showing improvements of 14.4% in semantic alignment, 44.0% in distributional divergence, and 35.7% in error magnitude.

Metric	Our Model			Baseline		
	August	February	Combined	August	February	Combined
Cosine Similarity ↑	0.8263	0.7989	**0.8127**	0.7090	0.7233	**0.7117**
KL Divergence ↓	0.3314	0.3411	**0.3362**	0.6118	0.5311	**0.5968**
MSE ↓	0.0220	0.0228	**0.0224**	0.0355	0.0311	**0.0346**

Table 4.8.: Performance comparison between Baseline and Our Model for multimodal experiments. The Baseline predictions are obtained by averaging the outputs from the two unimodal Baseline models. Higher values are better for ↑, and lower values are better for ↓. Bold values indicate combined scores.

Our model achieves a combined Cosine Similarity of 0.8127, outperforming the Baseline by **14.4%** ($\Delta = +0.1010$). Month-wise improvements show a 16.6% gain in August and a 10.5% gain in February, demonstrating stability across different time frames.

The KL Divergence reduction of **44.0%** ($\Delta = 0.2605$) reflects a substantial enhancement in probability distribution alignment. The largest reduction occurs in August (45.8%), closely followed by February (35.8%), reinforcing the model's robustness in learning more calibrated probability distributions.

For MSE, our model achieves a **35.7%** reduction ($\Delta = 0.0124$), reducing the overall error magnitude compared to the Baseline. Improvements are stable across months, with a 38.0% reduction in August and 26.3% in February, further validating the model's ability to generalize well across different periods.

4.6. Sensitivity to Datasets and Random Initialisation

All experiment configurations described in Section 3.7 (e.g., hyperparameters, tuning strategies, and architectures) were **identically repeated** for **both datasets** (*August* and *February*) and for each of the **two random seeds** (42 and 7). This setup ensures that any observed performance differences can be attributed primarily to *dataset-specific* factors or *random initialization*, rather than differences in training procedures.

4.6.1. Comparison Across Datasets

Table 4.9 presents aggregated metrics for each modality (averaged over the two seeds) on the August vs. February data. We see that August data tends to yield higher performance across text-only, image-only, and multimodal models. This aligns with prior research on climate change discourse on social media, which indicates that August tends to experience heightened engagement and visual content related to climate activism and extreme weather events (Mooseder et al. 2023).

Modality	Dataset	Cosine Similarity	KL Divergence	MSE
Text-Only	August	0.7998	0.3698	0.0243
Text-Only	February	0.7852	0.3763	0.0249
Image-Only	August	**0.8005**	0.3763	0.0250
Image-Only	February	**0.7732**	0.3982	0.0263
Multimodal	August	**0.8120**	0.3549	0.0234
Multimodal	February	**0.7973**	0.3704	0.0246

Table 4.9.: Comparison of mean performance on August vs. February data. For each month, the results are averaged across all experiments, including both random seeds (42 and 7), to provide a comprehensive performance summary for each modality (text-only, image-only, and multimodal).

Observations:

- **Consistent Patterns:** For every modality, performance is higher on August data in terms of Cosine Similarity and lower for error-based metrics (MSE, KL Divergence). This aligns with prior research showing that climate change discourse in August is more visually engaging, focusing on protests and environmental disasters (Mooseder et al. 2023).
- **Statistical Significance:** We performed an independent samples t-test comparing performance on August vs. February data.
  - **Cosine Similarity:** August data shows significantly higher values ($t = 34.52$, $p < 0.0001$).
  - **MSE:** August data exhibits significantly lower values ($t = -12.49$, $p < 0.0001$).
  - **KL Divergence:** August data also demonstrates significantly lower values ($t = -10.40$, $p < 0.0001$).
- **Seasonal Discourse Variations:** Climate change discussions in February tend to focus on scientific reports and policy debates, whereas August discussions emphasize activism and extreme weather, leading to higher interaction and sentiment consistency (Mooseder et al. 2023).
- **Bot Activity and Polarization:** Research indicates that automated social media accounts (bots) play a larger role in August, amplifying climate activism and scepticism, contributing to increased polarization and engagement and thus the volume of data (Mooseder et al. 2023).

- **Stable Ordering:** In both datasets, multimodal models outperform text-only, which in turn outperform image-only.

4.6.2. Effect of Random Initialisation

Modality	Seed	Cosine Similarity	KL Divergence	MSE
Text-Only	42	0.7994	0.3646	0.0243
Text-Only	7	0.7855	0.3816	0.0249
Image-Only	42	0.7884	0.3831	0.0257
Image-Only	7	0.7854	0.3914	0.0257
Multimodal	42	0.8023	0.3562	0.0238
Multimodal	7	0.7959	0.3691	0.0242

Table 4.10.: Comparison of mean performance with seeds 42 vs. 7. For each seed, the results are averaged across all experiments, including data from both months, to provide a comprehensive performance summary for each modality (text-only, image-only, and multimodal).

To check how initialization affects outcomes, each configuration was also trained using two random seeds. Table 4.10 presents aggregated metrics for each modality (averaged over both datasets) for seeds 42 vs. 7.

Observations:

- **Minor Variations:** Although seed 42 sometimes slightly outperforms 7, differences remain modest.
- **Two-Seeds Constraint:** While more seeds could provide additional robustness checks, these two indicate that performance differences due to initialization are small and do not affect overall conclusions.

In summary, training on August vs. February data produces larger, statistically significant performance differences than using different seeds. Although we only used two seeds, the minimal performance gap between them suggests that our primary findings are not unduly affected by random initialization. These results will be further discussed in Chapter 5.

4.7. Comparative Analysis: Performance Metrics, Model Configurations, and Hyperparameter Impact

This section provides a unified view of performance across text, image, and multimodal models. We begin with an overview of top and bottom performers (Table 4.11), followed by detailed insights into the role of hyperparameters in each modality.

4.7.1. Overall Performance Across Modalities

Table 4.11 summarizes the highest and lowest performers, revealing that multimodal approaches consistently outperform single-modality setups.

Model	Score	CosineSim	KLDiv	MSE	LR	Dropout	Epochs	Key Features
Top Performers (Multimodal)								
RoBERTa-Large	2.454	0.813	0.336	0.022	1e-5	0.5	5	3-Layer MLP, Frozen
RoBERTa-Base	2.452	0.812	0.337	0.022	5e-6	0.3	2	GELU, Frozen CLIP
RoBERTa-Large	2.450	0.811	0.339	0.023	1e-5	0.5	5	GELU, Frozen
Bottom Performers (Single-Modality)								
RoBERTa-Large	2.281	0.760	0.450	0.029	1e-5	0.3	10	Text Only, Tuned
CLIP	2.374	0.788	0.395	0.025	5e-6	0.5	5	Image Only, Frozen

Table 4.11.: Top and bottom performers across text, image, and multimodal settings.

Key Observations from Table 4.11

- **Multimodal Superiority**: All top performers combine RoBERTa (text) and CLIP (image) with partial freezing strategies (e.g., frozen RoBERTa, CLIP, or both). Their Cosine Similarity (0.811–0.813) surpasses single-modality models by 3–7% (score difference: 6.9% for text-only, 3.3% for image-only).
- **Balanced Regularization**: Top configurations use moderate learning rates (1e-5 or 5e-6) and dropout (0.3–0.5), achieving optimal trade-offs between capacity and generalization.
- **Efficient Convergence**: Multimodal models converge within 2–5 epochs. Extending training to 10 epochs degrades tuned single-modality performance (e.g., text-only Cosine Similarity drops 3.5%).

4.7.2. Hyperparameter Impact: Text-Only Models

Experiments on text-only RoBERTa models (Section 3.6) reveal critical hyperparameter interactions:
Key Trends:

- **Freezing Encoders**: Frozen models outperform tuned counterparts by 3.7% in score (2.4318 vs. 2.3455). Stability in pre-trained representations prevents overfitting, especially with longer training (10 epochs).
- **Learning Rate Sensitivity**: Optimal performance occurs at lr=1e-5 for frozen models. Lower rates (5e-6) require fewer epochs to match performance.
- **Epoch Management**: Tuned models degrade sharply beyond 5 epochs (score drops 5.1% at 10 epochs), highlighting the necessity of early stopping.

Tuning	Optimizer	Learning Rate	Epochs	Dropout	CosineSim ↑	KLDiv ↓	MSE ↓	Score ↑
Frozen	Adam	1e-5	10	0.3	**0.8028**	**0.3478**	**0.0232**	**2.4318**
		5e-6	10	0.3	0.8015	0.3498	0.0233	2.4285
		5e-6	5	0.3	0.8016	0.3518	0.0233	2.4265
		5e-6	2	0.3	0.8010	0.3481	0.0232	2.4297
		1e-5	2	0.3	0.8019	0.3529	0.0233	2.4256
		1e-5	5	0.3	0.8015	0.3500	0.0233	2.4281
Tuned	Adam	1e-5	5	0.3	0.7773	0.4048	0.0270	2.3455
		5e-6	5	0.3	0.7861	0.3891	0.0259	2.3711
		5e-6	2	0.3	0.7947	0.3527	0.0237	2.4182
		5e-6	10	0.3	0.7675	0.4385	0.0285	2.3005

Table 4.12.: Impact of learning rate, epoch count, optimizer, and freezing on text-only performance. The best configuration is in bold.

4.7.3. Hyperparameter Impact: Image-Only Models

Image-only CLIP models exhibit distinct hyperparameter dynamics compared to text:

Tuning	Optimizer	Learning Rate	Epochs	Dropout	CosineSim ↑	KLDiv ↓	MSE ↓	Score ↑
Tuned	Adam	5e-6	2	0.5	**0.7996**	**0.3583**	**0.0238**	**2.4174**
		5e-6	2	0.3	0.7990	0.3591	0.0240	2.4159
		5e-6	10	0.3	0.7743	0.4130	0.0270	2.3343
		1e-5	10	0.3	0.7700	0.4224	0.0276	2.3200
		1e-5	2	0.3	0.7960	0.3670	0.0246	2.4044
		1e-5	5	0.5	0.7812	0.3911	0.0261	2.3640
	AdamW	5e-6	2	0.3	0.7993	0.3580	0.0238	2.4174
		5e-6	2	0.5	0.7977	0.3612	0.0241	2.4124
		1e-5	5	0.3	0.7754	0.4085	0.0270	2.3399
Frozen	Adam	1e-5	10	0.3	**0.7996**	**0.3584**	**0.0238**	**2.4173**
		5e-6	2	0.3	0.7748	0.4174	0.0276	2.3298
		5e-6	2	0.5	0.7727	0.4222	0.0279	2.3226

Table 4.13.: Impact of learning rate, epoch count, optimizer, and freezing on image-only performance. The best configurations are in bold.

Key Trends:

- **Tuned Efficiency**: Short training (2 epochs) with lr=5e-6 and dropout=0.5 yields peak performance (2.4174). Extending to 10 epochs reduces scores by 3.4%.
- **Frozen Flexibility**: Frozen models achieve comparable scores (2.4173) but require higher learning rates (1e-5) and extended training (10 epochs).
- **Optimizer Parity**: Adam and AdamW show negligible differences (< 0.1% score variance) under identical hyperparameters.

4.7.4. Hyperparameter Impact: Multimodal Models

Multimodal configurations demonstrate synergistic benefits from modality fusion:

Model	Optimizer	Learning Rate	Dropout	Epochs	Fusion/MLP	Tuning	CosineSim ↑
RoBERTa-large	AdamW	1e-5	0.5	5	3-Layer MLP	Both frozen	**0.8126**
		1e-5	0.5	5	GELU / 3-Layer MLP	Both frozen	0.8114
		1e-5	0.5	5	-	Tuned	0.8085
		5e-6	0.5	2	-	Tuned	0.8085
		5e-5	0.5	5	Residual fusion	-	0.7858
		1e-5	0.3	5	-	Frozen Clip	0.7684
	Adam	5e-6	0.5	2	Residual fusion	-	0.8064
		1e-5	0.3	5	-	Tuned	0.7712
RoBERTa-base	AdamW	1e-5	0.5	5	-	Tuned	0.7679
		1e-5	0.5	5	GELU / 3-Layer MLP	Tuned	0.7725
		5e-6	0.3	2	GELU / 3-Layer MLP	Tuned	**0.8115**
		5e-6	0.5	5	-	Both frozen	0.8070
	Adam	5e-6	0.5	2	3-Layer MLP	Tuned	0.8094
		5e-6	0.3	5	-	Frozen Clip	0.7805
		1e-5	0.5	2	-	Frozen Clip	0.8079

Table 4.14.: Impact of learning rate, epochs, dropout, optimizer, fusion strategy, and freezing on multimodal performance. The best configurations are in bold.

Key Insights:

- **Freezing Efficacy**: Fully frozen RoBERTa-Large and CLIP achieve the highest score (2.4539), emphasizing the value of preserving pre-trained features. Partial tuning (e.g., RoBERTa-Base with GELU) narrows the gap to 0.04% (2.4522 vs. 2.4539).
- **Architectural Adaptability**: RoBERTa-Base competes with RoBERTa-Large when paired with GELU activations and tuned layers, despite 55% fewer parameters.
- **Dropout Stratification**: Larger models (RoBERTa-Large) require higher dropout (0.5), while Base variants perform optimally at 0.3.
- **Optimizer Robustness**: AdamW marginally outperforms Adam ($< 0.5\%$ difference), suggesting optimizer choice is secondary to freezing and learning rate.

4.7.5. Summary of Comparative Findings

The cross-modal analysis yields four principal conclusions:

- **Multimodal Synergy**: Fusion of text and image encoders with partial freezing elevates scores by 3–7% over single-modality models, achieving the highest alignment scores (Cosine Similarity > 0.81).

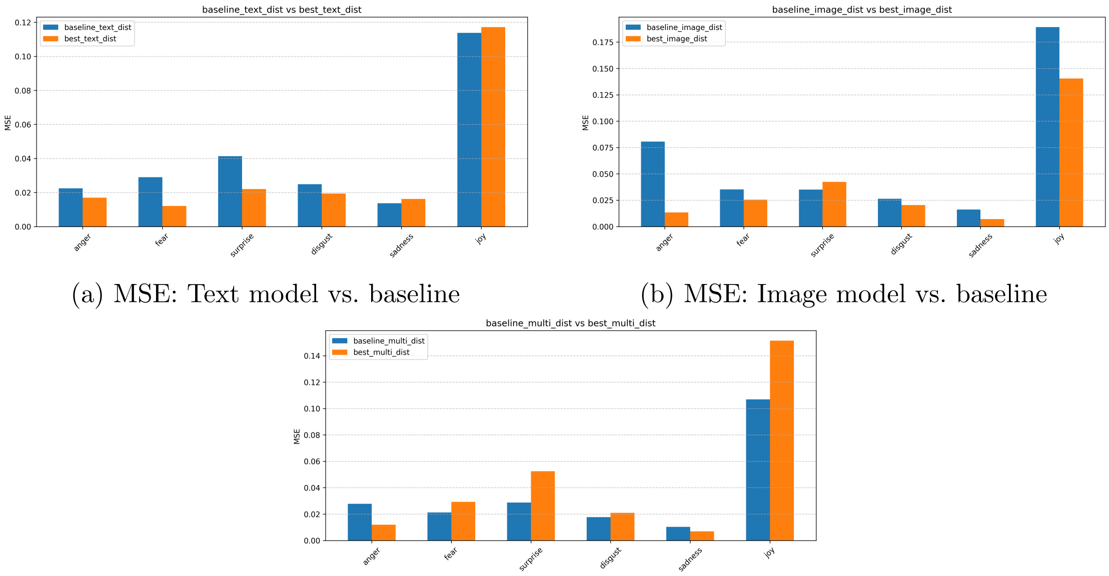
- **Modality-Specific Tuning**: Text models favour frozen encoders and moderate learning rates; image models benefit from brief tuning cycles; multimodal setups demand balanced regularization.

- **Efficient Training**: Optimal epoch counts are modality-dependent: 2–5 for multimodal, ≤ 5 for tuned single-modality, and ≤ 10 for frozen single-modality.

- **Parameter Efficiency**: Smaller architectures (RoBERTa-Base) can rival larger counterparts through activation function optimization (GELU) and targeted tuning.

These findings underscore the necessity of modality-aware hyperparameter strategies. Chapter 5 explores their theoretical implications and practical applications in depth.

4.8. Error Analysis



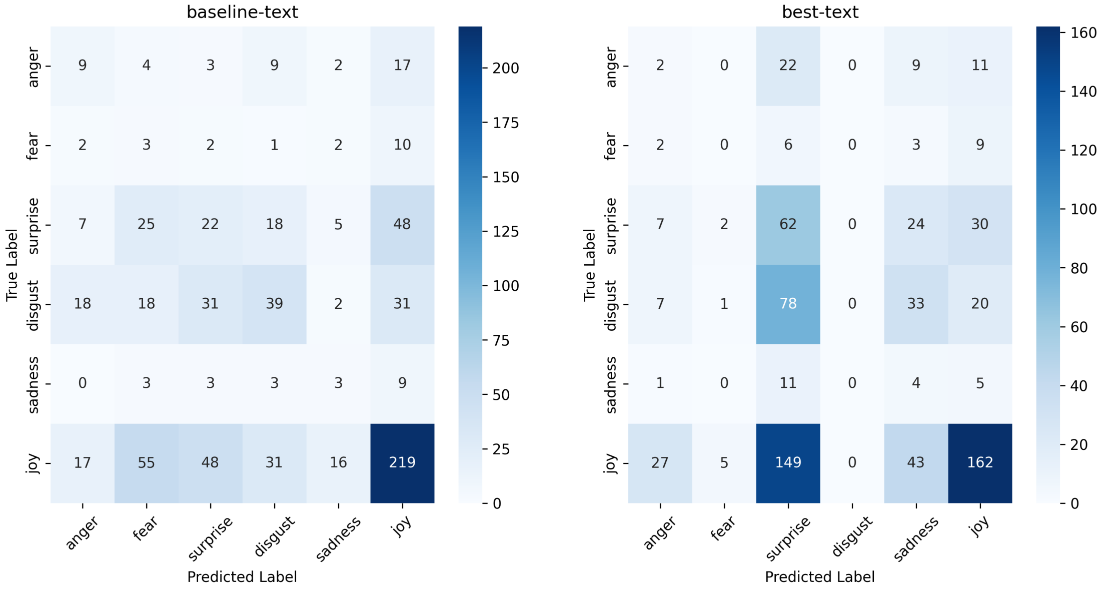
(a) MSE: Text model vs. baseline

(b) MSE: Image model vs. baseline

(c) MSE: Multimodal model vs. baseline

Figure 4.2.: Emotion-wise comparison of Mean Squared Error (MSE) across models and baselines. Emotion order - Anger, Fear, Surprise, Disgust, Sadness, Joy. Lower is better.

Despite significant gains in distribution metrics (Tables 4.6–4.8), our models (text-only, image-only, and multimodal) still struggle with several recurring error patterns. We classify these issues into three main categories in subsections 4.8.1–4.8.2.



(a) Baseline Text Model.

(b) Best Text Model.

Figure 4.3.: Confusion matrices for text models. Darker blues indicate higher prediction frequency.

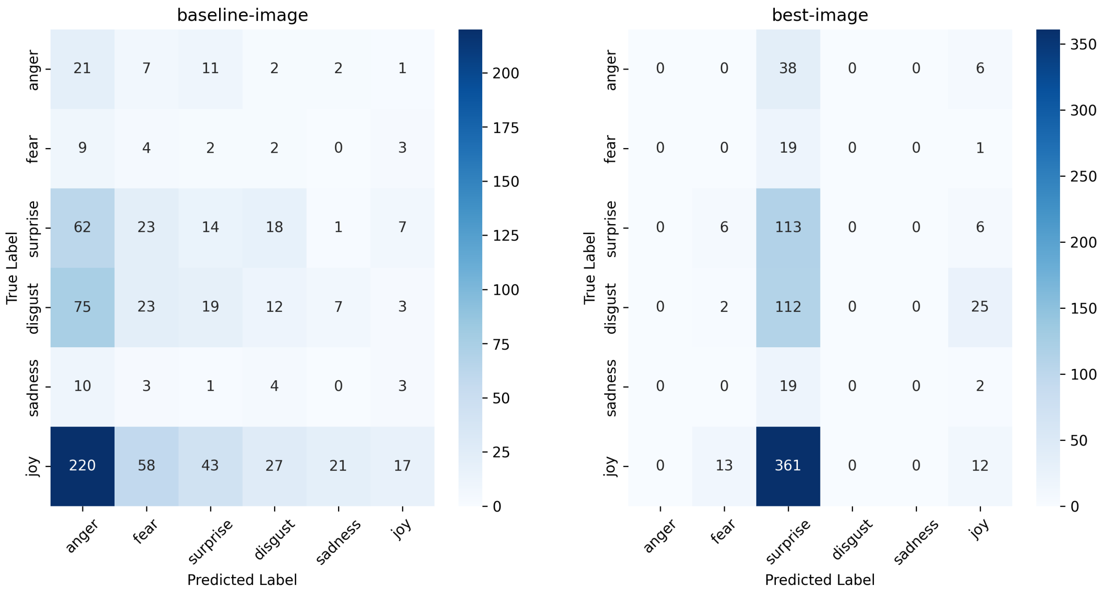
4.8.1. Cross-Modal Error Patterns

Our experiments reveal three broad error trends shared across text, image, and multimodal models:

- **Minority Class Collapse.** Rare emotions such as *sadness* or *fear* remain poorly recognized, with recall often below 15%. Text models frequently confuse *sadness* with *surprise* or *joy* (79% of errors), while image models can fail to detect *sadness* altogether (0% recall). Even though Mean Squared Error (MSE) drops by 35–64% for some rare classes, Figure 4.2 and the confusion matrices confirm these mistakes persist.

- **Modality-Specific Overfitting.** Optimizing for distribution metrics can harm instance-level accuracy. Text-based models “forget” *anger* (only 2 out of 44 correct) when improving other metrics. Image-only models overcompensate for *anger* bias, predicting *surprise* almost exclusively (113/125 correct on *surprise*, but 361 false positives). Multimodal fusion inherits both problems, absorbing the text model’s *sadness* mistakes (92% misclassifications) and the image model’s *surprise* bias.

- **Distribution-Instance Dissonance.** Large gains in cosine similarity (16–56% improvement) can mask argmax errors. For example, *joy* appears best predicted (162–219 correct in text models), but this partly stems from excessive *joy* pre-



(a) Baseline Image Model.

(b) Best Image Model.

Figure 4.4.: Confusion matrices for image-only models.

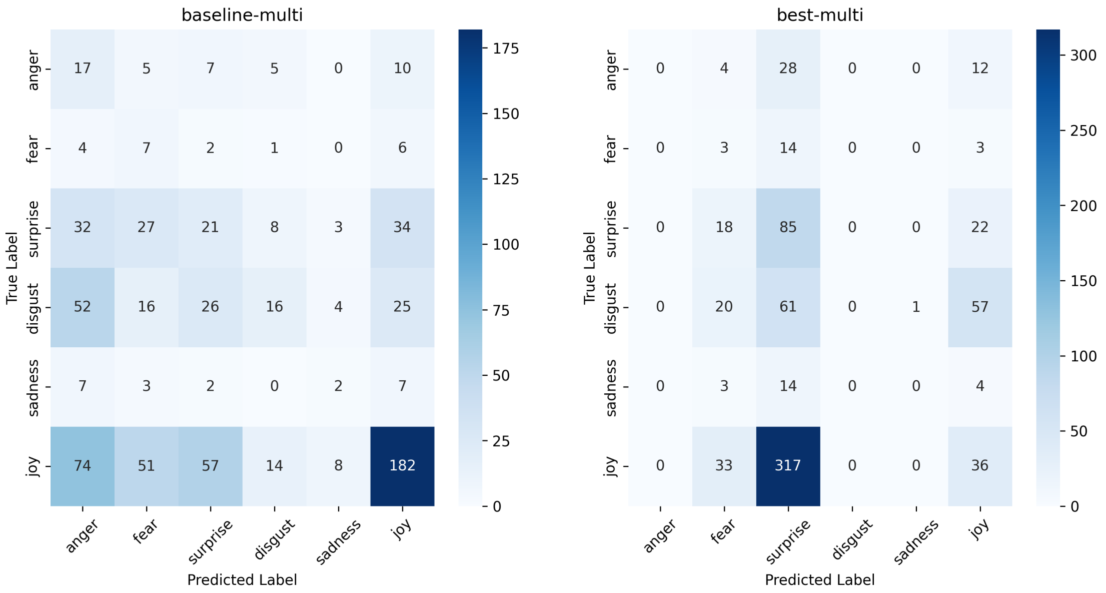
dictions (36–48% of all outputs). The inherent trade-off between distributional alignment and per-instance fidelity remains a core challenge.

4.8.2. Class-Level Analysis

To better understand model misclassifications, we closely examine both the confusion matrices (Figures 4.3–4.5) and the MSE trends (Figure 4.2; also see the per-class MSE values in Tables 4.6–4.8).

Text Model. From Table 4.6, the *best text model* achieves notable MSE reductions for *anger* (0.0224→0.0170) and *fear* (0.0290→0.0121) compared to the baseline. However, it slightly increases MSE for *sadness* (0.0137→0.0162) and *joy* (0.1138→0.1172). This trade-off is also reflected in Figure 4.3b (row 5: sadness), where only 3 of 18 sadness instances are correctly identified. The confusion matrix further reveals that *anger* is often misclassified as *surprise* (22 out of 44 anger instances).

Image Model. For image-only models (Table 4.6), the *best image model* significantly lowers MSE for *joy* (0.1892→0.1405) but worsens for *surprise* (0.0350→0.0424). In Figure 4.4b, the majority of misclassifications (361 false positives) come from mistakenly predicting *surprise*, which also damages *disgust* recognition (79% mislabeled as *surprise*). Hence, while distribution metrics improve overall, the confusion matrix illustrates a strong one-label bias.



(a) Baseline Multimodal.

(b) Best Multimodal.

Figure 4.5.: Multimodal confusion matrices.

Multimodal Model. Finally, the multimodal model (Table 4.8) shows mixed gains. The *best multimodal* run achieves lower MSE for *anger* (0.0278→0.0120) and *sadness* (0.0103→0.0069) but raises MSE for *surprise* (0.0287→0.0525) and *joy* (0.1070→0.1514). The confusion matrix in Figure 4.5b confirms that a large portion of *disgust* entries (61%) are absorbed into *surprise*, indicating the model struggles to differentiate negative emotions when trained to optimize overall distribution resemblance. Freezing text layers (to retain gains on *fear* and *anger*) can, unfortunately, lock in earlier classification mistakes, such as *sadness* → *joy* mislabels.

Findings

- **Improvements in Rare Classes Are Fragile.** While MSE for *anger* and *fear* generally decreases, confusion matrices reveal significant misclassifications into *surprise* or *joy*. Minor distribution changes can drastically alter these minority-class predictions.

- **Overproduction of Dominant Emotions.** In both the text and image models, *surprise* and *joy* are predicted too often (see darker cells in Figures 4.3–4.4). This skew helps match the global emotion distribution but harms per-instance accuracy.

- **Multimodal Fusion Conflicts.** Although combining text and image usually improves distribution metrics, the confusion matrices (Figures 4.5b) highlight how errors from each modality can reinforce each other rather than cancelling out.

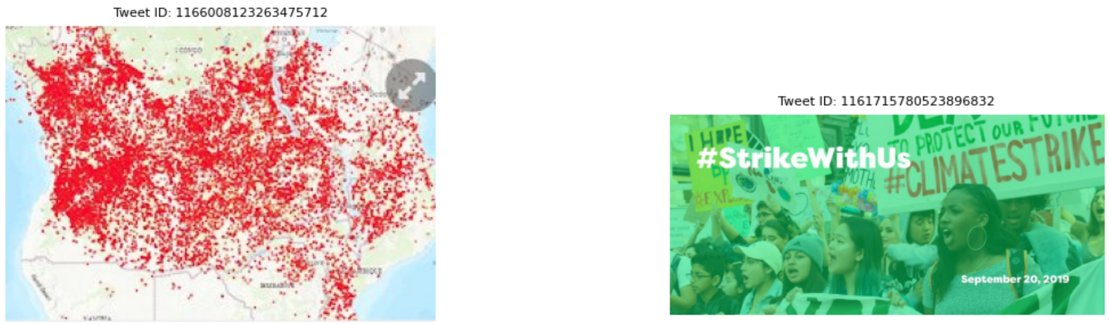
> Notably, *disgust* is frequently overridden by the image model's *surprise* bias, and *sadness* from text is often lost to *joy*.

In summary, while distribution-level metrics show progress, real-use scenarios require robust, per-instance predictions. Our analysis of the confusion matrices and MSE trends underscores the persistent difficulty in balancing minority-class accuracy with global distribution alignment. Addressing these issues will likely require new architectures and training strategies (e.g., multi-task losses, calibration techniques, or explicit rare-class focus) rather than pure metric optimization.

4.9. Qualitative Analysis

In this section, we present a detailed qualitative analysis of our emotion distribution predictions. Alongside each tweet snippet and its aggregated distribution of replies (the "gold" target), we also show samples of *individual* reply predictions. This provides insight into how each user response contributes to the final averaged (aggregated) target distribution and how the model might struggle or excel in capturing these varied emotional signals.

Poor Performance Cases



Tweet ID: 1166008123263475712

Tweet ID: 1161715780523896832

Figure 4.6.: Examples of high-error predictions (Case A on left, Case B on right).

Case A: Tweet ID 1166008123263475712

Text Snippet:

> #AlTenoorOverflow #Africafires #Angola #Zambia #Tanzania #Congo Links to the #SkirmishEvents statements in multiple languages: [URLs]

Aggregated Reply Distribution (1 reply):

{anger: 0.0694, fear: 0.4879, surprise: 0.0686,
disgust: 0.0421, sadness: 0.3046, joy: 0.0274}

Our Predictions (Text-only):

$$\{\text{anger: }0.1701,\quad \text{fear: }0.1532,\quad \text{surprise: }0.2239,$$
$$\text{disgust: }0.1151,\quad \text{sadness: }0.1629,\quad \text{joy: }0.1749\}$$

Our Predictions (Image-only):

$$\{\text{anger: }0.1181,\quad \text{fear: }0.2317,\quad \text{surprise: }0.2666,$$
$$\text{disgust: }0.1130,\quad \text{sadness: }0.1180,\quad \text{joy: }0.1526\}$$

Our Predictions (Multimodal):

$$\{\text{anger: }0.1089,\quad \text{fear: }0.2705,\quad \text{surprise: }0.2549,$$
$$\text{disgust: }0.0992,\quad \text{sadness: }0.1060,\quad \text{joy: }0.1606\}$$

Since only one user replied, the aggregated distribution essentially mirrors that single reply's predicted emotions. The user's response shows a dominant signal of *fear* and *sadness*:

> *"Enlightenment at a time when the world began to turn towards fires, it turns out that fires broke out in entire countries in sub-Saharan Africa..."*

Despite this clear emotional focus (with nearly half the probability mass on *fear*), the model's final (multimodal) prediction spread its probabilities more evenly, assigning:

- *Fear:* 27.0% (vs. 48.8% in replies),
- *Sadness:* 18.3% (vs. 30.5%),
- Extra probability mass on *Anger* and *Surprise*.

Notably, the text-only predictions (*fear*: 15.3%, *sadness*: 16.3%) and image-only predictions (*fear*: 23.2%, *sadness*: 11.8%) also dispersed their probability across other emotions like *anger* and *surprise*, failing to capture the strong peaks in *fear* and *sadness*. While the image-only model gave a slightly higher estimate for *fear* (23.2%), it still significantly underestimated both top emotions relative to the aggregated reply.

This mismatch highlights how **low-entropy replies** (favouring one or two emotions) can lead to larger KL divergence when the model avoids placing too much confidence in a single label. Our KL-based training objective often prevents the model from collapsing onto a single dominant peak—resulting in a more "flattened" (high-entropy) prediction distribution.

Case B: Tweet ID 1161715780523896832

Text Snippet:

> [USER] set sail today (sailboat) so..who's in to #StrikeWithUs (fire)(earth)? (world map)
> **Find Your Local Event** [URL] **Search Climate Actions** [URL] **Organize & Promote Your Own** [URL]
> GO (raised fist)(raised fist)(raised fist)(raised fist) #Fridays4Future [URL]

Aggregated Reply Distribution (1 reply):

{anger: 0.0498, fear: 0.0740, surprise: 0.6239,
disgust: 0.0973, sadness: 0.0866, joy: 0.0683}

Our Predictions (Text-only):

{anger: 0.1408, fear: 0.1396, surprise: 0.2146,
disgust: 0.1536, sadness: 0.1299, joy: 0.2212}

Our Predictions (Image-only):

{anger: 0.1063, fear: 0.2226, surprise: 0.3884,
disgust: 0.0765, sadness: 0.0701, joy: 0.1359}

Our Predictions (Multimodal):

{anger: 0.0844, fear: 0.2314, surprise: 0.4655,
disgust: 0.0664, sadness: 0.0569, joy: 0.0954}

There was only a single reply, predominantly expressing **surprise (62.4%)**. The user's short reaction suggests excitement or astonishment regarding the climate strike:

> *"p.s. this really matters for [. . .]"*

Our text-only model pushed more probability toward *joy* and *disgust*, while the image-only model improved the *surprise* estimate (38.8%) yet still fell below the gold distribution. Although *surprise* remained the highest predicted emotion in the final multimodal model (around 46.5%), it again did not reach the gold's strong single peak (62.4%).

As in Case A, our model's final prediction showed a more flattened distribution. Although *surprise* was still the highest predicted emotion, the model redistributed excess probability to *fear* and *anger*, increasing the KL divergence. Again, we see how a strong single-emotion reply can penalize the model's natural tendency to avoid overconfidence.

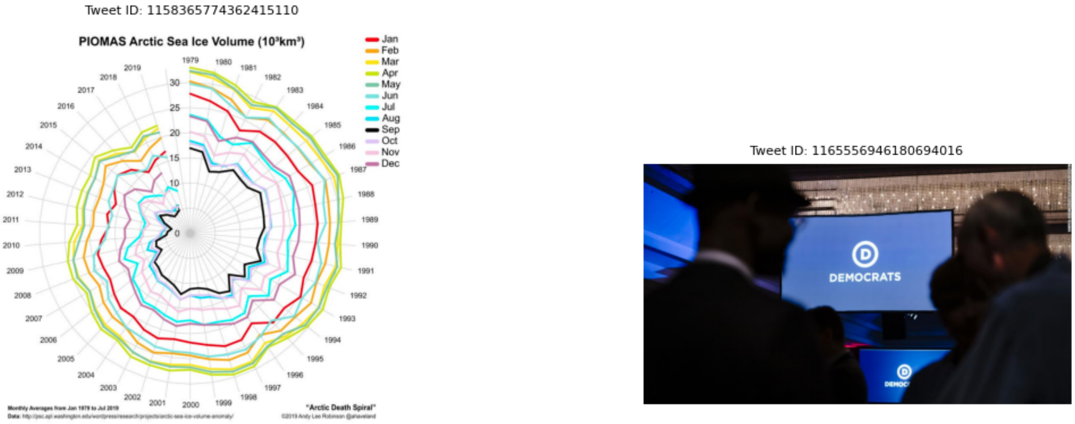


Figure 4.7.: Examples of well-aligned predictions (Case C on left, Case D on right).

Good Performance Cases

Case C: Tweet ID 1158365774362415110

Text Snippet:

> #Arctic sea ice volume broke another record minimum for July ... less than
> half
> of what it was 20 years ago. #climatechange #climatecrisis #dataviz

Aggregated Reply Distribution (multiple replies):

{anger: 0.0768, fear: 0.1475, surprise: 0.2777,
disgust: 0.1378, sadness: 0.1090, joy: 0.2511}

Our Predictions (Text-only):

{anger: 0.1557, fear: 0.1377, surprise: 0.2135,
disgust: 0.1306, sadness: 0.2068, joy: 0.1558}

Our Predictions (Image-only):

{anger: 0.1074, fear: 0.2742, surprise: 0.2789,
disgust: 0.1031, sadness: 0.1066, joy: 0.1298}

Our Predictions (Multimodal):

{anger: 0.0992, fear: 0.2918, surprise: 0.2525,
disgust: 0.0917, sadness: 0.1198, joy: 0.1450}

In total, there were *several distinct replies*, reflecting a wide range of sentiments. Some expressed concern (*fear, sadness*), others surprise or disgust at the data, while certain replies maintained an optimistic or grateful tone (*joy*):

> *"Great work—even if it highlights a worrying trend,"*
> *"The planet & all that inhabit it are in a world of trouble..."*
> *"Thank you very much, I have been producing it every month [...] hoping people will notice, think, and act..."*

This naturally yields a **higher entropy target distribution**, with each of the six emotions receiving non-trivial probability. Crucially, the KL-divergence loss encourages the model to match this spread.

Overall, both the text-only and image-only predictions show varied emotion assignments; text-only dedicates a larger fraction to *anger* and *sadness*, while image-only shifts more heavily to *fear* and *surprise.* However, by combining both modalities, the multimodal system manages to produce a distribution that remains relatively balanced across the top four or five emotions. Indeed, our multimodal system successfully produced a more balanced distribution—resulting in lower KL divergence and better performance overall.

Case D: Tweet ID 1165556946180694016

Text Snippet:

> Democratic National Committee votes against allowing 2020 candidates
> to participate in a climate change debate [URL]

Aggregated Reply Distribution (multiple replies):

$$\{\text{anger: } 0.2326, \quad \text{fear: } 0.0911, \quad \text{surprise: } 0.1422,$$
$$\text{disgust: } 0.2849, \quad \text{sadness: } 0.0885, \quad \text{joy: } 0.1607\}$$

Our Predictions (Text-only):

$$\{\text{anger: } 0.1660, \quad \text{fear: } 0.1510, \quad \text{surprise: } 0.1761,$$
$$\text{disgust: } 0.1281, \quad \text{sadness: } 0.1922, \quad \text{joy: } 0.1866\}$$

Our Predictions (Image-only):

$$\{\text{anger: } 0.1749, \quad \text{fear: } 0.1821, \quad \text{surprise: } 0.2258,$$
$$\text{disgust: } 0.0994, \quad \text{sadness: } 0.0946, \quad \text{joy: } 0.2232\}$$

Our Predictions (Multimodal):

$$\{\text{anger: } 0.2084, \quad \text{fear: } 0.1822, \quad \text{surprise: } 0.1753,$$
$$\text{disgust: } 0.0910, \quad \text{sadness: } 0.0857, \quad \text{joy: } 0.2575\}$$

Here, we see a diverse mixture of emotions across the replies—many users felt *anger* or *disgust* (finding the decision frustrating), a few expressed *surprise*, and others found *joy* or positivity in certain commentary:

> *“This is stupid,”*
> *“Why let us hear what each [candidate] has to contribute?”*
> *“I appreciate their effort in losing as many elections as possible...”*

Since multiple replies contributed to the final distribution, it exhibited relatively high entropy. Our text-only predictions emphasized *sadness* (19.2%) more than the aggregated replies, while our image-only model gave heavier weight to *fear* (18.2%) and *surprise* (22.6%). Ultimately, the multimodal approach balanced these signals, producing a smoothed distribution with strong peaks in *anger* (20.8%) and *joy* (25.8%), aligning well with the diverse nature of user replies.

Hence, in cases of **diverse or high-entropy user replies**, the model excels at mimicking the broad emotional composition, lowering KL divergence.

Illustration of Noisy or Poorly Written Replies

While many replies are coherent, we also observed *poorly written* or difficult-to-parse responses. Such messages can introduce errors or unusual probability spikes when aggregated. For instance, consider a few actual examples from our dataset (unedited):

> *“BASIL AMAELO????”*
> *“nibiu close eath”*
> *“dnc no commentsame trough”*

The model sometimes interprets these short, ambiguous, or grammatically unclear statements as expressing *anger* or *disgust*, based on certain keywords or negative connotations (e.g. “no comment” or random capital letters). Other times, the mention of a presumably serious topic (“energy production”) may inflate *fear* or *surprise.* When such replies are averaged with others, they can skew the final aggregated distribution and increase error—particularly if there are only a few total replies.

4.9.1. Key Findings

Impact of Reply Entropy on Model Predictions

The variability in human reply distributions plays a critical role in model alignment with soft labels:

- **High Entropy → Better Alignment:** When replies exhibit diverse emotional reactions (e.g., Cases C and D), the model effectively replicates the distribution, accommodating overlapping sentiments such as fear, disgust, sadness, and joy.

- **Low Entropy → Greater Divergence:** When replies predominantly express one or two emotions (e.g., Cases A and B), the model tends to over-distribute probabilities, leading to prediction "flattening" and higher KL divergence. This occurs despite strong human consensus on emotions like fear, sadness, or surprise.

Additionally, the training objective plays a crucial role in shaping the model's behaviour:

- The **KL divergence loss discourages overconfidence**, ensuring that when human responses are uncertain or divided, the model produces softer distributions.
- However, when replies overwhelmingly agree on a single emotion, the model often spreads its predictions too broadly, underestimating the true peak.

Text and Image Ambiguity: A Secondary Factor

While sarcasm, slang, or ambiguous imagery can contribute to misclassifications, these factors were not the primary drivers of errors. Instead, **the shape of reply distributions and the model's training objective exerted a stronger influence**, as confirmed by both numerical trends and qualitative inspection of user responses.

4.10. Conclusion

Fine-tuning provides marked improvements over zero-shot classification for both text and image modalities, significantly reducing KL divergence and increasing cosine similarity scores. The additional benefits from multimodal fusion confirm that integrating visual and textual information can yield more robust representations.

However, these gains also highlight emerging challenges—such as minority label collapse and over-reliance on dominant emotional categories—that warrant deeper exploration. In the next chapter, we delve into a comprehensive analysis of these issues, examining the interplay between distribution-level metrics and instance-level accuracy, as well as the broader implications for real-world emotion classification tasks.

5. Discussion

In this chapter, we delve into the challenges and outcomes of emotion classification to social media data—specifically climate-related posts where emotional context can be nuanced and diverse. We begin by examining the baseline performance of various large language models, both domain-specific (CardiffNLP RoBERTa models) and general-purpose (e.g., DeBERTa-v3-large), highlighting how domain adaptation can yield measurable improvements in Exact Match and Top-3 Accuracy.

Next, we discuss the motivation for consolidating eleven initially predicted labels into Ekman's six basic emotions (anger, fear, surprise, disgust, sadness, joy). This step enhances interpretability and consistency in a space often plagued by class imbalance and label ambiguity. We then present detailed experimental results showing that while distribution-level improvements (e.g., higher cosine similarity) are attainable through label aggregation and soft-label averaging, these gains do not necessarily translate to accurate per-instance predictions. Confusion matrices and Mean Squared Error (MSE) analyses reveal systemic biases across text-only, image-only, and multimodal models. Finally, we propose future directions—ranging from hybrid loss functions and dynamic fusion mechanisms to refined label aggregation methods—that aim to mitigate minority label collapse and address the tension between distribution alignment and real-world applicability.

By considering these findings and open challenges, this chapter sets the stage for improving zero-shot and minimally supervised approaches in emotion classification. The goal is to inform both researchers and practitioners about the multifaceted nature of model performance, ensuring that future efforts balance distribution-level fidelity with fine-grained predictive accuracy.

5.1. Analysis of Findings

5.1.1. Zero-shot performance

The results align with expectations, as the CardiffNLP RoBERTa models, pre-trained on Twitter data, outperformed general-purpose models in emotion classification. Their higher Exact Match and Top-3 Accuracy confirm the advantage of domain-specific pre-training for social media text.

Applying a confidence threshold of 0.9 improved precision but significantly reduced the number of retained predictions. The DeBERTa-v3-large-Zeroshot model had the highest precision (72.73%) but generated fewer confident predictions, indicating greater uncertainty. In contrast, CardiffNLP models retained the most high-confidence predictions (63%), demonstrating both reliability and coverage.

These findings reinforce that while zero-shot models benefit from domain adaptation, moderate Exact Match scores (65%) suggest they still struggle with emotional nuances. Moreover, the best zero-shot model i.e. CardiffNLP RoBERTa-Large, predicted over 11 different labels, leading to inconsistencies in label distribution and interpretation. To address this, we consolidated the predictions into a more structured and theoretically grounded framework.

5.1.2. Experimental Results: Distribution vs. Instance-Level Accuracy

Label Aggregation and Distribution Flattening

The manual mapping of emotions (e.g., merging *love*, *optimism*, and *trust* into *joy*) introduced semantic ambiguity, as evidenced by the confusion matrices (Figures **??**–**??**). For instance:

Overprediction of Merged Classes: *joy* and *surprise* (which absorbed multiple original labels) dominated predictions. For our best text-only model, *joy* had 162 correct predictions but 149 false positives (e.g., *anger*/*sadness* misclassified as *joy*). Qualitative examples highlight this confusion: a tweet about climate strikes (Case B) elicited replies dominated by *surprise* (62.4%), but the model redistributed probabilities to *fear* and *anger* (Figure 4.6), conflating semantically distinct emotions. Similarly, *surprise* (merged with anticipation) was overpredicted for *anger* (22/44 instances) and *sadness* (43/83).

Minority Class Collapse: Rare emotions like *fear* and *sadness* suffered catastrophic recall (< 10% in our best multimodal model), with *fear* predictions often collapsing into *surprise* or *joy*. For example, in Case A (Figure 4.6), a reply explicitly describing wildfires in sub-Saharan Africa as “fires broke out in entire countries” was assigned only 27% probability to *fear*—half the ground truth value (48.8%)—with excess mass allocated to *surprise* and *anger*. The MSE values for *sadness* (0.0069–0.0162) and *fear* (0.0120–0.0293) were consistently low, suggesting the model learned to minimize their contributions to the loss by predicting near-zero probabilities.

The fine-tuning setup with KL divergence loss exacerbated this by prioritizing smooth, distribution-wide fidelity over per-class accuracy. Case B (Figure 4.6) exemplifies this: the model’s prediction for a reply expressing 62.4% *surprise* was flattened to 46.5%, with the remaining probability spread across unrelated classes like *fear* and *anger*. Averaging labels across replies further flattened the target distributions, incentivizing models to avoid confident predictions for minority classes.

Modality-Specific Biases and Error Propagation

The MSE tables highlight modality-specific weaknesses:

Text Models: Achieved the lowest *anger* MSE (0.0169 vs. baseline 0.0224) but failed to classify *anger* instances (2/44 correct in our best text-only model). Case D illustrates this: while textual cues like "This is stupid" (expressing anger toward political decisions) were present, the text model conflated *anger* with *disgust* due to overlapping lexicon.

Image Models: Overpredicted *surprise* (361 false positives), likely due to CLIP's bias toward visually salient cues. In Case C (Figure 4.7), the Arctic ice melt graph image drove surprise overprediction (46.5%) despite replies emphasizing *fear*—a disconnect between visual salience and textual sentiment.

Multimodal Fusion: While our best multimodal model reduced *anger* MSE (0.0120 vs. baseline 0.0278), confusion matrices reveal it inherited text's overprediction of *surprise*. Case C demonstrates this: the fusion model assigned 27.7% probability to *surprise* (matching the aggregated replies) but misclassified *fear* (14.8% vs. 27.7% ground truth), reflecting unresolved cross-modal conflicts.

Metric Misalignment and Practical Implications

The discrepancy between cosine similarity (aggregate distribution alignment) and confusion matrices (instance-level errors) raises questions about metric suitability:

Cosine Similarity rewards global shape matching but masks critical misclassifications. Case C exemplifies this: high cosine similarity was achieved by matching the spread of *joy*, *surprise*, and *fear*, but the model failed to detect the true intensity of *fear* (14.8% vs. 27.7%).

MSE penalizes large deviations but is insensitive to class swaps. In Case D, the model's confusion between *anger* and *disgust* (both high-probability classes) resulted in low MSE despite misclassifications critical for tracking public sentiment.

Overfitting and Generalization Trends

The observed train and validation loss curves further confirm the overfitting tendency of tuned models as shown in Tables (**??**–4.14). Typically, when models are tuned, the training loss decreases to very low values, whereas frozen models exhibit a much higher loss that decreases more slowly. The validation loss curve shows the opposite effect, reinforcing the hypothesis that tuning leads to overfitting. For a detailed visualization of these trends, refer to the loss curve images in the Appendix chapter C.

For real-world applications (e.g., tracking *anger* in crisis responses), these metrics poorly reflect operational needs. A model with high cosine similarity could still fail to detect critical emotions.

5.2. Future Directions

To address the identified limitations, we propose six research directions combining methodological innovation with rigorous evaluation:

- **Hybrid and Weighted Loss Functions**
  - **Why**: Purely distribution-focused objectives (e.g., KL divergence) risk ”flattening” predictions and neglecting minority classes.
  - **How**: Combine these objectives with focal or class-weighted losses to emphasize labels like *sadness* or *fear*, which may be underrepresented even in zero-shot outputs.

$$\mathcal{L} = \alpha D_{KL}(P \parallel Q) + (1 - \alpha)\mathcal{L}_{\text{Class-weighted}} \tag{5.1}$$

- **Refined Label Generation and Aggregation**
  - **Why**: Relying solely on zero-shot models for soft labels can propagate upstream biases and simple aggregation strategies like averaging can magnify them (e.g., merging *anticipation* and *surprise*).
  - **How**: Consider more nuanced label aggregation (e.g., attention-weighted merging) to preserve emotional specificity, or incorporate partial human labeling to calibrate and refine zero-shot predictions.
- **Multi-Task and Disentangled Training**
  - **Why**: A single objective often cannot balance overall distribution alignment with per-class accuracy, especially when label quality is uncertain.
  - **How**: First optimize on the full, zero-shot-labeled dataset for broad coverage, then fine-tune on smaller, more carefully curated or partially human-annotated subsets to correct for minority-class underrepresentation.
- **Dynamic Fusion and Modality Gating**
  - **Why**: Text and image modalities provide complementary signals, but the confidence of zero-shot models may vary across modalities.
  - **How**: Implement gating mechanisms or attention-based weighting to override misleading cues in one modality when the other provides stronger evidence.

$$w_{\text{text}}, w_{\text{image}} = f_{\text{attention}}(s_{\text{text}}, s_{\text{image}}) \tag{5.2}$$

- **Enhanced Evaluation Protocols**
  - **Why**: Metrics that only measure distribution alignment can mask performance on minority classes and overestimate real-world suitability.

  - **How**: Pair distribution-based metrics (KL divergence, Earth Mover's Distance) with per-class indicators (macro-F1) to surface critical performance gaps.

- **Metadata and Bias Mitigation**
  - **Why**: Zero-shot models may reflect cultural or demographic biases in their pretrained distributions.
  - **How**: Incorporate metadata (e.g., annotator demographics or domain context) when available, to model and mitigate hidden biases within soft-label predictions.

By coupling zero-shot label generation with carefully chosen loss functions, gating strategies, and evaluation frameworks, future research can reduce the risk of "flattened" distributions and minority label collapse. This holistic approach is pivotal for modelling real-world emotional complexity, ensuring that both distribution-level and instance-level performance remains robust—despite the challenges of working with fully unlabelled datasets

6. Conclusion

6.1. Summary of the Thesis

Bibliography

Al-Halah, Z., A. Aitken, W. Shi, and J. Caballero (2019). Smile, be happy :) emoji embedding for visual sentiment analysis. In *IEEE International Conference on Computer Vision Workshops.*

Antypas, D., A. Ushio, F. Barbieri, L. Neves, K. Rezaee, L. Espinosa-Anke, J. Pei, and J. Camacho-Collados (2023). Supertweeteval: A challenging, unified and heterogeneous benchmark for social media nlp research.

Baccianella, S., A. Esuli, and F. Sebastiani (2010, May). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Bahdanau, D., K. Cho, and Y. Bengio (2014). Neural machine translation by jointly learning to align and translate. *CoRR abs/1409.0473.*

Baltrušaitis, T., C. Ahuja, and L.-P. Morency (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence 41*(2), 423–443.

Chen, L.-M., B.-X. Xiu, and Z.-Y. Ding (2022, June). Multiple weak supervision for short text classification. *Applied Intelligence 52*(8), 9101–9116.

Cheng, Z., X. Bu, Q. Wang, T. Yang, and J. Tu (2024, December). EEG-based emotion recognition using multi-scale dynamic CNN and gated transformer. *Scientific Reports 14*(1), 31319. Publisher: Nature Publishing Group.

Cho, K., B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov (2019). Unsupervised cross-lingual representation learning at scale. *CoRR abs/1911.02116.*

Dalal, N. and B. Triggs (2005). Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Volume 1, pp. 886–893 vol. 1.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255.

Deng, Y., Y. Li, S. Xian, L. Li, and H. Qiu (2024, July). Mual: enhancing multimodal sentiment analysis with cross-modal attention and difference loss. *International Journal of Multimedia Information Retrieval 13*(3), 31.

Devlin, J., M. Chang, K. Lee, and K. Toutanova (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR abs/1810.04805*.

Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby (2021). An image is worth 16x16 words: Transformers for image recognition at scale.

Ekman, P. (1992). Are there basic emotions?

Gera, A., A. Halfon, E. Shnarch, Y. Perlitz, L. Ein-Dor, and N. Slonim (2022, December). Zero-Shot Text Classification with Self-Training. In Y. Goldberg, Z. Kozareva, and Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, pp. 1107–1119. Association for Computational Linguistics.

Graves, A. and J. Schmidhuber (2005). Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, Volume 4, pp. 2047–2052 vol. 4.

He, K., X. Zhang, S. Ren, and J. Sun (2015). Deep residual learning for image recognition.

Hendrycks, D. and T. Dietterich (2019). Benchmarking neural network robustness to common corruptions and perturbations.

Hendrycks, D., K. Zhao, S. Basart, J. Steinhardt, and D. Song (2021). Natural adversarial examples.

Hochreiter, S. and J. Schmidhuber (1997, November). Long short-term memory. *Neural Comput. 9*(8), 1735–1780.

Hu, M., H. Han, S. Shan, and X. Chen (2019, June). Weakly Supervised Image Classification Through Noise Regularization. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11509–11517. ISSN: 2575-7075.

Krasin, I., T. Duerig, N. Alldrin, A. Veit, S. Abu-El-Haija, S. Belongie, D. Cai, Z. Feng, V. Ferrari, and V. Gomes (2016, 01). Openimages: A public dataset for large-scale multi-label and multi-class image classification.

Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, Red Hook, NY, USA, pp. 1097–1105. Curran Associates Inc.

Laurer, M., W. v. Atteveldt, A. S. Casas, and K. Welbers (2022, June). Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI. *Preprint*. Publisher: Open Science Framework.

Laurer, M., W. van Atteveldt, A. Casas, and K. Welbers (2023, December). Building Efficient Universal Classifiers with Natural Language Inference. arXiv:2312.17543 [cs].

LeCun, Y., Y. Bengio, and G. Hinton (2015, May). Deep learning. *Nature 521*(7553), 436–444.

Lewis, M., Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Li, J., D. Li, C. Xiong, and S. Hoi (2022). Blip: Bootstrapping language-image pretraining for unified vision-language understanding and generation.

Li, J., R. R. Selvaraju, A. D. Gotmare, S. Joty, C. Xiong, and S. Hoi (2021). Align before fuse: Vision and language representation learning with momentum distillation.

Li, L. H., M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang (2019). Visualbert: A simple and performant baseline for vision and language.

Li, S. and H. Tang (2024, November). Multimodal Alignment and Fusion: A Survey. arXiv:2411.17040 [cs].

Limami, F., B. Hdioud, and R. Oulad Haj Thami (2024, June). Contextual emotion detection in images using deep learning. *Frontiers in Artificial Intelligence 7*. Publisher: Frontiers.

Lin, T.-Y., M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár (2015). Microsoft coco: Common objects in context.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR abs/1907.11692*.

Liu, Z., Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo (2021). Swin transformer: Hierarchical vision transformer using shifted windows.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision 60*, 91–110.

Mahajan, D., R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten (2018). Exploring the Limits of Weakly Supervised Pretraining. pp. 181–196.

Mikolov, T., M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur (2010). Recurrent Neural Network Based Language Model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association*, INTERSPEECH 2010, pp. 1045–1048. ISCA.

Mooseder, A., C. Brantner, R. Zamith, and J. Pfeffer (2023). (Social) Media Logics and Visualizing Climate Change: 10 Years of #climatechange Images on Twitter. *Social Media + Society 9*(1), 20563051231164310. _eprint: https://doi.org/10.1177/20563051231164310.

Nguyen, D. Q., T. Vu, and A. Tuan Nguyen (2020, October). BERTweet: A pre-trained language model for English tweets. In Q. Liu and D. Schlangen (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online, pp. 9–14. Association for Computational Linguistics.

Pawłowski, M., A. Wróblewska, and S. Sysko-Romańczuk (2023, January). Effective Techniques for Multimodal Data Fusion: A Comparative Analysis. *Sensors 23*(5), 2381. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

Poria, S., E. Cambria, R. Bajpai, and A. Hussain (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion 37*, 98–125.

Prasse, K., S. Jung, I. B. Bravo, S. Walter, and M. Keuper (2023). Towards understanding climate change perceptions: A social media dataset. In *NeurIPS 2023 Workshop on Tackling Climate Change with Machine Learning.*

Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever (2021). Learning transferable visual models from natural language supervision.

Radford, A. and K. Narasimhan (2018). Improving language understanding by generative pre-training.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu (2023). Exploring the limits of transfer learning with a unified text-to-text transformer.

Sanh, V., L. Debut, J. Chaumond, and T. Wolf (2020). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Sariyanidi, E., H. Gunes, and A. Cavallaro (2015). Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence 37*(6), 1113–1133.

Shao, B., L. Doucet, and D. R. Caruso (2015). Universality versus cultural specificity of three emotion domains. *Journal of Cross-Cultural Psychology 46*, 229 – 251.

Simonyan, K. and A. Zisserman (2015). Very deep convolutional networks for large-scale image recognition.

Song, H., M. Kim, D. Park, Y. Shin, and J.-G. Lee (2022, March). Learning from Noisy Labels with Deep Neural Networks: A Survey. arXiv:2007.08199.

Soni, J., N. Prabakar, and H. Upadhyay (2024). Vision Transformer-Based Emotion Detection in HCI for Enhanced Interaction. In B. J. Choi, D. Singh, U. S. Tiwary, and W.-Y. Chung (Eds.), *Intelligent Human Computer Interaction*, Cham, pp. 76–86. Springer Nature Switzerland.

Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna (2015). Rethinking the inception architecture for computer vision.

Tan, H. and M. Bansal (2019). Lxmert: Learning cross-modality encoder representations from transformers.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2023). Attention is all you need.

Xiao, T., T. Xia, Y. Yang, C. Huang, and X. Wang (2015). Learning from massive noisy labeled data for image classification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2691–2699.

Xie, Q., M.-T. Luong, E. Hovy, and Q. V. Le (2020, June). Self-training with Noisy Student improves ImageNet classification. arXiv:1911.04252.

Yin, W., J. Hay, and D. Roth (2019). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *CoRR abs/1909.00161.*

Yosinski, J., J. Clune, Y. Bengio, and H. Lipson (2014). How transferable are features in deep neural networks?

Zeng, Z., W. Ni, T. Fang, X. Li, X. Zhao, and Y. Song (2022, May). Weakly Supervised Text Classification using Supervision Signals from a Language Model. arXiv:2205.06604.

Zhang, F., T. Zhang, Q. Mao, L. Duan, and C. Xu (2018). Facial expression recognition in the wild: A cycle-consistent adversarial attention transfer approach. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, New York, NY, USA, pp. 126–135. Association for Computing Machinery.

Zhang, X., J. J. Zhao, and Y. LeCun (2015). Character-level convolutional networks for text classification. *CoRR abs/1509.01626.*

Zhu, D., X. Shen, M. Mosbach, A. Stephan, and D. Klakow (2023, July). Weaker Than You Think: A Critical Look at Weakly Supervised Learning. In A. Rogers, J. Boyd-Graber, and N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada, pp. 14229–14253. Association for Computational Linguistics.

Zhu, T., L. Li, J. Yang, S. Zhao, and X. Xiao (2023). Multimodal emotion classification with multi-level semantic reasoning network. *IEEE Transactions on Multimedia 25*, 6868–6880.

A. Program Code / Resources

The source code and additional experimental results are available at the Github Repository https://github.com/tyagiprnv/masterthesis

B. Complete Experimental Results

Table B.1.: Complete multimodal experimental results

Experiment	Test/Cosine_Sim	Test/KL_Div	Test/MSE	Test/Cosine_Sim_Norm	Test/MSE_Norm	Test/KL_Div_Norm	Score
exp_adamw_roberta_large_lr1e-05_drop0.5_bigger_mlp_epochs5_both_frozen	0.8126	0.3363	0.0224	0.8126	0.9776	0.6637	2.4539
exp_adamw_roberta_base_lr5e-06_drop0.3_gelu_epochs2	0.8115	0.337	0.0223	0.8115	0.9777	0.663	2.4522
exp_adamw_roberta_large_lr1e-05_drop0.5_gelu_epochs5_both_frozen	0.8114	0.3385	0.0225	0.8114	0.9775	0.6615	2.4504
exp_adamw_roberta_base_lr5e-06_drop0.5_epochs2_frozen_clip	0.812	0.3393	0.0225	0.812	0.9775	0.6607	2.4502
exp_adamw_roberta_large_lr1e-05_drop0.3_gelu_epochs5_both_frozen	0.811	0.3383	0.0225	0.811	0.9775	0.6617	2.4501
exp_adamw_roberta_base_lr1e-05_drop0.3_epochs2_frozen_clip	0.8119	0.3396	0.0226	0.8119	0.9774	0.6604	2.4497
exp_adamw_roberta_base_lr5e-06_drop0.3_epochs2	0.8103	0.3384	0.0225	0.8103	0.9775	0.6616	2.4494
exp_adamw_roberta_base_lr5e-06_drop0.5_gelu_epochs2	0.8097	0.338	0.0224	0.8097	0.9776	0.662	2.4493
exp_adamw_roberta_large_lr1e-05_drop0.3_bigger_mlp_epochs5_both_frozen	0.8104	0.3387	0.0226	0.8104	0.9774	0.6613	2.4491
exp_adamw_roberta_base_lr1e-05_drop0.3_bigger_mlp_epochs2	0.8111	0.3395	0.0225	0.8111	0.9775	0.6605	2.4491
exp_adamw_roberta_large_lr5e-06_drop0.5_epochs2_frozen_clip	0.8097	0.3383	0.0224	0.8097	0.9776	0.6617	2.4489
exp_adamw_roberta_base_lr5e-06_drop0.5_bigger_mlp_epochs2	0.8094	0.3382	0.0224	0.8094	0.9776	0.6618	2.4487
exp_adamw_roberta_large_lr1e-05_drop0.3_epochs5_both_frozen	0.8106	0.3396	0.0226	0.8106	0.9774	0.6604	2.4485
exp_adamw_roberta_base_lr1e-05_drop0.5_epochs2_frozen_clip	0.8112	0.3401	0.0226	0.8112	0.9774	0.6599	2.4485

Continued on next page

Table B.1 Continued from previous page

Experiment	Test/Cosine_Sim	Test/KL_Div	Test/MSE	Test/Cosine_Sim_Norm	Test/MSE_Norm	Test/KL_Div_Norm	Score
exp_adamw_roberta_large_lr5e-05_drop0.3_hierarchical_epochs2_frozen_clip	0.8114	0.3407	0.0226	0.8114	0.9774	0.6593	2.4481
exp_adamw_roberta_base_lr5e-06_drop0.5_epochs2	0.8098	0.3392	0.0226	0.8098	0.9774	0.6608	2.4481
exp_adamw_roberta_base_lr1e-05_drop0.3_gelu_epochs2	0.8108	0.3402	0.0226	0.8108	0.9774	0.6598	2.448
exp_adamw_roberta_large_lr5e-06_drop0.5_gelu_epochs2	0.8096	0.3391	0.0226	0.8096	0.9774	0.6609	2.448
exp_adamw_roberta_large_lr5e-06_drop0.5_epochs2	0.8085	0.3381	0.0225	0.8085	0.9775	0.6619	2.4479
exp_adamw_roberta_base_lr1e-05_drop0.5_bigger_mlp_epochs2	0.8111	0.3406	0.0226	0.8111	0.9774	0.6594	2.4479
exp_adamw_roberta_base_lr1e-05_drop0.3_gelu_epochs5_both_frozen	0.8104	0.34	0.0226	0.8104	0.9774	0.66	2.4478
exp_adamw_roberta_large_lr5e-06_drop0.3_gelu_epochs5_both_frozen	0.8101	0.34	0.0226	0.8101	0.9774	0.66	2.4475
exp_roberta_large_lr1e-05_drop0.3_epochs5_both_frozen	0.8106	0.3406	0.0226	0.8106	0.9774	0.6594	2.4474
exp_adamw_roberta_base_lr1e-05_drop0.5_gelu_epochs2	0.8102	0.3404	0.0226	0.8102	0.9774	0.6596	2.4472
exp_roberta_large_lr1e-05_drop0.5_epochs5_both_frozen	0.8105	0.3408	0.0227	0.8105	0.9773	0.6592	2.447
exp_adamw_roberta_base_lr5e-06_drop0.3_epochs2_frozen_clip	0.8101	0.3406	0.0227	0.8101	0.9773	0.6594	2.4468
exp_roberta_base_lr5e-06_drop0.5_epochs2	0.8104	0.341	0.0226	0.8104	0.9774	0.659	2.4468
exp_adamw_roberta_base_lr1e-05_drop0.3_bigger_mlp_epochs2_frozen_clip	0.8094	0.3401	0.0227	0.8094	0.9773	0.6599	2.4467
exp_adamw_roberta_base_lr1e-05_drop0.3_bigger_mlp_epochs5_both_frozen	0.8094	0.3403	0.0227	0.8094	0.9773	0.6597	2.4464
exp_adamw_roberta_large_lr5e-06_drop0.3_bigger_mlp_epochs5_both_frozen	0.8094	0.3404	0.0227	0.8094	0.9773	0.6596	2.4464
exp_adamw_roberta_base_lr5e-06_drop0.3_bigger_mlp_epochs2_frozen_clip	0.8089	0.34	0.0226	0.8089	0.9774	0.66	2.4462
exp_adamw_roberta_base_lr5e-06_drop0.5_bigger_mlp_epochs2_frozen_clip	0.8097	0.341	0.0226	0.8097	0.9774	0.659	2.4461
exp_adamw_roberta_base_lr1e-05_drop0.5_gelu_epochs2_frozen_clip	0.809	0.3404	0.0226	0.809	0.9774	0.6596	2.446
exp_adamw_roberta_base_lr1e-05_drop0.5_gelu_epochs5_both_frozen	0.8098	0.3411	0.0227	0.8098	0.9773	0.6589	2.446

Continued on next page

Table B.1 Continued from previous page

Experiment	Test/Cosine_Sim	Test/KL_Div	Test/MSE	Test/Cosine_Sim_Norm	Test/MSE_Norm	Test/KL_Div_Norm	Score
exp_adamw_roberta_base_lr1e-05_drop0.5_bigger_mlp_epochs2_frozen_clip	0.8089	0.3403	0.0226	0.8089	0.9774	0.6597	2.446
exp_adamw_roberta_large_lr5e-06_drop0.5_gelu_epochs5_both_frozen	0.8098	0.3412	0.0227	0.8098	0.9773	0.6588	2.4459
exp_roberta_base_lr1e-05_drop0.3_epochs5_both_frozen	0.8098	0.3413	0.0226	0.8098	0.9774	0.6587	2.4459
exp_adamw_roberta_base_lr5e-06_drop0.5_gelu_epochs2_frozen_clip	0.809	0.3407	0.0226	0.809	0.9774	0.6593	2.4457
exp_roberta_base_lr1e-05_drop0.5_epochs5_both_frozen	0.81	0.3416	0.0226	0.81	0.9774	0.6584	2.4457
exp_adamw_roberta_base_lr5e-06_drop0.3_epochs2_frozen_roberta	0.8105	0.342	0.0228	0.8105	0.9772	0.658	2.4457
exp_adamw_roberta_large_lr5e-06_drop0.5_gelu_epochs2_frozen_roberta	0.8088	0.3409	0.0226	0.8088	0.9774	0.6591	2.4454
exp_adamw_roberta_large_lr5e-06_drop0.3_epochs5_both_frozen	0.8095	0.3418	0.0227	0.8095	0.9773	0.6582	2.445
exp_adamw_roberta_base_lr1e-05_drop0.5_epochs5_both_frozen	0.809	0.3414	0.0227	0.809	0.9773	0.6586	2.4449
exp_adamw_roberta_base_lr1e-05_drop0.3_epochs5_both_frozen	0.8088	0.3412	0.0227	0.8088	0.9773	0.6588	2.4449
exp_roberta_large_lr5e-06_drop0.5_epochs2_frozen_clip	0.8079	0.3405	0.0226	0.8079	0.9774	0.6595	2.4448
exp_adamw_roberta_large_lr1e-05_drop0.5_epochs5_frozen_roberta	0.8094	0.3419	0.0227	0.8094	0.9773	0.6581	2.4447
exp_adamw_roberta_large_lr1e-05_drop0.3_epochs2_both_frozen	0.8092	0.3418	0.0227	0.8092	0.9773	0.6582	2.4447
exp_adamw_roberta_large_lr5e-06_drop0.3_epochs2_frozen_clip	0.8075	0.3403	0.0226	0.8075	0.9774	0.6597	2.4446
exp_adamw_roberta_base_lr1e-05_drop0.3_epochs2_both_frozen	0.8103	0.3429	0.0228	0.8103	0.9772	0.6571	2.4446
exp_adamw_roberta_base_lr1e-05_drop0.5_bigger_mlp_epochs5_both_frozen	0.8088	0.3416	0.0227	0.8088	0.9773	0.6584	2.4445
exp_adamw_roberta_large_lr5e-06_drop0.5_epochs5_both_frozen	0.8095	0.3423	0.0227	0.8095	0.9773	0.6577	2.4445
exp_roberta_base_lr5e-06_drop0.3_epochs2	0.8096	0.3424	0.0227	0.8096	0.9773	0.6576	2.4445
exp_adamw_roberta_large_lr5e-06_drop0.5_bigger_mlp_epochs5_both_frozen	0.8092	0.3422	0.0227	0.8092	0.9773	0.6578	2.4443
exp_adamw_roberta_large_lr1e-05_drop0.5_epochs2_frozen_clip	0.8076	0.3407	0.0226	0.8076	0.9774	0.6593	2.4443

Continued on next page

Table B.1 Continued from previous page

Experiment	Test/Cosine_Sim	Test/KL_Div	Test/MSE	Test/Cosine_Sim_Norm	Test/MSE_Norm	Test/KL_Div_Norm	Score
exp_adamw_roberta_base_lr5e-06_drop0.3_gelu_epochs2_frozen_clip	0.8078	0.341	0.0227	0.8078	0.9773	0.659	2.4442
exp_adamw_roberta_large_lr5e-06_drop0.3_bigger_mlp_epochs2_frozen_roberta	0.8075	0.341	0.0226	0.8075	0.9774	0.659	2.4439
exp_roberta_large_lr1e-05_drop0.5_epochs2	0.8084	0.3418	0.0227	0.8084	0.9773	0.6582	2.4438
exp_adamw_roberta_large_lr5e-06_drop0.5_epochs2_frozen_roberta	0.8082	0.3417	0.0227	0.8082	0.9773	0.6583	2.4438
exp_roberta_large_lr5e-06_drop0.5_epochs2	0.8072	0.3408	0.0227	0.8072	0.9773	0.6592	2.4437
exp_adamw_roberta_base_lr1e-05_drop0.3_gelu_epochs2_frozen_clip	0.8084	0.3421	0.0227	0.8084	0.9773	0.6579	2.4436
exp_adamw_roberta_large_lr1e-05_drop0.5_epochs2_both_frozen	0.809	0.3428	0.0227	0.809	0.9773	0.6572	2.4434
exp_adamw_roberta_base_lr5e-06_drop0.3_bigger_mlp_epochs2	0.8077	0.3417	0.0228	0.8077	0.9772	0.6583	2.4433
exp_adamw_roberta_large_lr5e-06_drop0.3_gelu_epochs2_frozen_roberta	0.8077	0.3417	0.0227	0.8077	0.9773	0.6583	2.4432
exp_adamw_roberta_large_lr5e-05_drop0.3_hierarchical_epochs2_both_frozen	0.809	0.343	0.0229	0.809	0.9771	0.657	2.4432
exp_adamw_roberta_base_lr5e-06_drop0.5_epochs2_frozen_roberta	0.8098	0.344	0.0229	0.8098	0.9771	0.656	2.4429
exp_adamw_roberta_base_lr1e-05_drop0.5_epochs2_both_frozen	0.8101	0.3443	0.0229	0.8101	0.9771	0.6557	2.4428
exp_roberta_large_lr5e-06_drop0.3_epochs5_both_frozen	0.809	0.3434	0.0228	0.809	0.9772	0.6566	2.4428
exp_adamw_roberta_large_lr1e-05_drop0.5_gelu_epochs5_frozen_roberta	0.8086	0.343	0.0228	0.8086	0.9772	0.657	2.4428
exp_adamw_roberta_base_lr5e-06_drop0.3_gelu_epochs5_both_frozen	0.8086	0.3431	0.0228	0.8086	0.9772	0.6569	2.4426
exp_roberta_large_lr5e-06_drop0.5_epochs5_both_frozen	0.8089	0.3437	0.0229	0.8089	0.9771	0.6563	2.4423
exp_adamw_roberta_large_lr1e-05_drop0.3_bigger_mlp_epochs2_both_frozen	0.8078	0.3427	0.0228	0.8078	0.9772	0.6573	2.4422
exp_adamw_roberta_large_lr5e-05_drop0.5_hierarchical_epochs2_frozen_clip	0.81	0.345	0.0229	0.81	0.9771	0.655	2.4422
exp_roberta_large_lr5e-06_drop0.5_epochs2_frozen_roberta	0.807	0.3422	0.0227	0.807	0.9773	0.6578	2.4421
exp_adamw_roberta_base_lr5e-06_drop0.3_bigger_mlp_epochs5_both_frozen	0.8078	0.343	0.0228	0.8078	0.9772	0.657	2.442

Continued on next page

Table B.1 Continued from previous page

Experiment	Test/Cosine_Sim	Test/KL_Div	Test/MSE	Test/Cosine_Sim_Norm	Test/MSE_Norm	Test/KL_Div_Norm	Score
exp_roberta_large_lr1e-05_drop0.3_epochs2_both_frozen	0.8084	0.3436	0.0228	0.8084	0.9772	0.6564	2.4419
exp_roberta_base_lr5e-06_drop0.3_epochs5_both_frozen	0.8086	0.3439	0.0228	0.8086	0.9772	0.6561	2.4419
exp_adamw_roberta_base_lr1e-05_drop0.3_gelu_epochs2_frozen_roberta	0.8075	0.3433	0.0228	0.8075	0.9772	0.6567	2.4414
exp_adamw_roberta_base_lr5e-06_drop0.3_epochs5_both_frozen	0.8078	0.3436	0.0229	0.8078	0.9771	0.6564	2.4413
exp_adamw_roberta_large_lr1e-05_drop0.3_epochs2	0.8067	0.3428	0.0227	0.8067	0.9773	0.6572	2.4412
exp_roberta_base_lr1e-05_drop0.3_epochs2_frozen_clip	0.8081	0.3441	0.0228	0.8081	0.9772	0.6559	2.4411
exp_adamw_roberta_large_lr1e-05_drop0.5_gelu_epochs2_frozen_roberta	0.8072	0.3432	0.0229	0.8072	0.9771	0.6568	2.4411
exp_roberta_base_lr5e-06_drop0.5_epochs2_frozen_clip	0.8074	0.3435	0.0228	0.8074	0.9772	0.6565	2.4411
exp_adamw_roberta_large_lr5e-06_drop0.3_gelu_epochs2	0.8069	0.3432	0.0229	0.8069	0.9771	0.6568	2.4408
exp_adamw_roberta_large_lr5e-06_drop0.5_bigger_mlp_epochs2_frozen_clip	0.807	0.3434	0.0228	0.807	0.9772	0.6566	2.4408
exp_roberta_large_lr1e-05_drop0.5_epochs2_both_frozen	0.808	0.3445	0.0229	0.808	0.9771	0.6555	2.4407
exp_adamw_roberta_base_lr5e-06_drop0.3_gelu_epochs2_frozen_roberta	0.8064	0.3429	0.0227	0.8064	0.9773	0.6571	2.4407
exp_adamw_roberta_large_lr1e-05_drop0.5_bigger_mlp_epochs2_both_frozen	0.8081	0.3445	0.0229	0.8081	0.9771	0.6555	2.4407
exp_roberta_base_lr5e-06_drop0.5_epochs5_both_frozen	0.8083	0.3448	0.0229	0.8083	0.9771	0.6552	2.4406
exp_adamw_roberta_large_lr1e-05_drop0.5_epochs2	0.8061	0.3427	0.0228	0.8061	0.9772	0.6573	2.4406
exp_adamw_roberta_large_lr5e-06_drop0.3_bigger_mlp_epochs2	0.8062	0.3427	0.0229	0.8062	0.9771	0.6573	2.4406
exp_adamw_roberta_large_lr1e-05_drop0.5_bigger_mlp_epochs2_frozen_clip	0.8075	0.3441	0.0229	0.8075	0.9771	0.6559	2.4405
exp_adamw_roberta_large_lr1e-05_drop0.3_gelu_epochs2_frozen_roberta	0.8064	0.3431	0.0229	0.8064	0.9771	0.6569	2.4404
exp_adamw_roberta_large_lr5e-06_drop0.5_bigger_mlp_epochs2_frozen_roberta	0.8072	0.3447	0.0228	0.8072	0.9772	0.6553	2.4397
exp_roberta_base_lr5e-06_drop0.5_epochs2_frozen_roberta	0.8077	0.3452	0.0229	0.8077	0.9771	0.6548	2.4396

Continued on next page

Table B.1 Continued from previous page

Experiment	Test/Cosine_Sim	Test/KL_Div	Test/MSE	Test/Cosine_Sim_Norm	Test/MSE_Norm	Test/KL_Div_Norm	Score
exp_roberta_base_lr1e-05_drop0.5_epochs2_frozen_clip	0.8073	0.345	0.0229	0.8073	0.9771	0.655	2.4394
exp_adamw_roberta_large_lr1e-05_drop0.3_bigger_mlp_epochs2_frozen_clip	0.8077	0.3454	0.0229	0.8077	0.9771	0.6546	2.4394
exp_adamw_roberta_base_lr1e-05_drop0.3_gelu_epochs2_both_frozen	0.8069	0.3446	0.0229	0.8069	0.9771	0.6554	2.4393
exp_adamw_roberta_base_lr1e-05_drop0.3_bigger_mlp_epochs2_both_frozen	0.8064	0.3443	0.0229	0.8064	0.9771	0.6557	2.4391
exp_adamw_roberta_base_lr5e-06_drop0.5_epochs5_both_frozen	0.807	0.3451	0.0229	0.807	0.9771	0.6549	2.439
exp_adamw_roberta_base_lr5e-06_drop0.5_gelu_epochs5_both_frozen	0.8073	0.3454	0.023	0.8073	0.977	0.6546	2.4389
exp_adamw_roberta_base_lr1e-05_drop0.5_bigger_mlp_epochs2_frozen_roberta	0.8056	0.344	0.0228	0.8056	0.9772	0.656	2.4387
exp_adamw_roberta_large_lr5e-06_drop0.3_epochs2	0.8048	0.3432	0.0229	0.8048	0.9771	0.6568	2.4387
exp_roberta_large_lr5e-06_drop0.3_epochs2_frozen_clip	0.8054	0.3439	0.0229	0.8054	0.9771	0.6561	2.4386
exp_roberta_base_lr1e-05_drop0.3_epochs2_both_frozen	0.8072	0.3456	0.0229	0.8072	0.9771	0.6544	2.4386
exp_adamw_roberta_large_lr1e-05_drop0.3_epochs2_frozen_clip	0.805	0.3436	0.0229	0.805	0.9771	0.6564	2.4385
exp_adamw_roberta_large_lr5e-06_drop0.3_epochs2_both_frozen	0.8074	0.3462	0.0229	0.8074	0.9771	0.6538	2.4383
exp_roberta_base_lr5e-06_drop0.3_epochs2_frozen_clip	0.8063	0.3452	0.023	0.8063	0.977	0.6548	2.4382
exp_adamw_roberta_large_lr1e-05_drop0.5_epochs5_both_frozen	0.8045	0.3436	0.0228	0.8045	0.9772	0.6564	2.4381
exp_adamw_roberta_large_lr5e-06_drop0.3_bigger_mlp_epochs2_both_frozen	0.8072	0.3463	0.023	0.8072	0.977	0.6537	2.438
exp_adamw_roberta_large_lr5e-05_drop0.5_hierarchical_epochs2_frozen_roberta	0.8071	0.3463	0.023	0.8071	0.977	0.6537	2.4378
exp_adamw_roberta_base_lr5e-06_drop0.5_bigger_mlp_epochs5_both_frozen	0.8063	0.3457	0.023	0.8063	0.977	0.6543	2.4376
exp_roberta_base_lr5e-06_drop0.3_epochs2_frozen_roberta	0.8073	0.3467	0.0231	0.8073	0.9769	0.6533	2.4375
exp_roberta_large_lr1e-05_drop0.5_epochs2_frozen_clip	0.8059	0.3454	0.023	0.8059	0.977	0.6546	2.4375
exp_roberta_base_lr1e-05_drop0.5_epochs2_both_frozen	0.807	0.3466	0.023	0.807	0.977	0.6534	2.4374

Continued on next page

Table B.1 Continued from previous page

Experiment	Test/Cosine_Sim	Test/KL_Div	Test/MSE	Test/Cosine_Sim_Norm	Test/MSE_Norm	Test/KL_Div_Norm	Score
exp_adamw_roberta_large_lr5e-05_drop0.5_hierarchical_epochs2	0.8079	0.3474	0.0231	0.8079	0.9769	0.6526	2.4374
exp_adamw_roberta_large_lr1e-05_drop0.3_bigger_mlp_epochs2_frozen_roberta	0.8053	0.3449	0.023	0.8053	0.977	0.6551	2.4373
exp_roberta_large_lr5e-06_drop0.3_epochs2	0.8054	0.3451	0.023	0.8054	0.977	0.6549	2.4373
exp_roberta_base_lr1e-05_drop0.5_epochs2	0.8073	0.347	0.0231	0.8073	0.9769	0.653	2.4373
exp_adamw_roberta_large_lr5e-05_drop0.5_hierarchical_epochs2_both_frozen	0.8064	0.3463	0.023	0.8064	0.977	0.6537	2.4371
exp_adamw_roberta_base_lr1e-05_drop0.5_epochs2_frozen_roberta	0.8085	0.3482	0.0232	0.8085	0.9768	0.6518	2.437
exp_adamw_roberta_base_lr5e-06_drop0.3_epochs2_both_frozen	0.8087	0.3486	0.0231	0.8087	0.9769	0.6514	2.437
exp_adamw_roberta_large_lr1e-05_drop0.5_bigger_mlp_epochs2	0.806	0.3461	0.023	0.806	0.977	0.6539	2.4368
exp_adamw_roberta_large_lr5e-06_drop0.5_bigger_mlp_epochs2	0.8055	0.3458	0.0231	0.8055	0.9769	0.6542	2.4367
exp_adamw_roberta_large_lr1e-05_drop0.5_gelu_epochs2	0.8061	0.3465	0.0231	0.8061	0.9769	0.6535	2.4365
exp_adamw_roberta_large_lr1e-05_drop0.5_bigger_mlp_epochs2_frozen_roberta	0.8056	0.3461	0.023	0.8056	0.977	0.6539	2.4365
exp_adamw_roberta_base_lr5e-06_drop0.3_bigger_mlp_epochs2_frozen_roberta	0.8043	0.3451	0.0229	0.8043	0.9771	0.6549	2.4362
exp_adamw_roberta_base_lr1e-05_drop0.5_gelu_epochs2_both_frozen	0.8061	0.3469	0.0231	0.8061	0.9769	0.6531	2.4362
exp_roberta_large_lr5e-06_drop0.3_epochs2_both_frozen	0.8067	0.3475	0.0231	0.8067	0.9769	0.6525	2.4362
exp_adamw_roberta_base_lr1e-05_drop0.3_epochs2	0.8064	0.3473	0.0231	0.8064	0.9769	0.6527	2.436
exp_adamw_roberta_base_lr1e-05_drop0.5_gelu_epochs2_frozen_roberta	0.8054	0.347	0.023	0.8054	0.977	0.653	2.4355
exp_adamw_roberta_base_lr1e-05_drop0.5_bigger_mlp_epochs2_both_frozen	0.8055	0.3471	0.0231	0.8055	0.9769	0.6529	2.4354
exp_adamw_roberta_base_lr1e-05_drop0.3_bigger_mlp_epochs2_frozen_roberta	0.8038	0.3455	0.023	0.8038	0.977	0.6545	2.4353
exp_adamw_roberta_large_lr5e-06_drop0.5_epochs2_both_frozen	0.8073	0.349	0.0231	0.8073	0.9769	0.651	2.4352
exp_adamw_roberta_base_lr1e-05_drop0.3_epochs2_frozen_roberta	0.8058	0.3477	0.0232	0.8058	0.9768	0.6523	2.4349

Continued on next page

Table B.1 Continued from previous page

Experiment	Test/Cosine_Sim	Test/KL_Div	Test/MSE	Test/Cosine_Sim_Norm	Test/MSE_Norm	Test/KL_Div_Norm	Score
exp_roberta_base_lr1e-05_drop0.3_epochs2_frozen_roberta	0.8058	0.3481	0.0232	0.8058	0.9768	0.6519	2.4345
exp_roberta_large_lr5e-06_drop0.5_epochs2_both_frozen	0.8064	0.3488	0.0231	0.8064	0.9769	0.6512	2.4344
exp_adamw_roberta_large_lr5e-06_drop0.3_bigger_mlp_epochs2_frozen_clip	0.804	0.3466	0.0231	0.804	0.9769	0.6534	2.4343
exp_adamw_roberta_large_lr5e-06_drop0.5_bigger_mlp_epochs2_both_frozen	0.8079	0.3505	0.0231	0.8079	0.9769	0.6495	2.4342
exp_adamw_roberta_large_lr5e-05_drop0.3_hierarchical_epochs2	0.8064	0.3495	0.0233	0.8064	0.9767	0.6505	2.4336
exp_adamw_roberta_large_lr1e-05_drop0.3_bigger_mlp_epochs2	0.8052	0.3484	0.0232	0.8052	0.9768	0.6516	2.4335
exp_adamw_roberta_base_lr5e-06_drop0.5_epochs2_both_frozen	0.8085	0.3517	0.0233	0.8085	0.9767	0.6483	2.4335
exp_roberta_large_lr1e-05_drop0.5_epochs2_frozen_roberta	0.8032	0.3472	0.0232	0.8032	0.9768	0.6528	2.4328
exp_roberta_large_lr1e-05_drop0.5_epochs5_frozen_roberta	0.8053	0.3495	0.0233	0.8053	0.9767	0.6505	2.4326
exp_adamw_roberta_large_lr1e-05_drop0.3_gelu_epochs2_both_frozen	0.8036	0.3479	0.0232	0.8036	0.9768	0.6521	2.4325
exp_adamw_roberta_large_lr1e-05_drop0.3_gelu_epochs2	0.8051	0.3494	0.0233	0.8051	0.9767	0.6506	2.4324
exp_adamw_roberta_base_lr5e-06_drop0.3_bigger_mlp_epochs2_both_frozen	0.8057	0.3501	0.0232	0.8057	0.9768	0.6499	2.4324
exp_adamw_roberta_base_lr1e-05_drop0.5_epochs2	0.8038	0.3483	0.0232	0.8038	0.9768	0.6517	2.4323
exp_roberta_base_lr1e-05_drop0.3_epochs2	0.8052	0.3499	0.0232	0.8052	0.9768	0.6501	2.4321
exp_adamw_roberta_base_lr5e-06_drop0.3_gelu_epochs2_both_frozen	0.8059	0.3506	0.0233	0.8059	0.9767	0.6494	2.432
exp_adamw_roberta_large_lr1e-05_drop0.5_gelu_epochs2_both_frozen	0.8039	0.3488	0.0232	0.8039	0.9768	0.6512	2.432
exp_adamw_roberta_large_lr1e-05_drop0.5_epochs2_frozen_roberta	0.803	0.3479	0.0231	0.803	0.9769	0.6521	2.4319
exp_roberta_base_lr5e-06_drop0.3_epochs2_both_frozen	0.8057	0.3507	0.0232	0.8057	0.9768	0.6493	2.4318
exp_adamw_roberta_large_lr5e-06_drop0.3_epochs2_frozen_roberta	0.8024	0.3479	0.0231	0.8024	0.9769	0.6521	2.4313
exp_adamw_roberta_large_lr5e-05_drop0.3_hierarchical_epochs2_frozen_roberta	0.8049	0.3508	0.0235	0.8049	0.9765	0.6492	2.4306

Continued on next page

Table B.1 Continued from previous page

Experiment	Test/Cosine_Sim	Test/KL_Div	Test/MSE	Test/Cosine_Sim_Norm	Test/MSE_Norm	Test/KL_Div_Norm	Score
exp_adamw_roberta_large_lr1e-05_drop0.3_epochs2_frozen_roberta	0.8019	0.3482	0.0232	0.8019	0.9768	0.6518	2.4305
exp_roberta_large_lr1e-05_drop0.3_epochs2	0.8017	0.3486	0.0232	0.8017	0.9768	0.6514	2.4299
exp_roberta_base_lr1e-05_drop0.5_epochs2_frozen_roberta	0.8041	0.3509	0.0233	0.8041	0.9767	0.6491	2.4299
exp_adamw_roberta_base_lr5e-06_drop0.5_gelu_epochs2_frozen_roberta	0.8027	0.3496	0.0232	0.8027	0.9768	0.6504	2.4299
exp_roberta_base_lr5e-06_drop0.5_epochs2_both_frozen	0.8055	0.3525	0.0233	0.8055	0.9767	0.6475	2.4297
exp_adamw_roberta_large_lr5e-06_drop0.3_gelu_epochs2_both_frozen	0.8035	0.3506	0.0232	0.8035	0.9768	0.6494	2.4297
exp_roberta_large_lr1e-05_drop0.3_epochs2_frozen_roberta	0.8006	0.3477	0.0233	0.8006	0.9767	0.6523	2.4296
exp_adamw_roberta_base_lr5e-06_drop0.5_bigger_mlp_epochs2_frozen_roberta	0.8013	0.3503	0.0233	0.8013	0.9767	0.6497	2.4278
exp_roberta_large_lr1e-05_drop0.3_epochs2_frozen_clip	0.8003	0.3493	0.0233	0.8003	0.9767	0.6507	2.4277
exp_adamw_roberta_large_lr5e-06_drop0.5_gelu_epochs2_both_frozen	0.8048	0.3541	0.0233	0.8048	0.9767	0.6459	2.4273
exp_roberta_large_lr5e-06_drop0.3_epochs2_frozen_roberta	0.8012	0.3511	0.0234	0.8012	0.9766	0.6489	2.4267
exp_adamw_roberta_large_lr5e-06_drop0.5_gelu_epochs2_frozen_clip	0.8003	0.3505	0.0234	0.8003	0.9766	0.6495	2.4264
exp_adamw_roberta_base_lr5e-06_drop0.5_gelu_epochs2_both_frozen	0.8053	0.356	0.0236	0.8053	0.9764	0.644	2.4257
exp_adamw_roberta_base_lr5e-06_drop0.5_bigger_mlp_epochs2_both_frozen	0.8051	0.3561	0.0236	0.8051	0.9764	0.6439	2.4254
exp_adamw_roberta_large_lr1e-05_drop0.3_gelu_epochs5_frozen_roberta	0.8031	0.3558	0.0235	0.8031	0.9765	0.6442	2.4238
exp_adamw_roberta_large_lr5e-06_drop0.3_gelu_epochs2_frozen_clip	0.7972	0.3541	0.0237	0.7972	0.9763	0.6459	2.4193
exp_adamw_roberta_large_lr1e-05_drop0.3_gelu_epochs2_frozen_clip	0.7964	0.3566	0.0238	0.7964	0.9762	0.6434	2.416
exp_adamw_roberta_large_lr1e-05_drop0.5_gelu_epochs2_frozen_clip	0.796	0.3575	0.0239	0.796	0.9761	0.6425	2.4147
exp_adamw_roberta_large_lr5e-06_drop0.5_epochs5_frozen_roberta	0.8008	0.363	0.024	0.8008	0.976	0.637	2.4137
exp_adamw_roberta_large_lr1e-05_drop0.5_bigger_mlp_epochs5_frozen_roberta	0.7999	0.3626	0.0242	0.7999	0.9758	0.6374	2.4131

Continued on next page

Table B.1 Continued from previous page

Experiment	Test/Cosine_Sim	Test/KL_Div	Test/MSE	Test/Cosine_Sim_Norm	Test/MSE_Norm	Test/KL_Div_Norm	Score
exp_adamw_roberta_base_lr1e-05_drop0.5_bigger_mlp_epochs5_frozen_roberta	0.7968	0.3613	0.0242	0.7968	0.9758	0.6387	2.4112
exp_adamw_roberta_base_lr5e-06_drop0.5_bigger_mlp_epochs5_frozen_clip	0.798	0.3657	0.0244	0.798	0.9756	0.6343	2.4079
exp_adamw_roberta_large_lr5e-06_drop0.5_bigger_mlp_epochs5_frozen_roberta	0.799	0.3673	0.0243	0.799	0.9757	0.6327	2.4074
exp_adamw_roberta_large_lr1e-05_drop0.3_epochs5_frozen_roberta	0.7974	0.3665	0.0244	0.7974	0.9756	0.6335	2.4065
exp_roberta_large_lr5e-06_drop0.5_epochs5_frozen_roberta	0.7969	0.3665	0.0243	0.7969	0.9757	0.6335	2.4062
exp_roberta_base_lr1e-05_drop0.5_epochs5_frozen_roberta	0.7966	0.3674	0.0243	0.7966	0.9757	0.6326	2.4049
exp_adamw_roberta_large_lr5e-05_drop0.5_hierarchical_epochs5_both_frozen	0.7959	0.3679	0.0245	0.7959	0.9755	0.6321	2.4035
exp_adamw_roberta_base_lr1e-05_drop0.3_epochs5_frozen_roberta	0.7932	0.3714	0.0247	0.7932	0.9753	0.6286	2.3971
exp_adamw_roberta_base_lr5e-06_drop0.5_epochs5_frozen_clip	0.794	0.3721	0.0249	0.794	0.9751	0.6279	2.3971
exp_adamw_roberta_base_lr5e-06_drop0.5_gelu_epochs5_frozen_clip	0.7945	0.373	0.0248	0.7945	0.9752	0.627	2.3967
exp_adamw_roberta_large_lr5e-05_drop0.3_hierarchical_epochs5_both_frozen	0.7943	0.3734	0.0248	0.7943	0.9752	0.6266	2.3961
exp_adamw_roberta_base_lr1e-05_drop0.3_bigger_mlp_epochs5_frozen_roberta	0.7935	0.3754	0.0249	0.7935	0.9751	0.6246	2.3931
exp_roberta_base_lr5e-06_drop0.5_epochs5_frozen_clip	0.7918	0.3758	0.0251	0.7918	0.9749	0.6242	2.3908
exp_adamw_roberta_base_lr5e-06_drop0.3_epochs5_frozen_clip	0.7918	0.3764	0.0251	0.7918	0.9749	0.6236	2.3903
exp_adamw_roberta_base_lr5e-06_drop0.5_bigger_mlp_epochs5	0.7924	0.3777	0.0251	0.7924	0.9749	0.6223	2.3897
exp_adamw_roberta_large_lr5e-05_drop0.5_hierarchical_epochs5_frozen_roberta	0.7924	0.3789	0.0249	0.7924	0.9751	0.6211	2.3885
exp_roberta_large_lr1e-05_drop0.3_epochs5_frozen_roberta	0.7897	0.3769	0.025	0.7897	0.975	0.6231	2.3878
exp_roberta_large_lr5e-06_drop0.3_epochs5_frozen_roberta	0.7913	0.3794	0.025	0.7913	0.975	0.6206	2.3869
exp_adamw_roberta_large_lr5e-06_drop0.3_bigger_mlp_epochs5_frozen_roberta	0.7908	0.3789	0.025	0.7908	0.975	0.6211	2.3869
exp_adamw_roberta_base_lr5e-06_drop0.5_epochs5_frozen_roberta	0.79	0.3787	0.0252	0.79	0.9748	0.6213	2.3861

Continued on next page

Table B.1 Continued from previous page

Experiment	Test/Cosine_Sim	Test/KL_Div	Test/MSE	Test/Cosine_Sim_Norm	Test/MSE_Norm	Test/KL_Div_Norm	Score
exp_adamw_roberta_base_lr5e-06_drop0.3_bigger_mlp_epochs5_frozen_clip	0.791	0.38	0.0253	0.791	0.9747	0.62	2.3857
exp_roberta_base_lr5e-06_drop0.3_epochs5_frozen_clip	0.7902	0.3795	0.0253	0.7902	0.9747	0.6205	2.3854
exp_adamw_roberta_large_lr5e-06_drop0.5_gelu_epochs5_frozen_roberta	0.7898	0.3804	0.0251	0.7898	0.9749	0.6196	2.3843
exp_roberta_base_lr1e-05_drop0.3_epochs5_frozen_roberta	0.7873	0.3788	0.0251	0.7873	0.9749	0.6212	2.3834
exp_adamw_roberta_large_lr5e-06_drop0.5_epochs5_frozen_clip	0.7896	0.3815	0.0253	0.7896	0.9747	0.6185	2.3829
exp_adamw_roberta_large_lr5e-06_drop0.3_epochs5_frozen_roberta	0.7901	0.3825	0.0253	0.7901	0.9747	0.6175	2.3823
exp_adamw_roberta_large_lr5e-06_drop0.3_epochs5	0.7896	0.3833	0.0254	0.7896	0.9746	0.6167	2.3808
exp_adamw_roberta_base_lr5e-06_drop0.3_gelu_epochs5_frozen_clip	0.7894	0.3839	0.0255	0.7894	0.9745	0.6161	2.38
exp_adamw_roberta_base_lr5e-06_drop0.3_epochs5	0.7886	0.3839	0.0255	0.7886	0.9745	0.6161	2.3792
exp_adamw_roberta_large_lr5e-06_drop0.3_gelu_epochs5_frozen_roberta	0.7889	0.3853	0.0254	0.7889	0.9746	0.6147	2.3782
exp_adamw_roberta_large_lr5e-06_drop0.5_bigger_mlp_epochs5_frozen_clip	0.7875	0.385	0.0256	0.7875	0.9744	0.615	2.3769
exp_adamw_roberta_large_lr5e-06_drop0.5_epochs5	0.7894	0.3876	0.0255	0.7894	0.9745	0.6124	2.3763
exp_adamw_roberta_base_lr1e-05_drop0.5_gelu_epochs5_frozen_roberta	0.7877	0.3857	0.0256	0.7877	0.9744	0.6143	2.3763
exp_adamw_roberta_base_lr5e-06_drop0.5_gelu_epochs5	0.79	0.3888	0.0257	0.79	0.9743	0.6112	2.3755
exp_roberta_base_lr5e-06_drop0.5_epochs5	0.7887	0.3884	0.0258	0.7887	0.9742	0.6116	2.3746
exp_adamw_roberta_base_lr5e-06_drop0.3_gelu_epochs5_frozen_roberta	0.7888	0.3886	0.0257	0.7888	0.9743	0.6114	2.3745
exp_adamw_roberta_base_lr1e-05_drop0.5_epochs5_frozen_roberta	0.7847	0.3855	0.0256	0.7847	0.9744	0.6145	2.3736
exp_adamw_roberta_large_lr5e-05_drop0.3_hierarchical_epochs5_frozen_roberta	0.7869	0.3885	0.0257	0.7869	0.9743	0.6115	2.3726
exp_adamw_roberta_base_lr5e-06_drop0.3_bigger_mlp_epochs5_frozen_roberta	0.7872	0.3893	0.0257	0.7872	0.9743	0.6107	2.3722
exp_adamw_roberta_large_lr5e-06_drop0.5_gelu_epochs5	0.7888	0.3912	0.0256	0.7888	0.9744	0.6088	2.372

Continued on next page

Table B.1 Continued from previous page

Experiment	Test/Cosine_Sim	Test/KL_Div	Test/MSE	Test/Cosine_Sim_Norm	Test/MSE_Norm	Test/KL_Div_Norm	Score
exp_adamw_roberta_base_lr5e-06_drop0.5_bigger_mlp_epochs5_frozen_roberta	0.7856	0.3878	0.0259	0.7856	0.9741	0.6122	2.3719
exp_adamw_roberta_base_lr5e-06_drop0.5_epochs5	0.7868	0.3898	0.0261	0.7868	0.9739	0.6102	2.3709
exp_adamw_roberta_base_lr1e-05_drop0.3_gelu_epochs5_frozen_roberta	0.7861	0.3897	0.0258	0.7861	0.9742	0.6103	2.3705
exp_adamw_roberta_large_lr1e-05_drop0.3_bigger_mlp_epochs5_frozen_roberta	0.7878	0.3925	0.0257	0.7878	0.9743	0.6075	2.3696
exp_adamw_roberta_large_lr5e-05_drop0.5_hierarchical_epochs5	0.7858	0.3918	0.0257	0.7858	0.9743	0.6082	2.3684
exp_adamw_roberta_base_lr5e-06_drop0.3_bigger_mlp_epochs5	0.7865	0.3926	0.026	0.7865	0.974	0.6074	2.3679
exp_adamw_roberta_base_lr1e-05_drop0.5_epochs5	0.7855	0.3928	0.0261	0.7855	0.9739	0.6072	2.3665
exp_adamw_roberta_base_lr5e-06_drop0.5_gelu_epochs5_frozen_roberta	0.7848	0.3929	0.0261	0.7848	0.9739	0.6071	2.3658
exp_adamw_roberta_base_lr1e-05_drop0.5_bigger_mlp_epochs5	0.7866	0.395	0.0262	0.7866	0.9738	0.605	2.3653
exp_adamw_roberta_large_lr5e-06_drop0.3_epochs5_frozen_clip	0.7847	0.3934	0.026	0.7847	0.974	0.6066	2.3652
exp_adamw_roberta_large_lr5e-06_drop0.3_gelu_epochs5	0.7861	0.3958	0.0258	0.7861	0.9742	0.6042	2.3645
exp_adamw_roberta_large_lr5e-05_drop0.5_hierarchical_epochs5_frozen_clip	0.7866	0.3975	0.0258	0.7866	0.9742	0.6025	2.3633
exp_roberta_large_lr5e-06_drop0.5_epochs5_frozen_clip	0.7845	0.3954	0.0261	0.7845	0.9739	0.6046	2.363
exp_adamw_roberta_base_lr5e-06_drop0.3_epochs5_frozen_roberta	0.7833	0.3942	0.0262	0.7833	0.9738	0.6058	2.363
exp_adamw_roberta_large_lr5e-06_drop0.3_bigger_mlp_epochs5	0.7847	0.3957	0.0262	0.7847	0.9738	0.6043	2.3629
exp_roberta_large_lr5e-06_drop0.5_epochs5	0.7835	0.3947	0.0261	0.7835	0.9739	0.6053	2.3627
exp_adamw_roberta_large_lr5e-06_drop0.5_bigger_mlp_epochs5	0.7834	0.3957	0.0262	0.7834	0.9738	0.6043	2.3615
exp_roberta_base_lr5e-06_drop0.3_epochs5	0.784	0.3962	0.0262	0.784	0.9738	0.6038	2.3615
exp_adamw_roberta_large_lr5e-05_drop0.3_hierarchical_epochs5	0.7835	0.3962	0.0261	0.7835	0.9739	0.6038	2.3612
exp_roberta_base_lr5e-06_drop0.5_epochs5_frozen_roberta	0.7815	0.3942	0.0262	0.7815	0.9738	0.6058	2.3611

Continued on next page

Table B.1 Continued from previous page

Experiment	Test/Cosine_Sim	Test/KL_Div	Test/MSE	Test/Cosine_Sim_Norm	Test/MSE_Norm	Test/KL_Div_Norm	Score
exp_roberta_base_lr5e-06_drop0.3_epochs5_frozen_roberta	0.7831	0.3964	0.0262	0.7831	0.9738	0.6036	2.3605
exp_adamw_roberta_base_lr5e-06_drop0.3_gelu_epochs5	0.7846	0.3983	0.0263	0.7846	0.9737	0.6017	2.36
exp_roberta_base_lr1e-05_drop0.5_epochs5	0.7843	0.3992	0.0263	0.7843	0.9737	0.6008	2.3588
exp_adamw_roberta_large_lr5e-06_drop0.3_gelu_epochs5_frozen_clip	0.7826	0.3985	0.0263	0.7826	0.9737	0.6015	2.3578
exp_roberta_large_lr5e-06_drop0.3_epochs5	0.7829	0.4004	0.0264	0.7829	0.9736	0.5996	2.3561
exp_adamw_roberta_large_lr1e-05_drop0.5_epochs5	0.7829	0.4032	0.0264	0.7829	0.9736	0.5968	2.3534
exp_adamw_roberta_large_lr5e-06_drop0.3_bigger_mlp_epochs5_frozen_clip	0.7809	0.4016	0.0266	0.7809	0.9734	0.5984	2.3527
exp_adamw_roberta_base_lr1e-05_drop0.5_gelu_epochs5	0.7824	0.403	0.0269	0.7824	0.9731	0.597	2.3526
exp_adamw_roberta_base_lr1e-05_drop0.5_gelu_epochs5_frozen_clip	0.7823	0.404	0.0265	0.7823	0.9735	0.596	2.3518
exp_adamw_roberta_large_lr5e-06_drop0.5_gelu_epochs5_frozen_clip	0.7793	0.4016	0.0266	0.7793	0.9734	0.5984	2.351
exp_roberta_large_lr5e-06_drop0.3_epochs5_frozen_clip	0.7805	0.4052	0.0267	0.7805	0.9733	0.5948	2.3486
exp_adamw_roberta_base_lr1e-05_drop0.5_bigger_mlp_epochs5_frozen_clip	0.78	0.4057	0.027	0.78	0.973	0.5943	2.3473
exp_adamw_roberta_large_lr1e-05_drop0.3_bigger_mlp_epochs5	0.779	0.4074	0.0267	0.779	0.9733	0.5926	2.3449
exp_adamw_roberta_base_lr1e-05_drop0.5_epochs5_frozen_clip	0.7794	0.4074	0.0272	0.7794	0.9728	0.5926	2.3448
exp_adamw_roberta_large_lr5e-05_drop0.3_hierarchical_epochs5_frozen_clip	0.7793	0.4107	0.0268	0.7793	0.9732	0.5893	2.3418
exp_adamw_roberta_large_lr1e-05_drop0.5_bigger_mlp_epochs5	0.7787	0.4113	0.0271	0.7787	0.9729	0.5887	2.3403
exp_roberta_base_lr1e-05_drop0.5_epochs5_frozen_clip	0.7782	0.412	0.0272	0.7782	0.9728	0.588	2.339
exp_roberta_large_lr1e-05_drop0.5_epochs5	0.7784	0.4133	0.027	0.7784	0.973	0.5867	2.3381
exp_adamw_roberta_base_lr1e-05_drop0.3_bigger_mlp_epochs5_frozen_clip	0.7776	0.4134	0.0274	0.7776	0.9726	0.5866	2.3368
exp_adamw_roberta_large_lr1e-05_drop0.5_bigger_mlp_epochs5_frozen_clip	0.7778	0.4144	0.0273	0.7778	0.9727	0.5856	2.3362

Continued on next page

Table B.1 Continued from previous page

Experiment	Test/Cosine_Sim	Test/KL_Div	Test/MSE	Test/Cosine_Sim_Norm	Test/MSE_Norm	Test/KL_Div_Norm	Score
exp_adamw_roberta_base_lr1e-05_drop0.3_gelu_epochs5	0.7775	0.4158	0.0275	0.7775	0.9725	0.5842	2.3342
exp_roberta_base_lr1e-05_drop0.3_epochs5	0.7721	0.4119	0.0273	0.7721	0.9727	0.5881	2.3329
exp_adamw_roberta_large_lr1e-05_drop0.5_gelu_epochs5_frozen_clip	0.7754	0.4185	0.0274	0.7754	0.9726	0.5815	2.3295
exp_adamw_roberta_large_lr1e-05_drop0.5_epochs5_frozen_clip	0.7751	0.4184	0.0274	0.7751	0.9726	0.5816	2.3293
exp_adamw_roberta_base_lr1e-05_drop0.3_epochs5_frozen_clip	0.774	0.4191	0.0279	0.774	0.9721	0.5809	2.3271
exp_roberta_base_lr1e-05_drop0.3_epochs5_frozen_clip	0.7735	0.4191	0.0277	0.7735	0.9723	0.5809	2.3268
exp_adamw_roberta_base_lr1e-05_drop0.3_bigger_mlp_epochs5	0.7728	0.4184	0.0281	0.7728	0.9719	0.5816	2.3262
exp_adamw_roberta_large_lr1e-05_drop0.3_gelu_epochs5_frozen_clip	0.7748	0.4215	0.0276	0.7748	0.9724	0.5785	2.3256
exp_adamw_roberta_large_lr1e-05_drop0.5_gelu_epochs5	0.7725	0.4208	0.0275	0.7725	0.9725	0.5792	2.3241
exp_adamw_roberta_base_lr1e-05_drop0.3_gelu_epochs5_frozen_clip	0.773	0.4218	0.0279	0.773	0.9721	0.5782	2.3232
exp_roberta_large_lr1e-05_drop0.5_epochs5_frozen_clip	0.7731	0.4241	0.0278	0.7731	0.9722	0.5759	2.3212
exp_adamw_roberta_large_lr1e-05_drop0.3_epochs5_frozen_clip	0.7738	0.4261	0.0278	0.7738	0.9722	0.5739	2.3199
exp_adamw_roberta_base_lr1e-05_drop0.3_epochs5	0.7709	0.4235	0.0281	0.7709	0.9719	0.5765	2.3193
exp_adamw_roberta_large_lr1e-05_drop0.3_bigger_mlp_epochs5_frozen_clip	0.771	0.4261	0.0279	0.771	0.9721	0.5739	2.317
exp_adamw_roberta_large_lr1e-05_drop0.3_epochs5	0.7712	0.4281	0.0279	0.7712	0.9721	0.5719	2.3152
exp_roberta_large_lr1e-05_drop0.3_epochs5	0.7679	0.4264	0.0279	0.7679	0.9721	0.5736	2.3135
exp_adamw_roberta_large_lr1e-05_drop0.3_gelu_epochs5	0.7711	0.4344	0.0282	0.7711	0.9718	0.5656	2.3085
exp_roberta_large_lr1e-05_drop0.3_epochs5_frozen_clip	0.7684	0.4361	0.0285	0.7684	0.9715	0.5639	2.3037

Table B.2.: Complete image experimental results

Experiment	Test/Cosine_Sim	Test/KL_Div	Test/MSE	Test/Cosine_Sim_Norm	Test/MSE_Norm	Test/KL_Div_Norm	Score
exp_only_clip_lr5e-06_drop0.5_epochs2	0.7996	0.3583	0.0239	0.7996	0.9761	0.6417	2.4174
exp_adamw_only_clip_lr5e-06_drop0.3_epochs2	0.7993	0.358	0.0239	0.7993	0.9761	0.642	2.4174
exp_only_clip_lr1e-05_drop0.3_epochs10_frozen	0.7997	0.3585	0.0239	0.7997	0.9761	0.6415	2.4173
exp_only_clip_lr5e-06_drop0.3_epochs2	0.799	0.3591	0.024	0.799	0.976	0.6409	2.4159
exp_adamw_only_clip_lr5e-06_drop0.5_epochs2	0.7977	0.3612	0.0241	0.7977	0.9759	0.6388	2.4124
exp_adamw_only_clip_lr1e-05_drop0.5_epochs2	0.7969	0.3641	0.0244	0.7969	0.9756	0.6359	2.4083
exp_only_clip_lr1e-05_drop0.3_epochs5_frozen	0.7971	0.366	0.0244	0.7971	0.9756	0.634	2.4068
exp_adamw_only_clip_lr1e-05_drop0.3_epochs2	0.7969	0.3662	0.0246	0.7969	0.9754	0.6338	2.4061
exp_adamw_only_clip_lr1e-05_drop0.3_epochs5_frozen	0.7965	0.3671	0.0244	0.7965	0.9756	0.6329	2.4049
exp_only_clip_lr5e-06_drop0.3_epochs10_frozen	0.7963	0.3673	0.0245	0.7963	0.9755	0.6327	2.4045
exp_only_clip_lr1e-05_drop0.3_epochs2	0.796	0.367	0.0246	0.796	0.9754	0.633	2.4044
exp_only_clip_lr1e-05_drop0.5_epochs2	0.7956	0.3675	0.0246	0.7956	0.9754	0.6325	2.4035
exp_only_clip_lr1e-05_drop0.5_epochs5_frozen	0.7957	0.3696	0.0246	0.7957	0.9754	0.6304	2.4015
exp_adamw_only_clip_lr1e-05_drop0.5_epochs5_frozen	0.7957	0.3696	0.0246	0.7957	0.9754	0.6304	2.4015
exp_only_clip_lr5e-06_drop0.3_epochs5_frozen	0.7902	0.3844	0.0255	0.7902	0.9745	0.6156	2.3803
exp_adamw_only_clip_lr5e-06_drop0.3_epochs5_frozen	0.7902	0.3844	0.0255	0.7902	0.9745	0.6156	2.3803
exp_only_clip_lr5e-06_drop0.5_epochs5_frozen	0.7885	0.3885	0.0258	0.7885	0.9742	0.6115	2.3742
exp_adamw_only_clip_lr5e-06_drop0.5_epochs5_frozen	0.7885	0.3885	0.0258	0.7885	0.9742	0.6115	2.3742
exp_only_clip_lr1e-05_drop0.3_epochs2_frozen	0.7879	0.3892	0.0258	0.7879	0.9742	0.6108	2.3728

Continued on next page

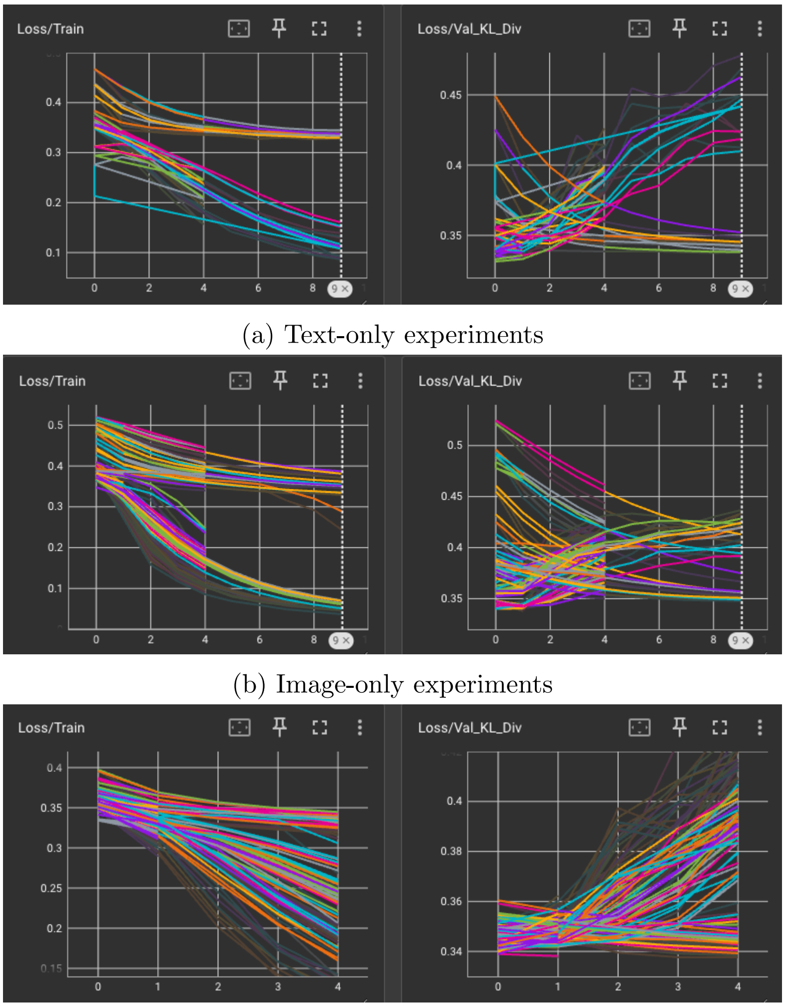
Table B.2 Continued from previous page

Experiment	Test/Cosine_Sim	Test/KL_Div	Test/MSE	Test/Cosine_Sim_Norm	Test/MSE_Norm	Test/KL_Div_Norm	Score
exp_only_clip_lr1e-05_drop0.5_epochs5	0.7812	0.3911	0.0261	0.7812	0.9739	0.6089	2.364
exp_only_clip_lr1e-05_drop0.5_epochs2_frozen	0.7851	0.3954	0.0262	0.7851	0.9738	0.6046	2.3635
exp_adamw_only_clip_lr1e-05_drop0.5_epochs2_frozen	0.7851	0.3954	0.0262	0.7851	0.9738	0.6046	2.3635
exp_adamw_only_clip_lr1e-05_drop0.3_epochs2_frozen	0.7835	0.3973	0.0262	0.7835	0.9738	0.6027	2.36
exp_adamw_only_clip_lr5e-06_drop0.3_epochs5	0.7813	0.3966	0.0263	0.7813	0.9737	0.6034	2.3585
exp_adamw_only_clip_lr5e-06_drop0.5_epochs5	0.7797	0.3957	0.0263	0.7797	0.9737	0.6043	2.3577
exp_adamw_only_clip_lr1e-05_drop0.5_epochs5	0.7795	0.3988	0.0265	0.7795	0.9735	0.6012	2.3543
exp_only_clip_lr5e-06_drop0.3_epochs5	0.7775	0.4027	0.0267	0.7775	0.9733	0.5973	2.348
exp_only_clip_lr5e-06_drop0.5_epochs5	0.7768	0.4036	0.0268	0.7768	0.9732	0.5964	2.3464
exp_adamw_only_clip_lr1e-05_drop0.3_epochs5	0.7754	0.4085	0.027	0.7754	0.973	0.5915	2.3399
exp_only_clip_lr5e-06_drop0.3_epochs10	0.7743	0.413	0.027	0.7743	0.973	0.587	2.3343
exp_only_clip_lr1e-05_drop0.3_epochs5	0.7738	0.4153	0.0273	0.7738	0.9727	0.5847	2.3312
exp_only_clip_lr5e-06_drop0.3_epochs2_frozen	0.7748	0.4174	0.0276	0.7748	0.9724	0.5826	2.3298
exp_adamw_only_clip_lr5e-06_drop0.3_epochs2_frozen	0.7748	0.4174	0.0276	0.7748	0.9724	0.5826	2.3298
exp_only_clip_lr5e-06_drop0.5_epochs2_frozen	0.7727	0.4222	0.0279	0.7727	0.9721	0.5778	2.3226
exp_adamw_only_clip_lr5e-06_drop0.5_epochs2_frozen	0.7727	0.4222	0.0279	0.7727	0.9721	0.5778	2.3226
exp_only_clip_lr1e-05_drop0.3_epochs10	0.77	0.4224	0.0276	0.77	0.9724	0.5776	2.32

Table B.3.: Complete text experimental results

Experiment	Test/Cosine_Sim	Test/KL_Div	Test/MSE	Test/Cosine_Sim_Norm	Test/MSE_Norm	Test/KL_Div_Norm	Score
exp_only_roberta_large_lr1e-05_drop0.3_epochs10_frozen	0.8028	0.3478	0.0232	0.8028	0.9768	0.6522	2.4318
exp_only_roberta_large_lr5e-06_drop0.3_epochs2	0.801	0.3481	0.0232	0.801	0.9768	0.6519	2.4297
exp_only_roberta_large_lr5e-06_drop0.3_epochs10_frozen	0.8015	0.3498	0.0233	0.8015	0.9767	0.6502	2.4285
exp_only_roberta_large_lr1e-05_drop0.3_epochs5_frozen	0.8015	0.35	0.0233	0.8015	0.9767	0.65	2.4281
exp_only_roberta_large_lr5e-06_drop0.3_epochs5_frozen	0.8016	0.3518	0.0233	0.8016	0.9767	0.6482	2.4265
exp_only_roberta_large_lr1e-05_drop0.3_epochs2_frozen	0.8019	0.3529	0.0233	0.8019	0.9767	0.6471	2.4256
exp_only_roberta_large_lr1e-05_drop0.3_epochs2	0.7947	0.3527	0.0237	0.7947	0.9763	0.6473	2.4182
exp_only_roberta_large_lr5e-06_drop0.3_epochs2_frozen	0.8	0.369	0.0242	0.8	0.9758	0.631	2.4068
exp_only_roberta_large_lr5e-06_drop0.3_epochs5	0.7861	0.3891	0.0259	0.7861	0.9741	0.6109	2.3711
exp_only_roberta_large_lr1e-05_drop0.3_epochs5	0.7773	0.4048	0.027	0.7773	0.973	0.5952	2.3455
exp_only_roberta_large_lr5e-06_drop0.3_epochs10	0.7675	0.4385	0.0285	0.7675	0.9715	0.5615	2.3005
exp_only_roberta_large_lr1e-05_drop0.3_epochs10	0.7597	0.4496	0.029	0.7597	0.971	0.5504	2.2812

C. Loss Curves



(a) Text-only experiments

(b) Image-only experiments

(c) Multimodal experiments (August only)

Figure C.1.: Train and Validation KL Divergence loss curves

Ehrenwörtliche Erklärung

Ich versichere, dass ich die beiliegende Bachelor-, Master-, Seminar-, oder Projektarbeit ohne Hilfe Dritter und ohne Benutzung anderer als der angegebenen Quellen und in der untenstehenden Tabelle angegebenen Hilfsmittel angefertigt und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen. Ich bin mir bewusst, dass eine falsche Erklärung rechtliche Folgen haben wird.

Declaration of Used AI Tools

Tool	Purpose	Where?	Useful?
ChatGPT	Rephrasing	Throughout	+
DeepL	Translation	Throughout	+
ResearchGPT	Summarization of related work	Sec. **??**	-
Dall-E	Image generation	Figs. 2, 3	++
GPT-4	Code generation	functions.py	+
ChatGPT	Related work hallucination	Most of bibliography	++

Unterschrift
Mannheim, den XX. XXXX 2024