**7 Text-Based Emotion Recognition and Detection: Mining Sentiments and Feelings**

Nataša Kovač, Faculty of Applied Sciences, University of Donja Gorica, Oktoih 1, Podgorica, 81000, Montenegro, https://orcid.org/ [0000-0002-6671-2938](https://orcid.org/0000-0002-6671-2938)

Hojjatollah Farahani, Department of Psychology, Faculty of Humanities, Tarbiat Modares University, Tehran, Iran, <https://orcid.org/0000-0002-9799-7008>

Peter Watson, MRC Cognition and Brain Sciences Unit, University of Cambridge, 15 Chaucer Road, Cambridge, CB2 7EF, United Kingdom, https://orcid.org/0000-0002-9436-0693

**Abstract:** In this chapter, the focus shifts to the profound influence of textual content on emotion recognition and detection. The chapter commences by underscoring the potency of words in conveying emotions, delving deep into the domain of textual emotion exploration. It then explores the critical role of natural language processing techniques in dissecting and understanding emotions encoded in text. The chapter introduces the EmText methodology, an advanced tool for sentiment analysis within the context of embodied emotion. Additionally, it presents Emsemitext, a specialised approach for deriving clinical insights from textual data. The chapter concludes by showcasing case studies that illuminate real-world applications of text-based emotion analysis, offering compelling evidence of its practicality and effectiveness.

**Keywords:** Exploring Emotion in Text; Natural Language Processing; Analysing Sentiments; Emsemitext for Clinical Insights

Emotions shape human experiences deeply, influencing decisions, relationships, health, and overall well-being. Among various ways to express emotions, textual data uniquely captures complex emotional differences through words and language structures. With the rapid growth of digital communication platforms, social media, and online interactions, analysing text has become an essential method for emotion recognition, significantly impacting psychology, health informatics, marketing, and artificial intelligence (AI). This integration aligns with the Artificial Psychology (PsAIchology) framing that connects AI methods and psychological science (Farahani et al., 2024). Textual data naturally contain emotional information, which can be explicit, through clearly stated emotional words (e.g., *happy*, *anxious*, *angry*), or implicit, conveyed through tone, context, metaphor, or sentiment. Emotional analysis of text requires robust computational methods to decode these intricate emotional expressions accurately. The rise of Natural Language Processing (NLP) techniques has greatly improved our ability to interpret and analyse emotional content in textual data.

Historically, text-based emotion recognition was mainly manual, with psychologists relying on qualitative analysis of diaries, interviews, and written correspondence. In contemporary psychological research, the integration of advanced computational techniques such as text mining and sentiment analysis has opened new methodological frontiers for understanding the nuances of human emotional expression within therapeutic and clinical contexts. The spoken words of patients captured during interviews are more than narrative data; they are psycholinguistic structures that encode affective states, cognitive patterns, and interpersonal dynamics. By applying text mining to transcribed interviews, psychologists can systematically detect latent themes, relational markers, and shifts in language usage that may remain elusive through traditional qualitative coding alone (Tausczik & Pennebaker, 2010). Such computational augmentation does not replace human interpretation but rather enriches it, enabling a multi-layered analysis of emotional tone, self-referential language, and the evolution of linguistic markers that correlate with psychological change.

Sentiment analysis extends this potential by quantifying emotional valence in patient discourse, thus transforming subjective narratives into measurable indicators of affective states. When integrated into a longitudinal therapeutic framework, sentiment analysis allows clinicians to observe whether emotional polarity (in terms of positivity, negativity, or ambivalence) shifts across treatment sessions (Mohammad, 2021). This quantitative insight aligns closely with the psychodynamic principle that language serves as an externalised expression of internal affective regulation (Ricardi, 2025). As patients progress through therapy, their lexical patterns often become less dominated by catastrophic or self-blaming terms and more characterised by agentic, hopeful expressions, phenomena that can be captured empirically through affective lexical modelling. This analytic complement to clinical intuition supports the growing call for data-informed clinical practice in psychotherapy research (Ingram & Siegle, 2009).

Beyond individual analysis, text mining methods facilitate the identification of collective response patterns across patient populations. Through natural language processing pipelines, large corpora of therapeutic session transcripts can be coded for thematic density, metaphorical content, or even markers of therapeutic alliance, offering psychologists a scalable means to test psychological theory against naturally occurring discourse (Pennebaker & Chung, 2011). Importantly, this approach underscores the ethical imperative of maintaining confidentiality while harnessing anonymised linguistic data to drive precision in clinical psychology (Alm, 2012). The fusion of computational and interpretative paradigms thus represents an epistemological evolution: the movement from purely hermeneutic understanding toward an integrative model where human empathy and machine learning together construct deeper insights into mental healing. By systematically analysing how language embodies transformation, psychologists can trace therapeutic progress in a more transparent and scientifically rigorous manner, transforming qualitative narratives into bridges between subjective experience and measurable change. The development of computational linguistics introduced quantitative approaches, enabling researchers to systematically evaluate large textual datasets. Early works by Pennebaker and Chung (2011) established linguistic markers linked to psychological states and health outcomes, forming a foundational knowledge that informs modern NLP-based emotional analysis.

Contemporary methods employ sophisticated NLP techniques, including sentiment analysis, emotion lexicons, and machine learning models. Sentiment analysis typically categorises text into positive, negative, or neutral sentiments, providing a foundational but coarse-grained approach to emotional analysis (Liu, 2022). Emotion lexicons such as NRC Emotion Lexicon (Mohammad & Turney, 2013), Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001), and Affective Norms for English Words (ANEW) (Bradley & Lang, 1999) provide comprehensive word-emotion mappings, enabling more detailed emotional insights. Machine learning and deep learning further advance this analysis by extracting emotional context from semantic embeddings, achieving unparalleled accuracy in emotional classification tasks (Devlin et al., 2019). For psychology-focused, Python-based pipeline patterns and regression modelling, see (Kovač et al., 2024).

Text-based emotion recognition has proven particularly insightful in clinical settings. Analysis of patient narratives, social media posts, or therapeutic dialogues reveals emotional patterns essential for mental health diagnostics and interventions. Recent studies demonstrate the efficacy of textual emotion detection in identifying early indicators of depression, anxiety, or suicidal ideation, underscoring its critical role in proactive mental health care (De Choudhury et al., 2013a). Integrating textual emotional data with physiological and behavioural measures further enriches emotional analysis, providing a comprehensive understanding of embodied emotional experiences. Multimodal approaches leverage synchronised text, voice, facial expressions, and physiological data, significantly enhancing emotion recognition accuracy (D'Mello & Kory, 2015).

Ethical considerations remain paramount in text-based emotional analysis. Text data often reveals deeply personal and sensitive emotional states, raising privacy concerns. Researchers must uphold stringent ethical standards, including participant consent, data anonymisation, transparency of data usage, and secure data handling practices (Jobin, Ienca, & Vayena, 2019).

To facilitate readers' hands-on exploration, public datasets like the Emotion Intensity in Tweets dataset (EmoInt) (Mohammad & Bravo-Marquez, 2017), the SemEval Sentiment Analysis datasets (Dimitrov et al., 2024), or publicly available tweets and Reddit discussions offer valuable resources for practical emotional analysis. These datasets enable robust methodological exploration, supporting reproducible and transparent research.

**7.1 The Power of Words: Exploring Emotion in Text**

Words are more than mere symbols arranged on a page or a screen; they carry emotional resonance, capture human experiences, and reveal deep psychological truths. Through words, we narrate our joys and sadness, express fears and hopes, and communicate frustrations and desires. The emotional power embedded in language provides a window into the human psyche, making text a rich source for understanding not just what we say, but how we feel. Emotion embedded in language is central to our identity, social interactions, and psychological health (Pennebaker & Chung, 2011).

Humans inherently use language as an expressive tool, instinctively encoding emotions within words. The process of writing or speaking involves not only cognitive but also emotional processes, linking internal emotional states with external linguistic expressions. Research demonstrates that linguistic choices, such as specific word usage, sentence structure, and narrative style, reflect underlying emotional and psychological states (Tausczik & Pennebaker, 2010). For example, a person experiencing anxiety might subconsciously use more tentative words like *maybe* or *perhaps*, while someone expressing sadness may use more first-person singular pronouns (*I*, *me*, *myself*) and negative emotion words like *loss*, *sad*, and *alone*. Through the systematic analysis of these language patterns, researchers can identify emotional states that may not be immediately obvious through other modes of communication. Emotions embedded in text can often be subtle, concealed in linguistic cues rather than being explicitly stated. Words carry emotional weight not only in their literal meaning but also in their connotations and contextual usage. For instance, the phrase *I'm fine* can either communicate genuine well-being or mask deeper emotional distress, depending on the broader linguistic setting. Likewise, metaphors and idiomatic expressions (*feeling blue*, *broken-hearted*, *on cloud nine*) show how language creatively captures complex emotional states.

Language patterns not only reflect immediate emotional reactions but can also reveal long-term psychological states and personality traits. Linguistic analysis has demonstrated potential in detecting depression, anxiety, stress, and even signs of suicidal ideation through textual data such as social media posts, therapy transcripts, or personal diaries (De Choudhury et al., 2013b). James Pennebaker's foundational work on expressive writing reveals the therapeutic power of emotional expression through text. He found that individuals who wrote openly about traumatic events experienced significant improvements in mental and physical health compared to those who wrote neutrally about mundane topics (Pennebaker, 1997). Writing about emotions, even without direct feedback, seems to organise and clarify thoughts, reducing stress and fostering emotional regulation.

Thus, the emotional significance of text has implications beyond mere analysis. It directly impacts psychological well-being. This connection between emotion, language, and health forms the theoretical foundation for numerous applications in clinical psychology, digital health interventions, and emotional AI. To systematically explore emotional content in language, researchers rely on frameworks categorising emotions along key dimensions or discrete categories:

* **Dimensional Models** classify emotions according to continuous dimensions, primarily valence (pleasantness vs. unpleasantness), arousal (intensity), and dominance (control or power). Russell’s Circumplex Model is a prime example, placing emotions in a two-dimensional circular space defined by valence and arousal (Russell, 1980).
* **Categorical Models** are found in discrete emotional theories, which identify specific emotions such as joy, sadness, anger, fear, surprise, and disgust. Theories by Ekman (1992) or Plutchik (1980a) are well-known categorical models used extensively in text emotion analysis.

Below is an illustrative Python example for quickly exploring emotional dimensions using the NRCLex package (Mohammad & Turney, 2013), providing an intuitive bar plot visualisation of emotional word distribution:

from nrclex import NRCLex

import matplotlib.pyplot as plt

text = "Today, I'm thrilled about the upcoming vacation but slightly anxious about flying."

# Analyze emotional content

emotion\_analysis = NRCLex(text)

emotion\_freq = emotion\_analysis.raw\_emotion\_scores

# Plot emotional distribution

plt.figure(figsize=(8, 4))

plt.bar(emotion\_freq.keys(), emotion\_freq.values())

plt.title("Distribution of emotions")

plt.xlabel("Emotion")

plt.ylabel("Frequency")

plt.tight\_layout()

plt.show()

Visualising textual emotion via word clouds helps uncover dominant emotional themes intuitively. Below is an example illustrating this method:

from wordcloud import WordCloud

text = "I felt joyful, inspired, and motivated after talking to my friend, though still slightly worried."

wordcloud = WordCloud(width=800, height=400, background\_color='white').generate(text)

plt.figure(figsize=(10, 5))

plt.imshow(wordcloud, interpolation='bilinear')

plt.axis("off")

plt.title("Word Cloud-emotions")

plt.tight\_layout()

plt.show()

A word cloud (Figure 1) produced by previous code instantly communicates the emotional landscape, capturing dominant emotions (e.g., joyful, motivated, inspired) that vividly portray a complex emotional moment.



Figure 1. Word cloud example

While text emotion analysis offers profound insights, ethical considerations must guide its practice. Analysing emotional content in text data must always respect privacy, consent, and confidentiality. Text data can reveal deeply personal psychological states, making anonymisation and ethical oversight non-negotiable components of research and practical applications (Jobin et al., 2019).

## 7.2 Natural Language Processing for Emotion Analysis

NLP is the set of methods that lets computers work with human language. In the context of emotion, NLP helps us move from unstructured text to structured evidence about what people may feel. A post, a review, or a line of dialogue becomes a set of clues. By combining linguistics with statistical learning, NLP turns those clues into interpretable signals about affect and mood (Jurafsky & Martin, 2025; Cambria et al., 2015).

Most projects follow a simple flow. We begin by tidying the text so that later steps do not get distracted by noise. Lowercasing, splitting into words, and reducing words to a base form prevent us from counting the same idea in many ways (Bird et al., 2009). Common words such as *the* or *and* are often removed because they rarely carry emotion. This stage does not detect feelings on its own. It prepares the ground so later methods can focus on meaningful content.

A straightforward way to estimate emotions is to look up words in a curated dictionary. Resources such as the NRC Emotion Lexicon map words to categories like joy, anger, fear, sadness, disgust, surprise, trust, and anticipation. If a text contains many joy words and very few anger words, we can say the balance of evidence points toward joy. Lexicons are transparent and easy to explain to non-specialists, and they work even when there is little training data. Their weakness is context. *Great* is usually positive, yet in *great, just perfect*, it can be sarcastic. Lexicons are best used as baselines and as tools for auditing what more complex models might be picking up (Mohammad & Turney, 2013). Many systems start with sentiment polarity because it tells us whether the overall tone is positive or negative. Rule-based tools like VADER work well on short informal texts and require no training data (Hutto & Gilbert, 2014). Polarity is useful, but emotions are richer than a single scale. A sentence can carry both excitement and fear at the same time. For that, we need models that read words in context.

The biggest advance in the last few years came from Transformer models that learn how words behave in many contexts during pretraining, then adapt to a specific task with a small amount of labelled data (Vaswani et al., 2017; Devlin et al., 2019; Rogers et al., 2021). These models are good at separating close feelings, for example, anger versus annoyance, because they consider the surrounding words and the way expressions are used together. In practice, many teams start with a public model and fine-tune it on their domain. The short snippet below sends a sentence to a pre-trained emotion classifier and returns label scores. It is meant to show the idea rather than to teach coding.

from transformers import pipeline

clf = pipeline("text-classification",

model="j-hartmann/emotion-english-distilroberta-base",

return\_all\_scores=True)

clf("I am excited about the launch but a little nervous")

Output:

[[{'label': 'anger', 'score': 0.0006416769465431571},

{'label': 'disgust', 'score': 0.0002330473653273657},

{'label': 'fear', 'score': 0.9930770397186279},

{'label': 'joy', 'score': 0.0027830556500703096},

{'label': 'neutral', 'score': 0.0005855631898157299},

{'label': 'sadness', 'score': 0.0017977369716390967},

{'label': 'surprise', 'score': 0.0008818271453492343}]]

The output lists emotions with probabilities that sum to one. You can read it as a distribution of plausible feelings expressed by the sentence.

Because emotion datasets are often unbalanced, accuracy alone can be misleading. We prefer macro-averaged F1, which treats each emotion equally, and per-class precision and recall, which reveal where a model confuses classes. When the system outputs probabilities, curves such as ROC and precision-recall help us understand performance at different decision thresholds. Visual tools like confusion matrices and label frequency plots make these ideas easy to read at a glance.

Words shift meaning with context. *Fine* can be contentment or sarcasm. Irony and humour often require knowledge that spans several sentences or a shared background. Domain shift is common. A model trained on tweets may not work well on clinical notes or customer support chat without additional tuning. Annotation is also noisy because people sometimes disagree about what a text conveys. These challenges are active research topics, and they are the reason we combine simple, explainable methods with stronger learned models.

Inferring feelings from language touches on privacy and dignity. Good practice includes clear consent for data use, minimal collection, transparency about how the model was trained, and regular checks for bias in both data and outputs (Hovy & Spruit, 2016; Blodgett et al., 2020; Jobin, Ienca, & Vayena, 2019). In sensitive settings, we recommend human oversight and aggregate reporting rather than individual classification.

**7.3 EmText: Analysing Sentiments in Embodied Emotion**

We use the term Emotionally embodied Text (EmText) to emphasise that textual expressions can reflect not only appraisals but also bodily and affective states, an approach grounded in theories of embodied/grounded cognition and embodied emotion (Barsalou, 2008; Niedenthal, 2007; Niedenthal & Maringer, 2009). Evidence linking emotional language to neural and bodily processes further motivates this view (Saxbe et al., 2013; Giraud et al., 2023), alongside multimodal datasets and reviews that connect physiology with emotion recognition (Park et al., 2020; Larradet et al., 2020).

EmText is a methodological approach designed to integrate linguistic analysis with embodied emotion frameworks. Rather than viewing text as simply informational or expressive, EmText views linguistic output as inherently tied to emotional bodily states, such as anxiety manifesting in stomach discomfort, happiness felt as warmth, or sadness experienced as heaviness in the chest. It bridges psychology, neuroscience, and linguistics, leveraging computational tools to identify deeper emotional experiences in textual data (Barsalou, 2008). Related machine-learning frameworks have been applied to affect-linked constructs beyond core emotions, such as love addiction, where features and explanations clarify predictive factors (Farahani et al., 2025). EmText relies heavily on the theory of embodied cognition, proposing that language and thought processes are deeply grounded in sensory-motor experiences. According to this framework, emotional words (*hurt*, *warm*, *heavy*) evoke corresponding physical sensations and motor responses in the reader or listener, facilitating richer emotional understanding beyond purely cognitive interpretation (Barsalou, 2008; Niedenthal et al., 2005).

For instance, reading *she felt crushed by sadness* activates neural patterns associated with physical heaviness, making emotion perception embodied rather than purely conceptual. Recognising these embodied metaphors and expressions through computational text analysis enhances the interpretive power of emotional AI applications. To analyse embodied sentiments in text, EmText methodologies typically involve:

* **Identifying Embodied Metaphors**: Detecting metaphorical language explicitly connected to bodily sensations.
* **Emotion-Sensation Lexicons**: Leveraging specialised lexicons linking emotions to physiological sensations (e.g., heat for anger, heaviness for sadness).
* **Contextual Embodied Emotion Modelling**: Using advanced NLP tools (e.g., transformers) to model emotional context and capture embodied emotional states.

Here’s a practical Python example demonstrating how to detect embodied emotional metaphors using simple keyword matching:

import re

text = "I felt heavy with grief but her smile warmed my heart."

# Define basic embodied metaphor patterns

embodied\_patterns = {

'sadness': ['heavy', 'crushed', 'weight'],

'happiness': ['warm', 'light', 'bright'],

'anger': ['hot', 'burning', 'fuming'],

'fear': ['cold', 'chilled', 'frozen']

}

embodied\_emotions = {}

for emotion, keywords in embodied\_patterns.items():

for kw in keywords:

if re.search(r'\b' + kw + r'\b', text.lower()):

embodied\_emotions[emotion] = embodied\_emotions.get(emotion, 0) + 1

print("Detected embodied emotions:", embodied\_emotions)

Output:

Detected embodied emotions: {'sadness': 1}

This simple method effectively identifies embodied emotional expressions, highlighting different emotional experiences embedded in linguistic metaphors (*heavy* for sadness, *warm* for happiness). Visualisation greatly aids in understanding how embodied emotions appear in text. Figure 2 presents one example implemented with PyPlutchik (Semeraro et al., 2021). Panel (A) displays the relative frequency of the eight primary emotions arranged on Plutchik’s wheel (Plutchik, 1980b). Each petal’s length is proportional to that emotion’s share (neutral is omitted). Panel (B) shows the primary dyads, i.e., pairs such as *love* (joy + trust), *optimism* (anticipation + joy), and *aggressiveness* (anger + anticipation), computed from the primaries and rescaled so the largest dyad reaches the outer ring. Together, these views make it easy to read both the dominant feelings in the corpus and the blended, embodied states that co-occur in language.

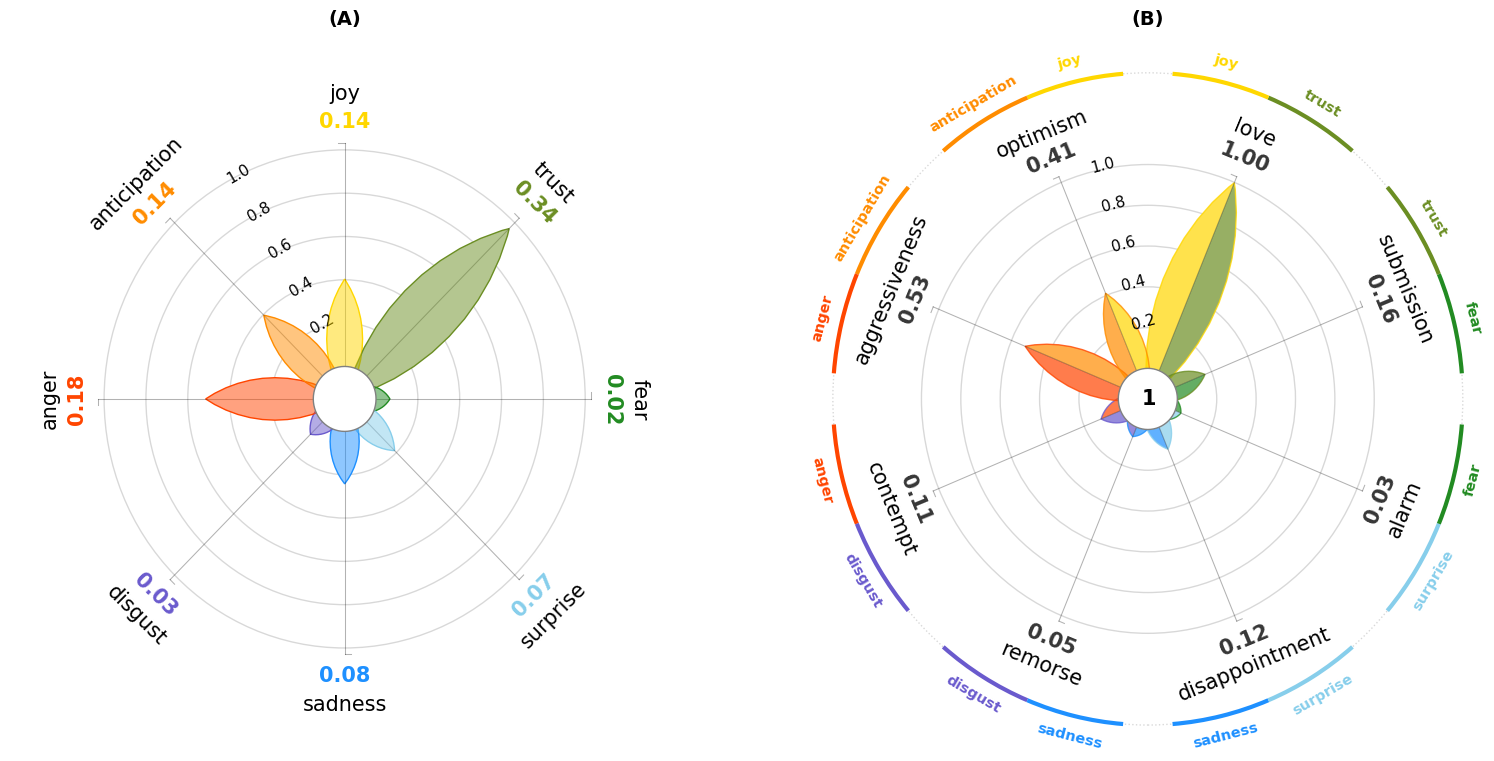


Figure 2. Embodied emotion profiles in text: (A) distribution of primary emotions on Plutchik’s wheel; (B) intensities of Plutchik’s primary dyads derived from those primaries.

Clinical psychology increasingly recognises the value of EmText methods. Analysis of therapy session transcripts can reveal embodied emotions associated with psychological conditions (e.g., anxiety manifesting as physical tension, sadness as heaviness). An EmText analysis might look for embodied emotional cues in clinical dialogues, highlighting critical therapeutic insights, as it is shown in Table 1.

|  |  |  |
| --- | --- | --- |
| Text Excerpt | Detected Emotion | Embodied Expression |
| I feel completely drained and heavy. | Sadness | heavy |
| My chest feels tight whenever worried | Anxiety | tight |
| Happiness feels warm, like sunlight. | Happiness | warm |

Table 1. Illustrative EmText cues in short clinical utterances: detected emotion and salient embodied expression

Clinical psychology increasingly recognises the value of EmText methods for reading embodied emotions in language. Short phrases in therapy transcripts often carry bodily metaphors that point to affective states. *Heavy* tends to co-occur with sadness, *tight* with anxiety, and *warm* with happiness. Table 1 shows how a simple EmText pass can surface these cues and present them in a compact form that a clinician can scan during or after a session. The goal is not diagnosis but faster noticing. When a transcript highlights repeated references to tightness, weight, or warmth, the therapist gains a structured view of how the client’s body enters the narrative.

Modern transformer models help with this task because they read words in context and learn how metaphors behave across many examples. The transformer architecture builds a representation of each token while attending to the rest of the sentence, which lets the model connect *heavy* with *sadness* in *my heart feels heavy*, and ignore it in unrelated uses (Vaswani et al., 2017). BERT introduced bidirectional pretraining that predicts masked words from both left and right context and improved many classification tasks, including emotion recognition (Devlin et al., 2019). RoBERTa refined this recipe through longer training and dynamic masking and often yields stronger results on downstream text classification (Liu et al., 2019). DistilBERT keeps most of BERT’s language understanding while being smaller and faster, which makes it practical for real-time assistance in clinics or call centres (Sanh et al., 2019). The Hugging Face ecosystem wraps these models in a single function call, which is why the short code example below returns a label and a confidence score for a clinical sentence (Wolf et al., 2020). Below is a simple demonstration of emotion classification using HuggingFace transformers:

from transformers import pipeline

classifier = pipeline("text-classification", model="bhadresh-savani/distilbert-base-uncased-emotion")

text = "My heart feels heavy with sadness."

result = classifier(text)

print("Emotion classification:", result)

Output:

config.json: 100% 768/768 [00:00<00:00, 73.9kB/s]

model.safetensors: 100% 268M/268M [00:23<00:00, 16.7MB/s]

tokenizer\_config.json: 100% 291/291 [00:00<00:00, 29.0kB/s]

vocab.txt: 232k/? [00:00<00:00, 4.78MB/s]

special\_tokens\_map.json: 100% 112/112 [00:00<00:00, 7.59kB/s]

Device set to use mps:0

Emotion classification: [{'label': 'sadness', 'score': 0.947920560836792}]

When you see {'label': 'sadness', 'score': 0.95}, read it as the model’s current best guess and its confidence. In practice, you would calibrate thresholds, evaluate per emotion, and fine-tune on domain text such as therapy notes before deployment.

Analysing embodied emotions from text carries ethical responsibilities. Emotional expressions, particularly those tied to embodied experiences, are deeply personal. Ethical guidelines must ensure sensitive handling, informed consent, anonymisation, and strict confidentiality when applying EmText analyses in research or practical applications (Jobin et al., 2019).

**7.4 Emsemitext for Clinical Insights**

This section introduces *Emsemitext*, our term for a practical approach that reads clinical language through the lens of the body. The idea is simple. People often describe feelings as bodily states. A client says *my chest feels tight*, *grief is heavy*, or *I feel warm and light again*. Emsemitext treats those phrases as signals, links them to established emotion categories, and then tracks how they change across time. The method draws on research in embodied cognition, metaphor, and clinical language, and it packages those lines of work into a workflow that a health team can adopt with modest technical support. The label is new in this book. The building blocks are not. Comparable supervised pipelines are also used across psychological health domains, including the classification of chronic pain outcomes (Kovač et al., 2025). A complementary line of work uses calibrated regression to predict internal shame, with XGBoost performing best and distress tolerance emerging as the strongest predictor (Kovač, Ratković, Farahani, & Watson, 2025b). Prior work shows that emotions are grounded in perception and action, that metaphors map bodily sensations to affective meaning, and that clinical text contains reliable markers of mental health status (Barsalou, 2008; Lakoff & Johnson, 2008; Mohammad & Turney, 2013; Cohn et al., 2004; Rumshisky et al., 2016; De Choudhury et al., 2013a).

Clinicians already listen for language that points to the body. A therapist notes repeated references to weight, tightness, heat, cold, or emptiness. A psychiatrist reads *numb* and *frozen* and wonders about dissociation. These cues are not diagnoses. They are patterns that help structure a conversation. Embodied expressions are common across languages and cultures, which means they can support cross-setting work when combined with local knowledge about idioms (Kövecses, 2003; Gibbs, 2005). Cognitive science offers a reason for their stability. When people talk about feelings, they recruit the same sensory and motor systems that help them navigate the world, so the vocabulary of the body becomes the vocabulary of emotion (Niedenthal, 2007; Barsalou, 2008).

Clinical text concentrates these signals. Therapy notes, patient journals, and peer support forums contain first-person narratives that mix symptoms, events, and feelings. Large-scale studies find that language use changes with depression and anxiety, including shifts in pronoun use, negations, and affect lexicon frequency (Rude et al., 2004; Coppersmith et al., 2014; Guntuku et al., 2017). When we add embodied phrases to this picture, we get a complementary channel that is intuitive for clinicians and easy to explain to patients. Emsemitext makes this channel explicit. Emsemitext is a **workflow** that links clinical language to embodied emotion cues and then aggregates those cues for monitoring and reflection. It is not a single algorithm. It is not a diagnostic tool. It is a way to organise textual evidence so that a team can ask better questions. The workflow rests on three pillars.

1. **A compact mapping between words and embodied cues.** For example, *heavy* tends to align with sadness, *tight* with anxiety or fear, *warm* with happiness or relief. The mapping is derived from published lexicons, clinical annotations, and small domain-specific extensions (Mohammad & Turney, 2013; Pennebaker et al., 2015).
2. **Contextual disambiguation.** The same word can be literal or metaphorical. *Carrying this heavy bag* is not the same as *my heart feels heavy*. Modern Transformer models help decide which sense is active, and they do so by reading the surrounding words and the speaker’s prior statements (Devlin et al., 2019; Liu et al., 2019; Rogers et al., 2021).
3. **Longitudinal aggregation.** A single session can be noisy. Patterns across several sessions are more informative. Emsemitext counts and normalises embodied cues per visit, then visualises trends so a clinician can see whether heaviness is fading or tightness is spiking after a medication change. This builds on work that tracks sentiment and symptom language over time in mental health records and social media (De Choudhury et al., 2013a; Rumshisky et al., 2016; Chancellor & De Choudhury, 2020).

Consider two short vignettes built from common patterns in the literature. A client in grief therapy repeatedly uses *heavy*, *weighed down*, and *drained*. Across eight sessions, the total frequency of heaviness words drops, while words of warmth and lightness begin to appear. Even without a probability model, the pattern supports the client’s report that activities feel less effortful. In a different case, a client with panic symptoms uses *tight*, *pressure*, and *grip* around the chest. The language spikes after a stressful work event, then recedes after a breathing protocol and schedule changes. Neither pattern proves recovery or relapse. Each gives the team a shared set of observations to discuss. These vignettes echo findings that bodily metaphors are stable markers of affect, that therapy progress can be read in narrative texture, and that text-based monitoring can complement clinical judgment when used with care (Kövecses, 2003; Angus et al., 1999; Cohn et al., 2004).

The core of Emsemitext is the mapping from expressions to embodied cues and then to emotion categories. The safest way to build it is to start from established resources and adapt. The NRC Emotion Lexicon links words to eight basic emotions and has been validated across tasks and languages (Mohammad & Turney, 2013). LIWC provides psychologically grounded dictionaries that have seen decades of use in clinical and health research, including categories for body and sensation (Pennebaker et al., 2015). These resources do not focus on embodied phrases, but they cover many affective expressions and provide a template for annotation guidelines.

A practical annotation protocol looks like this. Begin with a seed list of candidate phrases such as *heavy*, *light*, *tight*, *open*, *warm*, *cold*, *numb*, *empty*, *shaky*, and *grounded*. Collect short spans from your own corpus that contain these words. Ask two or more annotators with clinical knowledge to judge whether the span uses the word metaphorically to describe feeling states or literally to describe the physical world. When the usage is metaphorical, ask the annotators to link the span to one or more emotions. Measure agreement with Cohen’s kappa or Krippendorff’s alpha and resolve disagreements in discussion. Keep the label set small. The goal is to capture robust signals rather than all details. This process follows well-tested practice in clinical NLP and metaphor annotation (Neumann et al., 2019; Shutova, 2010; Demner-Fushman et al., 2009).

Two safeguards help maintain quality. First, separate detection from interpretation. Mark that a phrase is an embodied cue, then record the emotion label, and keep notes on context. Second, revalidate the mapping on a new sample every few months, since vocabulary shifts with time and with the population you serve. Drift monitoring is a standard requirement in clinical prediction models and applies here as well (Saria et al., 2018).

When the same word can be literal or metaphorical, context decides. Transformer models learn context during pretraining and have strong results on metaphor detection, emotion classification, and clinical note tasks after fine-tuning (Devlin et al., 2019; Liu et al., 2019; Chalkidis et al., 2020; Neumann et al., 2019). For Emsemitext, we use them for two tasks. First, we filter out literal uses that would inflate counts. Second, we recover emotion when the cue word is absent but the pattern is clear. A sentence like *it takes everything to get out of bed* often conveys heaviness even if the word *heavy* never appears. Model outputs are not facts. They are hypotheses that guide attention. Team members should spot-check a small sample every cycle and confirm that the reasoning still makes sense.

Model choice depends on constraints. DistilBERT is small and works on a laptop. RoBERTa is stronger in many classification tasks. ClinicalBERT style models learn from notes and can pick up domain phrasing more quickly, although they require careful governance because of the data used to pretrain them (Alsentzer et al., 2019; Sanh et al., 2019; Liu et al., 2019). Regardless of the model, calibration and per-class metrics matter because embodied cues are sparse and imbalanced. Reports should include macro-F1 and per emotion precision and recall rather than only overall accuracy, echoing recommendations in clinical NLP evaluation (Roberts et al., 2022).

Emsemitext aims to provide views that fit into a clinical hour. Three summaries tend to work well:

1. A one-page view that lists the most frequent embodied cues for the session, a brief selection of supporting quotes, and a small bar showing change since the last visit. This mirrors common practice in psychotherapy notes and supports reflective supervision.
2. A line plot for each cue, normalised per thousand words, with markers for significant events, for example, a medication change or a life stressor. This type of timeline appears in prior work on symptom tracking in notes and helps detect sudden shifts that warrant a conversation (Rumshisky et al., 2016).
3. A compact summary of co-occurrence between cues, for example, heaviness with numbness or tightness with racing thoughts. The goal is to reveal patterns across sessions that match known clinical constellations such as anxious depression or grief with anger. Co-occurrence analysis is standard in corpus linguistics and public health text mining, and it adapts well here (Berry & Kogan, 2010).

All displays must flag uncertainty. If the model is not confident that *tight* is metaphorical, the card should mark it as tentative or exclude it from counts. This keeps trust high and fosters a culture of cautious interpretation.

Emsemitext is a complement to clinical practice, not a replacement. It is most useful in four settings:

* When caseloads are large, a weekly scan for sudden increases in heaviness or tightness can prompt a check-in. Prior work shows that such text-based alerts are feasible in health systems when combined with clinician oversight (Rumshisky et al., 2016).
* Many therapies invite clients to write between sessions in order to monitor progress. Aggregating embodied cues in those entries can reveal improvement that is not yet visible in standard scales. Narrative transformation has long been used as evidence of change, and Emsemitext provides a quantitative mirror for that tradition (Angus et al., 1999).
* A card that displays a client’s own words next to a simple trend line can support collaborative planning. People often find it easier to discuss their language than a raw score, which improves engagement.
* Teams can ask population questions, such as which embodied cues are common in patients who drop out early, or how seasonal stressors shift the mix of heaviness and tightness. Population analysis has already improved service design in digital mental health, and the same approach applies here when data governance is strong (Chancellor & De Choudhury, 2020).

Inferring internal states from text is sensitive. Clinical text includes personal stories and protected health information. Responsible use requires attention to governance at each step. Data should be minimised, stored securely, and de-identified whenever possible. Access should be logged and limited to the care team. Outputs should never be used to deny services or benefits. They should be used to structure conversations and to suggest that a clinician take a closer look. These recommendations align with guidance for digital mental health analytics and clinical NLP deployments (Benton et al., 2017; Chen et al., 2022; Roberts et al., 2022).

Bias is a practical risk. If a mapping is built from one group’s language, it may not capture the metaphors used by other groups. Teams should test for differential performance across demographic slices and invite local stakeholders to review the cue list. Community involvement improves construct validity and has become standard in ethical AI projects in health (Gebru et al., 2021; Chen et al., 2022). Transparency is equally important. A short public document that describes what is collected, why it is collected, and how it is used allows clients to give informed consent. For research, institutional review and clear protocols are mandatory.

Emsemitext focuses on language. Many patients show emotion with posture, breath, or silence. Text cannot capture all of that, and it should not try. Future work will link text with other low-burden signals, for example, heart rate variability or simple ecological momentary assessments, always with explicit consent and clear opt-out options. Another direction is multilingual support. Idioms vary, and so do metaphors, but cross-linguistic work suggests that basic mappings such as weight for sadness and warmth for happiness are widespread. Careful local validation will still be needed (Kövecses, 2003; Mohammad, 2012).

Finally, models drift. Vocabulary changes. Service delivery changes. The safest way to keep Emsemitext useful is to build a habit of regular review. Set a reminder every quarter to sample a few dozen outputs and ask whether the reasoning still makes sense. If not, retrain, revise the cue list, and document the change. This is ordinary model stewardship, and it belongs in any clinical AI project (Saria et al., 2018; Roberts et al., 2022).

**7.5 Case Studies in Text-Based Emotion Analysis**

Emotion in text is rarely explicit. It surfaces in word choice, in the balance of positive and negative cues, and when people talk to each other, in how one turn responds to the last. This section examines three widely used corpora that capture these facets at different levels of complexity: short single-label posts from the Twitter Emotion dataset (Saravia et al., 2018), multi-label comments with overlapping feelings from the GoEmotions dataset (Demszky et al., 2020), and conversational utterances where emotions evolve over turns found in the MELD dataset (Poria et al., 2018). MELD is used here as a text-only stand-in for EmotionLines.

Across all three, the unit of analysis is a short span of language stored as text. What differs is how emotion is annotated and what additional structure is available. Twitter Emotion assigns exactly one label per instance drawn from a small set of basic emotions. This makes it a clean testbed for single-label classification and for inspecting which categories are easily separated by surface cues. GoEmotions allows multiple labels to co-occur on the same text (for example, admiration and joy), expanding the label space and requiring evaluation to account for overlap and class imbalance. MELD preserves the utterance-level label but adds two integers, dialog\_id and turn\_id, that place each line inside a conversation. Those fields enable a second kind of inference: not only *what emotion is expressed here?* but also, *given the current state, what tends to come next?*

The goals in this section are accordingly threefold: (1) to quantify how much signal simple lexical features carry for emotion recognition in short texts, and to make the resulting error patterns legible; (2) to handle multi-label annotation properly by treating prediction as a set decision rather than a single choice, and to use metrics that reflect both common and rare emotions; (3) to read emotion in context by estimating transition tendencies across dialogue turns and illustrating how local dynamics shape short emotional arcs.

Methods are kept deliberately comparable. A classical baseline uses TF-IDF features and a linear classifier (Logistic Regression as the mainline, with LinearSVC as a margin-based contrast). A compact transformer baseline (DistilBERT) is fine-tuned on the same splits to show what contextual representations add without changing anything else. For single-label tasks, performance is summarised with accuracy and macro-F1 and interpreted through confusion matrices and ROC analyses. For multi-label data, evaluation shifts to micro- and macro-averaged F1 and precision–recall behaviour. For conversations, a row-normalised transition matrix summarises how emotions flow from one turn to the next; a short timeline from a single dialogue makes those flows concrete.

Preprocessing is minimal and transparent. Text is used as provided, with Unicode normalisation and standard tokenisation inside the vectorizers or tokeniser. Where label vocabularies differ by name but not meaning, they are unified (e.g., MELD’s neutral is treated as no\_emotion, and joy as happiness), so comparisons remain consistent. All experiments are seeded for reproducibility and rely on the datasets library, which downloads and caches corpora locally. The following Python snippet shows only the structural idea: each dataset is mapped into a consistent dataframe with the fields needed for its task. Additionally, this code shows the exact schema used in the analysis.

# unified schema examples (illustrative)

from datasets import load\_dataset

import pandas as pd

# Twitter Emotion: single-label

tw = load\_dataset("dair-ai/emotion")

twitter\_train = pd.DataFrame({"text": tw["train"]["text"], "label": tw["train"]["label"]})

# GoEmotions: multi-label (binarized later)

go = load\_dataset("go\_emotions", "simplified") # or a prefiltered split you selected

go\_train = pd.DataFrame({"text": go["train"]["text"], "labels": go["train"]["labels"]})

# MELD (text-only): single-label + dialogue structure

meld\_urls = {

"train": "https://raw.githubusercontent.com/declare-lab/MELD/master/data/MELD/train\_sent\_emo.csv",

"dev": "https://raw.githubusercontent.com/declare-lab/MELD/master/data/MELD/dev\_sent\_emo.csv",

"test": "https://raw.githubusercontent.com/declare-lab/MELD/master/data/MELD/test\_sent\_emo.csv",

}

meld = load\_dataset("csv", data\_files=meld\_urls)

def norm\_emotion(s):

s = str(s).lower().strip()

return {"neutral":"no\_emotion","joy":"happiness"}.get(s, s)

meld\_train = pd.DataFrame({

"dialog\_id": meld["train"]["Dialogue\_ID"],

"turn\_id": meld["train"]["Utterance\_ID"],

"text": meld["train"]["Utterance"],

"label\_name":pd.Series(meld["train"]["Emotion"]).map(norm\_emotion)

})

The Twitter corpus contains short, self-contained posts labelled with a single emotion from a small vocabulary (sadness, joy, love, anger, fear, surprise). Because each tweet receives exactly one label, the task is a classic single-label classification problem; results are easy to interpret, and the errors are often intuitive.

The first thing to notice is the class balance. **Figure 3** shows a long tail: joy is the most common category, followed by sadness; surprise is rare. This matters for how we read all subsequent results. With few examples, a model will see fewer lexical patterns for surprise during training, and evaluation metrics that average over classes (e.g., macro-F1) are more informative than plain accuracy.

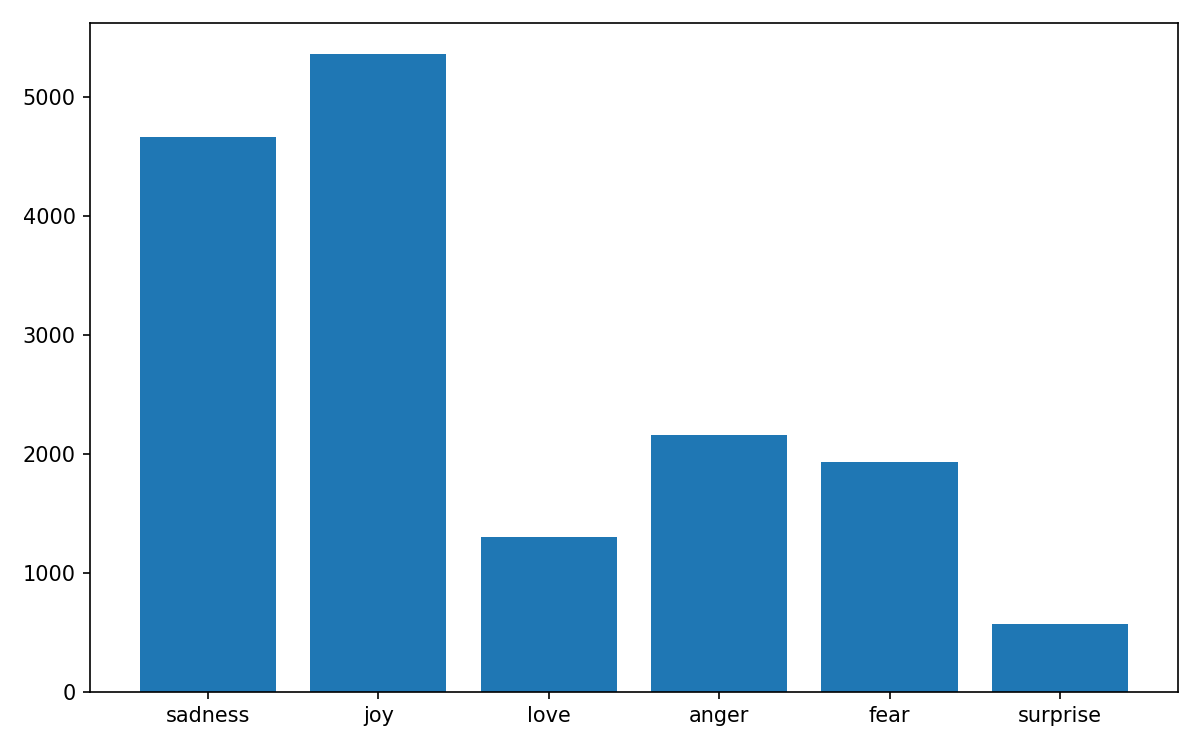
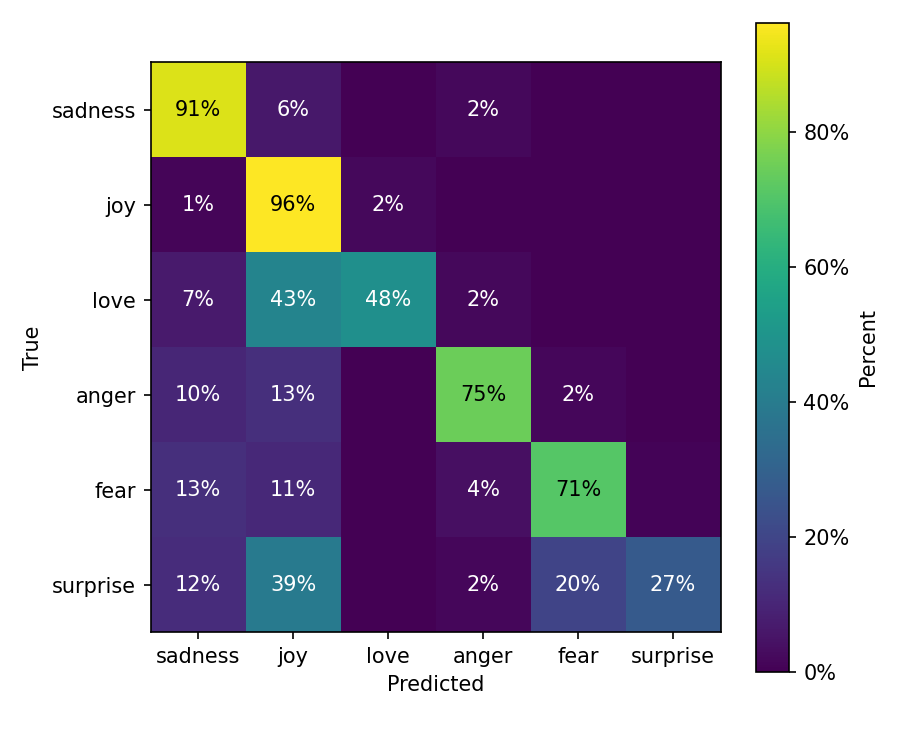


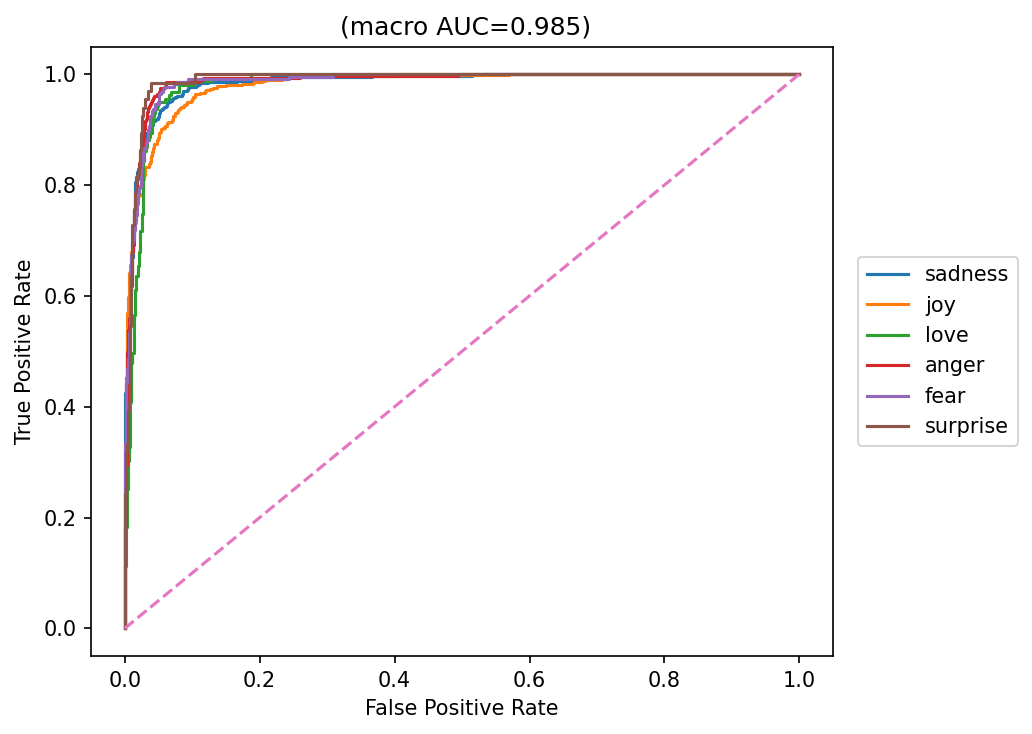
Figure 3. Twitter Emotion label distribution in the training split

Trained on bag-of-words features with TF-IDF, the model produces an almost diagonal confusion matrix (Figure 4). The rows are normalised to 100%, so each percentage shows how a true class is distributed across predictions. Joy is the easiest class, with ≈ 96% correct. Sadness is also high at ≈ 91%. Mid-tier performance appears for anger (≈ 75%) and fear (≈ 71%). Love splits across love (≈ 48%) and joy (≈ 43%), reflecting lexical overlap between affectionate and joyful wording. Surprise is the hardest label, with only ≈ 27% correct and substantial spillover into joy (≈ 39%) and fear (≈ 20%), which is consistent with its contextual and low-frequency nature.



**Figure 4.** Row-normalised confusion matrix for a TF-IDF + Logistic Regression classifier for Twitter Emotion

Threshold-free separability tells the same story. The ROC overlay in **Figure 5** plots each emotion against the rest; curves crowd the top-left, and the macro-AUC is ≈ 0.985. High AUC means that, for most classes, a simple linear model can rank true positives above negatives with few trade-offs. In practice, this confirms that much of Twitter’s emotional signal lives in easily learned lexical patterns (hashtags, intensifiers, sentiment words) rather than long-range context.



**Figure 5.** Twitter Emotion ROC curves for a TF-IDF plus Logistic Regression baseline

A lightweight transformer fine-tune provides an instructive counterpoint. **Figure 6** shows the row-normalised confusion matrix after a short DistilBERT training run on the same split. The model produces a confusion matrix that is almost entirely diagonal. This means that each row answers the question: *Given all actual instances of this emotion, how did the model classify them?* The diagonal cells for sadness (≈ 96%), joy (≈ 95%), and anger (≈ 94%) are very high. This indicates that these categories are effectively captured by surface cues such as keywords, intensifiers, and sentiment-bearing expressions. Fear is slightly lower (≈ 90%), and two systematic confusions appear: love is sometimes predicted as joy (≈ 14%), and surprise splits between fear (≈ 21%) and joy (≈ 12%), with only ≈ 62% correct. Both patterns make linguistic sense. Love and joy share positive valence and overlapping vocabularies; short tweets often lack enough context to separate *I love this* from *this makes me happy*. Surprise frequently arrives via punctuation and interjections that can also accompany alarm, so fear and surprise are not cleanly separable in short text.

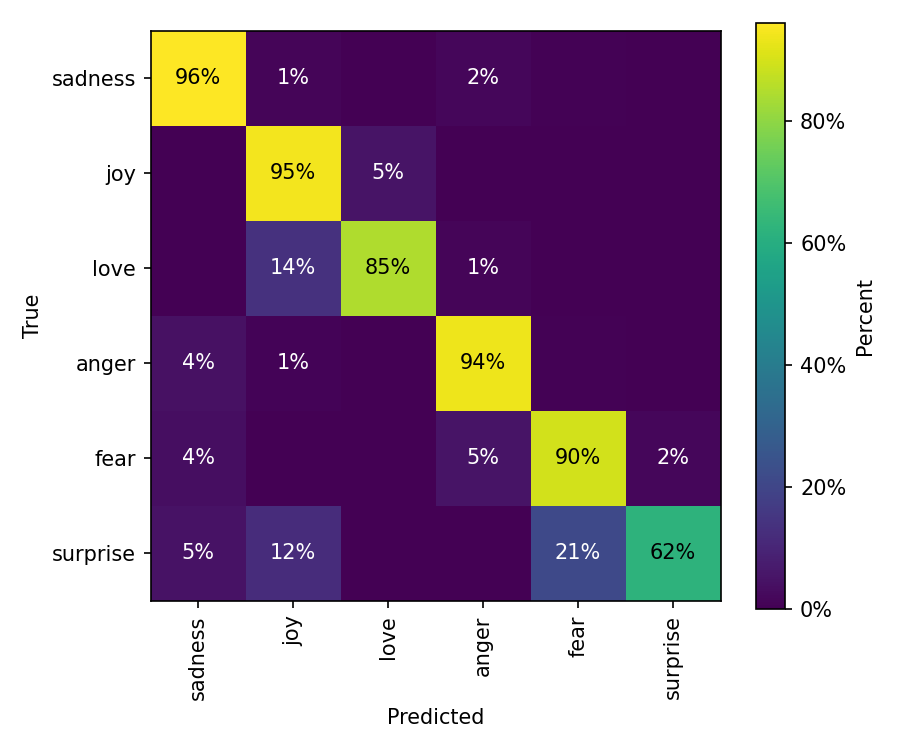


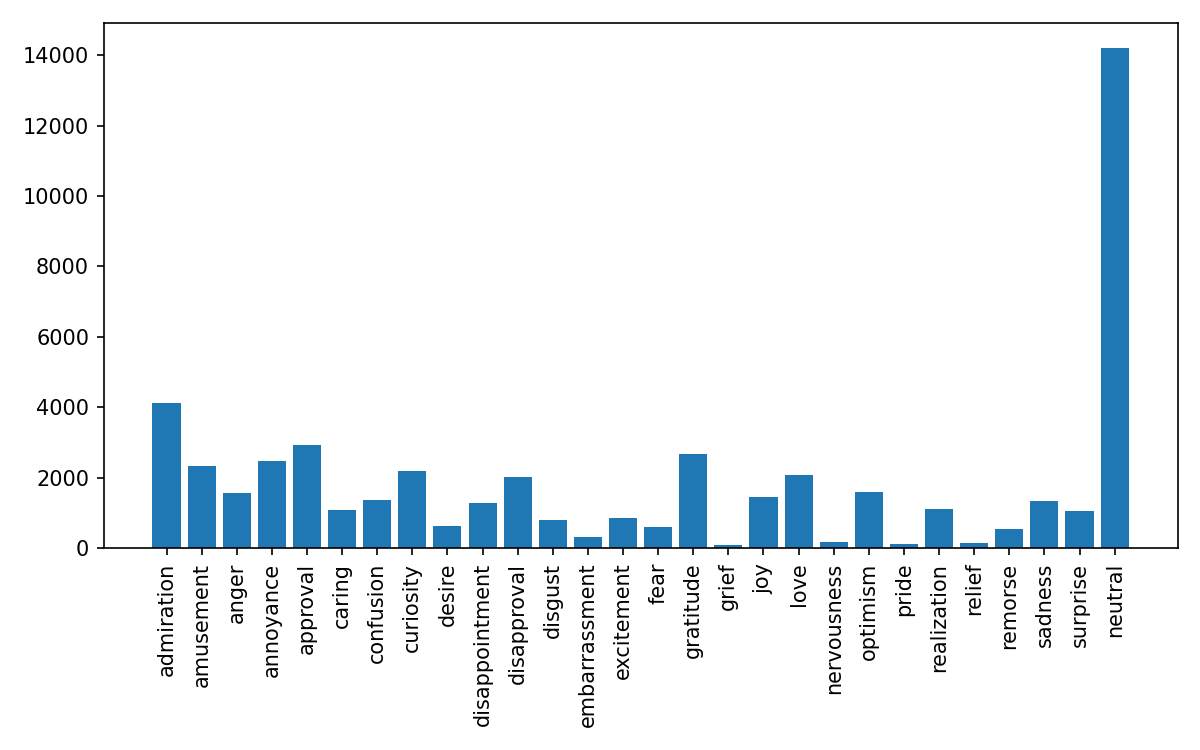
Figure 6. Twitter Emotion Transformer confusion matrix row normalised

TF-IDF is a strong baseline, but in our runs, a small fine-tuned transformer surpasses it (compare Figure 4 and Figure 6). Accuracy along the diagonal rises for anger (≈ 75% → 94%), fear (≈ 71% → 90%), love (≈ 48% → 85%), and sadness (≈ 91% → 96%), while joy stays high (≈ 96%). Surprise remains the hardest class because it is rare (see Figure 3), yet performance more than doubles (≈ 27% → 62%). Residual confusion still flows from *surprise* into joy (~12%) and fear (~21%), reflecting lexical ambiguity and class imbalance rather than a lack of contextual capacity. The transformer’s contextual representations disambiguate overlaps such as *love* vs. *joy* and pick up cues that TF-IDF misses, while imbalance keeps *surprise* fragile; class-weighted training or rebalanced sampling would likely trim the remaining spillover.

Taken together, these four figures sketch a coherent picture of emotion in short social posts. Imbalance shapes the problem; lexical features already separate most classes extremely well; love versus joy and surprise versus fear mark the natural fault lines; and transformer gains are not automatic at a small scale. This establishes a clear baseline before we turn to settings where the task itself is harder. These include cases with multiple overlapping labels and the presence of conversational context.

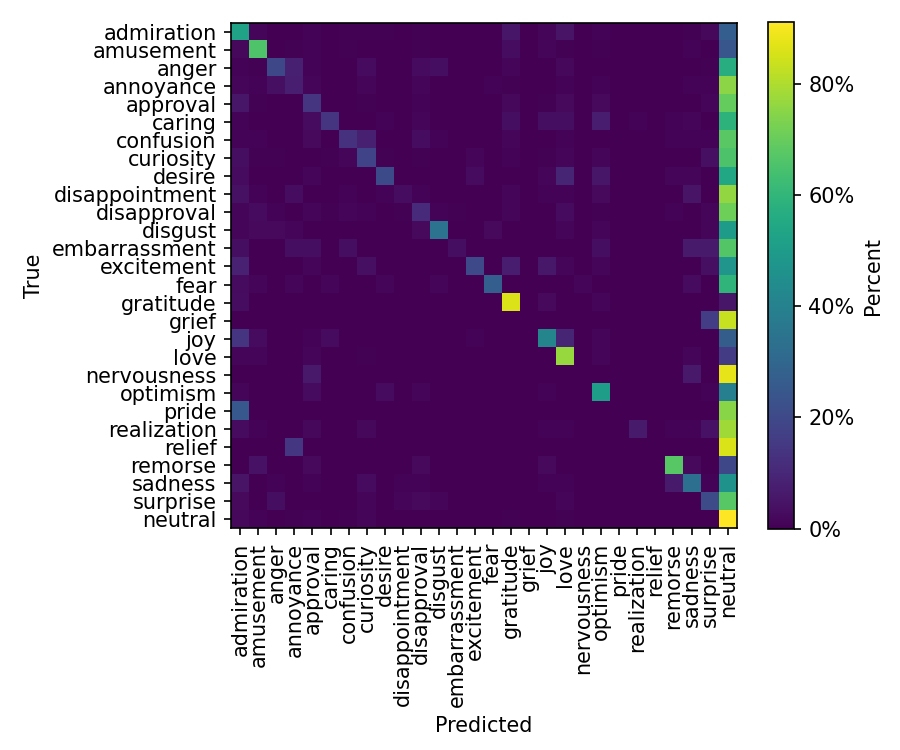
GoEmotions widens the lens from a single, mutually exclusive label to **overlapping emotions**. A single comment can be tagged with several feelings (e.g., admiration and joy), plus a frequent neutral tag. This small change in annotation reshapes both modelling and evaluation: instead of choosing one class, a model must **decide a set** of labels; instead of accuracy, we read **precision–recall behaviour** and class-balanced F-scores.

The first constraint is the label landscape itself. **Figure 7** illustrates the uneven distribution of examples across the 28 categories. Neutral dominates the corpus. Positive social signals such as admiration, approval, and gratitude are common. Many others, such as pride, nervousness, and embarrassment, are tiny. This long tail creates two pressures: the model learns very sharp detectors for frequent categories, and it is tempted to default to neutral when evidence is thin.



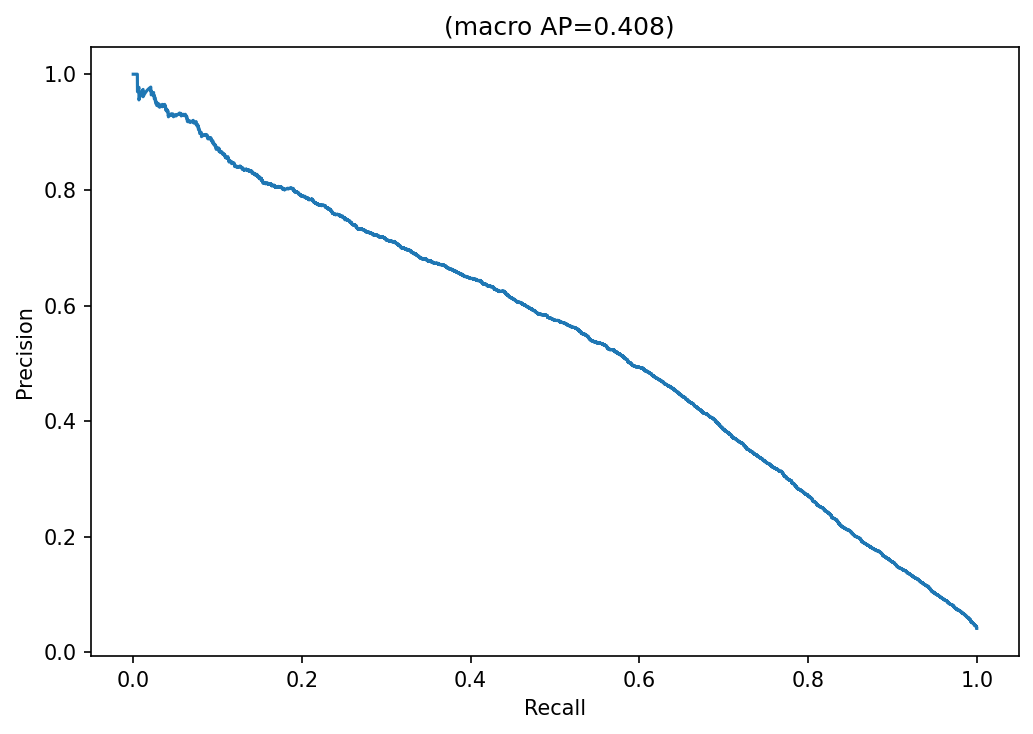
**Figure 7.** GoEmotions label frequency across the corpus

Because a comment can carry multiple labels, a standard confusion matrix is not entirely suitable. We include one nonetheless as a **top-1 visualisation**, selecting each example’s most confident predicted label to produce **Figure 8**. Interpret it as a qualitative map rather than a strict scorecard. The bright diagonal for neutral reflects its scale and lexical distinctiveness; gratitude and optimism also display relatively clear diagonals, likely influenced by formulaic phrases (*thank you*, *I hope…*). Off-diagonal smears indicate natural ambiguities: admiration merges with approval and joy; annoyance with anger and disapproval; sadness with disappointment, grief, and remorse. Rare, psychologically specific categories (embarrassment, nervousness, pride) have minimal signal and spread broadly into neighbours or revert to neutral. The inventory is comprehensive, but much of it exists within a low-data regime where surface cues are subtle and context is important.



**Figure 8.** GoEmotions confusion matrix after projecting the multi-label task to a single top-1 label and row normalizing

For evaluation that respects multi-label reality, we collapse all one-vs-rest decisions into a **micro-averaged precision–recall curve**. **Figure 9** climbs near the top-left at small recall (the model’s highest-confidence detections are usually correct) and then falls steadily as recall increases, reflecting the cost of retrieving rarer, harder labels. The title reports a **macro average precision ≈ 0.408**, which treats every emotion equally. This reveals how the long tail lowers the overall average, exactly the behaviour observed in the top-1 heatmap.



**Figure 9.** GoEmotions precision–recall curve aggregated across labels for a TF-IDF plus one-vs-rest Logistic Regression model

A minimal schematic of what we compute looks like this (the point is the evaluation, not the particular model):

# Y\_test: (n\_samples, n\_labels) binary matrix

# decision\_function -> scores; sigmoid -> probabilities per label

from scipy.special import expit

Y\_score = expit(ovr\_logreg.decision\_function(X\_test))

# micro-PR for the multi-label task

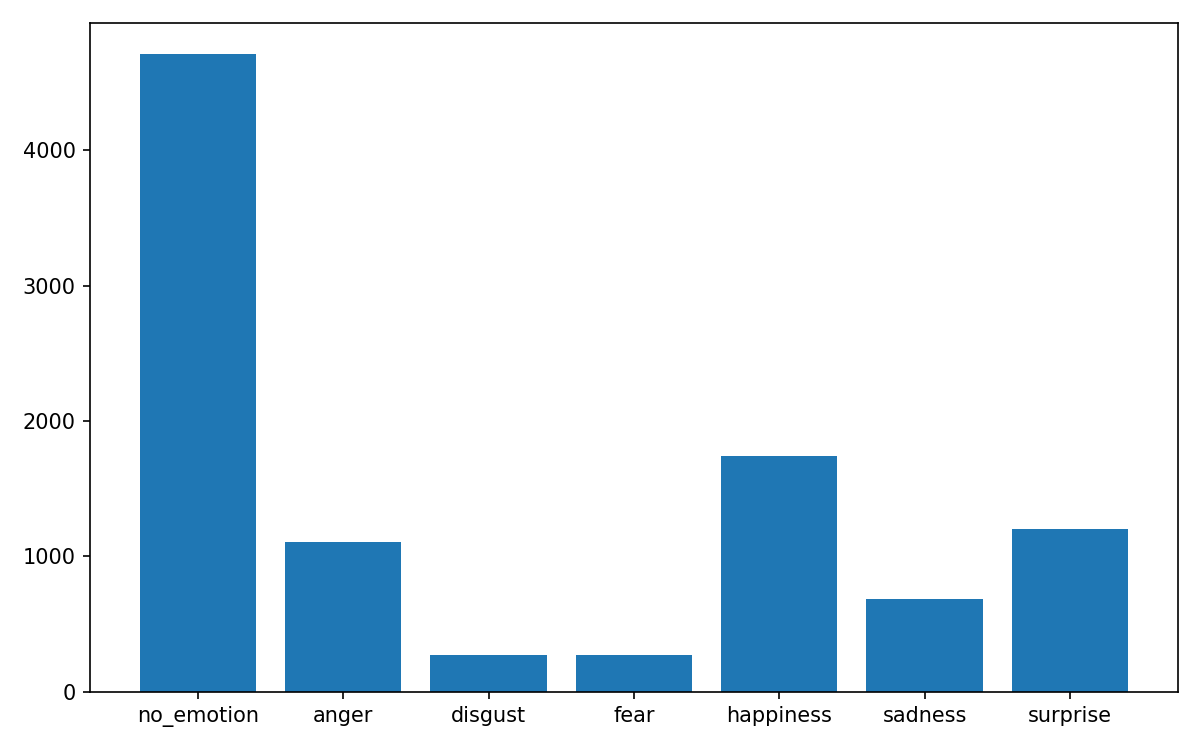
from sklearn.metrics import precision\_recall\_curve, average\_precision\_score

p, r, \_ = precision\_recall\_curve(Y\_test.ravel(), Y\_score.ravel())

macro\_ap = average\_precision\_score(Y\_test, Y\_score, average="macro")

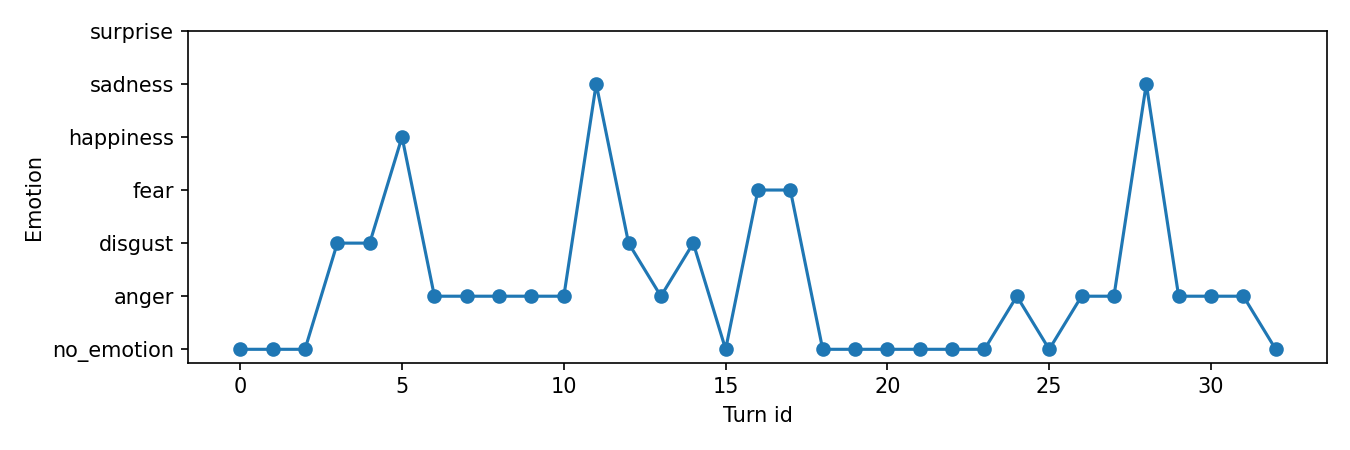
Interpreting the three figures together gives a clear picture of the dataset rather than just the model. The frequency profile (**Figure 7**) explains why neutral and a handful of social-approval emotions are easy and why specificity is hard. The top-1 confusion **(Figure 8)** lets us see which categories blur together when we force a single choice. The micro-PR curve **(Figure 9)** then quantifies the set-prediction task itself: strong precision at low recall, steady degradation as we chase the tail. Improvements, therefore, hinge on better **thresholding per label** (not a single 0.5 cut), **reweighting or focal losses** to lift rare classes, and models that can borrow context beyond the sentence when the wording alone is too polite, too brief, or too subtle to carry the feeling.

MELD adds structure that the previous corpora lack: each utterance belongs to a dialogue (dialog\_id) and has an order (turn\_id). The label inventory is compact, no\_emotion, anger, disgust, fear, happiness, sadness, surprise, but the distribution is not. **Figure 10** shows a heavy skew toward no\_emotion, with happiness the next most frequent and the rest far smaller. This imbalance sets the baseline expectation: the model will see ample neutral language and relatively little evidence for the rarer negative states.



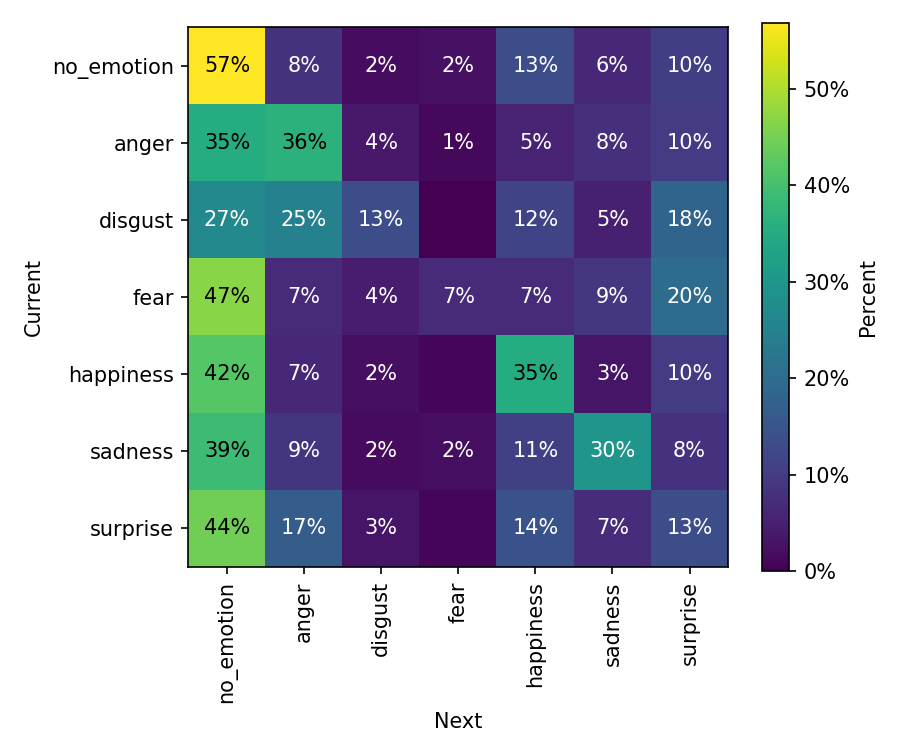
**Figure 10.** MELD label distribution in the training split

Two plots make the conversational setting tangible. The short trajectory in **Figure 11** traces one dialogue turn by turn: long neutral stretches punctuated by brief spikes to happiness or surprise.



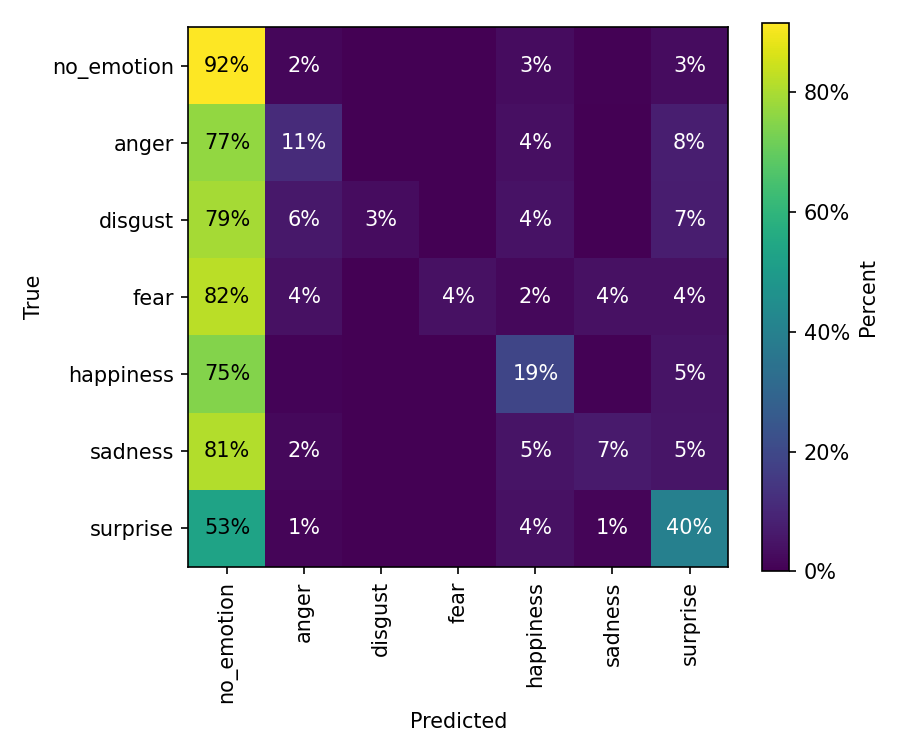
**Figure 11.** Example dialogue timeline in MELD

Aggregating across all conversations, the **row-normalised transition matrix** in **Figure 12** summarises *what tends to come next*. Rows are current emotion, and columns are next emotion with percentages per row. From no\_emotion, the most likely next step is to remain neutral (≈ 57%), with non-trivial flows into happiness (~13%) and surprise (~10%). Negative emotions show persistence but also leakage toward neutrality; for example, sadness → sadness is about 30%, but sadness → no\_emotion is even more common (~39%). These flows are exactly what a turn-by-turn system must model if it is to use context rather than treat utterances in isolation.



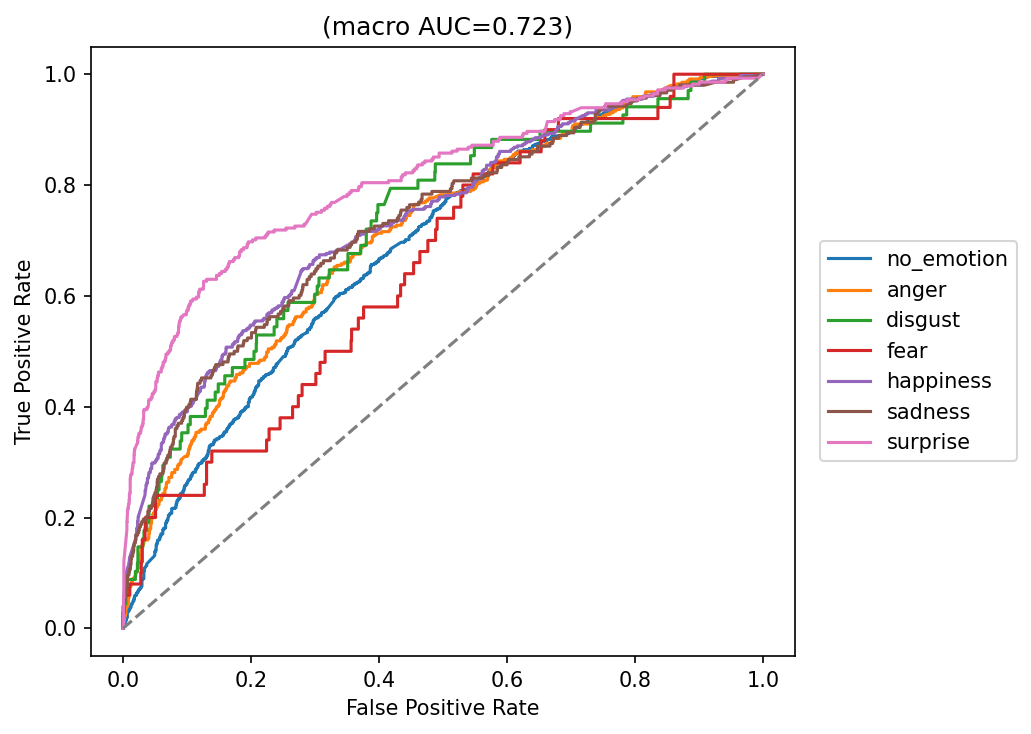
**Figure 12.** Dialogue emotion transition matrix in MELD

On single-utterance classification, the TF-IDF + Logistic Regression baseline (Figure 13) yields overall accuracy around 0.54 and macro-F1 around 0.26. Class recall is highly uneven. Performance is highly skewed toward no\_emotion (92% recall) while most other classes are frequently mapped to neutral (anger 77%, disgust 79%, fear 82%, happiness 75%, sadness 81%, surprise 53% to neutral), with true class recalls on the diagonal of 11%, 3%, 4%, 19%, 7%, and 40% respectively. This is consistent with short, context-poor turns and strong label imbalance.



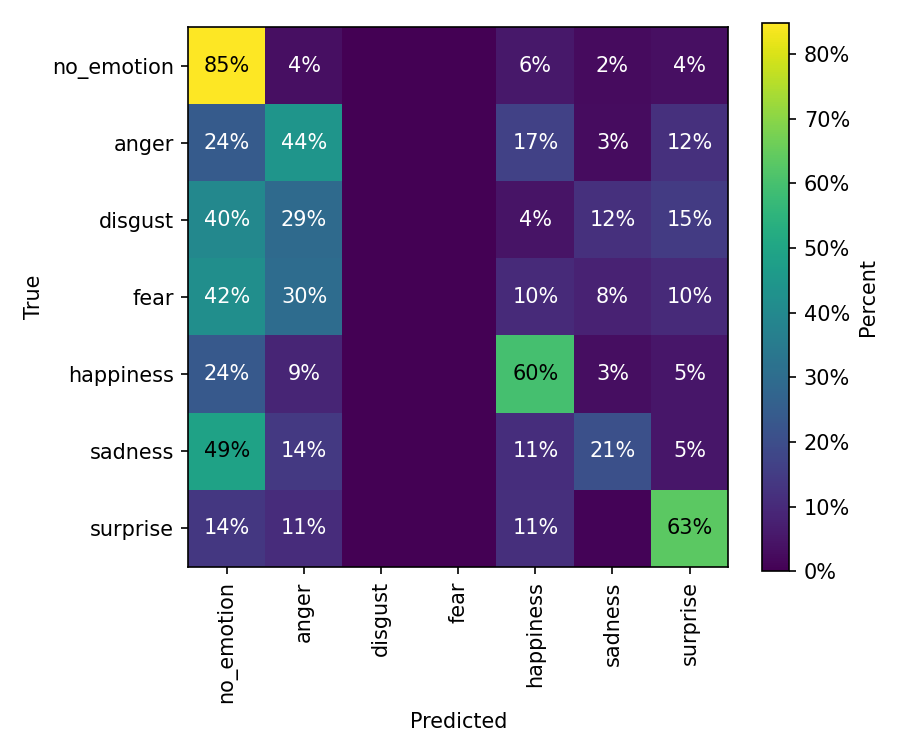
**Figure 13.** Confusion matrix for the TF-IDF Logistic Regression baseline on MELD

Threshold-free separability is modest. The ROC overlay in **Figure 14** clusters well below the perfect top-left corner; the macro-AUC of **≈ 0.723** is respectable for broad categories but far from the near-separation seen on Twitter. In other words, lexical cues alone only partly separate emotions once we move into dialogue.



**Figure 14.** ROC curves by class for MELD with a Logistic Regression baseline

A small transformer (DistilBERT) trained on the same split changed coverage (compare Figure 13 and Figure 15) for most classes but leaves clear asymmetries. Neutral drops to ~85% compared with the TF-IDF baseline, while happiness rises to ~60% and surprise improves to ~63%. Anger sits near ~44%, fear remains low at ~10%, and disgust is rarely recovered (~3–4%), with many instances drifting into *no\_emotion* or neighbouring negatives. The net effect is a sharper diagonal for happiness and surprise, little relief for the rarer negatives, and a modest reduction for neutral. This is evidence that context helps some positive states, but the class imbalance and short utterances still drive confusion.

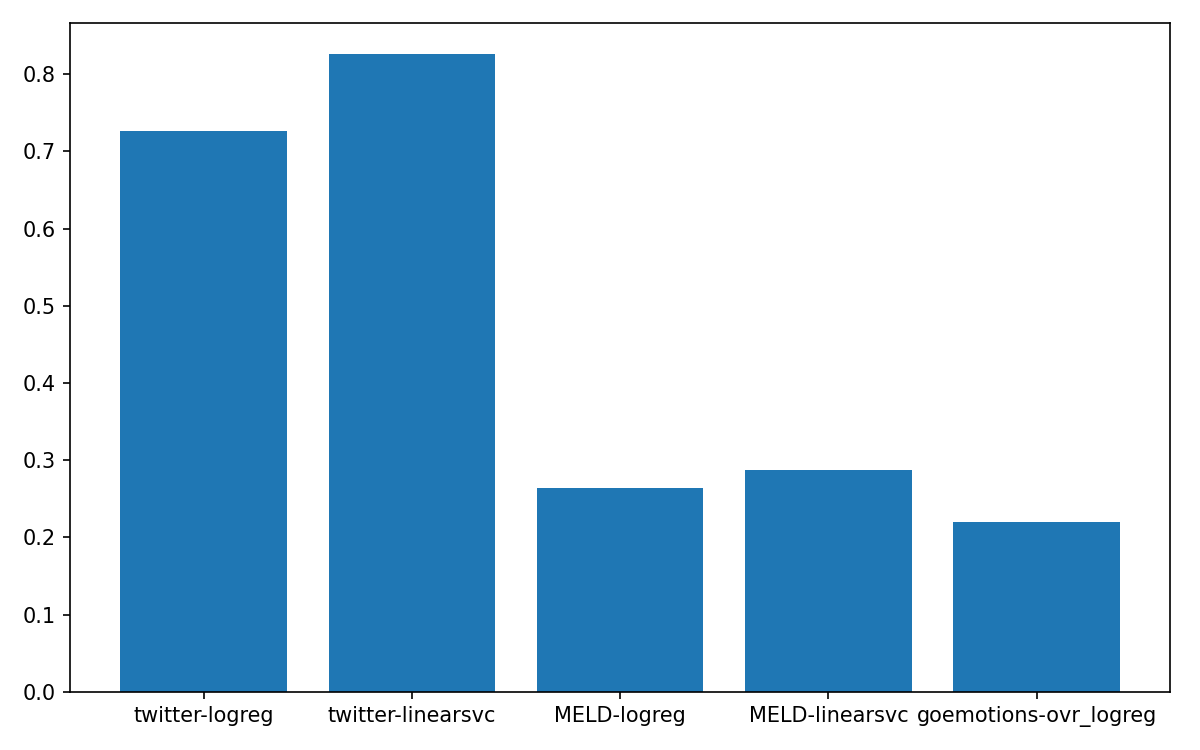


**Figure 15.** Confusion matrix for a small DistilBERT fine-tune on MELD

Taken together, the six figures outline the challenge that conversation brings. Label skew and brevity bias models toward neutrality; negative states require more context to distinguish; and even when a model scores well overall, macro-averaged metrics and the transition view reveal where it struggles. The practical implication is clear: dialogue-aware systems need both better representations and sequence-level objectives that reward getting the flow right, not just the next label in isolation.

The three case studies trace a gradient in difficulty. On short, single-label tweets, the signal is overwhelmingly lexical and simple linear models soak it up. Macro-F1 is high for Twitter (Figure 16 shows ≈ 0.73 for Logistic Regression and ≈ 0.83 for LinearSVC), and the ROC overlay is close to the top-left corner (Figure 5). The row-normalised confusion (Figure 4) is nearly diagonal, with two recurrent fault lines: love frequently bleeds into joy, and surprise often drifts toward fear and joy. A small transformer fine-tune improves rather than hurts in this regime, sharpening most diagonals and boosting the hardest class: surprise rises to ≈ 62% recall while fear reaches ≈ 90% (Figure 6).

GoEmotions changes the evaluation lens rather than the domain. Many comments carry several emotions, so any single-label summary understates the difficulty. The frequency plot in Figure 7 shows why this matters. Neutral dominates, and several positive labels, such as admiration and approval, are common, while many others live in the long tail. The top-1 confusion in Figure 8 is therefore a proxy, not a verdict. It still helps us see where boundaries blur. Admiration often appears near approval, annoyance often sits near anger, and several rare categories scatter into neutral when the signal is faint. For an overall picture, the precision–recall curve in Figure 9 is the most honest summary. The curve is macro-averaged with an area of about 0.408, which means high precision at very small recall, followed by a smooth decline as the model tries to retrieve low-frequency labels. This behaviour is exactly what the long-tail distribution in Figure 7 would suggest. The macro-F1 bar in Figure 16, around 0.22 for our baseline, is low by design. Macro-averaging gives every emotion the same weight as neutral and gratitude, so it exposes how imbalance, label overlap, and subtle cues govern this dataset.



**Figure 16.** Macro-F1 summary across datasets and models

MELD pushes the problem into context. Utterances are short, the label distribution is skewed toward no\_emotion (Figure 10), and identical words can map to different states depending on the previous turn. Macro-F1 falls accordingly (≈ 0.26–0.29 in Figure 16), and threshold-free separability settles near a macro-AUC of ≈ 0.723 (Figure 14). The transition matrix (Figure 12) shows why: conversations often remain in, or revert to, neutral; many negative states flow into no\_emotion on the next turn. Both the classical baseline (Figure 13) and the small transformer (Figure 15) reflect that inertia. No\_emotion drops from ≈ 92% on the diagonal in the TF-IDF baseline (Figure 13) to ≈ % with the transformer (Figure 15). Happiness, surprise and anger improved its performance. The single-dialogue timeline (Figure 11) is a micro-view of the same story, with long, calm stretches punctuated by short spikes.

A few practical implications follow naturally: (1) **match your evaluation to the task** and remember that confusion matrices and ROC overlays tell most of the truth for single-label data while micro/macro-PR and per-label thresholds matter in multi-label settings; (2) **lexical baselines remain strong** for short posts and transformers help most when you either scale training or give them context they can use; (3) class imbalance is not just a minor issue. It directly influences what the model learns and what readers may misunderstand if only accuracy is reported; (4) in dialogue, **sequence matters**: the transition view is not cosmetic; it is the target for systems that need to understand how emotions move across turns.

This chapter’s figures give a compact map of those realities. Twitter shows how far simple features go. GoEmotions shows why evaluation must change when feelings overlap. MELD shows why context and dynamics belong in the objective. The next steps can incorporate whether intensity-aware corpora, such as DENS, or mental-health contexts like Vent. This would make those themes sharper rather than overturn them: better thresholds for the tail, better calibration, and models that reward getting the flow of emotion right, not just the next label in isolation.

**References**

Alm, C. O. (2012). The role of affect in the computational modeling of natural language. *Language and Linguistics Compass*, *6*(7), 416-430.

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*.

Angus, L., Levitt, H., & Hardtke, K. (1999). The narrative processes coding system: Research applications and implications for psychotherapy practice. *Journal of clinical psychology*, *55*(10), 1255-1270.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*(1), 617-645.

Benton, A., Coppersmith, G., & Dredze, M. (2017, April). Ethical research protocols for social media health research. In *Proceedings of the first ACL workshop on ethics in natural language processing* (pp. 94-102).

Berry, M. W., & Kogan, J. (Eds.). (2010). *Text mining: applications and theory*. John Wiley & Sons.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Blodgett, S. L., Barocas, S., Daumé Iii, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings* (Vol. 30, No. 1, pp. 25-36). Technical report C-1, the center for research in psychophysiology, University of Florida.

Cambria, E., & Hussain, A. (2015). Sentic computing. *Cognitive Computation*, *7*(2), 183-185.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, *3*(1), 43.

Chen, H., Gomez, C., Huang, C. M., & Unberath, M. (2022). Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. *NPJ digital medicine*, *5*(1), 156.

Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological science*, *15*(10), 687-693.

Coppersmith, G., Dredze, M., & Harman, C. (2014, June). Quantifying mental health signals in Twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality* (pp. 51-60).

D'mello, S. K., & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM computing surveys (CSUR)*, *47*(3), 1-36.

De Choudhury, M., Counts, S., & Horvitz, E. (2013b, April). Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 3267-3276).

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013a). Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media* (Vol. 7, No. 1, pp. 128-137).

Demner-Fushman, D., Chapman, W. W., & McDonald, C. J. (2009). What can natural language processing do for clinical decision support?. *Journal of biomedical informatics*, *42*(5), 760-772.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).

Dimitrov, D., Alam, F., Hasanain, M., Hasnat, A., Silvestri, F., Nakov, P., & Da San Martino, G. (2024, June). Semeval-2024 task 4: Multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)* (pp. 2009-2026).

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, *6*(3-4), 169-200.

Farahani, H., Watson, P., Bezan, T., Kovač, N., Winter, L. C., Blagojević, M., ... & Jiménez, P. (2024). PsAIchology: An Intelligent Direction in Psychological Sciences. In *10th International Scientific Conference Technics, Informatics and Education-TIE 2024*. Faculty of Technical Sciences Čačak, University of Kragujevac.

Farahani, H., Watson, P., Kovač, N., Sheykhangafshe, F. B., Azadfallah, P., Allahyari, A., ... & Chesli, R. R. (2025). Unraveling the Complexity of Love Addiction Using Machine Learning Algorithms: The Influence of Positive and Negative Affect, Interpersonal Needs, and Self-Hate. In *International Handbook of Love: Transcultural and Transdisciplinary Perspectives* (pp. 1-22). Cham: Springer Nature Switzerland.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86-92.

Gibbs Jr, R. W. (2005). *Embodiment and cognitive science*. Cambridge University Press.

Giraud, M., Marelli, M., & Nava, E. (2023). Embodied language of emotions: Predicting human intuitions with linguistic distributions in blind and sighted individuals. *Heliyon*, *9*(7).

Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, *18*, 43-49.

Hovy, D., & Spruit, S. L. (2016). The social impact of natural language processing. In *The 54th Annual Meeting of the Association for Computational Linguistics Proceedings of the Conference, Vol. 2 (Short Papers)*. Association for Computational Linguistics.

Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media* (Vol. 8, No. 1, pp. 216-225).

Ingram, R. E., Siegle, G. J., & Steidtmann, D. (2009). Methodological issues in the study of depression. *Handbook of depression*, *2*, 69-92.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature machine intelligence*, *1*(9), 389-399.

Jurafsky, D., & Martin, J. H. (2025, August 24). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, with language models (3rd ed.) [Online manuscript].

Kovač, N., Ratković, K., Farahani, H., & Watson, P. (2024). A practical applications guide to machine learning regression models in psychology with Python. *Methods in Psychology*, *11*, 100156.

Kovač, N., Ratković, K., Farahani, H., & Watson, P. (2025b). Machine learning regression models for internal shame. Acta Psychologica, 260, 105721.

Kovač, N., Ratković, K., Watson, P., Farahani, H., & Bagheri Sheykhangafshe, F. (2025). Machine learning classification models for predicting chronic pain. *Current Psychology*, 1-14.

Kövecses, Z. (2003). *Metaphor and emotion: Language, culture, and body in human feeling*. Cambridge University Press.

Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*. University of Chicago press.

Larradet, F., Niewiadomski, R., Barresi, G., Caldwell, D. G., & Mattos, L. S. (2020). Toward emotion recognition from physiological signals in the wild: approaching the methodological issues in real-life data collection. *Frontiers in psychology*, *11*, 1111.

Liu, B. (2022). *Sentiment analysis and opinion mining*. Springer Nature.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mohammad, S. (2012). # Emotional tweets. In *\* SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)* (pp. 246-255).

Mohammad, S. M. (2021). Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text. In *Emotion measurement* (pp. 323-379). Woodhead Publishing.

Mohammad, S. M., & Bravo-Marquez, F. (2017). WASSA-2017 shared task on emotion intensity. *arXiv preprint arXiv:1708.03700*.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, *29*(3), 436-465.

Neumann, M., King, D., Beltagy, I., & Ammar, W. (2019). ScispaCy: fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*.

Niedenthal, P. M. (2007). Embodying emotion. *science*, *316*(5827), 1002-1005.

Niedenthal, P. M., & Maringer, M. (2009). Embodied emotion considered. *Emotion Review*, *1*(2), 122-128.

Niedenthal, P. M., Barsalou, L. W., Winkielman, P., Krauth-Gruber, S., & Ric, F. (2005). Embodiment in attitudes, social perception, and emotion. *Personality and social psychology review*, *9*(3), 184-211.

Park, C. Y., Cha, N., Kang, S., Kim, A., Khandoker, A. H., Hadjileontiadis, L., ... & Lee, U. (2020). K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data*, *7*(1), 293.

Pennebaker, J. W. (1997). Writing about emotional experiences as a therapeutic process. *Psychological science*, *8*(3), 162-166.

Pennebaker, J. W., & Chung, C. K. (2011). Expressive writing: Connections to physical and mental health. In H. S. Friedman (Ed.), *The Oxford handbook of health psychology* (pp. 417–437). Oxford University Press.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, *71*(2001), 2001.

Plutchik, R. (1980a). A general psychoevolutionary theory of emotion. In *Theories of emotion* (pp. 3-33). Academic press.

Plutchik, R. (1980b). Emotion: Theory, research, and experience. vol. 1: Theories of emotion.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2018). Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Ricardi, F. (2025). From symptom to symbol. Language in psychotherapy. *Medical Research Archives*, *13*(2).

Roberts, K., Demner-Fushman, D., Voorhees, E. M., Bedrick, S., & Hersh, W. R. (2022). Overview of the TREC 2022 Clinical Trials Track. In *TREC*.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2021). A primer in BERTology: What we know about how BERT works. *Transactions of the association for computational linguistics*, *8*, 842-866.

Rude, S., Gortner, E. M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, *18*(8), 1121-1133.

Rumshisky, A., Ghassemi, M., Naumann, T., Szolovits, P., Castro, V. M., McCoy, T. H., & Perlis, R. H. (2016). Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Translational psychiatry*, *6*(10), e921-e921.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Saravia, E., Liu, H. C. T., Huang, Y. H., Wu, J., & Chen, Y. S. (2018). CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 3687-3697).

Saria, S., Butte, A., & Sheikh, A. (2018). Better medicine through machine learning: what’s real, and what’s artificial?. *PLoS medicine*, *15*(12), e1002721.

Saxbe, D. E., Yang, X. F., Borofsky, L. A., & Immordino-Yang, M. H. (2013). The embodiment of emotion: language use during the feeling of social emotions predicts cortical somatosensory activity. *Social cognitive and affective neuroscience*, *8*(7), 806-812.

Semeraro, A., Vilella, S., & Ruffo, G. (2021). PyPlutchik: Visualising and comparing emotion-annotated corpora. *Plos one*, *16*(9), e0256503.

Shutova, E. (2010, June). Automatic metaphor interpretation as a paraphrasing task. In *Human language technologies: the 2010 annual conference of the North American chapter of the association for computational linguistics* (pp. 1029-1037).

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, *29*(1), 24-54.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38-45).