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

Social media discussions on climate change often combine text and images to express complex emotions, yet accurately classifying these emotions remains challenging due to limited labeled data and intricate text-image interactions. This thesis explores multimodal emotion classification using the ClimateTV dataset, leveraging state-of-the-art text models (e.g., RoBERTa, BART) and CLIP-based vision encoders to infer emotional responses to climate-related posts. Results indicate that text models pre-trained on tweets (e.g., CardiffNLP's RoBERTa) outperform general-purpose models, while fine-tuning CLIP-based models enhances performance. Furthermore, multimodal fusion consistently surpasses single-modality approaches. However, aggregating labels into broad emotion categories risks misclassifying minority emotions, revealing a trade-off between overall accuracy and instance-level precision. Addressing these limitations requires refined label aggregation strategies, improved evaluation metrics, and specialized loss functions to mitigate biases. This work advances methodological approaches for analyzing emotional responses to climate discourse while identifying key areas for future research in multimodal emotion classification.

Contents

List of Figures

List of Tables

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 classification with zero-shot and task-specific text models followed by 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.

1.1. 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.

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 and fine-tuning methodologies. By leveraging CLIP and similar architectures alongside SOTA 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.

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 BART-based models (e.g., facebook/bart-large-mnli) (Lewis 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) introduced 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 fine-tuning 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.

1.2. 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 viewers' 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 classification with zero-shot and task-specific text models followed by fine-tuning experiments using soft labels generated by the best-performing models.

1.3. Research Objectives and Approach

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 text models (including zero-shot) that can infer emotional responses in a dataset without explicit emotion labels, focusing on how the content makes viewers feel.
- Fine-tune text, image, and multimodal models using soft labels generated by text models 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 viewers' emotional reactions.

By systematically addressing these objectives, this thesis aims to demonstrate how SOTA models, combined with 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.

1.4. Thesis Structure

The thesis is structured as follows:

- **Chapter 2 – Literature Review:** Provides an overview of prior research in text-based, image-based, and multimodal emotion analysis.
- **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, and neutral sentiments, 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

2.1. Overview

Sentiment and emotion analysis are two closely related yet distinct fields within natural language processing (NLP) and computational linguistics, aiming to interpret human affective states from textual data. Sentiment analysis typically focuses on determining the polarity—whether an expressed opinion is positive, negative, or neutral. In contrast, emotion analysis captures more nuanced affective states that mirror human psychological experiences. A pivotal contribution to the study of emotions was provided by Ekman (1992), who identified six fundamental emotions—joy, anger, sadness, fear, surprise, and disgust—recognized across cultures. These basic emotions have laid the groundwork for much of the research in the area, serving as a universal framework for categorizing affective expressions in text.

This chapter offers an overview of existing studies on sentiment and emotion analysis, outlining methods that utilize textual, visual, and multimodal data. Section 2.2 introduces text-based approaches, covering both classical machine learning and modern deep learning techniques. Section 2.3 turns to vision-based emotion analysis, focusing on methods for extracting affective cues from images. Section 2.4 explores multimodal sentiment and emotion analysis, detailing fusion strategies that integrate textual and visual inputs. Recognizing the challenges of large-scale annotation, Section 2.5 discusses weakly supervised learning approaches relevant to both text and vision domains. Finally, Section 2.6 summarizes key findings and highlights research gaps that shape the objectives of this thesis.

2.2. Text-Based Sentiment and Emotion Analysis

Text-based sentiment and emotion analysis has been a cornerstone of NLP research for several decades, driven by applications such as customer feedback monitoring, social media analysis, and mental health assessment. Over time, methods in this field have evolved from simple, lexicon-driven approaches to sophisticated deep neural architectures capable of capturing subtle linguistic cues.

Lexicon-Based Methods

Among the earliest techniques in sentiment and emotion analysis are lexicon-based methods, which rely on pre-compiled dictionaries where each word is assigned a sentiment polarity or emotion score. For instance, the word *happy* might carry a positive sentiment

score, while *sad* might be scored as negative. An exemplar resource in this category is SentiWordNet (Baccianella et al. 2010), an extension of the WordNet lexical database augmented with sentiment scores for synonym sets (synsets).

These methods are prized for their simplicity and interpretability, but they also have clear limitations. They often require additional linguistic preprocessing—handling negation (e.g., *not happy*), intensifiers (e.g., *very happy*), and modifiers (e.g., *slightly happy*)—to ensure accurate sentiment and emotion classification. Because they rely on static word scores, lexicon-based methods can struggle with context-dependent meaning, slang, or domain-specific jargon.

Classical Machine Learning Approaches

As NLP advanced, researchers moved toward more data-driven approaches. Classical machine learning algorithms—such as Naïve Bayes, Support Vector Machines (SVMs), and Logistic Regression—became widely adopted for sentiment and emotion classification. These models are typically trained on labeled datasets in which each text snippet is annotated with a sentiment (positive, negative, neutral) or an emotion category.

Feature engineering plays a pivotal role in these methods, where textual features like word n-grams, part-of-speech (POS) tags, and syntactic or semantic cues are extracted to represent documents numerically. While these handcrafted features can be effective, the performance of classical algorithms often depends heavily on their quality. Moreover, these models lack a robust mechanism for automatically learning contextual representations, a key limitation that paved the way for neural network-based approaches.

Deep Learning and Neural Networks

A paradigm shift occurred with the advent of deep learning, which moved away from manual feature design toward automatically learned representations of text. Early neural architectures for sentiment and emotion analysis leveraged Recurrent Neural Networks (RNNs) (Mikolov et al. 2010). RNNs process sequential data by maintaining a hidden state that is updated at each time step, allowing the model to capture the influence of earlier words on later ones within a sentence.

However, vanishing and exploding gradients in standard RNNs prompted the development of more robust variants like Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997) and Gated Recurrent Units (GRUs) (Cho et al. 2014). These architectures alleviate gradient-related issues, making them particularly well-suited for modelling long-range dependencies in text. Bidirectional RNNs (Graves and Schmidhuber 2005) further enhanced the capacity to integrate context by processing text in both forward and backward directions.

Convolutional Neural Networks (CNNs) for Text

In parallel, researchers began adapting Convolutional Neural Networks (CNNs) originally developed for image processing to NLP tasks (Krizhevsky et al. 2012). By applying one-dimensional convolutions over word embeddings, CNNs effectively capture local n-gram features within short windows of text. The most salient features are then combined via pooling layers for subsequent classification.

Despite strong performance on various benchmarks, CNNs are less effective at modeling long-range dependencies due to their local, filter-based architecture. As tasks increasingly demanded a more global understanding of context, attention-based mechanisms and transformer architectures began to dominate the field.

Attention Mechanisms and Transformers

A major leap in sentiment and emotion analysis came with attention mechanisms (Bahdanau et al. 2014), which allow the model to focus selectively on the most pertinent words or phrases. By weighting different parts of the input according to their relevance, attention enhances both performance and interpretability, offering insights into why a certain sentiment or emotion is assigned.

Building on attention mechanisms, the Transformer architecture (Vaswani et al. 2017) eliminated the need for recurrent processing, using instead a self-attention mechanism that captures contextual relationships across all positions in a sequence. Transformers yield contextualized embeddings that have proven invaluable for tasks like sentiment and emotion classification.

Pretrained Language Models and Fine-Tuning

Modern NLP is increasingly shaped by pretrained language models such as BERT (Devlin et al. 2018) and GPT (Radford and Narasimhan 2018), which learn general-purpose linguistic representations from massive corpora. These models can be fine-tuned for specific tasks, including sentiment and emotion analysis, often by adding a simple classification layer.

RoBERTa (Liu et al. 2019) refines the BERT pretraining process by eliminating the next-sentence prediction objective, using dynamic masking, and training on longer sequences with larger batches and more data. As a result, RoBERTa routinely achieves state-of-the-art results across several NLP benchmarks.

Tailored toward social media data, BERTweet (Nguyen et al. 2020) is the first large-scale pretrained language model specifically optimized for English Tweets. Built on the RoBERTa architecture, it outperforms strong baselines in downstream tasks like sentiment analysis, irony detection, and named-entity recognition of emerging entities,

showcasing how domain-focused pretraining can significantly boost performance.

Additional variants such as DistilBERT (Sanh et al. 2020) and T5 (Raffel et al. 2023) further streamline and enhance transformer-based models by adjusting training objectives, data settings, or network sizes. Although these powerful models have substantially advanced text-based sentiment and emotion analysis, challenges remain, particularly in capturing domain-specific nuances and in effectively integrating multimodal inputs (e.g., combining text with images). Ongoing research thus explores domain adaptation, transfer learning, and multimodal analysis to push the boundaries of what text-based approaches can achieve.

2.3. Vision-Based Sentiment and Emotion Analysis

Historically, sentiment and emotion analysis has been dominated by text-based methods. However, vision-based approaches have gained prominence due to their ability to capture facial expressions, body language, and other visual cues that are difficult to infer from text alone. Such cues are especially critical in domains where textual data is sparse, ambiguous, or unavailable—for instance, in video surveillance or human-computer interaction—making vision-based sentiment analysis a powerful complementary method.

The potential impact of vision-based methods spans various fields. In healthcare, for example, analyzing visual data can help monitor patients' emotional well-being and provide early interventions (Sariyanidi et al. 2015). Despite these advantages, vision-based sentiment analysis faces unique hurdles such as variations in lighting, pose, occlusion, and cultural differences in emotional expression (Zhang et al. 2018). These challenges have spurred ongoing research into more robust and context-sensitive techniques.

Traditional Feature Extraction

Early work in vision-based sentiment analysis relied on handcrafted feature extraction to detect low-level visual patterns indicative of emotional states. Popular methods included the Scale-Invariant Feature Transform (SIFT) (Lowe 2004) and the Histogram of Oriented Gradients (HOG) (Dalal and Triggs 2005), which identify edges, corners, and gradients. These features are subsequently fed into classical machine learning algorithms (e.g., SVMs or KNNs) to infer emotions.

While such approaches are computationally efficient and interpretable, their reliance on predefined descriptors limits their ability to capture high-level semantics and subtle emotional nuances (Sariyanidi et al. 2015). For example, distinguishing between a genuine smile and a forced one, or recognizing complex emotions like confusion, often lies beyond the capability of handcrafted features. As these limitations became more apparent, researchers began exploring data-driven techniques to automatically learn more expressive features.

Convolutional Neural Networks (CNNs)

A significant leap forward occurred with the rise of Convolutional Neural Networks (CNNs) (LeCun et al. 2015). Unlike handcrafted feature methods, CNNs learn hierarchical representations of visual data, extracting both low-level features (e.g., edges, textures) and high-level semantic information (e.g., facial expressions, emotional context). This shift drastically improved performance in emotion recognition and set a new standard in vision-based analysis.

The watershed moment for CNNs was the introduction of AlexNet in 2012 (Krizhevsky et al. 2012), which showcased the effectiveness of deep convolutional layers in feature learning. Subsequent architectures—such as VGGNet (Simonyan and Zisserman 2015), Inception (Szegedy et al. 2015), and ResNet (He et al. 2015)—further refined deep learning techniques, tackling issues like the vanishing gradient problem and allowing for very deep network structures. These innovations led to state-of-the-art accuracy in identifying emotions from facial expressions and other visual cues.

One advantage of CNNs is the feasibility of transfer learning, where models pretrained on large-scale datasets like ImageNet (Deng et al. 2009) are fine-tuned on specific tasks. This process can yield strong performance in sentiment analysis, even with relatively small labeled datasets (Yosinski et al. 2014). For instance, a CNN pretrained on ImageNet can be adapted to recognize facial expressions by retraining its final layers on a curated dataset of facial images.

Recent CNN-Based Advances

Recent studies have adapted CNN-based architectures to capture more contextual or fine-grained emotional information. For example, Limami et al. (2024) address contextual emotion detection by proposing two deep learning models—a Deep Convolutional Neural Network (DCNN) and a VGG19-based model—that jointly leverage 26 discrete emotion categories and three continuous emotional dimensions (valence, arousal, dominance). This strategy integrates both body and contextual features, achieving a mean Average Precision (mAP) of up to 79.60%. Their work underscores the importance of context in interpreting subtle emotional states.

Another novel approach is SmileyNet (Al-Halah et al. 2019), which exploits emoji-based embeddings to overcome the limited size of sentiment datasets. By gathering 4 million images from Twitter, annotated with emojis, the authors trained a deep network to predict these emojis, thereby creating sentiment-aligned embeddings. This embedding outperformed traditional object-based representations (e.g., ImageNet) in visual sentiment analysis and fine-grained emotion classification. Furthermore, hybrid architectures that combine Multi-Scale Dynamic 1D CNNs with Gated Transformers have been shown to be highly effective in EEG-based emotion recognition, demonstrating the value of fusing spatial-spectral features with global dependencies (Cheng et al. 2024).

Despite their success, CNNs also come with notable challenges. They can be computationally expensive to train, especially as depth increases (e.g., ResNet (He et al. 2015)), and they sometimes struggle to capture long-range global relationships within an image. These limitations have paved the way for alternative architectures, particularly those built on transformers.

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. 2017). 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.

2.4. 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.

2.4.1. Multimodal Models

A variety of multimodal models have been proposed to address the challenges of sentiment and emotion analysis. These approaches often combine textual and visual information to capture complementary cues that improve recognition accuracy. Below, we highlight four representative models that illustrate different strategies for integrating and aligning multimodal data.

VisualBERT

VisualBERT (Li et al. 2019) is a transformer-based architecture designed for vision-language tasks. It unifies visual and textual inputs within the same transformer encoder, leveraging BERT's bidirectional attention mechanism to process image regions and text tokens jointly. VisualBERT has shown strong performance in tasks such as visual question answering and image captioning. However, it relies on aligned image-text pairs during training, which constrains its flexibility in scenarios where perfectly paired multimodal data may be limited.

MuAL: Multimodal Sentiment Analysis with Cross-Modal Attention and Difference Loss

MuAL (Deng et al. 2024) addresses the need for robust cross-modal representations specifically in sentiment analysis. By incorporating cross-modal attention and a difference loss to reduce the gap between image and text representations, MuAL outperforms traditional unimodal methods and demonstrates strong transfer-learning capabilities. Notably, it remains effective even when its pre-trained parameters are frozen, highlighting its potential for real-world applications where computational efficiency and generalization are crucial.

MULSER: Multi-Level Semantic Reasoning Network

While some multimodal approaches focus on broad sentiment categories (e.g., positive vs. negative), MULSER (Zhu et al. 2023) targets fine-grained emotion classification, such as differentiating between "happiness" and "love." The model employs graph attention networks to build multi-level graphs for images (object-level, global-level, and joint regional-global) and word-level graphs for text. A cross-modal attention fusion module then integrates the enriched visual and textual features. The resulting framework achieves state-of-the-art accuracy and F1 scores, underscoring the importance of sophisticated semantic reasoning and cross-modal interaction for nuanced emotion detection.

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 labelled 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.

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

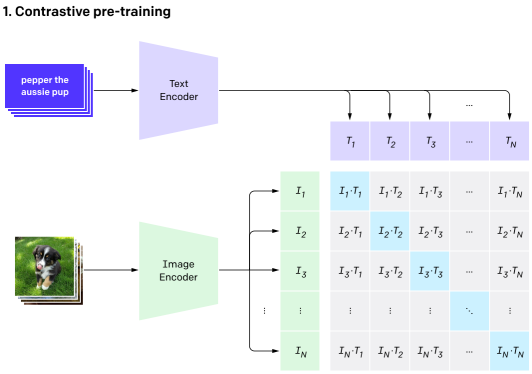


Figure 2.1.: Contrastive Pre-training

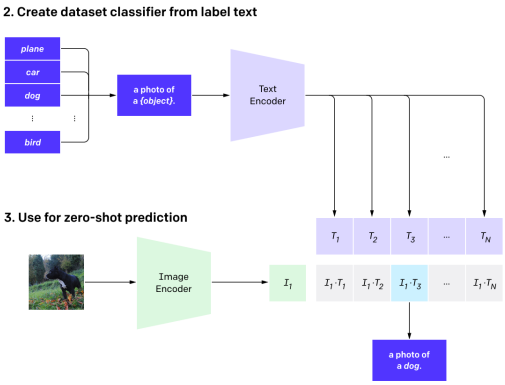


Figure 2.2.: Creation of dataset classifier and final prediction

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

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.

Through these examples, we see how multimodal models vary in their approaches to data alignment, representation learning, and zero-shot inference. However, each faces common challenges in effectively combining heterogeneous sources of information. In the next section, we review broader fusion strategies that address these challenges more generally.

2.4.2. Fusion Strategies

In addition to designing specialized architectures, researchers have explored general strategies for effectively integrating multiple data streams. This topic has garnered increasing attention due to the growing availability of diverse data sources, including text, images, audio, and video. Two key surveys offer comprehensive overviews of multimodal alignment and fusion techniques:

- **Pawlowski et al. (2023):** In "Effective Techniques for Multimodal Data Fusion: A Comparative Analysis" (Pawłowski et al. 2023), the authors systematically evaluate the performance of different fusion paradigms, ranging from early to late

fusion, and highlight the trade-offs in computational cost and representational power.

- **Li et al. (2024):** In "Multimodal Alignment and Fusion: A Survey" (Li and Tang 2024), the authors discuss alignment mechanisms, such as cross-attention and co-attention, as well as the role of contrastive objectives in bridging modalities. The survey also addresses practical considerations like data imbalance and missing modalities.

These analyses underline the importance of aligning representations from each modality before fusing them, as well as the potential benefits of modular architectures that can be tailored to domain-specific needs. By leveraging strategies drawn from these studies, researchers can build models that not only detect broad sentiments but also distinguish subtle emotional nuances from noisy, unstructured real-world data.

Fusion Techniques and Their Applications

Researchers have proposed various techniques to integrate data from multiple modalities effectively. In particular, two influential studies by Pawlowski et al. and Li & Tang provide comprehensive analyses of fusion strategies, each emphasizing different facets of multimodal data integration.

Pawlowski et al. (2023). 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 when 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 at the decision level, making it robust against modality-specific noise.

In contrast, **early fusion** integrates modalities by concatenating their embeddings at the input level, which can be advantageous when modalities are highly interdependent. However, Pawlowski et al. observe that early fusion often underperforms compared to late fusion. This pattern is notably evident in the MovieLens datasets, where the merged embeddings sometimes lead to information loss or redundancy, especially if one modality is less informative.

The **sketch representation** technique, which maps each modality into a lower-dimensional common space via hash functions, offers a memory-efficient alternative. While it underperforms on classification accuracy relative to the other methods, its scalability makes it attractive for large-scale systems (e.g., recommendation engines). Pawlowski et al. underscore the importance of selecting a fusion approach that aligns with task requirements, modality impact, and memory constraints.

Li & Tang (2024). Building on these ideas, Li & Tang propose a broader taxonomy by categorizing fusion strategies into **encoder-decoder fusion**, **kernel-based fusion**, **graphical fusion**, and **attention-based fusion**. Each method aims to exploit inter-modal relationships to improve performance on tasks that combine text, images, audio, and other data sources.

Encoder-Decoder Fusion. In this approach, each modality is first processed by a separate encoder to produce latent representations, which are then merged by a decoder. This strategy is particularly suited for tasks like image captioning and video summarization, where the goal is to generate coherent outputs (e.g., natural language descriptions) from multiple input types (Baltrušaitis et al. 2019).

Attention-Based Fusion. Attention mechanisms have gained widespread popularity due to their ability to selectively weight the importance of features across modalities. Li & Tang highlight the effectiveness of models like ALBEF (Align Before Fuse) and BLIP (Bootstrapped Language-Image Pretraining) (Li et al. 2021, 2022), which first align modalities before fusing them in a joint latent space. These techniques are especially valuable in social media analytics and emotion recognition, where subtle interactions among text, images, and audio can significantly influence the outcome (Poria et al. 2017).

Alignment Challenges and Conclusion

Despite these advancements, aligning heterogeneous data sources remains a critical challenge. Li & Tang distinguish between **explicit alignment** methods (e.g., Dynamic Time Warping, Canonical Correlation Analysis) and **implicit alignment** approaches (e.g., attention mechanisms, GANs, and VAEs). Explicit methods are especially useful for tasks requiring precise temporal or spatial synchronization (e.g., video-audio alignment), whereas implicit methods learn a shared latent space that can adapt to incomplete or noisy inputs.

Both Pawlowski et al. and Li & Tang underscore the need for robust, scalable frameworks to handle the growing volume and complexity of multimodal datasets. Challenges such as modal feature misalignment, varying data quality, and high computational costs persist. Furthermore, the limited contribution of certain modalities (as observed with visual data in the MovieLens case) highlights the importance of selecting the most informative data sources. Future research will benefit from standardized benchmarks, akin to GLUE in NLP, and the continued development of adaptive methods—such as attention-based and graphical fusion techniques—to effectively integrate diverse data modalities at scale.

2.5. 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.

2.5.1. 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.

2.5.2. 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.

2.6. 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 ClimateTV dataset introduced in Chapter 1, this chapter outlines the methodology for analyzing climate change perceptions using social media data. The approach involves two key components: (1) **Benchmarking zero-shot and task-specific text models**, and (2) **Fine-tuning text and image models with soft labels** to enhance the capture of nuanced emotions in tweets and their associated images. Additionally, weakly supervised learning approaches are explored to enhance classification performance. This structured approach ensures addressing key challenges associated with analyzing multimodal social media data, such as linguistic variability, noisy inputs, and the subjective nature of climate change discourse. The methodology is structured as follows:

1. **Data Preparation**: The ClimateTV dataset is pre-processed to ensure high-quality text and image inputs. This includes filtering, language detection, tweet selection criteria, and additional preprocessing steps to standardize data for model evaluation.

2. **Evaluation Metrics**: Metrics are established to systematically assess the performance of zero-shot and task-specific models as well as fine-tuned models using soft labels, ensuring a fair comparison.

3. **Benchmarking Zero-Shot and Task-Specific Text Models**: A range of pre-trained text models, both zero-shot and task-specific for emotion classification, are evaluated to identify the most effective approach.

4. **Fine-Tuning with Soft Labels**: The best-performing text model is leveraged to generate soft labels, which are then used to fine-tune additional text and image models, preserving nuanced emotional signals inferred from tweet replies.

5. **Weakly Supervised Learning**: Weakly supervised techniques are explored to enhance classification performance.

3.2. Data Preparation

3.2.1. Data Filtering

1. **Temporal Filtering**: To capture temporal variations in climate change discourse, this work focuses on the months of **February and August 2019** from ClimateTV.

2. **Language Filtering**:
    - **Primary Tweets**: The `papluca/xlm-roberta-base-language-detection` model from Hugging Face (papluca 2022) 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 ClimateTV dataset to English.

3.2.2. Pre-processing

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).

3.2.3. Final Dataset Statistics

Table 3.1 presents the final dataset statistics after pre-processing the data between February and August 2019.

Metric	February 2019	August 2019
Total Tweets	1,774	7,769
English Tweets	1,673	7,345
Replies	12,616	51,147
Images	1,673	7,345

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

This pre-processing 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 assess ranking performance for pre-trained models (including both zero-shot and task-specific models) and those that measure distributional accuracy when models are fine-tuned with soft labels.

3.3.1. Evaluation Metrics for Zero-Shot and Task-Specific Models

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): This metric assigns a weighted score to each prediction, giving higher importance to correct labels that appear earlier in the ranked list. Specifically, if a correct label is found at rank r, it receives a score of $\frac{1}{\log_2(r+1)}$. This formulation ensures that models are rewarded for ranking relevant classes higher, thereby emphasizing the importance of correctly prioritizing relevant predictions.

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 with Soft Labels Evaluation Metrics

The following metrics evaluate how closely the predicted probability 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:

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

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

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

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

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 the metrics described above. 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. Benchmarking Zero-Shot and Task-Specific Text Models

This section outlines the approach to evaluating and adapting text models for emotion classification in climate-change discourse. Two main categories of models are considered: (1) **Zero-shot models**, which are general-purpose models applied to this task without prior fine-tuning, and (2) **Task-specific models** that had been pre-fine-tuned for emotion detection.

3.4.1. Dataset and Annotation

To establish a ground-truth evaluation dataset, **99 English replies** are sampled from the corpus and manually assigned a single emotion label to each, following Ekman's six basic emotions—*anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise*. The label distribution is summarized in Table 3.2.

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

Table 3.2.: Distribution of Manual Emotion Labels

3.4.2. Models

The five models benchmarked are chosen for either their zero-shot classification capabilities or their task-specific fine-tuning on Twitter emotion detection:

1. `facebook/bart-large-mnli` (Lewis et al. 2019)
   - A BART model trained for Natural Language Inference (NLI).
   - Used in a **zero-shot** classification setting, employing an entailment-based approach (Yin et al. 2019).
2. `MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7` (Laurer et al. 2022)
   - A multilingual DeBERTa model trained on cross-lingual NLI tasks.
   - Enables **zero-shot** classification across different languages.
3. `MoritzLaurer/deberta-v3-large-zeroshot-v2.0` (Laurer et al. 2023)
   - A high-performing DeBERTa model fine-tuned specifically for robust **zero-shot** classification.
4. `cardiffnlp/twitter-roberta-base-emotion-latest` (Antypas et al. 2023)
   - Built on a **RoBERTa-base** architecture.
   - Fine-tuned for **emotion detection on Twitter**, aligning well with our dataset.
   - Part of the **SuperTweetEval** benchmark, trained on 154M tweets.
5. `cardiffnlp/twitter-roberta-large-emotion-latest` (Antypas et al. 2023)
   - A **RoBERTa-large** variant of the above model with greater capacity for contextual reasoning.
   - Also fine-tuned on Twitter data for emotion detection.

3.4.3. Label Mapping to Ekman's Six Emotions

The CardiffNLP RoBERTa models originally predict emotions across 11 categories (e.g., love, trust, optimism, anticipation). However, to align with the manually annotated dataset, which follows Ekman's six basic emotions (anger, disgust, fear, joy, sadness, surprise), a mapping is applied to consolidate the predictions into these six categories. This transformation is essential for a direct and meaningful evaluation against the annotated dataset.

Table 3.3 illustrates the mapping process, where semantically similar emotions are grouped based on shared affective characteristics. For example, love, trust, and optimism are combined under joy due to their common positive valence. Similarly, anticipation, which often carries an element of unpredictability, is merged with surprise.

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.

3.4.4. Evaluation Setup

Each model was evaluated in two ways:

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

2) Secondary Evaluation (Confidence Filtering) Low-confidence predictions were filtered out by retaining only those with a confidence score above 0.9. This step measured:

- the proportion of samples that met the threshold, and
- the accuracy of those high-confidence predictions.

3.5. Fine-Tuning with Soft Labels

This section presents the methodology for fine-tuning unimodal (text-only, image-only) and multimodal (text & image) models using soft labels. The best performing text model is leveraged to produce label distributions across all replies linked to each parent tweet. For each original tweet, the reply-level probability distributions is aggregated into a single soft label signal by averaging 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 enables the model to learn from the aggregated soft label distributions while ensuring that the emotional nuances embedded in the reply-level probability distributions are retained. 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 3.5.2.

3.5.1. Architectural Overview

In both cases i.e. Unimodal (Text or Image) and Multimodal, A standard MLP is attached as described below to the encoder for classification, unless specified otherwise.

Projection Head (MLP)

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

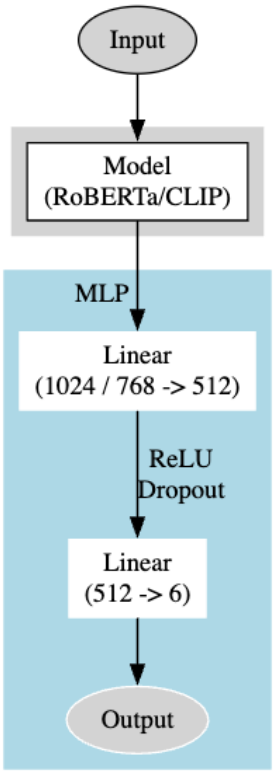


Figure 3.1.: Single Modality Architecture

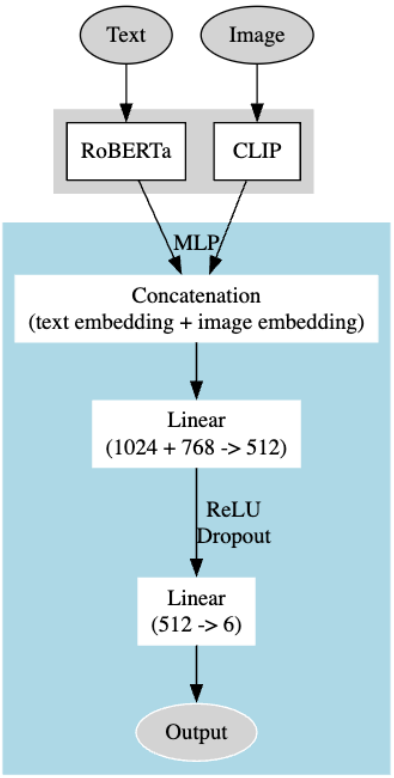


Figure 3.2.: Multimodal Fusion Architecture

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, Many fine-tuned variations are tested. The details are:

- **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), performed along with staggered-unfreezing.

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

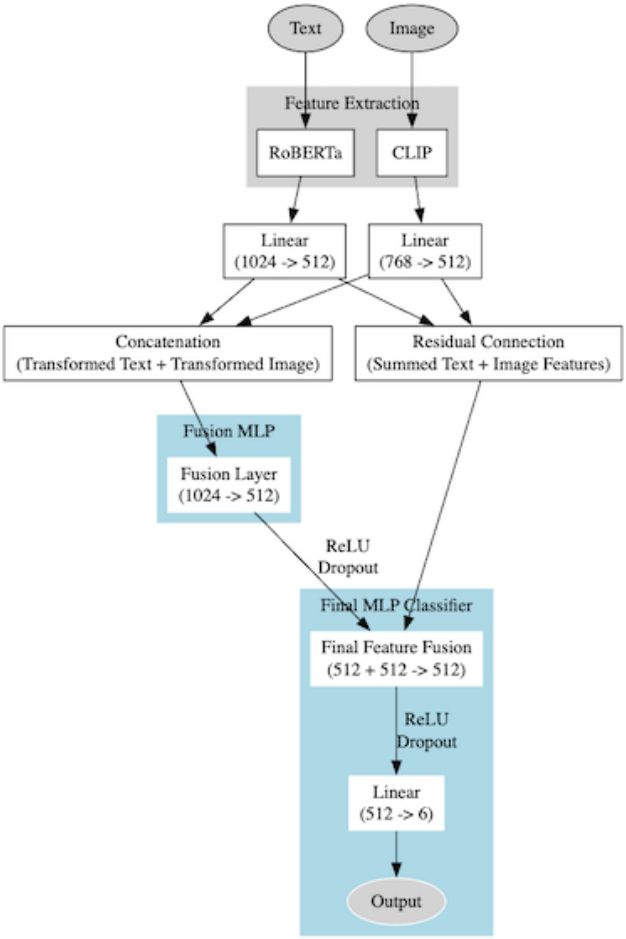


Figure 3.4.: Multimodal Residual fusion architecture.

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

This approach of fine-tuning contrasts with baseline models, which in this study are as follows:

- For the **text-based approach**, the baseline model is the task-specific **CardiffNLP RoBERTa-Large**.
- For the **image-based approach**, the baseline model is **CLIP ViT-L/14**, which is used in a zero-shot setting.
- For the **multimodal approach**, the baseline predictions are obtained by averaging the outputs from these two unimodal baseline models.

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

3.5.2. 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

- **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).

- **Data Split:** A 70-20-10 split was used for training, validation, and testing, respectively.

Reproducibility Protocol

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

3.6. Weakly Supervised Learning

In this section, two weakly supervised approaches were explored to improve model performance. The first approach builds upon (Gera et al. 2022), which proposed a self-training framework to enhance zero-shot classification by iteratively fine-tuning models on their own high-confidence predictions. The second approach focuses on loss re-weighting, where we integrate confidence scores into the loss function to prioritize reliable pseudo-labels while mitigating the impact of noise.

3.6.1. Zero-Shot Classification Boost with Self-training

As outlined in Section 2.5, (Gera et al. 2022) proposed a self-training framework. 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

- 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))* were used as described in the original work.
- Weak supervision signals were derived solely from class names and unlabelled text, mirroring the zero-shot setup.

Self-Training Protocol

- **Pseudo-Label Generation:** For each dataset, pseudo-labels by selecting the model's most confident predictions were generated (threshold: $\tau = 0.9$, as in the original study).
- **Token Masking:** The token masking heuristic to mask tokens with high semantic similarity to class names (using cosine similarity over SBERT embeddings) was implemented.
- **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).

The framework was adapted 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).

3.6.2. Loss Re-weighting

To mitigate label noise from weak supervision, a confidence-based loss re-weighting strategy was implemented as described in sub-section 2.5.1. 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. Model output probabilities were used as confidence scores.
- **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**: BERT-base model was fine-tuned 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.

4. Experimental Results

4.1. Overview

This chapter provides a comprehensive evaluation of the climate change emotion classification pipeline, covering zero-shot baselines, fine-tuned text models, image models, and multimodal models. The analysis begins with benchmarking several pre-trained language models to identify the strongest text-based baseline. Fine-tuning, particularly when guided by soft label distributions, is shown to significantly enhance performance across distribution-oriented metrics such as Cosine Similarity, KL Divergence, and MSE. A deeper dive into text-only, image-only, and multimodal approaches follows, highlighting the impact of modality fusion on performance, along with key considerations in hyperparameter tuning and encoder-freezing strategies.

Model sensitivity to dataset sizes is also examined, comparing tweets from August and February. Results indicate that data volume has a more pronounced effect on performance than random initialization. A comparative analysis of hyperparameters further reveals the critical influence of learning rate, dropout, epoch count, and encoder-freezing strategies on each modality's results.

To provide additional context for these numerical findings, the chapter concludes with an error analysis and a qualitative examination of specific tweets. The discussion highlights challenges in predicting minority emotions such as sadness and disgust, the tendency to overproduce dominant categories like surprise and joy, and difficulties in capturing low-entropy reply distributions. Finally, results from a weakly supervised self-training approach, effective on certain benchmark datasets but yielding inconsistent benefits in this domain, underscore the complexity of real-world climate change discourse.

4.2. Benchmarking Zero-Shot and Task-Specific Text Models

This section presents the performance of five pre-trained text 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.

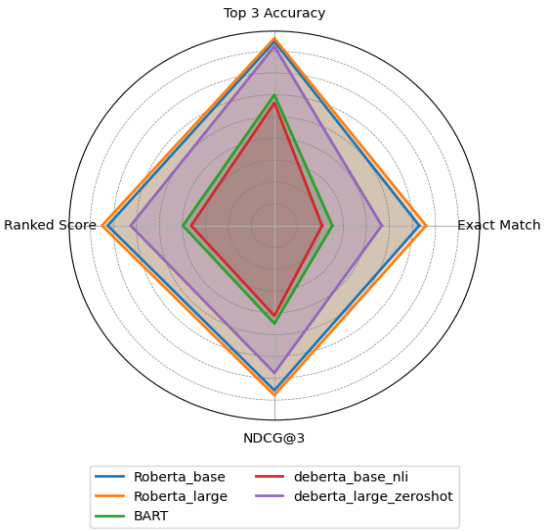


Figure 4.1.: Primary Evaluation: Performance comparison of text models based on key evaluation metrics.

4.2.1. Model Performance

Table 4.1 and Figure 4.1 summarize the results. The CardiffNLP RoBERTa-Large, 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 CardiffNLP RoBERTa-base, also performed well, outperforming the more general-purpose BART-large-MNLI and DeBERTa-based models in all metrics.

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.: Primary Evaluation: Performance of the text models on the annotated subset ($n = 99$).

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 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.

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

The CardiffNLP RoBERTa-large 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.

Based on both primary and secondary evaluation metrics outlined in Section 3.4.4, the CardiffNLP RoBERTa-large model demonstrates the highest overall performance, making it the most effective model.

4.3. Fine-Tuning with Soft Labels

4.3.1. Text-Based Results

Table 4.3 summarizes the results for our best text-based experiment, showing significant improvements across all metrics.

Metric	Our Model			Baseline		
	August	February	Average Score	August	February	Average Score
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 4.3.: Performance comparison between Baseline and Our Model. The Baseline is the CardiffNLP RoBERTa-Large. Higher values are better for ↑, and lower values are better for ↓. Bold values indicate averaged scores.

Our model achieves an average 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.

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.

4.3.2. Image-Based Results

Table 4.4 presents the results for our best image-based experiment.

Metric	Our Model			Baseline		
	August	February	Average Score	August	February	Average Score
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 4.4.: 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 averaged scores.

Our model achieves an average 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).

4.3.3. Multimodal Results

Table 4.5 summarizes results for our best multimodal experiment.

Our model achieves an average Cosine Similarity of 0.8127, outperforming the Baseline by **14.4%** ($\Delta = +0.1010$). Month-wise improvements show a 16.6% gain in August

Metric	Our Model			Baseline		
	August	February	Average Score	August	February	Average Score
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 4.5.: 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 averaged scores.

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.

4.4. Sensitivity to Datasets and Random Initialisation

All experiment configurations described in Section 3.5.2 (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.

4.4.1. Comparision Across Datasets

Table 4.6 presents aggregated metrics for all experiments for each modality (averaged over the two seeds) on the August vs. February data. We observe that models perform significantly better on August data across all modalities, likely due to the larger volume of available data (4.3x more than February). The increased data availability improves model generalization and stabilizes predictions, leading to higher Cosine Similarity and lower error-based metrics (MSE, KL Divergence).

Prior research suggests that climate change discourse on social media follows a cyclical pattern, with engagement spiking around major focusing events such as COP conferences, climate protests, and extreme weather events. However, the higher data availabil-

ity in August does not necessarily indicate heightened engagement but rather a larger presence of climate-related content overall, possibly due to greater discussion of summer weather patterns and environmental conditions. 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 4.6.: 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:

- **Impact of Data Volume on Model Performance:** Across all modalities, August data consistently leads to higher Cosine Similarity scores and lower error-based metrics (MSE, KL Divergence). The 4.3x larger dataset size provides models with a richer and more diverse representation of climate discourse, allowing for better generalization.
- **Seasonal Trends in Climate Change Discourse:** While engagement on climate change tends to peak in the late months of the year (October–December) due to global climate summits and activism, August appears to have a larger overall volume of discussions. This may be driven by extreme weather reporting, environmental tourism, and increased discussions around summer climate anomalies.
- **Stable Ordering Across Modalities:** In both datasets, multimodal models outperform text-only models, which in turn outperform image-only models, reinforcing the importance of leveraging both textual and visual cues.
- **Statistical Significance:** An independent samples t-test comparing performance on August vs. February data confirms significant differences:
  - **Cosine Similarity:** August data shows significantly higher values ($t = 34.52$, $p < 0.0001$), likely due to greater data availability stabilizing predictions.
  - **MSE:** August data exhibits significantly lower values ($t = -12.49$, $p < 0.0001$), suggesting lower variability in sentiment expression due to more consistent discussions.

– **KL Divergence:** August data also demonstrates significantly lower values ($t = -10.40$, $p < 0.0001$), indicating better model certainty in predictions.

4.4.2. Effect of Random Initialisation

To understand how initialization affects outcomes, each configuration was also trained using two random seeds. Table 4.7 presents aggregated metrics for each modality (averaged over both datasets) for seeds 42 vs. 7.

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 4.7.: 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).

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.

4.5. Comparative Analysis

This section provides a unified view of performance across text, image, and multimodal models. The analysis begins with an overview of top and bottom performers (Table 4.8), followed by detailed insights into the role of hyperparameters in each modality.

4.5.1. Overall Performance Across Modalities

Table 4.8 summarizes the highest and lowest performers, revealing that multimodal approaches consistently outperform single-modality setups.

Model	Ranking-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
Top Performers (Single-Modality)								
RoBERTa-Large	2.431	0.802	0.347	0.023	1e-5	0.3	10	Text Only, Frozen
CLIP	2.417	0.799	0.358	0.023	5e-6	0.5	2	Image Only, Tuned
Bottom Performers (Multimodal)								
RoBERTa-Large	2.308	0.771	0.434	0.028	1e-5	0.3	5	GELU
RoBERTa-Large	2.303	0.768	0.436	0.028	1e-5	0.3	5	Frozen CLIP
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 4.8.: Top and bottom performers across text, image, and multimodal settings.

Key Observations from Table 4.8

- **Multimodal Superiority (on Average)** The top multimodal models achieve higher cosine similarity scores (0.811, 0.812, and 0.813) compared to the bottom single-modality models (0.760 for text-only and 0.788 for image-only). However, the top-performing single-modality models (RoBERTa-Large text-only: 0.802 and CLIP image-only: 0.799) are not far off from the top multimodal configurations. This suggests that not all multimodal configurations are superior, and poor multimodal fusion strategies (e.g., inappropriate freezing/tuning combinations) can reduce performance below strong single-modality baselines. Despite this, the best multimodal models still consistently outperform their best single-modality counterparts, reinforcing the benefit of using both text and image data when optimized properly.

- **Efficient Convergence** Multimodal models converge in 2–5 epochs while still achieving top-tier performance. The single-modality text-only model that remains frozen and trains for 10 epochs performs well (cosine similarity = 0.802). However, when fine-tuned for 10 epochs, the text-only model's cosine similarity drops to 0.760, a decrease of over 5%. A similar trend is observed in image-only models, where longer training and fewer frozen layers lead to slight but notable drops in performance. This indicates that careful selection of epoch count and freezing/-tuning strategies is particularly important for single-modality setups.

4.5.2. Hyperparameter Impact: Text-Only Models

Experiments on text-only RoBERTa models reveal critical hyperparameter interactions:

Tuning	Optimizer	Learning Rate	Epochs	Dropout	CosineSim ↑	KLDiv ↓	MSE ↓	Ranking-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 4.9.: Impact of learning rate, epoch count, optimizer, and freezing on text-only performance. The best configuration is in bold.

Key Trends from Table 4.9:

- **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.

4.5.3. Hyperparameter Impact: Image-Only Models

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

Key Trends from Table 4.10:

- **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.

Tuning	Optimizer	Learning Rate	Epochs	Dropout	CosineSim ↑	KLDiv ↓	MSE ↓	Ranking-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 4.10.: Impact of learning rate, epoch count, optimizer, and freezing on image-only performance. The best configurations are in bold.

4.5.4. Hyperparameter Impact: Multimodal Models

Multimodal configurations demonstrate synergistic benefits from modality fusion:

Model	**Optimizer**	**Learning Rate**	**Dropout**	**Epochs**	**Fusion/MLP**	**Tuning**	**CosineSim ↑**
RoBERTa-large & CLIP	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 & CLIP	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 4.11.: Impact of learning rate, epochs, dropout, optimizer, fusion strategy, and freezing on multimodal performance. The best configurations are in bold.

Key Trends from Table 4.11:

- **Freezing Efficacy**: Fully frozen RoBERTa-Large and CLIP achieve the highest cosine similarity score (0.8126), emphasizing the value of preserving pre-trained features. Partial tuning (e.g., RoBERTa-Base with GELU) narrows the gap to 0.13% (0.8126 vs. 0.8115).
- **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.

These insights highlight the importance of balancing freezing, dropout, and fusion strategies to maximize multimodal model efficiency.

4.5.5. Summary of Comparative Findings

- **Multimodal Performance vs. Single-Modality**
  Across all experiments, the highest average cosine similarities (0.811–0.813) are achieved by well-optimized multimodal models. While this confirms the general benefit of combining text and image features, there are cases where strong single-modality models (e.g., RoBERTa-Large text-only at 0.802 or CLIP image-only at 0.799) can match or even exceed weaker multimodal configurations (0.768–0.771). In other words, merely blending modalities does not guarantee superior results; effective hyperparameter tuning (e.g., learning rate, dropout, freezing strategies) is essential. Nevertheless, the top-tier multimodal approaches maintain a decisive edge over the best purely text-based or image-based models.

- **Importance of Freezing and Regularization**
  Freezing pre-trained encoders consistently shows notable benefits for both single- and multi-modal setups. By preserving robust representations, frozen models avoid overfitting during extended training (e.g., 10 epochs), producing up to a 3.7% higher score compared to fully tuned counterparts in text-only experiments. Similarly, multimodal models that freeze both RoBERTa-Large and CLIP achieve the highest cosine similarity (0.8126). In contrast, excessive fine-tuning, especially over more epochs, tends to degrade performance for text-only and image-only models alike.

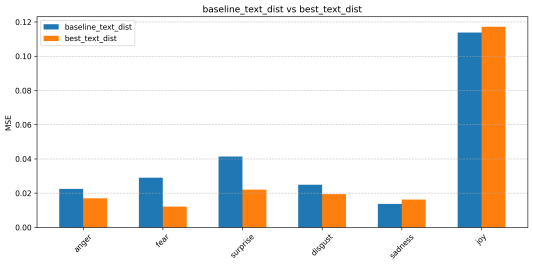
- **Convergence Efficiency and Epoch Management**
  A consistent trend across text, image, and multimodal experiments is that peak performance often emerges within just 2–5 epochs. Notably, text-only models that remain frozen and train for 10 epochs can still perform well (cosine similarity = 0.802). However, extending to 10 epochs while fine-tuning leads to sharp declines (down to 0.760). Image-only experiments also show that short training (2 epochs) with a tuned approach and low learning rate (5e-6) yields top scores, whereas 10 epochs often result in 3–5% performance drops. Early stopping and careful epoch management are therefore critical, particularly in single-modality contexts.

- **Architectural and Optimizer Observations**

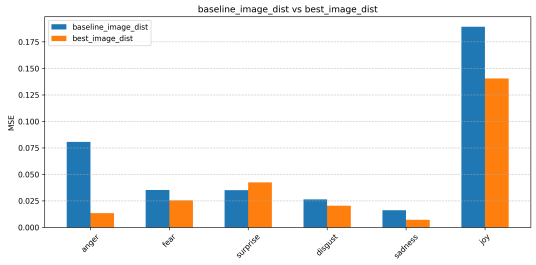
- **Model Size and Dropout:** Larger architectures (e.g., RoBERTa-Large) typically need higher dropout (0.5) to stabilize training, whereas smaller ones (RoBERTa-Base) perform best with dropout around 0.3. Despite having fewer parameters, RoBERTa-Base can still match or closely approximate RoBERTa-Large's performance if paired with effective activation functions (e.g., GELU) and partial tuning.
- **Optimizer Choice:** Both Adam and AdamW yield comparable results (less than 0.5% difference), indicating that while AdamW is marginally better in some setups, freezing strategies, learning rates, and dropout rates have a much larger overall impact on performance.

Overall, the findings emphasize that well-structured multimodal models, particularly those that freeze pre-trained encoders and employ moderate regularization, offer the highest cosine similarity scores, surpassing even the strongest text-only or image-only baselines. Hyperparameter selection (learning rate, dropout) and training procedures (epoch count, freezing vs. tuning) emerge as the most influential factors across all model types.

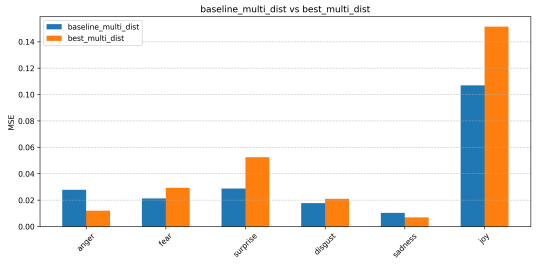
4.6. Error Analysis



(a) MSE: Our Text Model vs. Baseline



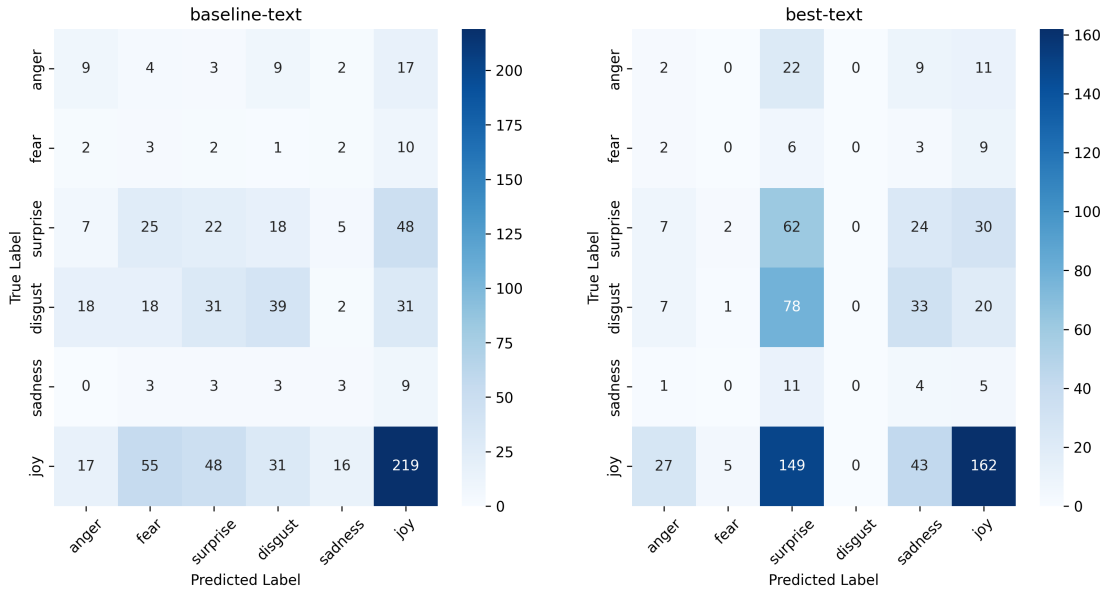
(b) MSE: Our Image Model vs. Baseline



(c) MSE: Our 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 4.3–4.5), 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 4.6.1–4.6.2.



(a) Baseline Text Model.

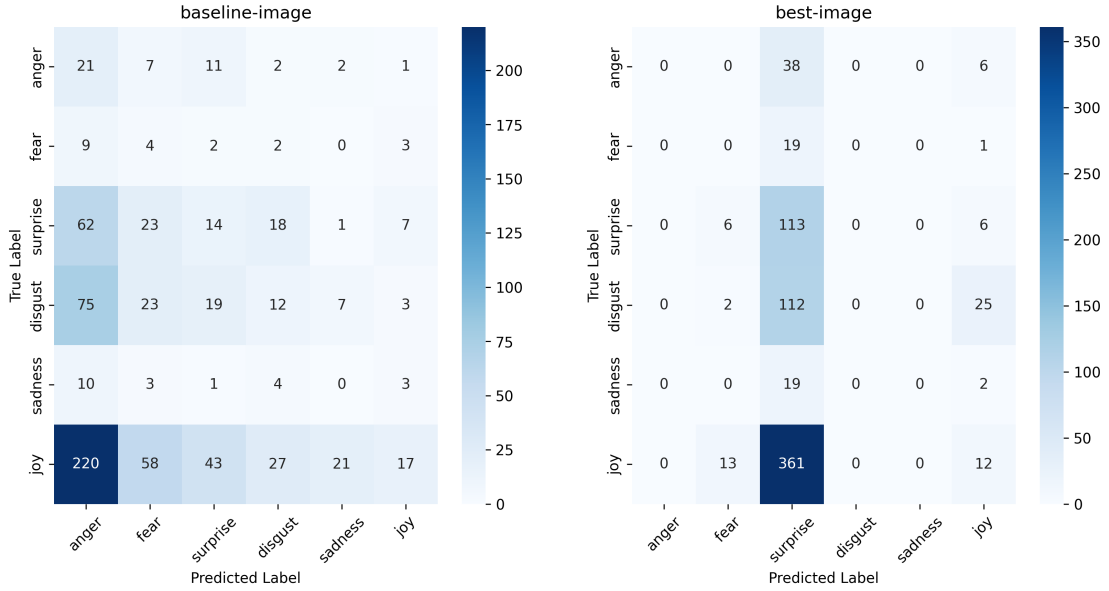
(b) Our Text Model.

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

4.6.1. 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) (Figure 4.3b), while image models can fail to detect sadness, disgust and anger altogether (0% recall) (Figure 4.4b). 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) (Figure 4.3b) when improving other metrics. Image-only models overcompensate for anger bias, predicting surprise almost exclusively (113/125 correct on surprise, but 361 false positives) (Figure 4.4b). Multimodal fusion inherits both problems, absorbing the text model's sadness mistakes (92% misclassifications) (Figure 4.5b) and the image model's surprise bias.



(a) Baseline Image Model.

(b) Our Image Model.

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

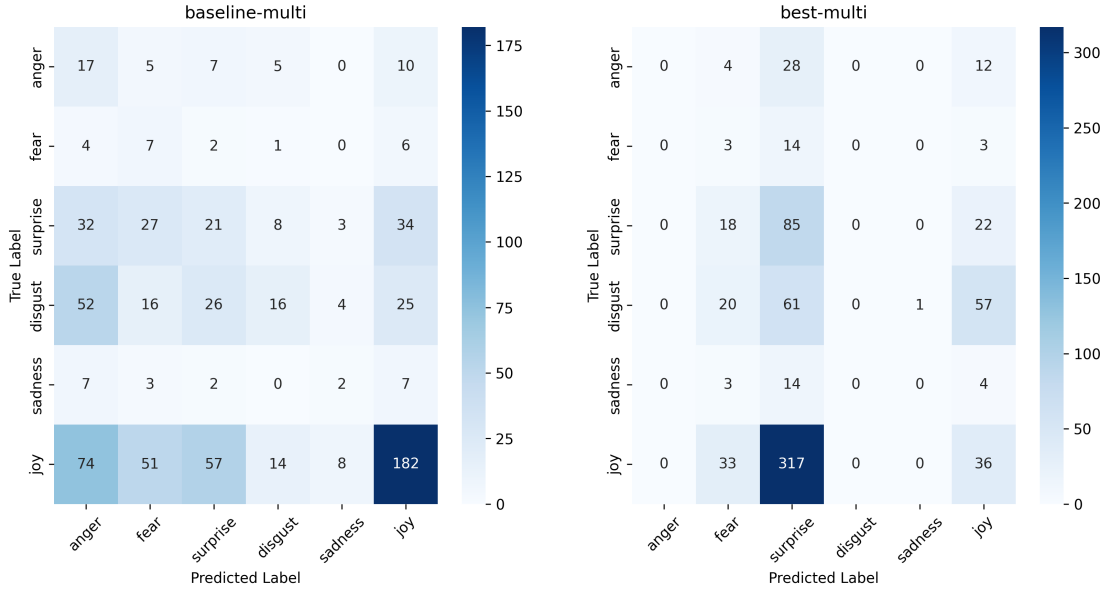
- **Distribution-Instance Dissonance.** Large gains in cosine similarity (16–56% improvement) can mask argmax errors. For example, joy appears best predicted (162–219 correct in text models), but this partly stems from excessive joy predictions (36–48% of all outputs) (Figure 4.3b). The inherent trade-off between distributional alignment and per-instance fidelity remains a core challenge.

4.6.2. 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).

Text Model. From Figure 4.2a, our 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 4 of 21 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 (Figure 4.2b), our 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



(a) Baseline Multimodal.

(b) Our Multimodal.

Figure 4.5.: Multimodal confusion matrices.

as surprise). Hence, while distribution metrics improve overall, the confusion matrix illustrates a strong one-label bias.

Multimodal Model. Finally, the multimodal model (Figure 4.2c) shows mixed gains. Our 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.

4.7. Qualitative Analysis

In this section, a detailed qualitative analysis of predicted emotion distributions is presented. Alongside each tweet snippet, its aggregated distribution of replies (the soft label), and model predictions, the samples of individual replies are also shown. This provides insight into how the model might struggle or excel in capturing these varied emotional signals.

Poor Performance Cases

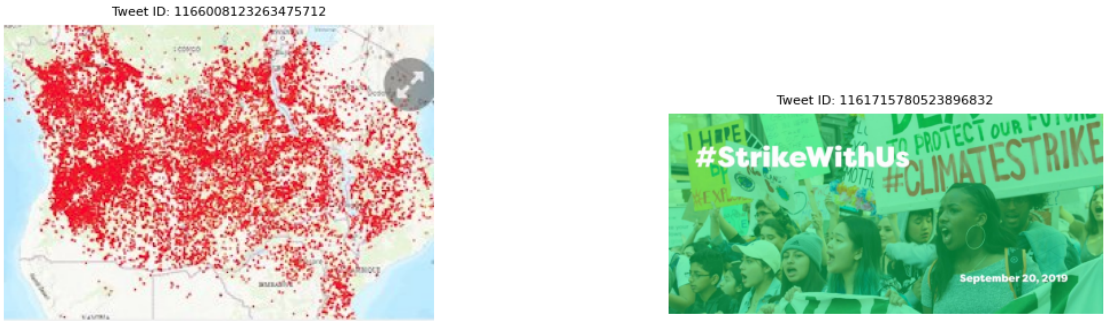


Figure 4.6.: Examples of high-error predictions (Case A on left, Case B on right).

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):

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

Our Predictions (Image-only):

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

Our Predictions (Multimodal):

{anger: 0.1089, fear: 0.2705, surprise: 0.2549,
disgust: 0.0992, sadness: 0.1060, 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), our multimodal model's final 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.

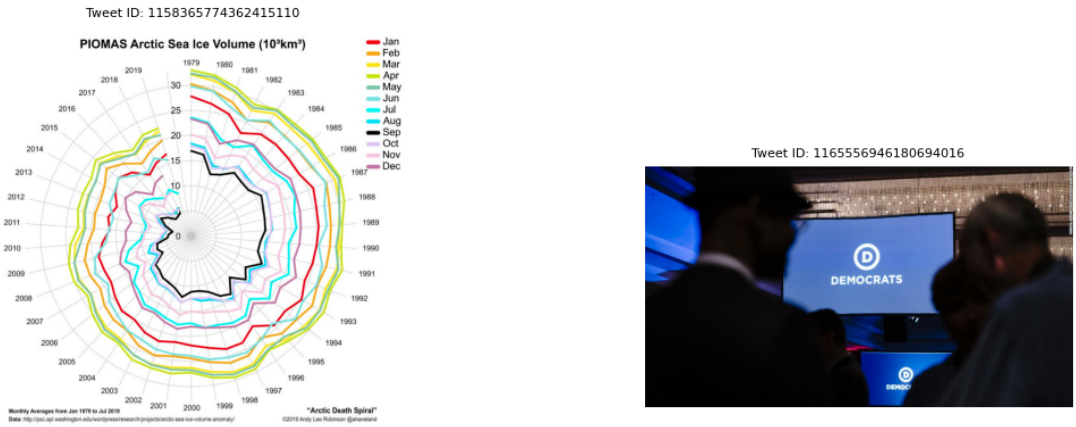


Figure 4.7.: Examples of well-aligned predictions (Case C on left, Case D on right).

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):

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

Our Predictions (Text-only):

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

Our Predictions (Image-only):

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

Our Predictions (Multimodal):

{anger: 0.2084, fear: 0.1822, surprise: 0.1753,
disgust: 0.0910, sadness: 0.0857, 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.

4.7.1. 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.

4.8. Weakly Supervised Learning Results

4.8.1. 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 4.12 presents the accuracy results on three datasets: AG (Zhang et al. 2015), ISEAR (Shao et al. 2015), 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 4.12.: 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.

4.8.2. 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 4.13 provides per-class accuracy comparisons. Additionally, the classification reports in Table 4.14 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 4.13.: Per-class accuracy comparison.

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.

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 4.14.: Classification report comparison before and after training.

4.9. Summary

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 discussion 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

5.1. Overview

This chapter examines how zero-shot and fine-tuned models classify emotions in climate change discourse on Twitter, emphasizing the value of task-specific pretraining. Comparisons between zero-shot models and CardiffNLP RoBERTa model underscore that alignment with social media data significantly boosts performance.

Beyond text-only methods, the chapter explores image-based and multimodal approaches. Visual cues often add valuable context such as protest imagery yet can introduce biases (e.g., an overemphasis on surprise). Evaluation metrics like cosine similarity, KL divergence, and MSE sometimes mask crucial misclassifications, underscoring the need for tailored measurement protocols. Overfitting remains a challenge, notably in fine-tuned and multimodal setups, prompting careful regularization and tuning strategies.

Finally, weakly supervised techniques, including self-training and loss re-weighting, yield modest gains but cannot fully resolve class imbalance or the subtleties of real-world emotional expression. The chapter concludes by proposing future research directions—improved label aggregation, multi-task training, dynamic fusion, and more robust evaluation—to address persistent gaps in classifying complex climate discourse.

5.2. Analysis of Findings

5.2.1. Zero-Shot vs. Task-Specific Text Models

The CardiffNLP RoBERTa models, pre-trained on Twitter data for emotion classification, consistently outperformed general-purpose zero-shot models across evaluations, underscoring the critical role of task-specific pretraining. Notably, the RoBERTa-large variant emerged as the strongest performer when considering both quantitative metrics and confidence-based filtering. Its superior Exact Match (EM) and Top-3 Accuracy highlighted the advantages of domain-specific adaptation, particularly in capturing the linguistic nuances and informal tone inherent to social media text.

A confidence threshold analysis (0.9) revealed a precision-coverage trade-off: while DeBERTa-v3-large-Zeroshot achieved the highest precision (72.73%), it produced far fewer confident predictions, reflecting greater model uncertainty. In contrast, the CardiffNLP RoBERTa-large model retained 63% of high-confidence predictions, striking an optimal

balance between reliability and coverage. This dual evaluation which combines ranking metrics with confidence filtering solidified RoBERTa-large's superiority, demonstrating that task-specific training is essential for robust performance in emotion classification.

The evaluation's limited annotated dataset (99 samples) introduced challenges due to class imbalance (e.g., 33 anger vs. 10 surprise labels). Such skew risks inflating EM scores if models over-predict majority classes. To address this, Top-3 Accuracy and ranking metrics (Ranked Score, NDCG@3) were prioritized, enabling nuanced assessment of label-ranking accuracy in ambiguous cases. Despite these constraints, the CardiffNLP RoBERTa models maintained consistent performance, further validating their domain-specific training.

5.2.2. Experimental Results

Text-Based Results

Table 4.3 demonstrates that additional fine-tuning on climate-related data significantly improves the performance of CardiffNLP RoBERTa-Large in text-only scenarios. Comparing Cosine Similarity, KL Divergence, and MSE reveals that the refined model not only fits the overall distribution of emotions more precisely (indicated by large reductions in KL) but also achieves a substantial 43% decrease in MSE. These results suggest that aligning the text encoder with climate-related language helps the model capture nuanced emotional cues, whether the tone is anger toward political inaction, fear of climate disasters, or joy in positive environmental developments.

Use of the Same Model for Label Generation and Fine-Tuning. It is worth noting that we employed the same CardiffNLP RoBERTa-Large model to both generate and later learn from the soft-label distributions. Specifically, we ran inference on each tweet's replies, aggregated those multiple outputs into a single emotion probability distribution, and used that distribution as the supervisory target for fine-tuning on the original tweet text. This raises the potential concern of circular reasoning, where the model is trained to reproduce its own predictions. While we do not have definitive evidence to fully rule out this possibility, two aspects suggest that this approach may still be beneficial. First, the reliance on reply-level inference injects genuine diversity and contextual depth into the labelling process: the aggregated labels are not simply a single inference from the original tweet, but rather a composite of varied emotional responses within the conversation thread. Second, the process of domain-specific fine-tuning reshapes the model parameters to handle climate-related language and context, potentially allowing it to go beyond merely memorizing its own outputs. This is an aspect that could be explored further to better understand its implications. Despite these considerations, the substantial gains in KL Divergence and MSE (Table 4.3) suggest that this strategy effectively leverages the model's strengths in emotion detection while adapting it to the nuanced climate domain.

Image-Based Results

In Table 4.4, fine-tuning CLIP ViT-L/14 on climate-related images yields marked gains in matching annotated user emotions. Compared to the zero-shot baseline, Cosine Similarity increases by over 50%, while MSE drops by more than 60%. However, without proper regularization, fine-tuning risks overfitting: climate imagery often includes protest signs, infographics, or disaster photos that may lead the model to latch onto superficial visual cues (e.g., color or composition) rather than underlying emotional context. Techniques such as moderate learning rates and partial freezing proved essential to stabilize these improvements.

Multimodal Results

Table 4.5 shows that combining textual and visual signals consistently outperforms single-modality approaches. Across Cosine Similarity, KL Divergence, and MSE, multimodal models better align with ground-truth emotion distributions. Month-wise analyses (e.g., August vs. February) demonstrate stable improvements, suggesting that text–image fusion can capture complementary information: language-based emotional content plus the visual impact (e.g., protest crowds, melting icebergs). Nonetheless, not all multimodal setups automatically surpassed the strongest unimodal baselines; careful tuning and fusion strategies were necessary to avoid overshadowing textual cues with dominant visual features (or vice versa).

Residual fusion with staggered unfreezing resulted in one of the lowest-performing multimodal approaches (Table B.1). The added complexity of residual connections may have hindered effective feature integration, amplifying discrepancies between text and image representations. Additionally, staggered unfreezing, intended to facilitate gradual adaptation, may have introduced instability, leading to overfitting on early-stage features before achieving a meaningful joint representation. These findings highlight the need for a more structured fusion and training strategy to balance multimodal contributions without excessive variance.

Despite this, the rationale for staggered unfreezing remains valid. CLIP, optimized for image-text alignment, may not fully capture emotional nuances, whereas RoBERTa is better suited for sentiment tasks. Unfreezing CLIP earlier allows image features to better align with the emotional space of text while keeping RoBERTa frozen longer preserves its strong generalization. However, unfreezing CLIP too soon may have destabilized multimodal alignment, while sequential fine-tuning may have led to suboptimal convergence. Future work could explore adaptive fine-tuning schedules to mitigate these issues while preserving the intended benefits.

Label Aggregation and Distribution Flattening

The manual mapping of emotions (e.g., merging love, optimism, and trust into a single joy category) introduced semantic ambiguity, as evidenced by the confusion matrices in

Figures 4.3–4.5. Two main issues emerged:

- **Overprediction of merged classes:** Joy and surprise (each absorbing multiple original labels) dominated predictions. In the best text-only model, joy had 162 correct predictions but also 149 false positives, sometimes misclassifying anger/sadness as joy. Qualitative analysis highlights this confusion: in Case B, a tweet about climate strikes elicited replies dominated by surprise (62.4%), but the model redistributed probabilities to fear and anger (Figure 4.6), conflating semantically distinct emotions. Likewise, surprise (merged with anticipation) was overpredicted for anger (22/44 instances) and sadness (43/83).
- **Minority class collapse:** Rare emotions such as fear and sadness suffered catastrophic recall ($< 10\%$ in the best multimodal model), with fear often collapsing into surprise or joy. For instance, 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 of fear (less than half its ground truth of 48.8%), with the excess probability allocated to surprise and anger. Persistently low MSE values for sadness (0.0069–0.0162) and fear (0.0120–0.0293) indicate the model minimized their contributions by predicting near-zero probabilities.

Using KL divergence as the loss function exacerbated these issues by emphasizing smoothness across the distribution rather than per-class accuracy. In Case B (Figure 4.6), the model's prediction for a reply expressing 62.4% surprise was flattened to 46.5%, with the remainder assigned to unrelated classes like fear and anger. Averaging labels across replies also diluted target distributions, discouraging confident predictions for these minority classes.

Qualitative Analysis

Detailed examples (Section 4.7) reveal how different reply entropies affect model performance:

- **High-entropy replies:** When users express a mixture of emotions, the model performs better because the training objective anticipates broad distributions.
- **Low-entropy replies:** When most users converge on one emotion, the model typically underestimates the true peak, hesitating to allocate very high probability to a single label.

Additionally, noisy or ambiguous replies (e.g., those containing slang or incomplete sentences) can skew distributional predictions, particularly when the total number of replies is small. This underscores the importance of stringent data filtering and text-cleaning strategies in real-world applications.

MSE values reveal modality-specific weaknesses:

- **Text models:** Achieved the lowest anger MSE (0.0169 vs. 0.0224 baseline) but often missed anger instances (only 2/44 correctly identified). In Case D, the text model conflated anger ("This is stupid") with disgust due to overlapping vocabulary, despite clear textual cues.
- **Image models:** Overpredicted surprise (361 false positives), likely due to CLIP's bias toward visually salient cues. In Case C (Figure 4.7), an Arctic ice melt graph caused overestimation of surprise (46.5%), even though replies emphasized fear, revealing a mismatch between visual salience and textual sentiment.
- **Multimodal fusion:** Although the best multimodal model reduced anger MSE (0.0120 vs. 0.0278), confusion matrices indicate it inherited text models' tendency to overpredict surprise. In Case C, the fusion model assigned 27.7% probability to surprise (matching aggregated replies) but missed the correct intensity of fear (14.8% vs. 27.7% ground truth), highlighting lingering cross-modal conflicts.

Metric Misalignment and Practical Implications

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

- **Cosine Similarity** rewards distributional shape matching but can mask severe misclassifications. In Case C, the model attained high cosine similarity by balancing joy, surprise, and fear proportions, yet failed to identify the true magnitude of fear (14.8% vs. 27.7%).
- **MSE** penalizes large deviations yet remains insensitive to label confusion. In Case D, anger and disgust were both predicted at high probabilities, leading to low MSE despite critical misclassification.

Overfitting and Generalization Trends

Train and validation loss curves (Tables 4.9–4.11) confirm the tendency of tuned models to overfit. While their training losses quickly decrease to very low levels, frozen models typically exhibit higher, more gradually declining losses. Conversely, validation losses often rise for tuned models, reinforcing the overfitting hypothesis. The Appendix (Chapter C) provides plots of these loss curves.

5.2.3. Weakly Supervised Learning

Zero-Shot Classification Boost with Self-training

In replicating the study from Gera et al. (2022) (Table 4.12), self-training provided consistent improvements on classic datasets (AG, ISEAR), demonstrating that iterative pseudo-labelling can refine zero-shot entailment models. These gains, however, did

not universally extend to the new ClimateTV dataset. While BART benefited significantly (up to +9.1% on ClimateTV), DeBERTa and RoBERTa saw no improvement or plateaued performance. This discrepancy suggests that the domain-specific complexity of climate change discourse, potentially involving technical terms, political nuances, or varied emotional expressions, poses challenges that self-training alone may not overcome for certain architectures.

Loss Re-weighting

Fine-tuning with loss re-weighting (Table 4.13 and 4.14) yielded a modest global accuracy improvement (+3.03%). Classes such as anger and surprise experienced the most benefit, reflecting that these classes can be bolstered by redistributing gradient emphasis away from dominant categories like joy. However, some classes like disgust deteriorated, highlighting that re-weighting can inadvertently downplay a minority class if it is not carefully calibrated. Additionally, the confidence of predictions dropped slightly on average, indicating that while the model made more correct classifications for certain labels, it became less certain overall, likely a symptom of a more evenly distributed training focus among various classes.

5.3. Future Directions

To address the identified limitations, we propose 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 Aggregation**
  - **Why**: Relying solely on soft labels generated from pre-trained models can propagate 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 labelling to calibrate and refine 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.

By coupling 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 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

This thesis set out to explore how state-of-the-art text and image models can enhance our understanding of the emotional impact of climate-related content on social media, particularly in the absence of explicit emotion labels. Our investigation demonstrated that domain-specific pre-training, such as CardiffNLP RoBERTa, provides a notable advantage over general-purpose language models for analyzing emotionally charged social media text.

Beyond text-based insights, our findings underscore the potential of multimodal fusion: integrating textual and visual features consistently improved emotion classification performance compared to unimodal approaches. However, this fusion also introduced new challenges, such as an overestimation of emotions like surprise when processing visually striking images. These challenges highlight the need for careful model calibration to balance distribution-wide fidelity with per-instance accuracy, particularly when working with inherently noisy social media data.

To address these limitations, we proposed various strategies, including hybrid loss functions, refined label aggregation techniques, and adaptive fusion mechanisms. By refining these approaches, future research can develop more robust and context-aware emotion classification systems. Such advancements will empower researchers, policymakers, and activists to better understand and respond to public sentiment surrounding climate change, ultimately fostering more effective communication and engagement strategies.

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 pre-training 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.

papluca (2022). xlm-roberta-base-language-detection. https://huggingface.co/papluca/xlm-roberta-base-language-detection.

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 (2017). 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

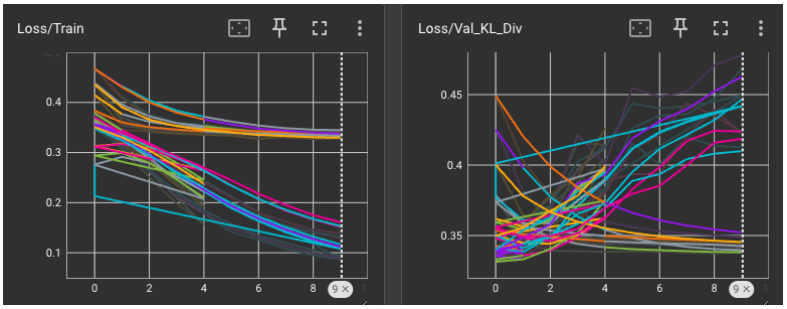
Table B.2 Continued from previous page

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

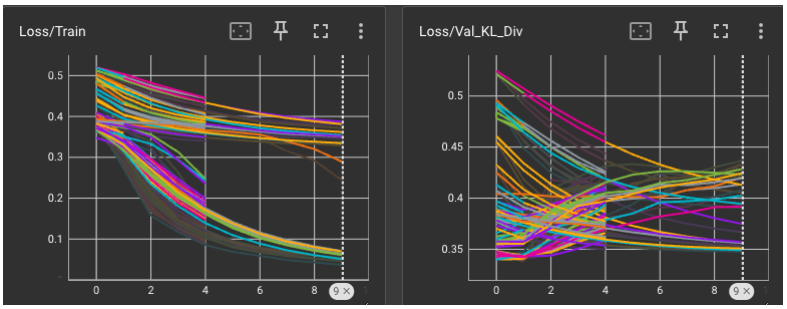
Table B.3.: Complete Text experimental results

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

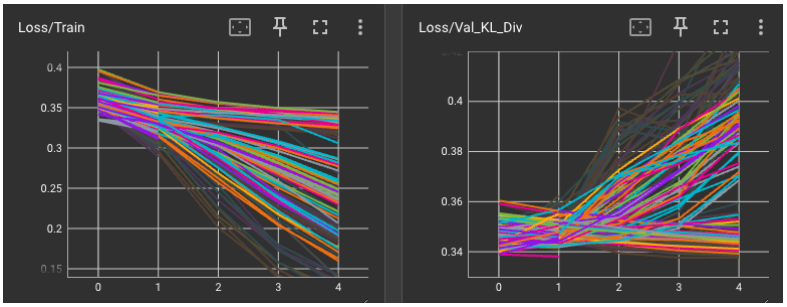
C. Loss Curves



(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?
GPT-4o	Rephrasing	Throughout	++
Deepseek-R1	Summarization of related work	Chap. 2	+
GPT-4o	Code generation	For visualisations	++
GPT-4o	Latex table formatting	Throughout	++

Unterschrift
Mannheim, den 14.02.2024