How Are Emotions Expressed in Literary Fiction, and Can Language Models Detect Them?

Anonymous ACL submission

Abstract

This paper investigates how emotions are expressed in 19th-century Danish and Norwegian literature and whether contemporary language models can detect them. We introduce a linguistically and culturally grounded annotation scheme distinguishing conceptual, expressive, and non-verbal emotion expressions. Applying this scheme, we construct a multi-label dataset of sentences from the MeMo corpus, annotated with nine emotion categories based on Plutchik's theory. We evaluate seven Danish and Norwegian pre-trained language models on this task and propose a novel Plutchik-aware Soft-F1 metric that accounts for affective proximity between emotion categories. Our results show that while models like DFM-Large achieve strong performance on standard metrics, they still struggle with overlapping and subtle emotional expressions common in literary texts. The study highlights the challenges of operationalizing emotion theories in NLP and the importance of interdisciplinary approaches to modeling affect in historical and narrative domains.

1 Introduction

Since the millennium change, the humanities have experienced what has been widely referred to as *the affective turn*—an intensified focus on emotions as central to human experience, cultural practices, and historical development (Clough and Halley, 2007). Parallel to this, emotion analysis (EA) has become an increasingly prominent area of research in natural language processing (NLP) over the past decade (Plaza-del Arco et al., 2024a). Despite these converging interests, there remains a lack of robust theoretical and methodological frameworks that explicitly and effectively combine psychological, linguistic, literary, and cultural theory with computational approaches.

This paper addresses this interdisciplinary gap by developing a theoretical framework for understanding how emotions are expressed in literary texts and examining whether contemporary language models can detect them. We present three main contributions: (1) We introduce an annotated dataset of 19th-century Danish and Norwegian literary texts labeled with emotion categories informed by psychological, linguistic, literary and cultural theory; (2) We evaluate the performance and generalization capabilities of large language models for classifying emotions in long and complex texts; (3) We propose a theoretical framework that brings together insights from psychology, linguistics, literary theory, and NLP to support future research on the cultural and historical formations of emotions.

2 Related work

Emotion analysis in NLP. Emotion analysis is a rapidly growing field within NLP. It has been applied to various types of textual material, including social media data (Mohammad and Kiritchenko, 2015), news articles (Staiano and Guerini, 2014), customer reviews (De Geyndt et al., 2022), transcribed conversations (Creanga and Dinu, 2024), literary fiction (Kim and Klinger, 2018), and poetry (Haider et al., 2020; Konle et al., 2024).

Despite its well-established presence in the field, emotion analysis remains a complex task with significant challenges (Plaza-del Arco et al., 2024a). In particular, two major challenges persist. First, the theoretical conceptualization of adapting psychological theories to examine linguistic and cultural data. Second, the operationalization of this theorization into methods suitable for computational analysis. This paper seeks to address these challenges by proposing a robust theoretical and methodological framework for emotion analysis that can be applied to cultural analysis.

Computational analysis of Scandinavian literary texts. Recent work has applied NLP meth-

ods to 19th-century Scandinavian literature. These studies have produced pre-trained language models for historical texts (Al-Laith et al., 2024a), developed sentiment classification approaches tailored to literary corpora (Allaith et al., 2023), and investigated euphemism detection (Al-Laith et al., 2025a) and gendered affect in female-authored fiction (Degn et al., 2025). Additional work includes the analysis of sonic representations (Al-Laith et al., 2024b) and annotated direct speech (Al-Laith et al., 2025b). Our contribution builds on this foundation by introducing a multi-label dataset and annotation scheme for emotion categories, and proposing a Plutchik-aware Soft-F1 evaluation metric that accounts for semantic similarity between emotion labels in literary classification tasks.

3 Theory of Emotions

We are modeling basic emotions, which have been characterized as a) discrete, b) detectable across cultures, and c) intertwined in different and complex forms depending on their cultural mediation (Donovan et al., 2025). Specifically, we rely on Plutchik's theory of emotions, which proposes eight primary emotions serving as the foundation of all human emotional experiences (Plutchik, 2001): Joy, Trust, Fear, Surprise, Sadness, Disgust, Anger, and Anticipation. Plutchik understands these emotions as evolutionary adaptations that aid human survival. For instance, love and emotional attachment advance pair bonding, reproduction, and parental investment (Plutchik, 2001). This theory of primary emotions is visualized as a cone – either intact or unfolded – called a "wheel of emotions". See Figure 1.

The circle displays the degrees of similarity between the emotions, e.g. disgust is most similar to the two neighboring emotions on the wheel, anger and sadness, and most contrary to the opposing emotion on the wheel, trust. The cone's vertical dimensions represent three levels of intensity, e.g. anger is the middle intensity with annoyance being the less intense experience, and rage being the more intense experience. In the unfolded model the emotions in the blank spaces between the primary emotions are the primary dyadic emotions, which are mixtures of two of the primary emotions, e.g. contempt is a mixture of anger and disgust and love is a mixture of joy and trust.

Plutchik's model shares similarities with other psychological models on basic emotions (e.g.

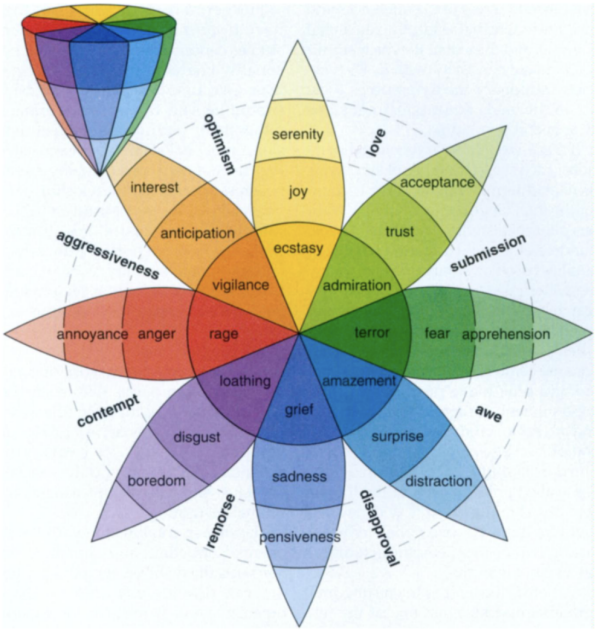

Figure 1: Plutchik's Wheel of Emotions. Source: (Plutchik, 2001).

Harmon-Jones et al. (2016)) has been widely used in both psychology and NLP, and most emotion analysis work is built upon it or similar categorical models of emotions (Plaza-del Arco et al., 2024a). However, applying the model to annotate and computationally analyze text presents important challenges that need consideration.

4 Annotating Emotions in Text

When working with emotion analysis of literary texts it is important to consider the transferability of psychological theory—based on emotions as they are expressed in human interaction—to the classification of texts, where emotions are conveyed through linguistic, aesthetic, and cultural compositions. A premise of emotion analysis is thus that emotions can become intersubjectively perceivable through many different mediums. Cultural theorists, historians of emotions, and literary scholars have done extensive work on describing how emotions interrelate with and can be understood as cultural practices (Cvetkovich, 2003; Ngai, 2005; Scheer, 2012; Ahmed, 2014; Greiner, 2014).

Building on this body of work, we argue that literary texts can be understood as sites of emotional practices because they do not merely represent emotions but actively participate in shaping them. Literary scholars have shown how such texts articulate characters' intellectual and emotional lives and mediate affective experience through tone, form,

and genre, thereby offering privileged insight into how emotions shape and are shaped by particular life-worlds and historical contexts (Nussbaum, 2008; Ngai, 2005; Felski, 2008).

However, these literary studies have not been particularly explicit in addressing how emotions are concretely expressed through the medium of literary fiction—namely, language. To better understand this, we draw on the linguistic theories of expressiveness (Foolen, 1997) and modality (Jensen, 1997). The linguistic examination of emotions has neither been extensively developed nor systematically pursued. Foolen (1997) presents a theory of linguistic emotionality that serves as our key inspiration. Drawing on scholars such as Karl Bühler and Roman Jakobson, Foolen argues that language fulfills multiple functions, one of which is emotionality—or what he terms "expressiveness."

Expressiveness, however, is just one of the ways emotions can be conveyed. Foolen (1997) distinguishes two primary modes of emotional communication: verbal and non-verbal. Non-verbal emotional communication encompasses behavior, facial expressions, bodily reactions (e.g., blushing, shivering), and non-verbal vocalizations (e.g., crying, laughing, screaming). Verbal expressions of emotion, according to Foolen, can be categorized as either conceptual or expressive. Conceptual communication of emotions refers to utterances in which the propositional content explicitly concerns emotions, as seen in a sentence such as *Anger and sorrow alternated especially in the lady's soul*, where *anger* and *sorrow* are integral to the proposition. Expressive communication of emotions, on the other hand, is more implicit. Drawing on the linguistic theory of modality, expressiveness can be understood as an additional layer of information that conveys the sender's stance toward the proposition (Jensen, 1997). For instance, *It damn well isn't irrelevant, Father!* conveys the sender's anger through the curse word *damn*, as well as the direct address of the recipient (*Father!*), yet without an explicit propositional reference to anger.

While Foolen distinguishes between verbal (conceptual or expressive) and non-verbal emotional communication, our data reveals a form of communication that bridges the two: the verbal depiction of non-verbal emotional behavior. For example, *A cold shiver ran down the student's spine* or *Eriksen's laughter came in short, rapid bursts* do not contain any verbal communication of emotions, but the texts still communicate emotions by describing non-verbal behavior–cold shivers and laughter–which can be intersubjectively interpreted as expressions of specific emotions i.e. fear and joy, respectively. Based on this linguistic framework we can define three ways in which emotions are expressed in our data: 1) Conceptual communication 2) Expressive communication, and 3) Depictions of non-verbal communication.

This interdisciplinary combination of insights from psychology, cultural theory, literary theory, and linguistics in the study of emotions and their expression in literary texts provides a robust foundation to examine such a complex phenomenon as emotions. We do not propose that our conceptualization of emotions and their literary expressions is universal in either cultural, historical, or epistemological terms. Rather, we have sought to create a theoretical foundation of emotions that is historically and culturally specific and attends to our specific research questions.

Challenges in annotating emotions. Some of the challenges associated with annotating emotions cannot be fully resolved but we have sought to make them more transparent by for instance acknowledging the inherent biases in annotation. Given that the experience of emotions is subjective and shaped by various intersections of social categorizations—including age, gender, racialization, class, nationality, and profession—emotion annotation cannot always be considered a purely linguistic analysis. Instead, it is likely to be influenced by factors such as domain expertise, implicit bias, and the socio-cultural valorization of certain behaviors and language use.

In this study, we elucidate the challenges of annotating emotions using a multi-label dataset, meaning that one data point can have multiple labels, as this approach better reflects how emotions function in literary texts, where one utterance can express a mix of more than one primary emotion in line with Plutchik's description of the dyadic mixtures of the primary emotions (Plutchik, 2001).

Importantly, although we have emphasized in the annotation guidelines that the annotation should focus on classifying the emotion conveyed by the text sample rather than the emotion experienced by the reader, certain cases may arise where the text sample can be interpreted as depicting multiple and sometimes conflicting emotions. This variation depends on whether the annotation considers the emotional experience of a particular char-

acter or the narrator's emotional stance toward the narrative. For instance, the sentence *Tight—like the rope around a hanged man's neck—her arm lay across his throat* can be classified as expressing either anger or fear, depending on whether the annotation foregrounds the experience of the female aggressor or the male victim. Such complexity is an inherent quality of literary texts and cannot be eliminated; nevertheless, it should be acknowledged as a limitation of emotion analysis as a methodological approach to literary examination.

5 Dataset

5.1 Main corpus

We use the MeMo corpus (Bjerring-Hansen et al., 2022), which contains 859 Danish and Norwegian novels from the last 30 years of the 19th century.[1] This dataset, referred to as the main corpus, comprises over 64 million tokens and serves as a valuable resource for exploring the emotional features and shifts of the era. It should be noted that written Norwegian and written Danish were virtually identical until 1907 (Vikør, 2022).

5.2 Annotated Sub-Corpus

To ensure that our emotion classifier is both accurate and reliable, we create a carefully annotated sub-corpus extracted from the main corpus. The main corpus is parsed in sentences before 3,018 random sentences are extracted. 2,532 sentences are included in the training data, and 486 sentences are included in the test data.

For constructing the training and testing sets, we first select all samples from the most recent year to be included in the testing set to ensure that the model is evaluated on "future" data relative to the training set. Additionally, to ensure temporal coverage, we randomly select one sample from each earlier year to add to the testing set. This approach guarantees representation across all years. Together, these samples were adjusted to comprise approximately 15% of the total selected dataset. The remaining 85% of the data was used to form the training set.

As bigger structures of text such as paragraphs or chapters would most likely express too many different emotions to meaningfully annotate as one expression, we carry out the annotation at the level

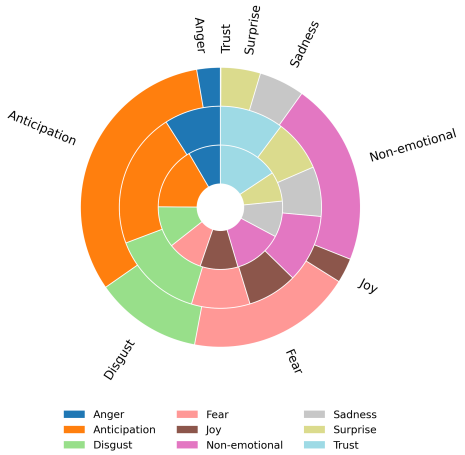

Figure 2: Emotion Class Distribution in annotations for Training (inner), Testing (middle) and predictions on the Main Corpus (outer).

of the sentence. Even though the contextual information is limited when working at the sentence level, often making it harder to infer what specific emotions are expressed in the text, we find that the sentence level is the most optimal compromise between level of contextual information and emotional specificity.

5.3 Annotation process

The annotation was conducted by three literary scholars across academic seniority, all native Danish speakers with domain knowledge in 19th-century Scandinavian literature .[2]

Annotation guidelines. To address the challenges described in section 4, we develop clear annotation criteria to ensure consistency and accuracy in identifying the emotional expressions.[3]

Annotation results. The annotated dataset comprises a total of 3,017 samples, with 2,531 (84.00%) allocated for training and 486 (16.00%) reserved for testing. The distribution of emotional content across these samples reveals notable differences between the two sets.

In the training set, approximately 15.53% of samples were labeled as "No Emotion" , while the majority—61.68% —were assigned a single emotion class . A significant proportion (22.80%) contained two or more emotion classes , indicating complex or overlapping annotations in some cases.

In contrast, the testing set showed a much higher prevalence of single-emotion samples , accounting

[1] Released with Creative Commons Attribution 4.0 license: https://huggingface.co/datasets/MiMe-MeMo/Corpus-v1.1.

[2] Appendix C provides detailed annotator information.

[3] See Appendices A and B.

	Training		Testing	
	Count	%	Count	%
No Emotion	393	15.53%	45	9.26%
1 Emotion Class	1,561	61.68%	400	82.30%
2 Emotion Classes	577	22.80%	41	8.44%
Total Sentences	2,531	84.00%	486	16.00%
Total words	40,546	–	7,610	–
Unique words	7,243	–	2,156	–
AVG word per sentence	16.02	–	15.66	–

Table 1: Statistics of the annotated corpus, showing sentence and emotion category distribution in training and testing sets.

for 82.30% of all test samples. Only 9.26% were labeled as "No Emotion," and just 8.44% contained multiple emotion classes. Figure 2 shows the distribution of emotion labels across the training set and testing set. Emotions like Anticipation and Non-emotional are more frequent, while others such as Surprise and Sadness appear less often.

These distributions reflect the complexity and variability in human emotion labeling and provide important context for evaluating inter-rater agreement and model performance. Table 1 shows more details about the annotation.

Agreement. The inter-annotator agreement (IAA) analysis demonstrates generally consistent annotations across multiple evaluation methods. When examining agreement for each emotion category separately using Cohen's Kappa, the results indicate moderate to substantial agreement, with an overall average of 0.62. Similarity-based metrics further support this consistency, with an average cosine similarity of 0.71 and an average Jaccard similarity of 0.65 across all samples. Additionally, when considering the binary distinction between emotional and non-emotional content, agreement remains strong, with an average pairwise Cohen's Kappa of 0.70 and a Fleiss' Kappa of 0.70. These findings collectively suggest reliable annotation behavior among the raters. Figure 3 shows the agreement per emotion class among the three annotators on the testing set.

6 Experiment and Results

6.1 Experimental Setup

We fine-tuned several Danish language models on our annotated emotion classification dataset. We split the training data into 90% for training and 10% for validation, and the test set comprised the same texts annotated separately by each expert. Training was performed with a fixed learning rate of 5e-5, a batch size of 32, and a maximum of 20 training epochs.

6.2 Pre-trained Language Models

We evaluated the performance of several Danish and Scandinavian pre-trained language models on our emotion classification task, each differing in architecture, training corpus, and language scope. DanskBERT[4] is a RoBERTa-base model trained extensively on Danish corpora and serves as a strong baseline for general Danish NLP tasks. Danish BERT BotXO[5] is a BERT-base model trained from scratch on contemporary Danish text using a WordPiece tokenizer, optimized specifically for monolingual Danish understanding. MeMo-BERT-3[6] is based on XLM-RoBERTa and further fine-tuned on Danish and Norwegian literature from the late 19th century to better capture historical and literary linguistic nuances. DanBERT[7] follows the standard BERT-base architecture and was trained on 2 million Danish sentences, targeting general-purpose Danish modeling. ScandiBERT[8] is a multilingual BERT model trained on five Scandinavian languages, including Danish, and offers cross-lingual robustness. DFM-Large[9] is a BERT-large model integrated into the Sentence-Transformers framework and optimized for semantic similarity tasks through mean pooling and ranking-based objectives. Lastly, NB-BERT-base[10] is a Norwegian BERT-base model trained on 200 years of Bokmål and Nynorsk texts, included in our evaluation for cross-lingual comparison with related languages.

6.3 Emotion Classification

We frame emotion classification as a multi-label text classification task, where the presence or absence of each emotion label is predicted independently as a binary classification problem. As stated in the previous sections, the dataset contains nine emotion categories: Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise, Trust, and Non-emotional.

[4] https://huggingface.co/vesteinn/DanskBERT

[5] https://huggingface.co/Maltehb/danish-bert-botxo

[6] https://huggingface.co/MiMe-MeMo/MeMo-BERT-03

[7] https://huggingface.co/alexanderfalk/danbert-small-cased

[8] https://huggingface.co/vesteinn/ScandiBERT

[9] https://huggingface.co/KennethEnevoldsen/dfm-sentence-encoder-large-exp2-no-lang-align

[10] https://huggingface.co/NbAiLab/nb-bert-base

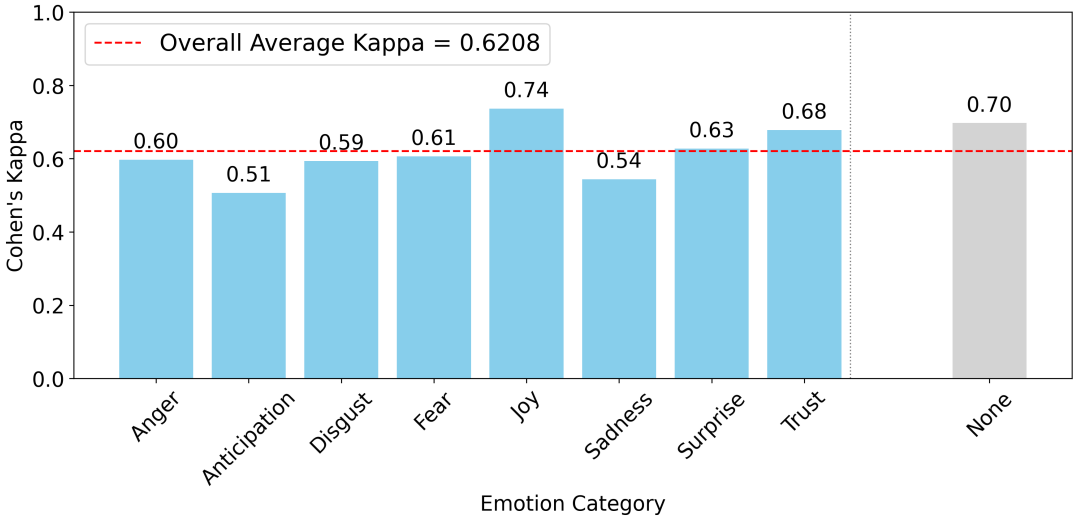

Figure 3: Agreement per Emotion Class (Cohen's Kappa).

Evaluation metrics. Emotion classification involves a high degree of semantic overlap between categories, especially when grounded in Plutchik's emotion wheel. On this wheel, emotions are positioned according to affective proximity, such that those adjacent or nearby (e.g., Anger and Disgust) are semantically more similar than those located further apart. However, standard evaluation metrics such as macro-F1 and micro-F1 treat all misclassifications equally, applying the same penalty to both adjacent and distant emotion pairs. This limitation is particularly problematic for fine-grained, multi-label emotion classification.

To address this issue, we introduce a novel metric called **Plutchik-aware Soft-F1**, which assigns partial credit to predictions that are semantically close to the ground truth. We assign angular positions to each emotion based on Plutchik's emotion wheel. The semantic similarity between two emotions is defined as a function of their angular distance (ranging from 0 to π radians), which is then linearly normalized to the range [0, 1]. This yields a similarity matrix used in our evaluation metric. In addition, we treat Non-emotional as a special class outside the Plutchik wheel. A fixed penalty is applied when it is confused with emotional classes, reflecting its semantic distance from affective states.

Plutchik-aware Soft-F1. Given multi-label ground-truth and predicted vectors $\mathbf{y}_i, \hat{\mathbf{y}}_i \in \{0,1\}^C$, we define the true label set $T_i = \{c \mid \mathbf{y}_i[c] = 1\}$ and the predicted set $\hat{P}_i = \{c \mid \hat{\mathbf{y}}_i[c] = 1\}$.

Let $\mathcal{N} = \{\text{Non-emotional}\}$ denote the set of non-affective label that is not part of Plutchik's emotion wheel. To incorporate semantic proximity between emotions, we define a similarity score $S_{t,p} \in [0,1]$ between any two labels t and p as:

$$S_{t,p} = \begin{cases} 1, & t = p \\ 1 - \dfrac{\theta_{t,p}}{\pi}, & t \notin \mathcal{N},\ p \notin \mathcal{N} \\ 0.5, & t \in \mathcal{N},\ p \notin \mathcal{N} \\ & \text{or } p \in \mathcal{N},\ t \notin \mathcal{N} \\ 1, & t \in \mathcal{N},\ p \in \mathcal{N} \end{cases}$$

where $\theta_{t,p} \in [0, \pi]$ is the angular distance (in radians) on Plutchik's emotion wheel. [11]. The average of these similarities yields the soft precision (SP_i) and soft recall (SR_i) for instance i:

$$\text{SP}_i = \frac{1}{|\hat{P}_i|} \sum_{p \in \hat{P}_i} \max_{t \in T_i} S_{p,t}$$

$$\text{SR}_i = \frac{1}{|T_i|} \sum_{t \in T_i} \max_{p \in \hat{P}_i} S_{t,p}$$

The per-instance soft-F1 is then:

$$\text{SF1}_i = \begin{cases} 1, & \text{if } |T_i| = |\hat{P}_i| = 0 \\ 0, & \text{if } |T_i| = 0 \text{ or } |\hat{P}_i| = 0 \\ \dfrac{2\,\text{SP}_i\,\text{SR}_i}{\text{SP}_i + \text{SR}_i}, & \text{otherwise} \end{cases}$$

[11] See Appendix D.

Finally, the overall **Plutchik-aware Soft-F1** is computed as:

$$\text{Soft-F1} = \frac{1}{N}\sum_{i=1}^{N}\text{SF1}_i$$

Model Performance. As shown in Table 2, DFM-Large achieves the best performance in terms of macro-F1 on both the validation and test sets, demonstrating its strong effectiveness for Danish multi-label emotion classification. While DFM-Large also performs competitively under the Soft-F1 metric, DanskBERT slightly surpasses it on the test set in this regard. This suggests that although DFM-Large is better at exact label matching, DanskBERT's predictions may be more semantically aligned with the ground truth. These results highlight the value of incorporating semantic similarity into evaluation, as Soft-F1 can better capture near-miss predictions that still carry meaningful emotional content.

Model	Macro-F1		Soft-F1	
	Val	Test	Val	Test
Danish BERT BotXO	0.40	0.42	0.58	0.59
DanBERT	0.21	0.30	0.42	0.47
DanskBERT	0.45	0.51	0.67	**0.73**
MeMo-BERT-3	0.50	0.50	0.65	0.65
ScandiBERT	0.41	0.44	0.58	0.58
DFM-Large	**0.54**	**0.53**	**0.71**	0.70
NB-BERT-base	0.45	0.52	0.68	0.67

Table 2: Validation and Test Performance of Danish models on Multi-label Emotion Classification.

To provide a more fine-grained analysis of model performance, we evaluate classification results at the level of individual emotion categories. Table 6 in Appendix E reports Macro-F1 scores for each emotion, based on the best-performing model overall, DFM-Large. We observe that Joy, Non-emotional, and Trust achieve the highest scores, indicating that these classes are relatively easier for the model to identify. In contrast, emotions such as Disgust, Sadness, and Anger exhibit notably lower F1 scores, suggesting greater difficulty in accurate classification. These differences likely reflect varying degrees of semantic clarity. This per-class variation reinforces the need for Soft-F1 metric, especially for emotion labels that are easily confused or semantically adjacent on the Plutchik wheel.

6.4 Sentiment Classification

While fine-grained emotion classification captures nuanced emotional expressions, it often suffers from class imbalance and semantic ambiguity, especially among similar emotion categories. To investigate whether a coarser and semantically clearer categorization would improve robustness and generalization, we introduce a sentiment polarity classification task. This task allows us to assess whether converting emotion labels into broader sentiment categories yields more stable and interpretable results, and serves as a complementary evaluation to the emotion task.

Mapping Rules. To enable sentiment polarity classification, we mapped fine-grained emotion labels into three sentiment classes: Positive, Negative, and Neutral. Specifically, emotions such as joy and trust are categorized as Positive; anger, disgust, fear, and sadness as Negative; and anticipation, surprise, and non-emotional as Neutral. This mapping simplifies the classification task and allows the model to focus on general sentiment trends rather than nuanced emotional states.

Model Performance. As Table 3 shows, among all compared models, DFM-Large exhibits the strongest performance on both validation and test sets. This indicates that DFM-Large remains a robust choice for sentiment classification. Compared to the more challenging emotion classification task, sentiment classification yields higher scores across all models, as expected due to the reduced label space and clearer class boundaries.

Model	Macro-F1	
	Val	Test
Danish BERT BotXO	0.59	0.61
DanBERT	0.47	0.54
DanskBERT	0.64	0.67
MeMo-BERT-3	0.64	0.69
ScandiBERT	0.62	0.64
DFM-Large	**0.69**	**0.75**
NB-BERT-base	0.64	0.70

Table 3: Macro-F1 scores of Danish Pre-trained Models on Sentiment Classification.

7 Classifier-assisted Corpus Analysis

We use the top-performing model, DFM-Large, to classify all unlabeled segments in the main corpus, which consists of over 3.7 million sentences. Figure 2 presents the distribution of emotion classes

predicted from the main corpus. We then analyze the distribution of emotion categories by author gender, as shown in Figure 4. This analysis reveals several notable biases in the correlations between gender, canonization, genre, and emotion—two of which we will highlight.

First, our analysis indicates a prominent gender-based difference in the expression of disgust. Disgust appears in 14.72 % of the sentences authored by women, compared to 12.06 % of those authored by men. This difference highlights nuanced interrelations between gender and emotional expression during a literary period in which patriarchal structures and gender roles were highly debated and polarized topics (Degn et al., 2025).

The analysis suggests that the female authors in particular express experiences of engaging with what is intolerable. Literary scholar Sianne Ngai has argued that "ugly feelings"—i.e., emotions deemed immoral or unproductive, such as disgust or envy—are connected to experiences of suppression and marginalization (Ngai, 2005). Thus, our results speak to contemporary theories on the cultural politics of emotions.

Second, when examining the texts with the highest levels of sadness and surprise, a clear divide emerges between highbrow and lowbrow fiction. The texts scoring highest on these two emotions are largely historical novels, with titles such as *Crozier and Royal Spire*, *Queen and Servant*, and *The Scottish Captain or The Sepoy Bride*. This pattern offers insights into the role that specific emotions played in the period's increasing differentiation of elite and popular literary culture (Bjerring-Hansen and Rasmussen, 2023). Sub-genres were characterized by certain formal and thematic features but also by distinct affective economies.

Although still early in scope, our analysis points to promising directions for future research on the interplay between aesthetics, emotions, and socio-historical contexts.

8 Conclusion

We have introduced an interdisciplinary and theoretically grounded annotation scheme for emotion analysis, integrating insights from psychology, linguistics, and literary studies to account for conceptual, expressive, and non-verbal dimensions of emotion. Building on this framework, we constructed a multi-label dataset of sentences from the Memo corpus, annotated with nine emotion

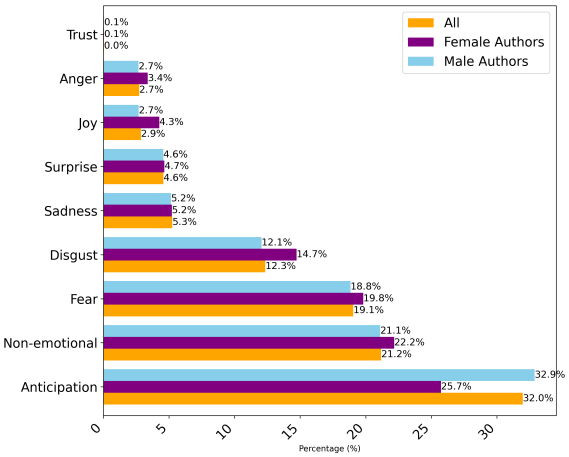

Figure 4: Emotion Distribution by Author Gender

categories derived from Plutchik's theory.

Using this dataset, we evaluated the performance of seven Danish and Norwegian pre-trained language models, and proposed a novel Plutchik-aware Soft-F1 metric that accounts for the affective proximity between emotion categories. Our results show that while models such as DFM-Large perform strongly on conventional metrics, they continue to struggle with the nuanced and overlapping emotional expressions typical of literary texts. These findings underscore the importance of theoretically informed annotation and evaluation strategies in emotion-oriented NLP, especially when applied to complex cultural texts.

Leveraging the best-performing model, we automatically annotated our main corpus. Subsequent analysis revealed significant biases in the relationships between emotion, gender, social status, and genre. For instance, we observed a higher prevalence of disgust in texts by women, contributing nuance to existing scholarship on gender and the cultural politics of emotion. Furthermore, the prominence of sadness and surprise in historical novels suggests that genre and social status are closely linked to affective structures.

Our results demonstrate the potential of combining NLP with interpretive cultural analysis. Our work highlights how emotion analysis can enrich the study of literary and emotional histories, offering scalable, yet nuanced, tools for exploring literary fiction's archive of feelings.

Limitations

While our study offers a novel interdisciplinary framework for emotion analysis in literary texts,

several limitations should be noted. First, the emotional categories are based on Plutchik's model, which, although widely used, imposes a fixed set of emotion labels that may not fully capture the complexity or cultural specificity of emotional expressions in 19th-century Scandinavian literature (De Bruyne et al., 2022; Plaza-del Arco et al., 2024b,a). This raises questions about the universality and granularity of emotion taxonomies when applied to historical and literary contexts.

Second, although our annotation guidelines were designed to balance conceptual and expressive emotion types, annotation remains subjective and shaped by annotators' cultural and academic backgrounds. While we provide transparency by describing annotator demographics, the limited diversity may influence the interpretation of emotions, particularly in ambiguous or polysemous literary expressions.

Third, our annotated corpus is limited in size and drawn exclusively from Danish-language texts, primarily novels. This restricts the generalizability of our findings across genres, time periods, and languages. Future work could extend this approach to poetry, drama, and multilingual corpora to evaluate cross-linguistic and cross-genre validity.

Fourth, our classifier evaluation, although enhanced by the Plutchik-aware Soft-F1 metric, still relies on discrete labels and does not yet incorporate soft or probabilistic annotations. This may oversimplify the often fluid and overlapping nature of emotional expression in literary texts.

Finally, while we integrate linguistic and cultural theory into the annotation and analysis pipeline, our framework does not yet model context beyond the sentence level. Emotions in literature often unfold across broader narrative structures. Future work should explore paragraph- or chapter-level modeling and discourse-aware architectures.

References

Sara Ahmed. 2014. *The Cultural Politics of Emotion*, 2 edition. Edinburgh University Press.

Ali Al-Laith, Alexander Conroy, Jens Bjerring-Hansen, and Daniel Hershcovich. 2024a. Development and evaluation of pre-trained language models for historical Danish and Norwegian literary texts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4811–4819, Torino, Italia. ELRA and ICCL.

Ali Al-Laith, Alexander Conroy, Jens Bjerring-Hansen, Bolette Pedersen, Carsten Levisen, and Daniel Hershcovich. 2025a. Dying or departing? euphemism detection for death discourse in historical texts. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1353–1364, Abu Dhabi, UAE. Association for Computational Linguistics.

Ali Al-Laith, Alexander Conroy, Kirstine Nielsen Degn, Jens Bjerring-Hansen, and Daniel Hershcovich. 2025b. Annotating and classifying direct speech in historical Danish and Norwegian literary texts. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 1–7, Tallinn, Estonia. University of Tartu Library.

Ali Al-Laith, Daniel Hershcovich, Jens Bjerring-Hansen, Jakob Ingemann Parby, Alexander Conroy, and Timothy R Tangherlini. 2024b. Noise, novels, numbers. a framework for detecting and categorizing noise in Danish and Norwegian literature. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3344–3354, Miami, Florida, USA. Association for Computational Linguistics.

Ali Allaith, Kirstine Degn, Alexander Conroy, Bolette Pedersen, Jens Bjerring-Hansen, and Daniel Hershcovich. 2023. Sentiment classification of historical Danish and Norwegian literary texts. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 324–334, Tórshavn, Faroe Islands. University of Tartu Library.

Jens Bjerring-Hansen, Ross Deans Kristensen-McLachlan, Philip Diderichsen, and Dorte Haltrup Hansen. 2022. Mending fractured texts: A heuristic procedure for correcting ocr data. *Digital Humanities in the Nordic and Baltic Countries Publications*, 4(1):177–186.

Jens Bjerring-Hansen and Sebastian Ørtoft Rasmussen. 2023. Litteratursociologi og kvantitative litteraturstudier: Den historiske roman i det moderne gennembrud som case. *Passage - Tidsskrift for litteratur og kritik*, 38(89):171–189.

Patricia Ticineto Clough and Jean Halley, editors. 2007. *The Affective Turn: Theorizing the Social*. Duke University Press.

Catalin Creanga and Liviu P. Dinu. 2024. Isds-nlp at semeval-2024 task 10: Transformer-based neural networks for emotion recognition in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*.

Ann Cvetkovich. 2003. *An Archive of Feelings: Trauma, Sexuality, and Lesbian Public Cultures*. Series Q. Duke University Press, Durham, NC.

Luna De Bruyne, Pranaydeep Singh, Orphee De Clercq, Els Lefever, and Veronique Hoste. 2022. How

language-dependent is emotion detection? evidence from multilingual BERT. In *Proceedings of the 2nd Workshop on Multi-lingual Representation Learning (MRL)*, pages 76–85, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ellen De Geyndt, Orphée De Clercq, Cynthia Van Hee, Els Lefever, Pranaydeep Singh, Olivier Parent, and Veronique Hoste. 2022. Sentemo: A multilingual adaptive platform for aspect-based sentiment and emotion analysis. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA 2022)*, pages 51–61.

Kirstine Degn, Jens Bjerring-Hansen Nielsen, Ali Al-Laith, and Daniel Hershcovich. 2025. Unhappy texts?: A gendered and computational rereading of the modern breakthrough. *Scandinavian Studies*, 97(1):1–24.

R. Donovan, A. Johnson, A. de Roiste, and R. O'Reilly. 2025. Investigating the relationships between basic emotions and the big five personality traits and their sub-traits. *Journal of Personality*. Epub ahead of print.

Rita Felski. 2008. *Uses of Literature*. Blackwell Manifestos. Blackwell Publishing, Malden, MA; Oxford.

Ad Foolen. 1997. The expressive function of language: Towards a cognitive semantic approach. In René Dirven and Susanne Niemeier, editors, *Language of Emotions: Conceptualization, Expression, and Theoretical Foundation*, pages 15–32. John Benjamins Publishing Company, The Netherlands.

Ulrich Greiner. 2014. *Schamverlust: vom Wandel der Gefühlskultur*, 2. edition. Rowohlt, Reinbek bei Hamburg.

Thomas Haider, Daniel Opherk, Lucas Lange, and Steffen Eger. 2020. Po-emo: Conceptualization, annotation, and modeling of aesthetic emotions in german and english poetry. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 1659–1667.

Cindy Harmon-Jones, Brock Bastian, and Eddie Harmon-Jones. 2016. The discrete emotions questionnaire: A new tool for measuring state self-reported emotions. *PLOS ONE*, 11(8):e0159915.

Eva Skafte Jensen. 1997. Modalitet og dansk. *NyS, Nydanske Sprogstudier*, 23:9–24.

Evgeny Kim and Roman Klinger. 2018. Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 1345–1359.

Leonard Konle, Merten Kröncke, Simone Winko, and Fotis Jannidis. 2024. Connecting the dots. variables of literary history and emotions in german-language poetry. *Journal of Computational Literary Studies*, 2(1).

Saif M. Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326.

Sianne Ngai. 2005. *Ugly Feelings*. Harvard University Press, Cambridge, UNITED STATES.

Martha C. Nussbaum. 2008. Democratic citizenship and the narrative imagination. *Teachers College Record*, 110(13):143–157.

Flor Miriam Plaza-del Arco, Alba A. Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 2024a. Emotion analysis in NLP: Trends, gaps and roadmap for future directions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5696–5710, Torino, Italia. ELRA and ICCL.

Flor Miriam Plaza-del Arco, Amanda Cercas Curry, Susanna Paoli, Alba Cercas Curry, and Dirk Hovy. 2024b. Divine LLaMAs: Bias, stereotypes, stigmatization, and emotion representation of religion in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4346–4366, Miami, Florida, USA. Association for Computational Linguistics.

Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350.

Monique Scheer. 2012. Are emotions a kind of practice (and is that what makes them have a history)? a bourdieuian approach to understanding emotion. *History and Theory*, 51(2):193–220.

Jacopo Staiano and Marco Guerini. 2014. Depeche mood: A lexicon for emotion analysis from crowd-annotated news. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 427–433.

Lars S. Vikør. 2022. Rettskrivingsreform i store norske leksikon på snl.no. Accessed: 2025-04-08.

A Annotation Guidelines

Table 4 presents the guidelines given to annotators.

B Category Descriptions

Table 5 lists the emotion categories and their descriptions

Guideline	Description
A	The annotation shall aim to classify the emotion conveyed by the text sample rather than the emotion experienced by the annotator. For instance, *She did not understand where she found the strength to say it.* might make a (feminist) reader feel 'joy,' but should be labeled 'surprise' as that is the emotion the text presents.
B	Each sentence must be labeled with at least one of the nine defined labels (see Table 5 in Appendix B). If a sentence expresses multiple primary emotions, it can be annotated with multiple labels, which are considered ahierarchical.
C	If determining the emotional expression of a sentence is difficult, the annotator should select the most fitting category. The 'Non-emotional' category is only used for utterances which do not express any emotional information.
D	The eight primary emotion categories are: Joy, Trust, Fear, Surprise, Sadness, Disgust, Anger, and Anticipation. The categories are named after the primary emotions but encompass all three levels of intensity. This means that a sentence should be labeled 'Trust' if it expresses either admiration, trust, or acceptance, 'Anger' if it expresses either annoyance, anger, or rage, and so forth. Additionally, the ninth category, 'Non-emotional,' applies to non-emotional utterances as described in Table 5 in Appendix B.
E	Since experiences of emotions are closely linked to language (Scheer, 2012), Danish translations of the three intensity levels for each emotion are included in the category descriptions, as the corpus consists of Danish texts.

Table 4: Annotation guidelines

C Annotator Information

All three annotators are university-educated and -employed literary scholars, who identify as white and possess Danish citizenship. Regarding the gender identity and age of the annotators, one is a woman in her twenties, one is a man in his thirties, and the last is a man in his forties. How these intersections of identity categories specifically influence the annotation of emotions is a very complex question to answer, but by stating them explicitly we hope to underscore that the social position of the annotator is likely to influence the experience of emotions in multiple and indiscrete ways.

D Emotion Wheel Angles

Figure 5 shows the angular arrangement of the eight primary emotions in Plutchik's emotion theory. These angles serve as the foundation for computing semantic similarity in the Plutchik-aware Soft-F1 metric.

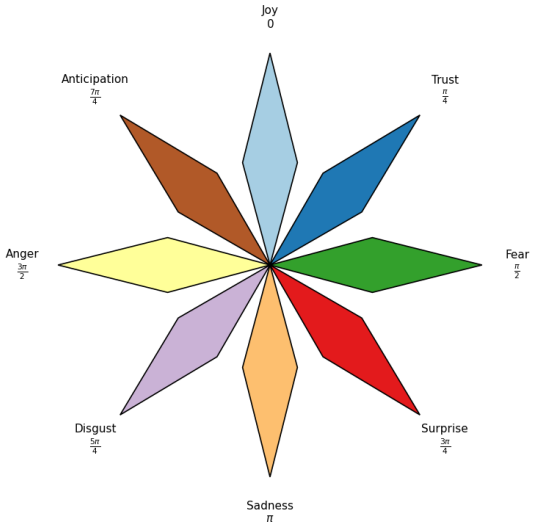

Figure 5: Visualization of Plutchik's emotion wheel, showing each emotion's position in radians.

E Per-class F1 Scores on DFM-Large

Table 6 shows per-class F1 scores on DFM-Large.

Emotion Category	Description	Example
Joy	Sentences expressing serenity, joy, or ecstasy (Danish: ubekymrethed, glæde, ekstase).	*— he concluded with a small smile.*
Trust	Sentences expressing acceptance, trust, or admiration (Danish: accept, tillid, beundring).	*Oh, no problem, my friend!*
Fear	Sentences expressing apprehension, fear, or terror (Danish: ængstelse, frygt, rædsel).	*It was as if her heart stopped beating just at the thought of it.*
Surprise	Sentences expressing distraction, surprise, or amazement (Danish: distraktion, overraskelse, forbløffelse).	*In his perplexity over this, he nearly failed*
Sadness	Sentences expressing pensiveness, sadness, or grief (Danish: vemodighed, bedrøvelse, sorg).	*I'm sad that they got so little out of that trip to Hersted.*
Disgust	Sentences expressing boredom, disgust, or loathing (Danish: kedsomhed, afsky, had).	*I don't care to be told it either.*
Anger	Sentences expressing annoyance, anger, or rage (Danish: irritation, vrede, raseri).	*It damn well isn't irrelevant, Father!*
Anticipation	Sentences expressing interest, anticipation, or vigilance (Danish: interesse, forventning, agtpågivenhed).	*"Well, Geert," said one of the listeners, "you handled your affairs well and received, I imagine, great thanks from the steward of the realm?"*
Non-emotional	Sentences that do not express any emotions, defined as utterances that do not communicate any emotional information through conceptual functions, expressive functions, or depictions of non-verbal communication.	*A week has passed since that evening.*

Table 5: Emotion category descriptions with examples

Class	Macro-F1
Joy	0.65
Non-emotional	0.62
Trust	0.58
Anticipation	0.56
Fear	0.52
Surprise	0.48
Disgust	0.48
Sadness	0.44
Anger	0.40

Table 6: Per-class F1 Scores on DFM-Large.