From Fiction to Fact:
Fine-Grained Emotion Classification in COVID-19 Newspaper Discourse

Anonymous ACL submission

Abstract

This study examines how a computational literary studies (CLS) emotion classification framework can be adapted to analyze newspaper discourse on COVID-19. We developed and tested single-layer and dual-layer BERT models to classify emotions at two levels: 9 primary emotion families (Level 1) and 87 subcategories (Level 2). Using 7,498 sentences from German newspapers, data sparsity directed our focus to the 10 most common Level-2 emotions. Our results revealed varied model performances across emotion categories. The single-layer model exhibited more consistent performance and a stronger correlation with emotion frequency. In contrast, the dual-layer model excelled at distinguishing specific emotions like interest, curiosity, and hope, although with greater variability. Both models struggled to recognize more complex emotions such as LOVE, DISGUST, and AMBIVALENCE. Our results underscore the complexities and potential of automated emotion detection in media discourse, highlighting the need for domain-specific classification methods.

1 Introduction

Emotions are a fundamental part of human cognition and life. Their expression is particularly diverse in language (Schwarz-Friesel, 2007), being omnipresent not only in literary texts (Anz, 2007) but also playing a significant role in seemingly neutral genres such as news reports. However, the systematic analysis and classification of emotions in language remains elusive.

The goals of the present paper are to (i) evaluate an emotion annotation framework developed for fictional texts and apply it to a different genre, i.e. news reporting, and (ii) to develop a classifier for emotion annotation in news reporting and thus determine the extent to which our emotion annotation framework can be generalized to unseen data.

2 Analysis of Emotion in Language and Text

The study of emotions has a rich history across various disciplines, offering diverse conceptualizations. Aristotle defined 15 basic emotions (incl. desire, anger, fear, and joy), while later philosophers such as Descartes and Hume offered different taxonomies, ranging from two to six fundamental emotions (Süselbeck, 2019). In contemporary research, emotion classification mainly follows two approaches: structure-oriented and function-oriented (Schwarz-Friesel, 2017).

Structure-oriented classifications conceptualize emotions as innate and culture-independent and assume that certain emotions emerge from the architecture of the human brain (Damasio 1997, 2004). One example is Ekman's (1972, 1988, 1994) influential model of seven basic emotions—happiness, anger, sadness, fear, disgust, surprise, and contempt, whereas Plutchik's (1984) "wheel of emotions" refers to eight categories. Other notable structural-oriented frameworks list five (Oatley and Johnson Laird 1987), six (Argyle 1996), seven Scherer (1993) or ten (Izard 1992) emotion types.

Function-oriented classifications, by contrast, differentiate emotions based on their referential targets and situational conditions. These include distinctions between target-oriented and non-directional emotions, environment-, body-, or pleasure-related emotions, and categorizations of relational, empathy, and target emotions (Mees, 1985; Holodynski, 2006). Furthermore, linguistic research has explored the pragmatics of the expression of emotion across communicative contexts, as well as semantic analyses on the linguistic representations of emotions (Schwarz-Friesel, 2007). As the above discussion makes clear, there is no psychologically or linguistically motivated consensus as to how emotion in language should be analyzed. As a result, computational applications have pursued a variety of different emotion classifications.

A distinction can be made between the classification of positive versus negative statements and emotion detection, both of which have been explored with machine learning approaches (Ahman 2011; Perikos & Hatzilygeroudis 2016; Al-Baity et al. 2022; Machová et al. 2023; Maruf et al. 2024). Emotion analysis has in particular focused on social media texts due to their accessibility, processability, and abundance of data (Klinger et al., 2020; for an overview, see Acheampong et al., 2020; Liu, 2020; Peng et al., 2022).

Recent research has demonstrated the value of using BERT for emotion classification (Khurdula et al., 2024; Papadimitriou et al., 2024), which we will employ in our own analysis, alongside LSTMs. Moreover, while previous research has mostly focused on smaller emotion inventories (typically 10 or fewer), some recent research has adopted more fine-grained classifications, such as the 27 classes in Singh (2023), and 80 in Luca et al. (2024). We will follow this recent trend.

2.1 Analysis of Emotion in Computational Literary Studies

Emotions also play a fundamental role in literary texts. Since the 1990s, computational literary studies (CLS) have explored emotions in literature (Flüh, 2019; Winko, 2019), employing two main methodologies: lexicon-based (Strapparava & Valitutti, 2004; Taboada & Gillies, 2006; Brugmann & Fabrikant, 2014; Lehmann, Mittelbach & Schmeier, 2017) and machine-learning approaches (Schmidt et al., 2018, Konle et al. 2022, 2023). Lexicon-based methods determine emotional content based on predefined dictionaries, but struggle with context-specific meanings. By contrast, machine-learning approaches generalize emotional content recognition from annotated data, offering greater adaptability to context-dependent usage (Schmidt et al., 2018).

CLS and CL share methodological roots but diverge in objectives; CLS typically focuses on fictional texts, whereas CL is applied to everyday and factual language. Both fields, however, face the limitations of lexicon-based sentiment analysis, which necessitates the development of large language models (LLMs) fine-tuned for specific contexts and relying on extensive manual annotation (Borst et al., 2023).

2.2 A New Emotion Classification System

The discussion of previous research has shown that more fine-grained emotion classifications may be better suited for emotion analysis, but have so far been explored by only a limited number of studies. Our methodology is grounded in a text-centered, inductive approach to emotion classification. Starting with an exploratory annotation phase, we systematically identified and categorized emotion-bearing segments in fictional texts to develop a comprehensive tag set. This data-driven approach differs from traditional deductive frameworks that apply predetermined psychological or sociological categories. The resulting classification system captures the nuanced ways emotions manifest in fictional texts. In a next step, we examined (i) to what extent this emotion classification can be adapted to a different domain, i.e. news reporting, and (ii) to what extent this emotion classification can be generalized with the help of machine learning.

2.3 Emotions in Public Discourse

Newspapers play a critical role in shaping public discourse, oscillating between neutral reporting and emotionally charged narratives. However, even reporting aiming at neutrality and objectivity contains emotional language (Stenvall, 2014; Zappettini et al., 2021). This variety in journalistic styles makes newspapers an ideal testcase to explore how media narratives influence societal perspectives and reactions through automated analysis (Schmitz, 2016; Storjohann & Cimander, 2022). In particular, we focus on reporting on the COVID-19 pandemic, in which emotional language played an important role (Lemor and Montpetit, 2024; Zhunis et al., 2022) and which had great practical relevance in its influence on public attitudes and compliance with rules and restrictions aimed at mitigating the spread and impact of the pandemic (Généreux et al., 2021).

3. Aims

This study bridges computational literary studies (CLS) and computational linguistics (CL) by adapting an emotion classification system from literary texts to analyze newspaper articles about the COVID-19 pandemic. By adopting a detailed emotion classification system from CLS, we pursue two objectives:

(1) to test the automatic classification of an extensive category system originally developed for literary studies (domain adaptation)

(2) to enable a more multifaceted analysis of emotions that reveals subtle emotional nuances and intermediate states.

We aim to assess how well this framework captures nuanced emotional expressions in public discourse, distinguishing relevant from irrelevant emotions. This domain adaptation seeks to develop a detailed understanding of emotional expression in news media while evaluating generalization potential to unseen data. Our assumption that human emotions extend beyond basic types to include complex feelings like nostalgia, envy, and pride underpins this approach, aiming for a more comprehensive view of human communication.

4. Data and Methods

In the present study, we expand on previous work that involved pilot annotations of emotions in 10 literary texts to develop and operationalize a nuanced tagset for emotion categories (AUTHOR A).

4.1 Emotion Classification Scheme

Our hierarchical classification spans two levels: 9 Level-1 emotion families (in small capitals) and 87 Level-2 subcategories (in italics). Within subcategories, emotions are ordered by intensity (see Appendix 1).

The Level-1 categories consist of six basic emotions (LOVE, JOY/HAPPINESS, DISGUST, FEAR, GRIEF, and ANGER) and three additional categories (UNCATEGORIZED_POS, UNCATEGORIZED_NEG, AMBIVALENCE) to capture emotional states that extend beyond the basic emotions.

UNCATEGORIZED_POS and UNCATEGORIZED_NEG include emotions that, while clearly positive or negative, could not be definitively assigned to any basic emotion during initial annotation. AMBIVALENCE encompasses emotional states that simultaneously exhibit both positive and negative qualities.[1]

4.2 Data

Our data is drawn from reporting on the COVID-19 pandemic and consists of a random sample of 7,500 sentences drawn from 59 German newspapers sourced from Lexis Nexis, spanning from January 2020 to June 2022. All sentences contained at least one keyword from the semantic categories of vaccination, COVID-19 names, or non-pharmaceutical interventions (e.g., lockdowns). Two sentences were removed because they consisted of one word only, leaving 7,498 sentences to be annotated.

Three human raters independently annotated each sentence for 87 Level-2 emotions, noting for each emotion whether it was present, absent, or potentially present. Annotators were provided with comprehensive guidelines outlining annotation criteria and examples for identifying the target emotions in the dataset. The final dataset consisted of mean scores calculated from all three raters for Level-2 emotions. The frequency of the 87 Level-2 emotions across the dataset was highly uneven and followed a Zipf distribution, with a few very frequent and many fairly infrequent emotion subcategories. For analysis, we focused on the 33 Level-2 emotions that occurred at least 20 times (Sec. 5.1). However, to aggregate the emotions into their corresponding Level-1 categories, the maximum scores of all 87 Level-2 emotions were used, regardless of their frequency (Sec. 5.2 and 5.3).

To illustrate our data, we present examples for the three most frequently occurring Level-2 emotions. The most frequent emotion, *concern*, was found in expressions of worry about ongoing COVID impacts, as shown in:

1. "Corona hat dafür gesorgt, dass es in den kommenden Jahren angespannt bleibt", sagte Wandrey. ["COVID has ensured that the situation will remain tense in the coming years," Wandrey said.]

2. Wenn sich Menschen mit einer Bedrohung wie der Corona-Pandemie

[1] We also include an *uncertain* category for emotionally charged text passages that are not captured by our classification system.

konfrontiert fühlen, gelangen sie häufig zu zwei Erkenntnissen, analysiert Perel: Die Welt, wie wir sie kennen, geht gerade verloren. [When people are confronted by a threat like the COVID pandemic, they often realize two things, according to Perel: The world as we know it is currently being lost.]

The second most frequent emotion, *seriousness*, was identified in statements on protective measures:

3. Der CSU-Politiker betonte aber, dass die bestehende Kontaktbeschränkung und das Distanzgebot weiterhin gelten. [However, the CSU politician emphasized that existing contact restrictions and distance requirements continue to apply.]
4. ZWIESEL Evangelische Gemeinde: Ab dem morgigen Sonntag ist in der Kreuzkirche das Tragen einer FFP2-Maske während des Gottesdienstes Pflicht. [ZWIESEL Protestant Church: From tomorrow, Sunday, FFP2 masks are compulsory during the church services in the Kreuzkirche.]

Finally, *disapproval* emerged as the third most frequently annotated Level-2 emotion, typically in sentences containing criticism towards regulations or social behavior during the pandemic:

5. Die geplanten Ausgangsbeschränkungen zwischen 21 und fünf Uhr seien bei einer Inzidenz von 100 "ein unverhältnismäßiger und epidemiologisch unbegründeter Eingriff in die Freiheit" der Bürger. [The planned curfews between 9 p.m. and 5 a.m. at an incidence of 100 are "a disproportionate and epidemiologically unfounded encroachment on the freedom" of citizens.]
6. In Anbetracht steigender Covid-19-Zahlen in Japan mangele es Bach wohl an “normalem Menschenverstand”, sagte etwa ein Regierungsberater für Infektionskrankheiten. [In view of rising COVID-19 numbers in Japan, Bach seems to lack "basic common sense," a government advisor for infectious diseases said.]

4.3 Model choice

We tested various approaches to generalize emotion annotations in our data, using LSTM and BERT models. LSTM models with FastText and Word2Vec embeddings tailored for German newspaper data showed poor performance in comparison to pre-trained BERT models, prompting us to opt for the latter approach. BERT's deep contextual embeddings and bidirectional processing offer nuanced semantic understanding, suitable for small datasets like ours (Delvin et al., 2019; Garí Soler & Apidianaki, 2021). We explored two distinct BERT-based model configurations to investigate different approaches to emotion classification, a dual and a single-layer model. While the dual-layer BERT model aimed to capture both fine-grained Level-2 emotions and their superordinate Level-1 categories simultaneously, the single-layer BERT model focused exclusively on Level-1 categories to examine whether a simpler architecture might achieve comparable results.

We split the data into training (75%), validation (15%), and testing (15%) sets, with performance evaluated using precision, recall, and Cohen’s kappa, as well as Pearson correlations for emotion frequency relationships.

4.4 Dual-Layer BERT Model

For the dual-layer BERT model, we used the 'bert-base-german-cased' pre-trained model, which was trained on extensive German datasets, including Wikipedia (6 GB), OpenLegalData (2.4 GB), and news articles (3.6 GB), providing robust German language comprehension. The model’s dual output layers served distinct purposes: one layer predicted 33 Level-2 emotions, occurring at least 20 times in the data, while the other layer focused on the nine Level-1 emotion families.

The dual-layer design treated all emotion variables as statistically independent predictors, utilizing separate Dense layers with 33 and 10 neurons respectively for Level-1 and Level-2 emotions. Both layers employed the sigmoid activation function, which operated independently for each neuron in the layer. This architecture allowed a single sentence to be associated with multiple emotions simultaneously, effectively capturing the complexity of emotional expression.

4.5 Single-Layer BERT Model

Our second configuration was a simpler model focused on predicting only Level-1 emotions (emotion families). This model employed the same 'bert-base-german-cased' pre-trained architecture and featured a single Dense layer with ten neurons corresponding to the Level-1 emotions. The model processed binarized input data, with Level-

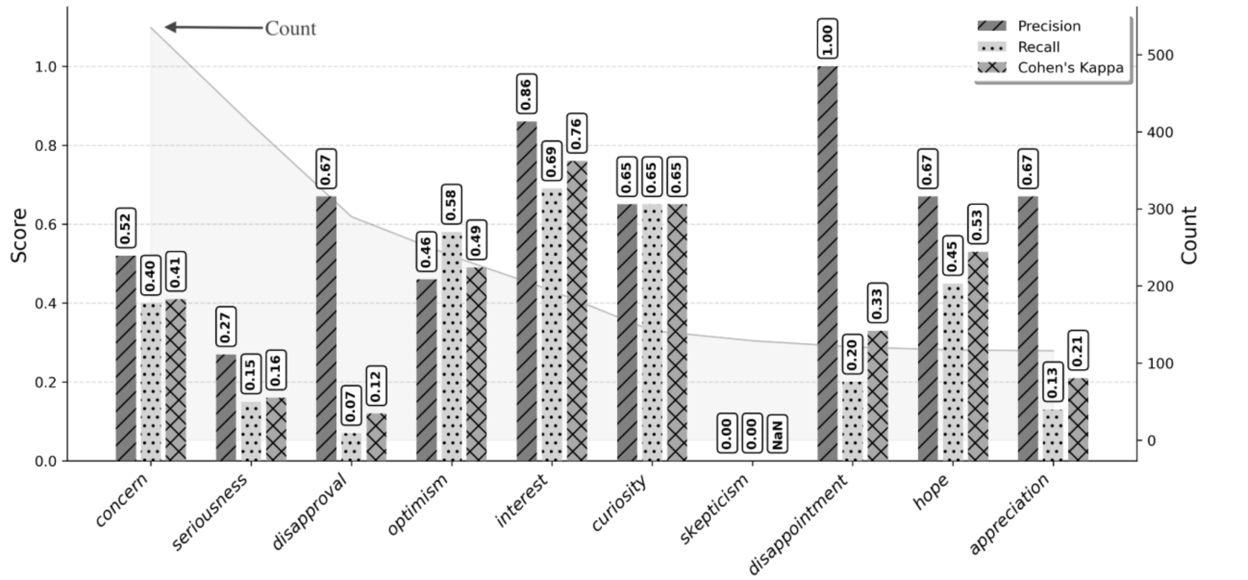

Figure 1: Dual-layer BERT model: precision, recall, Cohen's kappa (bars), and count (line and gray shading in background) for the ten most frequent Level-2 emotions.

1 emotions represented by the maximum scores of their associated Level-2 emotions. The use of the sigmoid activation function for each neuron maintained the independent treatment of variables.

4.6 Attempted Data Augmentation

Due to the uneven distribution of Level-2 emotions (see Sec. 4.2) and the resulting data scarcity, we explored options to augment the training data synthetically. We opted for a strategy of translating all training sentences into another language (both English and Russian) and back into German, which was done with GPT-4o mini. To prevent data leakage, sentences from the validation and testing sets were excluded from the augmentation process. The results showed slight precision and recall differences post augmentation, but no clear improvement, leading us to abandon this approach.

5. Results

5.1 Dual-Layer BERT Model: Level 2

Although the dual-layer BERT model was trained on 33 Level-2 emotions, robust metrics emerged only for the 10 most frequent emotions due to data sparsity in the remaining emotion classes (Fig. 1). The model demonstrated the most consistent and robust performance for *interest*, which showed the highest level of agreement with human annotators - a kappa of 0.76, along with strong precision and recall. *Curiosity*, *hope*, and *optimism* formed a second performance tier with balanced and moderately high scores.

However, the model's performance was notably uneven. Emotions like *concern* and *seriousness* showed moderate to low performance despite being the most frequent. *Disapproval*, *disappointment*, and *appreciation* revealed an interesting pattern of high precision but low recall. This asymmetry was particularly pronounced in *disappointment*, which reached perfect precision while capturing only a fifth of actual instances. The most striking limitation was observed for *skepticism*, which the model failed to identify entirely, resulting in zero precision and recall and an undefined Kappa score.

Overall, the model showed a conservative classification behavior with higher precision than recall values across emotions, and predominantly low-to-moderate kappa scores (below 0.50). Interestingly, correlation analysis revealed weak negative relationships between emotion frequency and all performance metrics (precision: $r = -0.25$, recall: $r = -0.03$, kappa: $r = -0.29$). However, none of these relationships were statistically significant. This indicates that emotion frequency was not meaningfully associated with model performance at Level 2.

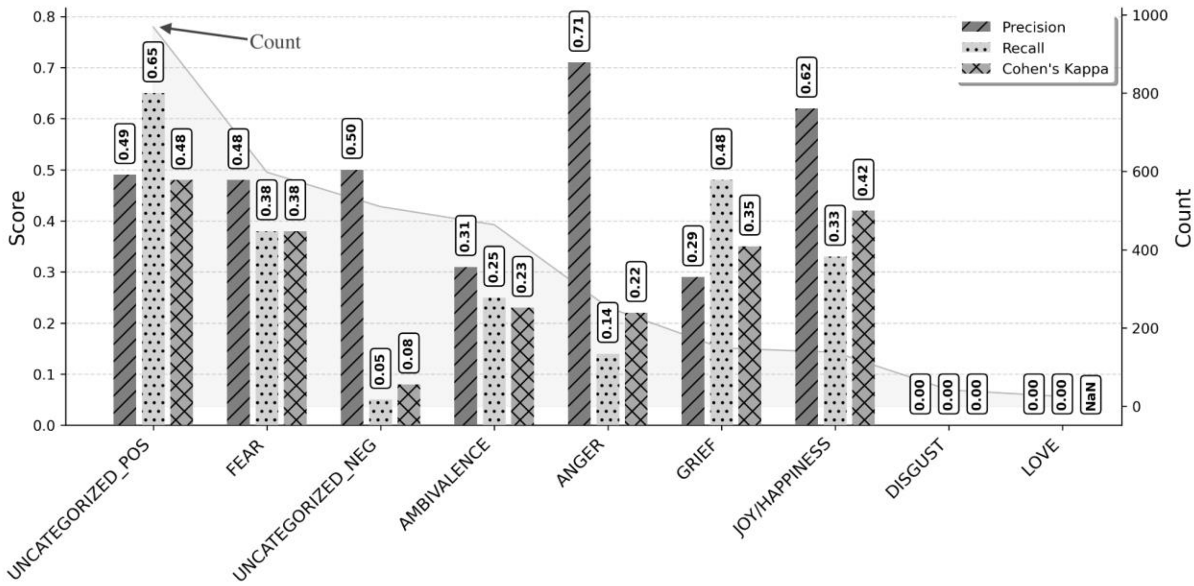

Figure 2: Dual-layer BERT model: precision, recall, Cohen's kappa, and count for Level-1 emotions.

5.2 Dual-Layer BERT Model: Level 1

At Level 1, performance was variable across categories (Fig. 2). The most prevalent category, UNCATEGORIZED_POS, achieved moderate performance with the highest kappa coefficient (0.65). FEAR, the second most frequent category, demonstrated comparable moderate performance levels.

While ANGER had the highest precision (71%), it showed notably low recall (14%). This asymmetric pattern between precision and recall was also evident in UNCATEGORIZED_NEG and JOY/HAPPINESS categories. Further, AMBIVALENCE consistently underperformed across all evaluation metrics. The model particularly struggled with DISGUST and LOVE categories, with LOVE showing zero recall and an undefined kappa value, indicating a complete absence of accurate predictions. In the Level-1 layer, the model generally prioritized accuracy over classification sensitivity, with UNCATEGORIZED_POS and GRIEF being notable exceptions. Correlation analysis revealed moderate positive (but not statistically significant) associations between emotion frequency and performance metrics (precision: r = 0.45, recall: r = 0.62, kappa: r = 0.45).

When compared to Level 2 (see Sec. 5.1), Level 1 demonstrated more uniform performance across categories. Both levels, however, exhibited conservative classification tendencies, manifesting in higher precision scores relative to recall. This effect was more pronounced in Level 2, which displayed more substantial precision-recall disparities (e.g., an 80% gap for *disappointment*) compared to the Level 1 model (e.g., a 57% gap for ANGER).

A notable distinction emerged in the relationship between emotion frequency and performance metrics across the two levels. While Level 2 exhibited weak negative correlations with frequency across all metrics, Level 1 showed moderate positive correlations. This pattern suggests that the increased granularity of emotion categories in Level 2 may have introduced complexities that counteracted potential benefits from higher frequency counts. However, it is important to note that none of these relationships achieved statistical significance.

5.3 Single-Layer BERT Model: Level 1

We finally turn to the single-layer BERT model, which exhibited varying performance across emotion categories (Fig. 3). It achieved balanced metrics for UNCATEGORIZED_POS, with moderate performance for FEAR. However, challenges arose with categories like AMBIVALENCE which had notably low recall (10%) and kappa (0.15) values, indicating substantial difficulties in classification accuracy. DISGUST and LOVE yielded zero or undefined values across all performance metrics.

Precision consistently exceeded recall, indicating reliably positive predictions but challenges in comprehensive emotion identification. Moderate-to-low kappa scores ranged from 0.15 to 0.52.

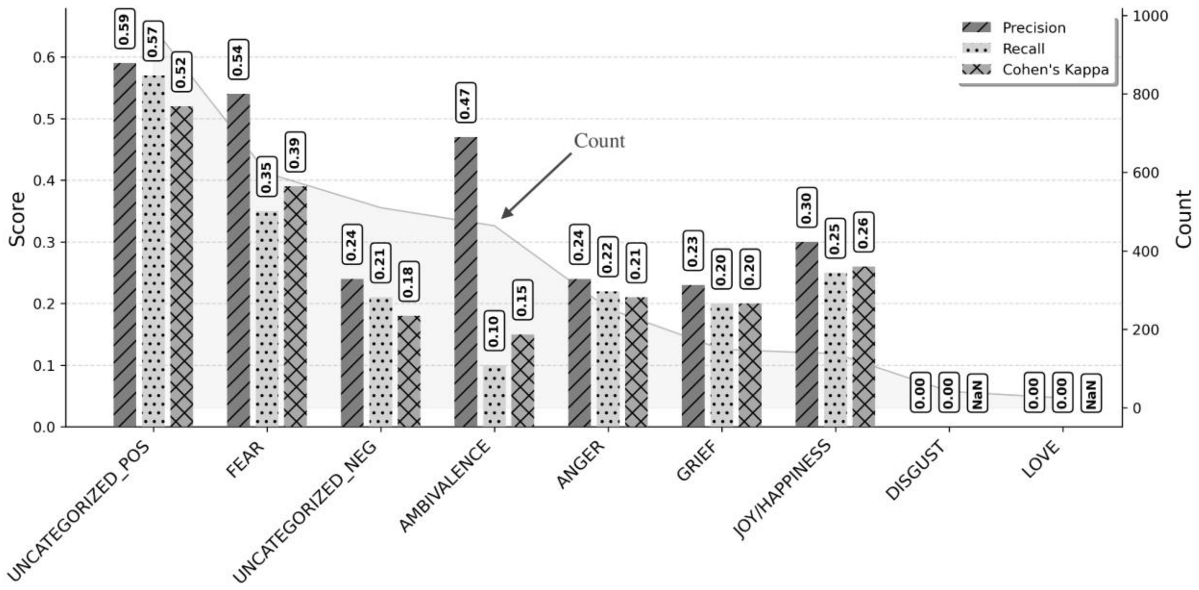

Figure 3: Single-layer BERT model: precision, recall, Cohen's kappa, and count for Level-1 emotions.

Strong, statistically significant correlations were found between emotion frequency and both precision ($r = 0.86$, $p = 0.0$) and recall ($r = 0.84$, $p = 0.01$), suggesting frequency greatly influenced performance.

5.4 Performance Comparison of Single-Layer and Dual-Layer BERT Models (Level 1)

Overall, the single- and dual-layer models revealed distinct classification patterns. Both effectively identified UNCATEGORIZED_POS, but the dual-layer model prioritized sensitivity. In addition, both models struggled with rare emotions like DISGUST and LOVE. Yet, the dual-layer model showed advantages with mid-frequency emotions like ANGER and JOY/HAPPINESS, outperforming the single-layer model, particularly for GRIEF.

A key distinction emerged in precision-recall trade-offs. The dual-layer model exhibited more extreme variations, exemplified by ANGER (71% precision, 14% recall) and UNCATEGORIZED_NEG (50% precision, 5% recall), while the single-layer model showed more moderate performance variations across categories. Although the single-layer model achieved slightly better agreement with human annotators for common emotions, the dual-layer model matched or exceeded these scores for medium-frequency emotions.

The single-layer model's performance benefited more from higher frequency, shown by strong positive correlations with precision and recall, whereas the dual-layer model's correlations were weaker and not statistically significant.

6. Discussion

This study set out to apply a fine-grained emotion classification system developed for literary texts to COVID-19 newspaper discourse. Overall, our findings indicate both promising advances and persistent challenges in adapting this framework to non-fiction texts.

6.1 Model Performance and Comparison

We observed that our dual-layer model excelled in distinguishing emotions such as ANGER and JOY/HAPPINESS, although it faced notable precision-recall trade-offs. Conversely, the single-layer model provided more consistent, balanced performance, effectively identifying UNCATEGORIZED_POS and FEAR. These differences suggest that increased model complexity can enhance sensitivity for certain emotions while potentially compromising overall balance.

6.2 Emotion Frequency, Linguistic Markers, and Granularity

The relationship between emotion frequency and classification performance varied across model architectures. Notably, the single-layer model showed a statistically significant positive correlation between emotion frequency and performance, whereas the dual-layer model did not exhibit significant frequency-performance correlations. This discrepancy may indicate that in more granular classifications, the presence of clear linguistic markers is more influential than mere frequency.

Similar limitations have been observed by Demszky et al. (2020), who analyzed 58,000 Reddit comments labeled with 27 emotions using a BERT-based model. They found that emotions with overt lexical markers (e.g., gratitude) were classified more successfully. Similarly, our findings suggest that in the dual-layer model, with its more granular classification, the presence of clear linguistic markers may be more important than mere frequency.

Comparing Level-1 and Level-2 classifications in the dual-layer model reveals insights into journalistic emotional granularity. Level-2 emotions showed a broader performance range and higher maximum kappa scores, but with greater variability. This result mirrors findings from Machová et al. (2023) suggesting that detecting multiple or weakly supported emotions remains a significant limitation in text-based emotion analysis.

6.3 Implications for Automated Emotion Detection

Both models encountered specific challenges, revealing key patterns in pandemic reporting. Both models struggled with AMBIVALENCE and *skepticism*, despite their potential relevance to COVID-19 news analysis. This finding underscores the challenge of detecting context-dependent emotions, which are often conveyed through subtle linguistic cues. This result aligns with Machová et al.'s (2023) findings, which highlighted the inherent challenge of modeling emotions that involve mixed or contradictory feelings as well as complex emotional expressions that depend on context or cultural understanding, including sarcasm, irony, idioms, and metaphors.

6.4 Generalization of Emotion Detection

Regarding generalization capability, both models exhibited conservative classification tendencies, resulting in higher precision but lower recall scores. This result suggests that while positive predictions generalize well to unseen data, many valid emotional instances may be overlooked.

In our study, 77 out of 87 original Level-2 categories were removed due to data sparsity, underscoring the challenges involved in fine-grained emotion classification. It is plausible to expect that a substantially larger dataset would have yielded more relevant data points, and thus better training data, for rare emotion classes. However, a Zipfian distribution of emotion categories may be inherent to the expression of emotion in language, and partly also context-dependent – the most frequent Level-2 emotions in our COVID-pandemic discourse data, namely *concern*, *seriousness*, *disapproval*, are unlikely to be the most frequent emotions in, say, romantic novels. Ultimately, an important obstacle to the automated detection of emotion in language may be its context-dependent nature. This conclusion is further supported by the lack of a clear relationship between frequency and model performance across emotion categories – more training data does not necessarily imply better model performance, perhaps due to the expression of some emotions being more context-dependent than that of others.

Taken together, these findings have important implications for the analysis of news discourse. Tasks requiring high-precision classification of specific emotions (e.g., *interest* in vaccine development) seem to benefit from a dual-layer architecture, despite its performance variability. Conversely, applications demanding stable and balanced performance across emotion categories (e.g., monitoring broad societal sentiment during vaccination campaigns—UNCATEGORIZED_POS, FEAR, UNCATEGORIZED_NEG) may find the single-layer model more suitable, albeit at the expense of granularity.

7. Conclusion

This study represents a pioneering effort to apply an emotion classification scheme from computational literary studies to newspaper discourse covering the COVID-19 pandemic. By utilizing dual-layer and single-layer BERT models we demonstrated both the potential and challenges of automated emotion identification in journalistic discourse. Our findings suggest that emotion classification is complex, with varying performance across different emotional categories and model architectures.

While our current study focused on German newspaper articles, we plan to extend this methodology to social media data (such as Sailunaz et al. 2018) and other linguistic contexts, thus refining our understanding of emotional communication across different communicative domains. The proposed methodological innovations and insights provide a foundation for more advanced computational approaches to emotion analysis in text.

8. Limitations

This study faced several methodological and practical constraints that should be considered when interpreting the results. First, significant class imbalance in the dataset posed a major challenge, necessitating the exclusion of 77 out of 87 Level-2 emotion categories due to insufficient representation. This substantial reduction in emotional granularity, while methodologically necessary, limited our ability to fully evaluate the effectiveness of fine-grained emotion classification in COVID-related newspaper articles. Targeted data collection strategies, such as selectively sampling articles likely to contain underrepresented emotions, can improve the representation of these categories.

The annotation process presented several interconnected challenges. The task of emotion identification requires high emotional literacy and a solid understanding of emotional nuances from annotators. Despite providing comprehensive guidelines, the inherent complexity of emotion recognition, particularly in journalistic text where emotions may be subtly expressed or implied, likely contributed to annotation inconsistencies. This challenge was particularly apparent in the classification of ambivalent emotions. This limitation could be addressed by developing more precise, domain-specific annotation guidelines, and implementing a multi-stage annotation process with cross-validation among annotators.

In addition, the granularity of our classification scheme, while theoretically comprehensive, proved challenging to implement in practice. The attempt to distinguish between 87 different emotional categories may have been overly ambitious for the newspaper domain, where emotional expression tends to be more restrained and less varied than in literary texts or social media posts. This suggests that a more domain-appropriate classification system might be necessary for analyzing journalistic content.

Finally, our decision to conduct annotation at the sentence level, while practical for implementation, may have limited our ability to capture emotional content that develops across multiple sentences or requires broader context for proper interpretation. Emotions in news articles often emerge through extended narrative development and contextual framing that may be better captured within paragraphs rather than sentences.

These limitations point to several potential improvements for future research: collecting more balanced datasets, developing more robust and domain-specific annotator training guidelines along with a multi-phase annotation approach, and potentially adjusting the granularity of emotion classification to better match the journalistic context. Additionally, exploring paragraph-level annotation might provide more insight into how emotions are conveyed in Coved-related news articles through extended context and narrative development.

Acknowledgements

ChatGPTo3-mini was used while editing this paper. Further acknowledgements will be added once the paper is accepted.

References

Al-Baity HH, Alshahrani HJ, Nour MK, Yafoz A, Alghushairy O, Alsini R, Othman M. Computational Linguistics Based Emotion Detection and Classification Model on Social Networking Data. *Applied Sciences*. 2022; 12(19):9680. https://doi.org/10.3390/app12199680

Acheampong, F. A., Wenyu, C., & Nunoo-Mensah, H. (2020). Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, *2*(7), e12189.

Ahmad, K., Workshop on Emotion, M., & EMOT. (2011). *Affective Computing and Sentiment Analysis: Emotion, Metaphor and Terminology*. Springer Science+Business Media B.V. https://doi.org/10.1007/978-94-007-1757-2.

Anz, T. (2007). Kulturtechniken der Emotionalisierung: Beobachtungen, Reflexionen und Vorschläge zur literaturwissenschaftlichen Gefühlsforschung. In K. Eibl, K. Mellmann, & R. Zymner (Eds.), *Im Rücken der Kulturen* (pp. 207–239). Paderborn: Schöningh.

Argyle, M. (1996). Körpersprache und Kommunikation. Paderborn: Jungfermann.

Borst, J., Klähn, J., & Burghardt, M. (2023). Death of the dictionary? – The rise of zero-shot sentiment classification. In *Proceedings of the Computational Humanities Research Conference* 2023 (pp. 303–319).

Bruggmann, A., & Fabrikant, S. I. (2014). Spatializing a digital text archive about history. In K. Janowicz, B. Adams, G. McKenzie, & T. Kauppinen (Eds.), *Workshop on Geographic Information Observatories 2014: Proceedings* (GIO 2014 / GIScience: 8, pp. 6–14). Aachen: CEUR Workshop Proceedings.

Damásio, A. R. (1997). Descartes' Irrtum. Fühlen, Denken und das menschliche Gehirn. München: List.

Damásio, A. R. (2004). Emotions and feelings. A neurobiological perspective. In A. S. R. Manstead, N. Frijda & A. Fisher (Eds.), Feelings and emotions. The Amsterdam Symposium (pp. 49–57). Cambridge: Cambridge University Press.

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). 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). Minneapolis, Minnesota: Association for Computational Linguistics.

Ekman, P. (1972). Universals and cultural differences in facial expression of emotion. In J. K. Cole (Eds.), Nebraska Symposium on Motivation (pp. 207–283). Lincoln: University of Nebraska Press.

Ekman, P. (1988). Gesichtsausdruck und Gefühl. 20 Jahre Forschung von Paul Ekman. Paderborn: Junfermann.

Ekman, P. (1994). The nature of emotions. Fundamental questions. New York: Oxford University Press.

Flüh, M. (2019). „Sentimentanalyse". In: *forTEXT. Literatur digital erforschen*. URL: https://fortext.net/routinen/methoden/sentimentanalyse [Access: 13. January 2025].

Garí Soler, A., & Apidianaki, M. (2021). Let's play mono-poly: BERT can reveal words' polysemy level and partitionability into senses. *Transactions of the Association for Computational Linguistics*, 9, pp. 825–844. https://doi.org/10.1162/tacl_a_00400

Généreux, M., David, M. D., O'Sullivan, T., Carignan, M. È., Blouin-Genest, G., Champagne-Poirier, O., ... & Roy, M. (2021). Communication strategies and media discourses in the age of COVID-19: an urgent need for action. *Health Promotion International*, *36*(4), 1178-1185.

Holodynski, M. (2006). Emotionen – Entwicklung und Regulation. Heidelberg: Springer-Medizin-Verlag.

Izard, C. E. (1992). Basic emotions, relations among emotions, and emotion-cognition relation. Psychological Review, 99, 561–565.

Khurdula, H. V., Pagutharivu, A., & Yoo, J. S. (2024, March). The Future of Feelings: Leveraging Bi-LSTM, BERT with Attention, Palm V2 & Gemini Pro for Advanced Text-Based Emotion Detection. In *SouthEastCon 2024* (pp. 275-278). IEEE.

Kim, E., & Klinger, R. (2019). A survey on sentiment and emotion analysis for computational literary studies. *Zeitschrift für digitale Geisteswissenschaften*. https://doi.org/10.17175/2019_008

Klinger, R., Kim, E., & Padó, S. (2020). Emotion analysis for literary studies. In N. Reiter, A. Pichler, & J. Kuhn (Eds.), *Reflektierte algorithmische Textanalyse* (pp. 237–269). Berlin/Boston: De Gruyter.

Konle, L., Jannidis, F., Kröncke, M., & Winko, S. (2022). Emotions and literary periods. In *DH Conference Abstracts*. Digital Humanities.

Konle, L., Kröncke, M., Winko, S., & Jannidis, F. (2023). Connecting the dots: Variables of literary history and emotions in German-language poetry. *Journal of Computational Literary Studies, 2*(1), 1–22. https://doi.org/10.48694/jcls.3604

Landweer, H., & Renz, U. (2008). Zur Geschichte philosophischer Emotionstheorien. In H. Landweer & U. Renz (Eds.), *Handbuch klassische Emotionstheorien* (pp. 1–18). Berlin, New York: de Gruyter.

Lehmann, J., Mittelbach, M., & Schmeier, S. (2017). *Quantifizierung von Emotionswörtern in Texten*. DARIAH-DE Working Papers Nr. 24. Göttingen: DARIAH-DE. https://doi.org/urn:nbn:de:gbv:7-dariah-2017-4-5

Lemor, A., & Montpetit, É. (2024). Exploring the role of uncertainty, emotions, and scientific discourse during the COVID-19 pandemic. *Policy and Society*, puae010.

Liu, B. (2020). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

Luca, M., Lopez, G., Longa, A., & Kaul, J. (2024). How are You Really Doing? Dig into the Wheel of Emotions with Large Language Models. In *2024 Artificial Intelligence for Business (AIxB)* (pp. 72-75). IEEE.

Machová, K., Szabóova, M., Paralič, J., & Mičko, J. (2023). *Detection of emotion by text analysis using machine learning. Frontiers in Psychology, 14.* https://doi.org/10.3389/fpsyg.2023.1190326

Maruf, A. A., Khanam, F., Haque, M. M., Jiyad, Z. M., Mridha, M. F., & Aung, Z. (2024). Challenges and opportunities of text-based emotion detection: A survey. *IEEE Access, 12,* pp. 18416–18450. https://doi.org/10.1109/ACCESS.2024.3356357

Oatley, K., & Johnson-Laird, P. (1987). Towards a cognitive theory of emotions. Cognition and Emotion, 1, 29–50.

Papadimitriou, O., Kanavos, A., Vonitsanos, G., Maragoudakis, M., Karkazis, P., & Mylonas, P. (2024, November). Enhancing Emotion Classification with a Hybrid BERT and CNN Architecture. In *2024 19th International Workshop on Semantic and Social Media Adaptation & Personalization (SMAP)* (pp. 156-161). IEEE.

Peng, S., Cao, L., Zhou, Y., Ouyang, Z., Yang, A., Li, X., ... & Yu, S. (2022). A survey on deep learning for textual emotion analysis in social networks. *Digital Communications and Networks*, *8*(5), 745-762.

Perikos, I., & Hatzilygeroudis, I. (2016). *Recognizing emotions in text using ensemble of classifiers. Engineering Applications of Artificial Intelligence, 51,* 191-201. https://doi.org/10.1016/j.engappai.2016.02.002

Plutchik, R. (1984). Emotions. A general psycho-evolutionary theory. In K. Scherer & P. Ekman (Eds.), Approaches to emotion (S. 197–200). Hillsdale: Erlbaum.

Sailunaz, K., Özeyer, T., Rokne, J., & Alhajj, R. (2018). Text-based analysis of emotion by considering tweets. In T. Özeyer & R. Alhajj (Eds.), *Machine learning techniques for online social networks* (pp. 219–236). Cham: Springer.

Scherer, K. (1993). On the nature and function of emotion: A component process approach. Theorien und aktuelle Probleme der Emotionspsychologie.

Schmidt, T., Burghardt, M., & Wolff, C. (2018). Herausforderungen für Sentiment Analysis bei literarischen Texten. In M. Burghardt & C. Müller-Birn (Eds.), *INF-DH 2018* (pp. 1–9). Bonn: Gesellschaft für Informatik e.V. https://doi.org/10.18420/infdh2018-16

Schmitz, U. (2016). 92. Kulturwissenschaftliche Orientierung in der Medienlinguistik. In L. Jäger, W. Holly, P. Krapp, S. Weber & S. Heekeren (Eds.), *Sprache - Kultur - Kommunikation / Language - Culture - Communication: Ein internationales Handbuch zu Linguistik als Kulturwissenschaft / An International Handbook of Linguistics as a Cultural Discipline* (pp. 901-908). Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110224504-093

Schwarz-Friesel, M. (2007). *Sprache und Emotionen*. UTB.

Singh, G., Brahma, D., Rai, P., & Modi, A. (2023). Text-based fine-grained emotion prediction. *IEEE Transactions on Affective Computing*.

Stenvall, M. (2014). Presenting and representing emotions in news agency reports: On journalists' stance on affect vis-à-vis objectivity and factuality. *Critical Discourse Studies*, *11*(4), 461-481.

Storjohann, P., & Cimander, L. (2022). Approaching the Covid-19 discourse through neologisms in public communication. In M. Jakosz & M. Kałasznik (Eds.), *Corona-Pandemie: Diverse Zugänge zu einem aktuellen Superdiskurs* (pp. 25–51). Göttingen: V&R.

Strapparava, C., & Valitutti, A. (2004). WordNet-Affect: An affective extension of WordNet. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, & R. Silva

(Eds.), P*roceedings of the 4th International Conference on Language Resources and Evaluation* (LREC: 4, Vol. 4, pp. 1083–1086). Paris.

Süselbeck, J. (2019). Sprache und emotionales Gedächtnis. Zur Konstruktion von Gefühlen und Erinnerungen in der Literatur und den Medien. In H. Kappelhoff, J. H. Bakels, H. Lehmann & C. Schmitt (Eds.), Emotionen. Ein interdisziplinäres Handbuch (pp. 282–295). Stuttgart: Metzler.

Taboada, M., Gillies, M. A., & McFetridge, P. (2006). Sentiment classification techniques for tracking literary reputation. In *LREC workshop: Towards computational models of literary analysis* (LREC: 5, pp. 36–43). Paris.

Winko, S. (2019). Literaturwissenschaftliche Emotionsforschung. In H. Kappelhoff, J.-H. Bakels, H. Lehmann, & C. Schmitt (Eds.), *Emotionen. Ein interdisziplinäres Handbuch* (pp. 397–407). Stuttgart: Metzler.

Zappettini, F., Ponton, D. M., & Larina, T. V. (2021). Emotionalisation of contemporary media discourse. A research agenda. *Russian Journal of Linguistics*, 25(3), 587-595.

Zhunis, A., Lima, G., Song, H., Han, J., & Cha, M. (2022, April). Emotion bubbles: Emotional composition of online discourse before and after the COVID-19 outbreak. In *Proceedings of the ACM Web Conference 2022* (pp. 2603-2613).

Appendix A: Emotion Classification Scheme

Category system for annotating emotions with nine supercategories and 87 subcategories. Within the subcategories, the emotion types are ordered from intense to less intense; positive emotions are red, negative emotions are blue, and emotions with an ambivalent valence are orange

LOVE: *affection, kindness, trust, intimacy, devotion, worship*

JOY/HAPPINESS: *contentment, pleasure, amusement, humor, (joyful) anticipation, enthusiasm, delight*

DISGUST: *weariness, reluctance, aversion, dislike, contempt*

FEAR: *concern, hesitancy, nervousness, creepiness, dread, terror, horror, consternation, panic*

GRIEF: *dejection, loneliness, sorrow, melancholy, despair, suffering*

ANGER: *disappointment, annoyance, indignation, resentment, rage, bitterness, hate*

UNCATEGORIZED_POS: *curiosity, appreciation, admiration, hope, pride, self-confidence, material desire, relief, interest, serenity, empathy, friendliness, gratitude, optimism, schadenfreude, helpfulness*

UNCATEGORIZED_NEG: *longing, masochism, confusion, aggression, nostalgia, impatience, disapproval, skepticism, greed/desire, perplexity, shame, remorse, jealousy, boredom, madness, compassion*

AMBIVALENCE: *courage, seriousness, astonishment, disregard, defiance, love-hate, being deeply moved, impulsiveness, reverence, humility, sadism, mockery, emotional coldness, vehemence*

uncertain: *[annotation]*