Caregivers Attitude Detection From Clinical Notes

Gaetano Manzo[1†], Leo Anthony Celi[2], Yasmeen Shabazz[2], Rory Mulcahey[1], Lorenzo Jaime Flores[2], Dina Demner-Fushman[1]
[1] Computational Health Research Branch, National Library of Medicine, Bethesda, Maryland, USA.
[2] Massachusetts Institute of Technology (MIT), Harvard Medical School, and the Beth Israel Deaconess Medical Center.

Abstract

Caregivers' attitudes impact healthcare quality and disparities. Clinical notes contain highly specialized and ambiguous language that requires extensive domain knowledge to understand, and using negative language does not necessarily imply a negative attitude. This study discusses the challenge of detecting caregivers' attitudes from their clinical notes. To address these challenges, we annotate MIMIC clinical notes and train state-of-the-art language models from the Hugging Face platform. The study focuses on the Neonatal Intensive Care Unit and evaluates models in zero-shot, few-shot, and fully-trained scenarios. Among the chosen models, $RoBERTa$ identifies caregivers' attitudes from clinical notes with an F1-score of 0.75. This approach not only enhances patient satisfaction, but opens up exciting possibilities for detecting and preventing care provider syndromes, such as fatigue, stress, and burnout. The paper concludes by discussing limitations and potential future work.

Introduction

Caregivers' attitudes toward patients impact healthcare quality and disparities.[1] Clinical notes can reveal language that perpetuates negative or positive attitudes, influencing clinicians' decision-making for patient care. Furthermore, analyzing clinicians' attitudes in their notes may catch preliminary markers of fatigue, stress, or burnout, preventing such syndromes.[2,3]

Detecting clinicians' attitudes in clinical notes is a challenging task because clinical notes contain highly specialized and domain-specific language that is often ambiguous and subject to interpretation. Clinicians' attitudes can be expressed explicitly or implicitly in clinical notes, making it difficult to identify them using traditional sentiment analysis methods. In contrast to social media, where emotional content is often expressed more directly, clinical notes contain many technical terms and abbreviations that require extensive domain knowledge to understand. Furthermore, negative language in clinical notes does not necessarily imply an adverse caregiver's attitude. For example, a clinician may use negative language to describe a patient's symptoms, which is necessary for accurate diagnosis and treatment. Therefore, a nuanced understanding of the context is essential to detect clinicians' attitudes in clinical notes. An analysis of sentiment in the nursing notes[4] highlights the challenges of using sentiment analysis techniques developed for social media in the healthcare domain. The authors found that the existing sentiment analysis tools were often misled by the negative language used in clinical notes and failed to identify clinicians' attitudes accurately. This underscores the need for specialized tools and approaches for detecting attitudes in clinical notes.

This study aims to detect caregivers' attitudes from their clinical notes. To this goal, our contributions are the following: (i) we annotate MIMIC clinical notes in collaboration with their owners to have a dataset that redefines the traditional sentiment in the sentiment analysis task. Indeed, since negative words can be used to define the patient's status without expressing caregivers' attitudes, creating a labeled dataset is essential for the success of this study. (ii) We train state-of-the-art language models available on the Hugging Face platform[5] for detecting clinicians' attitudes in clinical notes. (iii) Finally, we evaluate the selected models in zero-shot, few-shot, and fully-trained scenarios.

As the first step, we focus on the Neonatal Intensive Care Unit (NICU) and the top-rated sentiment analysis models on the Hugging Face platform. The findings highlight the importance of utilizing appropriate evaluation metrics and scenarios to accurately assess the model's performance. According to the results, $RoBERTa$[6] is the most suitable model for the given downstream task, with a peak performance of 0.7555 micro-average F1-score during testing.

[†] Correspondence to gaetano.manzo@nih.gov

The rest of the paper is organized as follows. We introduce the state-of-the-art for detecting caregivers' attitudes from clinical notes. Then, we explore the materials, methods, and settings, followed by the numerical evaluation in the Results section. The Discussion section provides the limitations and suggestions for further work. Finally, we recap the main findings in the Conclusion.

Related Literature

In recent years, sentiment analysis has gained significant attention in the healthcare domain due to its potential to improve patient outcomes and quality of care. This section introduces the state-of-the-art sentiment analysis approaches for detecting caregivers' emotions from clinical notes.

Several studies have examined the caregivers' attitudes in medical records and their impact on patient outcomes. Stigmatizing language was expected and may contribute to bias in healthcare settings, which can negatively impact patient outcomes, including access to care and quality of care.[1] Some authors[7] examined the stigma surrounding medical cannabis use among patients with post-traumatic stress disorder and found that stigmatizing language in healthcare settings may contribute to reluctance among patients to disclose their cannabis use. A cross-sectional study of 48,651 admission notes[8] found that stigmatizing language in hospital notes varied by medical condition and was more often used to describe non-Hispanic Black patients. An analysis of The Commonwealth Fund's 2001 Health Care Quality Survey, showed that satisfaction with health services was lower for Hispanics and Asians than for Blacks and Whites.[9] In this study, the lower satisfaction with the services could be explained by racial differences in the quality of patient-physician interactions.

These studies suggest that healthcare providers' attitudes impact patient perceptions and outcomes, particularly for patients from marginalized communities.

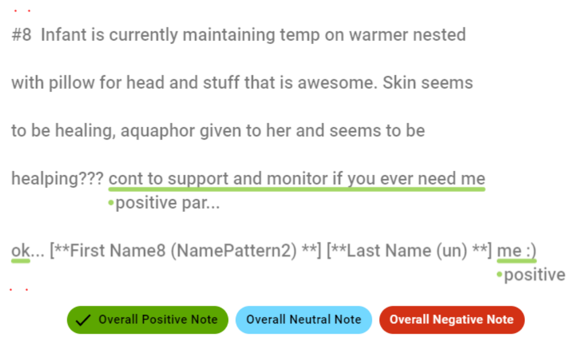
Several authors[10] analyzed caregivers' attitudes toward dementia patients in clinical notes. Caregivers expressed being overwhelmed, stressed, and experiencing emotions like frustration and sadness directed at patients, highlighting a need for interventions.[10] Another study[11] revealed mixed caregiver attitudes toward medication management, influenced by knowledge, complexity, and patient behavior. A pioneering study[12] examined links between language use and psychological traits, attempting to predict depression based on text markers and the Depression Anxiety Stress Scale-21.[13] Predictive models yielded weak recall and moderate precision, suggesting detectable markers but necessitating further analysis and architectures.[13]

Several studies employed machine learning to analyze sentiment in clinical notes within the MIMIC-III dataset.[14,15] These studies suggest sentiment analysis could identify higher readmission risk, despite limitations stemming from single sentiment lexicons and potential natural language processing errors. Collectively, these findings underscore the value of sentiment analysis in understanding caregivers' emotional experiences,[4,14–17] while urging further research to address limitations and explore potential impact of the caregiver's sentiment on the patients' outcomes.

To conclude our related work section, the following studies provide valuable insights and methodologies for understanding sentiment analysis mainly in the context of social networks, and they serve as relevant references for our current research. VADER is a rule-based sentiment analysis tool designed specifically for analyzing sentiment in social media texts.[18] While it is easy to implement and interpret, it may not capture complex contextual nuances and may struggle with words or phrases not present in the lexicon. To overcome this limitation, Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks were applied for sentiment analysis on Twitter data.[19] Additionally, an unsupervised joint aspect and sentiment model via Deep Learning were proposed to jointly learn aspect and sentiment representations from social media data.[15] Similarly, a Distant Supervision model investigates a distant supervision approach for emotion classification in social media by incorporating visual content (images and videos) in text.[15]

Methodology

In this section, we introduce the material and methods to detect caregivers' attitudes by analyzing their clinical notes. We first describe the MIMIC dataset, notes selection, and the annotation process. Then, we present the models selected



#8 Infant is currently maintaining temp on warmer nested

with pillow for head and stuff that is awesome. Skin seems

to be healing, aquaphor given to her and seems to be

healping??? cont to support and monitor if you ever need me

•positive par...

ok... [**First Name8 (NamePattern2) **] [**Last Name (un) **] me :)

•positive

✓ Overall Positive Note | Overall Neutral Note | Overall Negative Note

Figure 1: Example of a positive caregiver attitude from MIMIC clinical note (Subject ID 7919, Caregiver ID 17496, and Row ID 1791482).

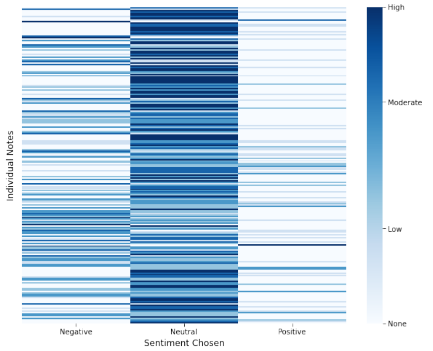


Figure 2: Heatmap of annotations for 223 clinical notes, with intense colors indicating higher agreement and lighter colors indicating lower agreement for each overall sentiment category.

leveraging the Hugging Face platform that allows users to share machine learning models and datasets.[5] Finally, we present the experimental setting from the raw notes to the evaluation of the results.

MIMIC database and Annotations

Medical Information Mart for Intensive Care (MIMIC) Clinical Database[20] is a freely-available database comprising de-identified health-related data hosted on the PhysioNet platform.[21] The database includes demographics, vital sign measurements at the bedside, laboratory test results, procedures, medications, imaging reports, mortality, and, most importantly for this work, clinical notes.

As a first step, we focus on the Neonatal Intensive Care Unit (NICU) patients who lived for at most one year. Please note that the cohort was selected after conferring with clinical note providers, as those clinical notes are reported in stressful conditions. We note this restriction as a limitation, and we intend to expand the cohort with further ICUs and patients in further studies. The selected notes were further pre-processed, removing duplicate notes, notes integrated into other notes, and notes missing caregiver id. The 223 resulting notes (69 caregivers, 39 patients, and 40 hospital admissions) were annotated by five annotators who selected one or more overall note sentiments (positive, negative, and neutral) and the respective note spans (positive or negative). We utilized the open-source text annotation tool Doccano, for our annotation process. An example of a positive caregiver attitude with the corresponding highlighted span is shown in Figure 1.

To ensure that the annotations of caregivers' attitudes are reliable and accurate, four different measures of agreement were used in this study: Fleiss' kappa,[22] Cohen's kappa,[23] Krippendorff's alpha,[24] and weighted F1-score. Fleiss' kappa is a statistic used to measure the degree of agreement among multiple annotators, while Cohen's kappa is used to measure the agreement between two annotators. Krippendorff's alpha is a measure of inter-rater reliability that considers the nominal, ordinal, and interval scales of data. Finally, F1-score is a popular metric commonly used to evaluate the performance of classification models. Results indicate a fair level of agreement among the five annotators. Fleiss' kappa, Cohen's kappa, and Krippendorff's alpha all produced agreement values in the range of 0.257-0.264, with an average agreement value of 0.260. In addition, the average agreement value obtained for F1-score is 0.632, with a minimum and maximum pair agreement of 0.453 and 0.853, respectively. These results suggest that the annotators were able to agree on the sentiment of the clinical notes to a reasonable degree. It should be noted that the minimum and maximum agreement thresholds were not specified for this study. However, assuming a minimum threshold of 0.2 and a maximum threshold of 0.8, the agreement values obtained in this study would be considered acceptable.[25] Overall, these results suggest that the agreement between annotators for sentiment analysis is reasonable and adequate for the task.

The heatmap in Figure 2 displays the annotations of 223 clinical notes, where warmer colors indicate high agreement and lighter colors indicate low agreement for each sentiment category. The results reveal that annotators reached a higher level of agreement for the neutral sentiment category compared to the positive and negative categories. This finding may suggest that the clinical notes contain more neutral content, making it easier for annotators to reach a consensus on this category. On the other hand, the lower agreement for positive sentiment could be due to the relative scarcity of positive expressions in clinical notes, or to the difficulty in identifying and agreeing on instances of positive sentiment.The annotated dataset is submitted to the PhysioNet platform as *'MIMIC III Neonatal Intensive Care Unit Caregivers Attitude Annotations'*.

Model Selection

Introduced in Vaswani et al.,[26] Transformer-based architecture has become extremely popular in natural language processing (NLP) for tasks such as text summarization, sentiment analysis, and beyond. The Transformer architecture is the basis for many NLP models, such as BERT,[27] GPT,[28] and T5.[29] Furthermore, the Hugging Face library,[5] which includes pre-trained models, provides easy-to-use, fine-tune, and test state-of-the-art models.

Transformer-based models have consistently outperformed their competitors across various NLP tasks, including Named Entity Recognition, translation, Question-Answering, and more. Consequently, opting for Transformer-based models in this paper is indispensable to ensure cutting-edge results and advancements in the field.[6,30–32] In this work, we used the most popular sentiment analysis models (i.e., best performance), both pre-trained and fine-tuned, available on Hugging Face:

- $DistilBERT$ is a distilled version of the BERT model, designed to be faster and more memory-efficient. It has been pre-trained on a large text corpus and fine-tuned on several sentiment analysis datasets.[30]
- $RoBERTa$ is a variant of the BERT model further optimized for pre-training. It has achieved state-of-the-art performance on several NLP benchmarks, including sentiment analysis.[6]
- $MiniLM$ is another variant of the BERT model optimized for pre-training. It has fewer parameters than BERT, making it faster and more memory-efficient.[31]
- $BLOOM$ is essentially similar to GPT (auto-regressive model for next token prediction). Still, it has been trained with 46 languages and 13 programming languages, outperforming other models in the Zero-shot setting.[32]

Please note that we intentionally left out models such as ClinicalBERT, BioBERT, and more since those models are not pre-trained for sentiment analysis tasks.

Setup

The goal of this experiment is to detect caregivers' attitudes using MIMIC clinical notes. To achieve this, we follow a pipeline from raw data to results, including data pre-processing, training, and fine-tuning of models based on the selected use case. Particularly, after collecting the data previously presented, removing all the unnecessary features (chart date, category, etc.), and keeping the content of the note and the respective identifier for caregivers and patients. Then, we train models on the annotated data via supervised learning techniques, where we instruct models on labeled data to predict whether a caregiver expresses emotions based on related clinical notes. Models' performance is evaluated on a held-out test set to determine accuracy, precision, recall, F1 score, and other performance metrics in the following scenarios: (I) Zero-shot learning is a technique that requires minimal human intervention and does not rely on data labeling. Instead, the models depend on previously trained concepts and additional existing data. In zero-shot learning, new categories are described at a high-level, allowing the models to relate them to existing categories learned during the pre-training step. (II) Few-shot or low-shot learning allows the selected models to tune their parameters using a small set of examples from new data. This method is used for optimizing the models' parameter-tuning with a low risk of overfitting. (III) In addition to the listed scenarios, we train our models using the standard split of training (60%), validation (20%), and testing (20%) datasets. The final step is to deploy the selected models. This can be done using the Hugging Face API, which allows users to input clinical notes and detect any caregiver's attitude within the notes. Section provides the links to the Hugging Face APIs and the pipeline implementations.

Evaluation

We use various average methods, including macro (M), micro (μ), and mean (m), to evaluate the results based on recall, precision, and F1-score. Our evaluation process consists of two parts: firstly, we evaluate the overall sentiment as positive, negative, or neutral. Secondly, we evaluate a binary scenario where the positive and negative results are combined into a single feature and compared against the neutral sentiment. This scenario is referred to as the *collapse* scenario, and we denote models evaluated under this scenario with c, such as $RoBERTa_c$.

Additionally, some models provide only positive and negative sentiment predictions in the zero-shot scenario. Therefore, we relabel the outliers of positive and negative labels into neutral to ensure consistency across all models. These models are denoted with n, such as $BLOOM_n$.

Results

In this section, we present the results of the selected models for detecting the caregivers' attitudes using the annotated data presented above. It is important to note that the performance of the models depends on the specific task and domain they are applied to. Therefore, it is crucial to evaluate the models under different scenarios and to select the appropriate model based on the specific requirements of the task.

We start by analyzing the zero-shot learning scenario.

Zero-shot

Table 1 presents a comprehensive evaluation of the selected sentiment analysis models based on different types of averages, including macro (M), micro (μ), and mean (m) metrics. The table compares the performance of $RoBERTa$, $DistilBERT$, $BLOOM$, and $MiniLM$, under the zero-shot scenario.

The first three rows of Table 1 provide the results of the sentiment analysis models based on their basic implementation. It is worth noting that these models use different labeling schemes. For instance, $RoBERTa$ uses positive, negative, and neutral labels, whereas DistilBERT, $BLOOM$, and $MiniLM$ use only positive and negative labels. In $DistilBERT_n$ and $BLOOM_n$, we relabel the outliers of positive and negative labels into neutral to ensure consistency across models. Please notice that $DistilBERT_n$ and $BLOOM_n$ are compared with $RoBERTa$, which does not require relabeling.

To extend the analysis, the last three rows of Table 1 collapse the positive and negative labels into a unique label to detect the attitude beside its sentiment. This approach allows us to assess the performance of sentiment analysis models in identifying the overall attitude of the clinical note, rather than focusing only on the polarity of the sentiment.

Finally, we avoid reporting the results of the $MiniLM$ model from the analysis. This is because the $MiniLM$ model provides the same output regardless of the input, making it unsuitable for the zero-shot scenario.

The $RoBERTa$ model outperforms all other models in every setting, which can be attributed to its pre-training on the sentiment analysis task in its basic implementation. In contrast, the pre-training of $DistilBERT$ and $BLOOM$ models focuses more on downstream tasks such as text translation and generation. Relabeling the $DistilBERT_n$ and $BLOOM_n$ models improves their accuracy, as demonstrated in Table 1. However, relabeling the models has an impact on their granularity. The sentiment of $RoBERTa_c$, $DistilBERT_c$, and $BLOOM_n$ models is collapsed into a single label, resulting in an accuracy gain but a loss of granularity.

Table 1: Zero-shot Evaluation: precision, recall, and F1-score metrics for different sentiment analysis models with various types of averages (Macro, Micro, and Mean). The best performing model in each scenario is highlighted in bold. Results with statistical significance at $p < 0.05$.

$Model$	$precision_M$	$recall_M$	$f1_M$	$precision_\mu$	$recall_\mu$	$f1_\mu$	$precision_m$	$recall_m$	$f1_m$
$RoBERTa$	0.7539	0.3683	**0.3567**	0.6604	0.6604	**0.6604**	0.6290	0.6604	**0.6443**
$DistilBERT$	0.1713	0.4436	0.2240	0.1927	0.1927	0.1927	0.0537	0.1927	0.0781
$BLOOM$	0.1092	0.4055	0.1702	0.1718	0.1718	0.1718	0.0415	0.1718	0.0665
$DistilBERT_n$	0.4647	0.3869	0.2750	0.2864	0.2864	0.2864	0.5986	0.2864	0.2740
$BLOOM_n$	0.3907	0.3473	0.2283	0.2708	0.2708	0.2708	0.5502	0.2708	0.2553
$RoBERTa_c$	0.7142	0.5178	**0.4711**	0.6604	0.6604	**0.6604**	0.6385	0.6604	**0.6489**
$DistilBERT_c$	0.4700	0.4763	0.3145	0.3177	0.3177	0.3177	0.5904	0.3177	0.2908
$BLOOM_c$	0.4400	0.4516	0.2993	0.3020	0.3020	0.3020	0.5508	0.3020	0.2769

Few-shot

Table 2 presents the results of the few-shot scenario, where only 10% of the clinical notes dataset (i.e., 22 clinical notes) is used to train the large-language models. In this scenario, 45% of the dataset is used for fine-tuning the models' parameters, while the remaining 45% is used for testing the models' performance. Unlike the zero-shot scenario, where the models are evaluated without any prior exposure to the target domain, all the models in the few-shot scenario have been trained using the appropriate labels (i.e., positive, neutral, and negative). Therefore, we do not relabel the models such as $DistilBERT_n$ and $BLOOM_n$ necessary in the zero-shot scenario. We still evaluate the collapse case, where a caregiver's attitude is detected beside the sentiment in $RoBERTa_c$, $DistilBERT_c$, $BLOOM_c$, and $MiniLM_c$.

Table 2: Few-shot Evaluation: precision, recall, and F1-score metrics for different sentiment analysis models with various types of averages (Macro, Micro, and Mean). The best performing model in each scenario is highlighted in bold. Results with statistical significance at $p < 0.05$.

$Model$	$precision_M$	$recall_M$	$f1_M$	$precision_\mu$	$recall_\mu$	$f1_\mu$	$precision_m$	$recall_m$	$f1_m$
$RoBERTa$	0.3297	0.3373	0.3208	0.6930	0.6930	0.6930	0.5958	0.6930	0.6326
$DistilBERT$	0.3306	0.3417	**0.3231**	0.7029	0.7029	**0.7029**	0.5979	0.7029	**0.6378**
$MiniLM$	0.2500	0.3066	0.2754	0.6831	0.6831	0.6831	0.5569	0.6831	0.6136
$BLOOM$	0.2465	0.3155	0.2768	0.7029	0.7029	**0.7029**	0.5491	0.7029	0.6166
$RoBERTa_c$	0.4806	0.4917	0.4643	0.6930	0.6930	0.6930	0.6060	0.6930	0.6341
$DistilBERT_c$	0.4959	0.4984	0.4695	0.7029	0.7029	0.7029	0.6152	0.7029	0.6402
$MiniLM_c$	0.5416	0.5176	**0.4988**	0.7128	0.7128	**0.7128**	0.6427	0.7128	**0.6577**
$BLOOM_c$	0.4697	0.4925	0.4474	0.7128	0.7128	**0.7128**	0.6006	0.7128	0.6332

As shown in Table 2, the performance of all the models increases compared to their zero-shot scenario, which is expected since the models have training data to rely on.

All models in Table 2 outperformed the zero-shot case in capturing caregivers' attitudes. Notably, $DistilBERT$ shows the highest performance, achieving a micro-averaged F1-score of 0.7029. This is unsurprising, as $DistilBERT$ has been extensively fine-tuned for sentiment analysis tasks. However, in the collapsed setting, where the caregiver's attitude is identified regardless of sentiment, $MiniLM_c$ and $BLOOM_c$ exhibited the best results, achieving a micro-averaged F1-score of 0.7128. This indicates that these models have the potential for even better performance with further fine-tuning on the clinical notes dataset. As shown in Table 2, the results highlight the importance of using appropriate evaluation metrics and scenarios to assess model performance. Moreover, given the low quantity of data in the few-shots scenario, $RoBERTa$ provides lower $f1_m$ than the zero-shot scenario. Overall, the evaluation of the models reveals that there is potential for further improvement. In light of this, we assess the models' performance below by increasing the number of training samples.

Full-trained

Table 3 presents the final evaluation results, where the selected models were trained on 60% of the available dataset (i.e., 133 clinical notes). These models were then fine-tuned and tested on 20% of the dataset each. As for the few-shot scenario, the same reasoning applies to this evaluation concerning the relabelling and the collapse cases.

Table 3: Full-Trained Evaluation: precision, recall, and F1-score metrics for different sentiment analysis models with various types of averages (Macro, Micro, and Mean). The best performing model in each scenario is highlighted in bold. Results with statistical significance at $p < 0.05$.

$Model$	$precision_M$	$recall_M$	$f1_M$	$precision_\mu$	$recall_\mu$	$f1_\mu$	$precision_m$	$recall_m$	$f1_m$
$RoBERTa$	0.4250	0.4083	**0.4043**	0.7333	0.7333	**0.7333**	0.6461	0.7333	**0.6793**
$DistilBERT$	0.3675	0.3982	0.3803	0.7111	0.7111	0.7111	0.6159	0.7111	0.6589
$MiniLM$	0.2500	0.3030	0.2739	0.6666	0.6666	0.6666	0.5500	0.6666	0.6027
$BLOOM$	0.2416	0.2929	0.2648	0.6444	0.6444	0.6444	0.5316	0.6444	0.5826
$RoBERTa_c$	0.6875	0.5946	**0.6011**	0.7555	0.7555	**0.7555**	0.7283	0.7555	**0.7169**
$DistilBERT_c$	0.6346	0.5795	0.5833	0.7333	0.7333	0.7333	0.6974	0.7333	0.7000
$MiniLM_c$	0.5750	0.5378	0.5286	0.7111	0.7111	0.7111	0.6566	0.7111	0.6654
$BLOOM_c$	0.4625	0.4810	0.4560	0.6666	0.6666	0.6666	0.5850	0.6666	0.6140

Despite the fact that increasing the number of training samples led to an improvement in the overall performance of the models in Table 3, $RoBERTa$ consistently outperformed the other models in all scenarios and metrics.

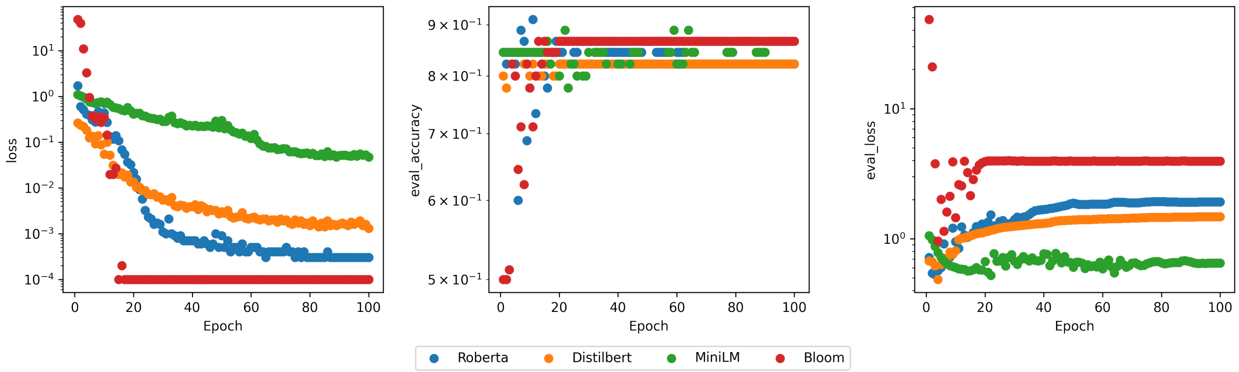


Figure 3: From left to right: training loss, validation F1-score, and validation loss over epochs in log scale.

It should be noted that $BLOOM$ delivers the fastest training in terms of epochs, the lowest loss, and the highest validation accuracy, as illustrated in Figure 3. Computational time was not a concern, as the pre-trained models were executed on 4 Tesla V100 32GB GPUs.

Discussion

The study aimed to detect caregivers' attitudes from their clinical notes and made several contributions, including annotating the MIMIC clinical notes, training state-of-the-art language models, and evaluating the selected models in zero-shot, few-shot, and fully-trained scenarios. The performance of the different sentiment analysis models evaluated depended on their pre-training and relabeling settings. In particular, $RoBERTa$, which has been pre-trained on sentiment analysis, outperformed the other models in every setting. However, relabeling $DistilBERT_n$ and $BLOOM_n$ improved their performance. Lastly, collapsing the sentiment into a single label resulted in a gain in accuracy but a loss of granularity.

The results highlight the superiority of $RoBERTa$ in sentiment classification due to the number of training steps, batch sizes, and additional pre-training data. Models like $BLOOM$ and $DistilBERT$ showed potential once trained.

$MiniLM$ also showed comparable performance, considering its smaller size and faster training time than larger models. Overall, the selected large language models were able to detect caregivers' attitudes from their clinical notes in the considered scenarios.

Our work has some limitations that should be considered in the future work. One limitation is the use of a single dataset (MIMIC) to train and evaluate the models. While MIMIC is a rich source of clinical notes, it may not be representative of other healthcare settings or populations. Future work could involve collecting data from multiple sources to increase the generalizability of the models. Additionally, the annotated dataset is relatively small and only covers a specific healthcare setting (NICU). A small dataset could limit the models' ability to learn complex patterns and generalize to other settings. Future work could involve collecting larger and more diverse datasets to improve the models' performance and generalizability. Furthermore, caregivers may express their attitudes explicitly, but it is also possible that their attitudes are implicit or communicated through nonverbal cues. Therefore, it would be interesting to investigate whether other sources of data, such as audio or video recordings, could provide additional insights into caregivers' attitudes.

Finally, the study did not investigate how the models' performance changes over time or how the models could be adapted to account for changes in attitudes. Therefore, future work could involve analyzing the trajectories of caregivers' attitudes over time and investigating how language models could be used to capture these changes.

Overall, this study enables the detection of caregivers' attitudes in healthcare reports, which can drastically impact caregivers' and patients' outcomes.

Resources

Please find here the link to the GitHub repository: $https://github.com/tanoManzo/mimic\_attitude$, and the link to the Hugging Face repository: $https://huggingface.co/tanoManzo/$.

Conclusions

The study successfully detected caregivers' attitudes from clinical notes using state-of-the-art language models in various scenarios, with RoBERTa performing best in zero-shot and few-shot scenarios. Future work could involve investigating other data sources, collecting larger and more diverse datasets, and analyzing the trajectories of caregivers' attitudes over time to improve the models' performance and generalizability.

Acknowledgements

This research was supported in part by the Intramural Research Program of the National Library of Medicine, National Institutes of Health, HHS.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Park J, Saha S, Chee B, Taylor J, and Beach MC. Physician use of stigmatizing language in patient medical records. *JAMA Netw Open*, 4(7):e2117052, 2021.

2. Nath S and Kurpicz-Briki M. Burnoutwords - detecting burnout for a clinical setting. *Machine Learning Techniques and Data Science*, 2021.

3. Merhbene G, Nath S, Puttick AR, and Kurpicz-Briki M. BurnoutEnsemble: Augmented intelligence to detect indications for burnout in clinical psychology. *Frontiers in Big Data*, 2022.

4. Waudby-Smith I, Tran N, Dubin JA, and Lee J. Sentiment in nursing notes as an indicator of out-of-hospital mortality in intensive care patients. *PLoS One*, 13(6):e0198687, 2018.

5. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, and Funtowicz M. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv preprint arXiv:1910.03771*, 2019.

6. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, and Stoyanov V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

7. Krediet E, Janssen DGA, Heerdink ER, Egberts TCG, and Vermetten E. Experiences with medical cannabis in the treatment of veterans with ptsd: Results from a focus group discussion. *European Neuropsychopharmacology*, 36:244–254, 2020.

8. Himmelstein G, Bates D, and Zhou L. Examination of stigmatizing language in the electronic health record. *JAMA Netw. Open*, 5(1):e2144967, 2022.

9. Saha S, Arbelaez JJ, and Cooper LA. Patient-physician relationships and racial disparities in the quality of health care. *Am. J. Public Health*, 93(10):1713–1719, 2003.

10. Brodaty H and Donkin M. Family caregivers of people with dementia. *Dialogues Clin. Neurosci.*, 11(2):217–228, 2009.

11. Lim RH and Sharmeen T. Medicines management issues in dementia and coping strategies used by people living with dementia and family carers: A systematic review. *Int. J. Geriatr. Psychiatry*, 33(12):1562–1581, 2018.

12. Havigerová JM, Haviger J, Kučera D, and Hoffmannová P. Text-based detection of the risk of depression. *Frontiers in Psychology*, 10, 2019.

13. Morales M, Scherer S, and Levitan R. A cross-modal review of indicators for depression detection systems. *Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality, ACL*, pages 10.18653/v1/W17–3101, 2017.

14. Gao Q, Wang D, Sun P, Luan X, and Wang W. Sentiment analysis based on the nursing notes on in-hospital 28-day mortality of sepsis patients utilizing the MIMIC-III database. *Comput. Math. Methods Med.*, 2021.

15. Zhu L, Xu M, Bao Y, Yu X, and Kong X. Deep learning for aspect-based sentiment analysis: a review. *PeerJ Comput Sci*, 8:e1044, 2022.

16. McCoy TH, Castro VM, Cagan A, Roberson AM, Kohane IS, and Perlis RH. Sentiment measured in hospital discharge notes is associated with readmission and mortality risk: An electronic health record study. *PLoS One*, 10(8):e0136341, 2015.

17. Mahmoudi E, Wu W, Najarian C, Aikens J, Bynum J, and VGV Vydiswaran. Identifying caregiver availability using medical notes with rule-based natural language processing: Retrospective cohort study. *JMIR Aging*, 5(3):e40241, 2022.

18. Hutto CJ and Gilbert EE. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM)*, 2014.

19. Ramadhani AM and Goo HS. Twitter sentiment analysis using deep learning methods. In *2017 7th International Annual Engineering Seminar (InAES)*, pages 1–4, 2017.

20. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035, 2016.

21. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, and Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):E215–20, 2000.

22. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

23. Cohen J. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

24. Krippendorff K. Content analysis: An introduction to its methodology. *Sage Publications, Inc*, 2004.

25. Landis JR and Koch GG. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

26. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, and Polosukhin I. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

27. Devlin J, Chang MW, Lee K, and Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

28. Radford A, Narasimhan K, Salimans T, and Sutskever I. Improving language understanding by generative pre-training. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22, 2018.

29. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, and Liu PJ. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

30. Sanh V, Debut L, Chaumond J, and Wolf T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

31. Wang W, Zhang F, Liu X, Sun X, Lin J, Xie H, Chen M, Chen Z, Zhang M, and Ji X. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *arXiv preprint arXiv:2002.10957*, 2020.

32. BigScience Workshop. Bloom: A 176b-parameter open-access multilingual language model, 2022.