

A Comparative Study of Healthcare Emotion and Sentiment Analysis using Machine Learning approach

Saroj Kushwah¹, Ajay kumar²

¹Lincoln University College Malaysia; ² Lincoln University College Malaysia

Email ID : ¹pdf.sarojkushwah@lincoln.edu.my; ²pdfsv.ajaykumar@lincoln.edu.my

Abstract: Healthcare-related communication is now available in many forms, including patient reviews, clinical notes, telemedicine conversations, and online feedback. These texts can reveal how patients respond to their experiences, but sentiment labels alone may not capture the full meaning of what is written. For example, a negative statement about a hospital service may reflect dissatisfaction, whereas a negative statement about a diagnosis may express fear or uncertainty.

In this study, we compare different approaches for identifying sentiment and emotion in healthcare text. The comparison covers four conventional machine-learning classifiers, namely Naive Bayes, Logistic Regression, Support Vector Machine, and Random Forest, together with Bi-LSTM, BERT, and ClinicalBERT. Healthcare datasets such as MIMIC-III, CADEC, and MedDialog are considered to represent different types of clinical and patient-authored text.

The evaluation is planned using accuracy, precision, recall, macro-F1, and class-wise results, along with factors such as computational cost, calibration, interpretability, and robustness. We also consider class imbalance and the differences between clinical documentation and patient-written text. The experimental runs are not yet complete, so no numerical results are presented as final findings. Instead, this paper describes the evaluation procedure and the comparisons that will be carried out. The resulting framework is intended to provide a consistent basis for future studies of sentiment and emotion analysis in healthcare communication, where model outputs can be considered alongside human judgment.

Keywords: Healthcare NLP; sentiment analysis; emotion recognition; machine learning; deep learning;

Introduction

The Healthcare communication is increasingly available in digital form through electronic health records, patient portals, telemedicine platforms, online reviews, and doctor–patient conversations. These texts contain more than medical information; patients also describe reactions such as satisfaction, frustration, fear, uncertainty, and trust. Understanding these expressions can provide useful information about patient experiences. Sentiment analysis usually assigns text to broad categories such as positive, negative, and neutral. However, two texts with the same sentiment may express very different emotions. For example, a complaint about a long waiting time and a statement expressing fear about a diagnosis may both be classified as negative, although their underlying meanings differ [1], [2].

Earlier sentiment-analysis methods commonly used word counts, n-grams, sentiment lexicons, and traditional classifiers [1], [2]. Neural models later learned representations directly from text, while transformer architectures improved the modeling of contextual information. BERT demonstrated the

effectiveness of bidirectional pertaining [3], and models such as BioBERT and ClinicalBERT extended this approach to biomedical and clinical language [4], [5].

This study compares classical machine-learning models, Bi-LSTM, BERT, and ClinicalBERT using a common evaluation procedure. The aim is to examine the trade-off between model complexity and performance across healthcare text. The study also considers differences between clinical notes, patient narratives, and doctor–patient conversations, since a model performing well on one type of text may not perform equally well on another.

In addition to classification performance, the evaluation considers class imbalance, interpretability, computational requirements, calibration, and privacy. These factors are important when healthcare text-analysis models are considered for practical use rather than only benchmark evaluation..

Related work

Research on sentiment analysis has traditionally focused on determining the overall polarity of written content. Pang et al. investigated supervised learning techniques for this purpose and provided early machine-learning baselines [2]. Liu’s work later covered a wider range of topics, including opinion mining and the identification of subjectivity in text [1].

Sentiment labels do not always provide enough detail to describe a person’s response. Emotion analysis addresses this limitation by assigning more specific emotional categories. The GoEmotions dataset is an example of fine-grained emotion classification research [7]. In healthcare communication, identifying emotions such as fear, anger, sadness, trust, or gratitude can add information that a simple polarity label may miss.

The language used in healthcare datasets varies with the source of the text. Clinical notes often contain medical abbreviations, specialized vocabulary, and incomplete sentences. MIMIC-III has been used extensively for research on clinical text [6]. CADEC contains patient reports related to adverse drug events [8], whereas MedDialog is built around doctor–patient conversations [9].

The development of domain-specific language models has also influenced healthcare NLP. BioBERT was designed to work with biomedical text, while ClinicalBERT was adapted to clinical language [4], [5]. Their effectiveness can depend on the dataset, task, and labels involved, so results from one healthcare text source may not directly apply to another.

Evaluation becomes more important when sentiment or emotion classes are unevenly distributed. Accuracy may hide poor performance on less frequent classes, making precision, recall, and F1-score useful complementary measures [11]. Interpretability is another practical concern. LIME can be used to examine which input features influence a model’s prediction, although such explanations should not be treated as evidence of clinical validity [12].

Table 1. Relevant studies and resources

Study / Resource	Domain	Main task	Approach	Relevance to this study
[1] Liu	General text	Sentiment	Survey / methods	Foundational sentiment concepts
[2] Pang et al.	General reviews	Polarity	Classical ML	Baseline supervised learning
[3] Devlin et al.	General NLP	Text classification	BERT	Transformer baseline
[4] Alsentzer et al.	Clinical NLP	Clinical text tasks	ClinicalBERT	Clinical-domain representation
[5] Lee et al.	Biomedical NLP	Biomedical text tasks	BioBERT	Domain-adapted pretraining
[7] Demszky et al.	Social/general text	Emotion	GoEmotions	Fine-grained emotion taxonomy
[8] Karimi et al.	Patient narratives	Adverse events	CADEC annotations	Patient-authored healthcare text
This work	Healthcare text	Sentiment + emotion	Comparative ML/DL/Transformers	Unified benchmark protocol

Key Contribution

The study is organized around a single comparison procedure rather than separate experiments designed independently for each model. The same general data preparation and evaluation principles are followed across the candidate models. Sentiment polarity and emotion recognition are treated as separate tasks so that their results can be examined independently.

The comparison includes four conventional classifiers, one recurrent neural network, and two transformer-based models. The conventional methods provide simple baselines, while Bi-LSTM, BERT, and ClinicalBERT allow the effect of learned contextual representations to be investigated.

Method, Experiments and Results

Workflow of proposed work

The analysis starts with healthcare text collected from suitable public or de-identified datasets. The raw text is checked before preprocessing so that duplicate records, incompatible labels, and other data-quality issues can be identified.

After preprocessing, the text is converted into a representation suitable for the selected model. TF-IDF and n-gram features are used for conventional models, while neural and transformer models operate on learned or contextual representations. The resulting representations are passed to separate sentiment and emotion classifiers.

The final stage compares the models using common evaluation metrics and examines the errors produced by each approach.

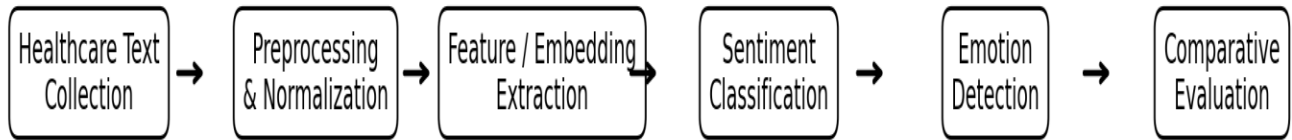


Figure 1. shows the main stages of the proposed workflow.

Dataset Strategy

The study considers healthcare resources that represent different forms of language. MIMIC-III is relevant to clinical documentation, CADEC provides patient-written adverse-drug-event narratives, and MedDialog contains doctor–patient conversations [6], [8], [9].

An emotion-focused dataset can be included when its annotation scheme is compatible with the emotion categories used in the study. The label definitions should be examined before combining or comparing datasets. Differences in annotation guidelines can otherwise make numerical comparisons misleading.

Each dataset should also be checked for duplicate records, class distribution, language, de-identification, licensing conditions, and privacy requirements. When label sets differ, the mapping between the original labels and the common evaluation labels should be documented rather than assumed.

Table 2. Candidate models for comparative evaluation

Model family	Candidate model	Representation	Strength	Limitation
Classical	Naïve Bayes	TF-IDF / n-grams	Fast, simple baseline	Weak contextual modeling
Classical	Logistic Regression	TF-IDF / n-grams	Strong linear baseline	Limited nonlinear context
Classical	SVM	TF-IDF / n-grams	Effective in sparse spaces	Training/inference tuning
Ensemble	Random Forest	TF-IDF / engineered features	Robust nonlinear baseline	Large sparse feature spaces
Deep learning	Bi-LSTM	Learned embeddings	Sequential context	Higher training cost
Transformer	BERT	Contextual embeddings	Strong contextual representation	Compute and memory cost
Clinical transformer	ClinicalBERT	Clinical contextual embeddings	Domain adaptation	Domain/data dependence

Preprocessing and Text Representation

The preprocessing stage includes duplicate removal, Uni-code normalization, whitespace correction, sentence segmentation, and tokenization. The purpose is to reduce noise without removing information that could affect the interpretation of a healthcare statement.

Stop-word removal is applied cautiously because words involved in negation can change the meaning of a sentence. Similarly, spelling normalization should not remove clinical abbreviations or domain-specific terms.

For the classical models, TF-IDF features with word and character n-grams are used as the main representation. These features provide a relatively simple way of representing the frequency and local structure of terms.

Bi-LSTM uses learned token representations and processes information in both directions of a sequence. BERT provides contextual token representations through transformer attention, while ClinicalBERT is included because clinical terminology and writing patterns can differ from general-domain text [3], [4].

Models Included in the Comparison

Experimental Protocol

The models should be evaluated using the same basic train, validation, and test procedure within each dataset. An 80:10:10 split can be used when the dataset is sufficiently large. For smaller datasets, stratified five-fold cross-validation may provide a more stable estimate.

Hyperparameters should be selected using training and validation data only. The final test set should remain separate until the model configuration has been fixed.

Healthcare datasets may contain uneven class distributions. Class weights or resampling can therefore be applied to the training data. The random seed, preprocessing configuration, model settings, and training parameters should be recorded so that the experiment can be reproduced.

When label definitions are sufficiently compatible, cross-dataset testing should also be performed. This can provide an indication of whether a model transfers beyond the source on which it was trained.

A. Evaluation Metrics

Accuracy is useful when class distributions are relatively balanced, but it can hide weak performance on minority classes. Precision, recall, and F1-score should therefore be re-reported for individual classes together with macro and weighted averages.

Macro-F1 gives equal importance to each class and is particularly useful when the distribution of sentiment or emotion labels is uneven [11].

Confusion matrices should be used to inspect common errors. For example, the analysis can identify whether fear is frequently confused with anxiety, or whether different negative emotions are being collapsed into a single class. For a deployment-oriented comparison, inference time, memory consumption, calibration, and interpretability should also be measured. LIME or similar methods can be used to inspect model decisions [12].

Experimental Results

At the present stage, the experiments have not been completed for all candidate models. As a result, this section does not present numerical scores as final findings. The values marked as “[RUN]” in Table 3 are intended to be filled after the corresponding model evaluations are finished.

Table 3. actual experimental comparison where [RUN] entries must be replaced by measured values before submission.

Model	Accuracy	Macro-Precision	Macro-Recall	Macro-F1	Relative cost
Naïve Bayes	[RUN]	[RUN]	[RUN]	[RUN]	Low
Logistic Regression	[RUN]	[RUN]	[RUN]	[RUN]	Low
SVM	[RUN]	[RUN]	[RUN]	[RUN]	Low–Medium
Random Forest	[RUN]	[RUN]	[RUN]	[RUN]	Medium
Bi-LSTM	[RUN]	[RUN]	[RUN]	[RUN]	Medium–High
BERT	[RUN]	[RUN]	[RUN]	[RUN]	High
ClinicalBERT	[RUN]	[RUN]	[RUN]	[RUN]	High

The expected behavior is that TF-IDF-based classifiers will offer simple and efficient baselines, while Bi-LSTM and trans-former models may be more effective when the classification depends on contextual relationships within the text. BERT and ClinicalBERT may also be useful for cases where general context or clinical terminology affects the prediction [3], [4]. These observations represent the hypotheses of the study.

They should be verified using the planned experiments rather than treated as measured performance differences. The entries marked “[RUN]” should be replaced only after the corresponding experiments have been completed.

Error Analysis and Ablation Plan

Model scores alone cannot explain why a classifier succeeds or fails. Errors should therefore be grouped into categories such as negation, uncertainty, abbreviations, mixed sentiment, implicit emotion, very short messages, long clinical narratives, and code-switching.

Ablation experiments can be used to examine the effect of individual components. Possible comparisons include TF-IDF against contextual representations, BERT against Clinical-BERT, and separate sentiment and emotion classifiers against a joint learning approach.

Results should also be compared across datasets. A model that performs well on one source but loses performance on another may be relying on source-specific language patterns rather than learning features that generalize across healthcare text.

Discussion

The expected comparison involves a practical trade-off between model complexity and performance. Classical methods are still useful when a system needs to be inexpensive, fast, and relatively easy to inspect. They also provide a necessary reference point for judging whether a more complex model actually provides a meaningful improvement.

Transformer models can represent context more effectively and may be useful when sentiment or emotion depends on words appearing earlier in a sentence or document. Their larger memory and computational requirements, however, can become important when models are deployed on limited hardware.

Clinical pre-training does not guarantee better results for every healthcare dataset. Clinical notes, patient reviews, and doctor–patient conversations use different styles of language. Consequently, a

model adapted to one source may still require additional evaluation before it is transferred to another [4], [5]. Emotion predictions should also be interpreted with care.

A model assigning a text to a “fear” category does not establish that the writer is clinically experiencing a psychological condition. Such predictions are better viewed as signals that may help organize communication, patient feedback, or other downstream tasks. Human judgment remains important for decisions that can affect patient care.

Privacy is another practical concern because healthcare text may contain sensitive information. Appropriate de-identification, access control, storage procedures, and rules for secondary use are required when working with such data [6], [8].

Fairness should also be examined before deployment. People may express similar emotions differently because of language, culture, communication style, or clinical context. Future evaluation should therefore consider subgroup performance, multilingual settings, cross-institution testing, and uncertainty in model predictions.

Conclusion

Healthcare text can express both general sentiment and more specific emotional responses. A polarity label alone may therefore miss useful differences between patient-authored and clinical text. Comparing several model families under a common procedure can help examine these differences.

The study outlines a comparison of classical machine-learning methods, Bi-LSTM, BERT, and ClinicalBERT. The planned evaluation covers data preparation, text representation, separate sentiment and emotion tasks, classification metrics, error analysis, computational cost, and interpretability.

One advantage of using a common evaluation procedure is that the models can be compared using the same general criteria instead of relying on results from unrelated experiments. Classical methods provide relatively efficient baselines, whereas transformer-based models can capture richer contextual information.

The experimental evaluation is still incomplete, so no model is claimed to be superior at this stage. Completing the benchmark, testing transfer between datasets, examining joint sentiment–emotion learning, and studying subgroup performance are important next steps. Input from healthcare experts will also be needed before such models are considered for practical deployment.

References

1. B. Liu, “Sentiment analysis and opinion mining,” *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, 2012.
2. B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment classification using machine learning techniques,” in *Proc. EMNLP*, 2002, pp. 79–86.
3. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
4. E. Alsentzer et al., “Publicly available clinical BERT embeddings,” in *Proc. 2nd Clinical NLP Workshop*, 2019, pp. 72–78.
5. J. Lee et al., “BioBERT: A pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

6. A. E. W. Johnson et al., "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, Art. no. 160035, 2016.
7. D. Demszky et al., "GoEmotions: A dataset of fine-grained emotions," in *Proc. ACL*, 2020, pp. 4040–4054.
8. S. Karimi, C. Wang, A. Metke-Jimenez, R. Gaire, and C. Paris, "CADEC: A corpus of adverse drug event annotations," *Journal of Biomedical Informatics*, vol. 55, pp. 73–86, 2015.
9. Z. Zeng et al., "MedDialog: Large-scale medical dialogue datasets," in *Proc. EMNLP: System Demonstrations*, 2020.
10. Y. Gu et al., "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 1, 2022.
11. M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.
12. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. KDD*, 2016, pp. 1135–1144.