BRIEF REPORT

OPEN

Negativity and Positivity in the ICU: Exploratory Development of Automated Sentiment Capture in the Electronic Health Record

Chris J. Kennedy, PhD[1,2]
Catherine Chiu, MD[3]
Allyson Cook Chapman, MD[4,5]
Oksana Gologorskaya, MS[6]
Hassan Farhan, MD[7]
Mary Han, MD[2]
MacGregor Hodgson, MD[2]
Daniel Lazzareschi, MD[2]
Deepshikha Ashana, MD, MBA, MS[8]
Sei Lee, MD[9,10]
Alexander K. Smith, MD, MPH[9,10]
Edie Espejo, MA[9]
John Boscardin, PhD[9,10]
Romain Pirracchio, MD, PhD[3]
Julien Cobert, MD[3,11]

OBJECTIVES: To develop proof-of-concept algorithms using alternative approaches to capture provider sentiment in ICU notes.

DESIGN: Retrospective observational cohort study.

SETTING: The Multiparameter Intelligent Monitoring of Intensive Care III (MIMIC-III) and the University of California, San Francisco (UCSF) deidentified notes databases.

PATIENTS: Adult (≥18 yr old) patients admitted to the ICU.

MEASUREMENTS AND MAIN RESULTS: We developed two sentiment models: 1) a keywords-based approach using a consensus-based clinical sentiment lexicon comprised of 72 positive and 103 negative phrases, including negations and 2) a Decoding-enhanced Bidirectional Encoder Representations from Transformers with disentangled attention-v3-based deep learning model (keywords-independent) trained on clinical sentiment labels. We applied the models to 198,944 notes across 52,997 ICU admissions in the MIMIC-III database. Analyses were replicated on an external sample of patients admitted to a UCSF ICU from 2018 to 2019. We also labeled sentiment in 1,493 note fragments and compared the predictive accuracy of our tools to three popular sentiment classifiers. Clinical sentiment terms were found in 99% of patient visits across 88% of notes. Our two sentiment tools were substantially more predictive (Spearman correlations of 0.62–0.84, p values < 0.00001) of labeled sentiment compared with general language algorithms (0.28–0.46).

CONCLUSION: Our exploratory healthcare-specific sentiment models can more accurately detect positivity and negativity in clinical notes compared with general sentiment tools not designed for clinical usage.

KEY WORDS: computer-assisted decision-making; critical care; critical care outcomes; natural language processing; sentiment analysis

Clinical decisions in the ICU, such as when to provide life-sustaining therapies, require synthesizing large amounts of data into an overall opinion or "sentiment" of a patient's clinical status and trajectory. The use of natural language processing (NLP) provides opportunities to analyze language for patterns. A subset of NLP, sentiment analysis, identifies language related to sentiment (1). Not all providers document sentiment explicitly, but subjectivity is often incorporated into notes. For example, if a provider documents "prognosis is poor" in a patient with multiple organ failure, the provider may be explicitly using objective measures (e.g., Sequential Organ Failure Assessment [SOFA] (2)) and implicitly calculating pretest probabilities of interventions that may be beneficial. Here, negative sentiment ("poor") reflects the clinical state and forecasted trajectory ("prognosis"). Negative sentiment is associated with increased hospital readmissions (3) and ICU mortality (4),

DOI: 10.1097/CCE.0000000000000960

KEY POINTS

Question: Can healthcare-specific provider sentiment be measured in an automated and accurate fashion in unstructured clinical notes?

Findings: Using different keywords-based and machine-learning approaches, negativity and positivity are common in clinical notes and have improved accuracy over existing tools.

Meaning: Healthcare-specific sentiment extraction from unstructured notes using keywords-based and machine-learning approaches is more accurate than existing nonspecific tools. This could lead to a deeper understanding of clinical decision frameworks and biases, and possibly interventions targeting provider behaviors.

but these methods rely on tools built from nonmedical dictionaries. When applied to the electronic health record, these algorithms prove poorly sensitive and specific and are inaccurate (5). Validated domain-specific (i.e., medical) sentiment tools are necessary to measure sentiment accurately in clinical settings.

We sought to explore how sentiment could be captured in notes. We first annotated notes for sentiment and performed preliminary validation of labels by comparing annotations for a subsample of notes across multiple labelers. We next addressed the development of automated sentiment detection in two complimentary ways: 1) a keywords-based approach and 2) a keywords-independent supervised learning approach. We hypothesized that negative sentiment is common and that domain-specific sentiment tools might perform better than general-purpose tools for identifying sentiment in clinical notes.

MATERIALS AND METHODS

The Beth Israel Deaconess Medical Center (BIDMC) and Massachusetts Institute of Technology (MIT) institutional review boards (IRB) approved the use of their deidentified dataset following data-use agreements. The use of University of California, San Francisco (UCSF) data was approved by the UCSF IRB (19-29429). We adhere to the TRIPOD reporting guidelines found in **Supplemental Digital Content** (http://links.lww.com/CCX/B236). The global framework for the study can be found in **eFigure 1** (http://links.lww.com/CCX/B236).

Data Sources and Study Populations

The keywords-based model does not rely on any specific dataset but we used Multiparameter Intelligent Monitoring of Intensive Care III (MIMIC-III) to train and validate the deep learning model. To determine the discriminative accuracy of our sentiment models, we then used a separate sample of notes from the MIMIC-III database (i.e., a sample of notes not used for deep learning training). For external validation, we used notes from ICU patients at UCSF from 2018 to 2019 (pre-COVID-19). MIMIC-III is a deidentified dataset developed by MIT and BIDMC and includes demographics, laboratory tests, and clinical outcomes linked to notes across 60,000 ICU admissions from 2001 to 2012 across 5 ICUs (6).

ICU notes per day per patient were the unit of analysis. We included patients greater than or equal to 18 years with greater than or equal to one ICU note from prespecified categories: physician, general, consult, nursing, respiratory, rehabilitation, and nutrition. Only final iterated nursing notes were included to prevent duplications. Data preprocessing and analysis were conducted using Python 3.8 (Python Software Foundation) and R 4.1.3 (R Foundation).

Sentiment keywords Generation and Consensus

Critical care experts (J.M.C., A.C., C.C.) participated in the generation of positive and negative keywords based on prespecified prompts. Individual meetings and group collaboration allowed multiple participants to share ideas, individually and then collectively. This was performed iteratively until consensus was achieved. The keywords list was finalized into a medical lexicon of 72 positive and 103 negative terms (**eTable 1**, http://links.lww.com/CCX/B236). Details on keywords generation are shown in Supplemental Digital Content (http://links.lww.com/CCX/B236). Our conceptualization of sentiment was purposely broad including sentiment related to patients and their clinical status as our initial goal was to show that domain-specific sentiment extraction was possible and valid.

TABLE 1.

Patient Demographics and Clinical Characteristics Across Multiparameter Intelligent Monitoring of Intensive Care and University of California, San Francisco Cohorts

Characteristics	Primary Dataset: MIMIC-III (*n* = 9,104)			External Dataset: UCSF (*n* = 1,123)		
	Total	ICU Survivor	ICU Nonsurvivor (or Discharge to Hospice)	Total	ICU Survivor	ICU Nonsurvivor (or Discharge to Hospice)
	n = 9,104	*n* = 8,073 (88.7%)	*n* = 1031 (11.3%)	*n* = 1,123	*n* = 793 (70.6%)	*n* = 330 (29.4%)
Age (mean [SD])	62.7 (16.6)	62.0 (16.7)	68.9 (14.2)	74.6 (7.9)	73.7 (7.3)	76.8 (8.7)
Self-identified gender						
Male	5,106 (56.1)	4,565 (56.5)	541 (52.5)	618 (55.0)	428 (54.0)	190 (57.6)
Female	3,998 (43.9)	3,508 (43.5)	490 (47.5)	505 (45.0)	365 (46.0)	140 (42.4)
Primary language category[a] (%)						
East or Southeast Asian	131 (1.4)	103 (1.3)	28 (2.7)	142 (12.6)	89 (11.2)	53 (16.1)
Continental European	272 (3.0)	237 (2.9)	35 (3.4)	29 (2.6)	13 (1.6)	16 (4.8)
Other/unknown	303 (3.3)	254 (3.1)	49 (4.8)	11 (1.0)	6 (0.8)	5 (1.5)
Spanish	193 (2.1)	179 (2.2)	14 (1.4)	43 (3.8)	29 (3.7)	14 (4.2)
English	8,205 (90.1)	7,300 (90.4)	905 (87.8)	898 (80.0)	656 (82.7)	242 (73.3)
Self-identified race and ethnicity[a] (%)						
Asian	219 (2.4)	188 (2.3)	31 (3.0)	218 (19.4)	147 (18.5)	71 (21.5)
Black	948 (10.4)	852 (10.6)	96 (9.3)	87 (7.7)	63 (7.9)	24 (7.3)
Hispanic/Latino	346 (3.8)	320 (4.0)	26 (2.5)	149 (13.3)	107 (13.5)	42 (12.7)
Multi/other	60 (0.6)	55 (0.7)	5 (0.5)	185 (16.5)	132 (16.6)	53 (16.1)
Unknown/declined	575 (6.3)	483 (6.0)	92 (8.9)	N/A	N/A	N/A
White	6,956 (76.4)	6,175 (76.5)	781 (75.8)	633 (56.4)	451 (56.9)	182 (55.2)
Clinical characteristics						
Mechanical ventilation (%)	4,231 (46.5)	3,505 (43.4)	726 (70.4)	155 (13.8)	95 (12.0)	60 (18.2)
Hemodialysis initiation (%)	2,098 (23.0)	1,737 (21.5)	361 (35.0)	109 (9.7)	40 (5.0)	69 (20.9)
Tracheostomy placement (%)	625 (6.9)	508 (6.3)	117 (1.3)	12 (1.1)	7 (0.9)	5 (1.5)
Length of stay in days (mean [SD])	9.3 (10.3)	9.1 (9.5)	10.7 (15.2)	3.5 (5.2)	4.7 (12.9)	3.9 (8.3)
Number of elixhauser comorbidities (MIMIC) (mean, SD) or van Walraven score (UCSF) (mean, SD)	2.88 (1.43)	2.85 (1.42)	3.11 (1.48)	13.7 (10.8)	11.1 (9.71)	20.0 (10.7)

MIMIC-III = Multiparameter Intelligent Monitoring of Intensive Care III, N/A = not applicable, UCSF = University of California, San Francisco.

[a]For the purposes of confidentiality, any group with fewer than 10 patients were merged into "Other."

Patients were categorized by those who survived their ICU stay and those who died in the ICU (or were discharged to hospice). *p* values are not included given these covariates are not the primary predictors of interest in this study.

Manual Sentiment Labeling

A total of 1,493 note fragments containing one or more sentiment keywords were manually labeled for sentiment using a Likert-style rating scheme of "very positive," "positive," "neutral," "negative," or "very negative" informed by theorization of a sentiment construct (details of conceptualization and details on note fragment generation are in **Supplemental Methods** http://links.lww.com/CCX/B236). An example sentence with labels is shown in **eFigure 2 (**http://links.lww.com/CCX/B236). We used labeled note fragments to validate the accuracy of sentiment measures and for training and testing the DeBERTa-v3 model.

Development of a Sentiment Score as the Primary Predictor

Two approaches for sentiment scores were performed: 1) using keywords from a lexicon of clinical sentiment ("keywords-based approach") and 2) a supervised machine-learning approach without keywords using deep learning based on the DeBERTa-v3 architecture ("keywords-independent approach") (7). The latter machine-learning model was evaluated because it does not rely on predefined keywords that could be subject to bias. For the keywords-based approach, the sentiment score was defined as (equation 1).

$$\text{Keyword sentiment score} = \frac{\text{No. of negative keywords}}{\text{No. of negative keywords} + \text{No. of positive keywords}} \quad (1)$$

This yielded a continuous sentiment measure ranging from 0 to 100%. For the DeBERTa-v3 approach, the sentiment score was estimated as average predicted sentiment across all sentences, whereby Negative = 0, Neutral = 1, Positive = 2. More negative sentiment represents a higher keywords sentiment score and a lower DeBERTa-v3 score. Details are shown in Supplemental Methods (http://links.lww.com/CCX/B236). The DeBERTa-v3 was trained and internally validated on MIMIC-III and we used UCSF notes for external validation.

Preliminary Sentiment Measure Validation

Ground truth labels used to train the DeBERTa-v3 approach were from a single annotator. To determine consistency of sentiment labeling across multiple annotators, we isolated 100 note excerpts and compared labeling across three blinded clinicians (M.H., M.H., D.L.) using a similar prompt. Sentiment was labeled on a five-value rating scale, using a labeling guide (Supplemental Methods, http://links.lww.com/CCX/B236). Agreement was evaluated using Krippendorf's alpha coefficient, treating sentiment labels as ordinal variables (8). Krippendorf's alpha does not erroneously increase when reviewers systematically disagree, unlike other measures like Cronbach's alpha (9).

Statistical Analysis

We compared our sentiment scores to alternative existing measures using the Spearman rank correlation coefficient for labeled note fragments. Alternative sentiment tools included Stanza (10), Sentimentr (11), and Pattern (12), which had the strongest associations with patient outcomes in one previous study (5). We also compared measures to one another on labeled data to determine similarity (or convergence). Spearman rank correlations and tests of equivalence (13) were calculated on the full set of labeled data except for DeBERTa-v3 measures—for those, only the test set was analyzed to counteract potential overfitting. For data visualizations, we used the ggplot2 (version 3.3.3) in R and "Plotly" package (version 4.14.3) in Python.

RESULTS

Characteristics of the MIMIC ICU Cohort

Sentiment term frequencies from MIMIC-III are shown in **eTable 2** (http://links.lww.com/CCX/B236). Most notes contained at least one sentiment term, and the presence of any keywords was most common in physician and discharge notes (96.6% and 95.6%, respectively). Nursing notes, discharge summaries, and rehabilitation notes had more positive than negative terms (63.9% vs 46.9%; 93.4% vs 84.8%; and 79.9% vs 60.6%, respectively). Consulting provider, nutrition, and respiratory notes had more negative than positive terms on average (85.1% vs 68.7%; 48.1% vs 32.7%; and 80.4% vs 51.3%, respectively). Primary team notes had roughly similar numbers of positive and negative keywords. Please see **Table 1** for patient and note characteristics across both cohorts.

Concordance and Convergence of Sentiment Tools

Concordance and convergence across sentiment classifiers compared with labeled note fragments and each other are shown in **Figure 1, *A*** and ***B***. Benchmark sentiment measures clustered together although exhibited

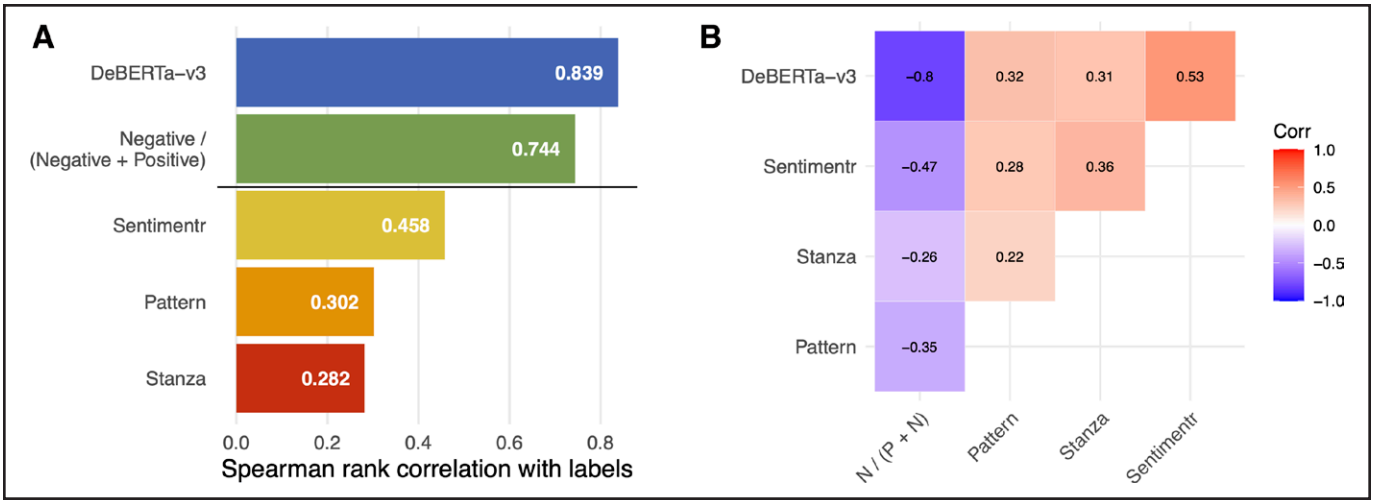

Figure 1. Validation of keywords-based sentiment and other common classifiers on labeled note excerpts for Multiparameter Intelligent Monitoring of Intensive Care III cohort. **A**, Spearman correlation (Corr) of each measure with ordinal, five-value sentiment labels. "Positive" represents the count of positive keywords in a note. Negative represents the count of negative keywords in a note. **B**, A correlogram whereby measures are sorted for display based on similarity from hierarchical clustering. For the keywords sentiment measures, observations without any sentiment keywords were given a neutral value of 0.5. The keywords sentiment score is inversely correlated with the DeBERTa-v3 sentiment score because a higher DeBERTa-v3 score represents a more positive sentiment, whereas for the keywords score, a higher score represents a more negative sentiment. Spearman Corr was used to validate our clinical sentiment measures with the ordinal labels of clinical sentiment recorded on the note excerpts. Each of the sentiment measures is a continuous score. DeBERTa = Decoding-enhanced Bidirectional Encoder Representations from Transformers with disentangled attention.

relatively low levels of correlation. Our novel measures clustered together and showed high levels of correlation, consistent with their similar performance at predicting the sentiment labels. The supervised learning model was most correlated (81%) with the manually labeled reference followed by the keywords-based algorithm (74%). Across nonmedical domain measures, Sentimentr had the best concordance with manual labels when compared with Pattern and Stanza. The correlations of our two sentiment measures were both significantly higher ($p < 0.00001$) than all nonmedical measures.

Interrater Reliability of Sentiment Annotations in a Subset of Notes

Interrater reliability was assessed using Krippendorf's alpha coefficient for a subset of note excerpts across three annotators in a blinded fashion. When using a five-level rating scale for sentiment, agreement was 76% (95% CI, 68–82%). When the rating scale collapsed to three levels (negative, neutral, and positive), agreement remained at 76% (95% CI, 67–83%).

DISCUSSION

We created ICU domain-specific sentiment scores to identify provider negativity and positivity in clinical notes. We found substantial variation in the number and type of sentiment keywords across different provider notes. When compared with a manually labeled reference, our sentiment scores demonstrated high correlation and were substantially more accurate than domain nonspecific algorithms. Further studies and validation can explore to what extent sentiment should be used for prognostication and how sentiment can be used for different clinical use cases. Future studies should explore whether early sentiment could be used to predict future patient/family and provider conflicts, provider moral distress or potentially used as triggers for palliative care interventions, especially if negative sentiment represents "perceptions of excessive care" (14).

This study should be characterized as exploratory. Further validation using mixed methods is required to understand how sentiment tools should and could be used (e.g., as a marker for perceptions of excessive care or for risk prediction). Our keywords-based classifiers can be iterated by incorporating concepts within unstructured notes and sentiment targets, to allow for more fine-grained sentiment. Further iterating annotation guidelines through additional cognitive interviews could improve agreement of labels and thus their validity, in the future. As described in the study

by Weissman et al (5), validation of a sentiment instrument (or any instrument), requires a demonstration of construct validity or how well the instrument measures the phenomenon of interest. Given the ambitious and subjective nature of capturing sentiment appropriately, more rigorous validation, particularly with predictive validation, is required before elevating our methods beyond exploratory.

We build on previous work demonstrating the use of sentiment to predict patient outcomes. Our study was not focused on risk prediction but instead introduces a medical domain-specific lexicon and deep learning approach to the literature. We also performed annotations of MIMIC-III and UCSF note segments that represent important contributions for future sentiment algorithm building. Other studies that used pre-existing sentiment classifiers to predict various outcomes (3, 4, 15, 16) are limited given their reliance on tools that are not specific to clinical medicine. These tools have substantial variability, demonstrate poor agreement, and have variable validity when applied to medical notes (5). Yet some success has been found with model-based measurement of psychiatric risk factor domains in clinical notes, bearing some similarity to this study (17, 18). We validated the most accurate tools from the study by Weissman et al (5) (Stanza, Sentimentr, and Pattern) on labeled note excerpts and our approach had improved convergent validity compared with other methods.

This study has limitations. Our methods could be enriched by incorporating additional labelers and medical ontologies. Currently, we do not include opinion targets in the keywords-based approach. MIMIC-III is from one hospital system and patients are predominantly White, English-speaking, and insured, limiting generalizability. Future validation should also incorporate varying cohort balances of genders, ethnicities, and races to ensure disparities are not reinforced in sentiment models. Although we performed external validation on more recent UCSF data that are also more diverse, additional comparisons of MIMIC-III data and UCSF data are required and future DeBERTa-v3 models should be trained on multiple datasets.

CONCLUSIONS

Our findings suggest that sentiment can be extracted from notes and is more accurate than existing nonmedical sentiment algorithms. Improved sentiment extraction from notes could lead to novel prognostic assessments, deeper understanding of decision frameworks, interventions targeting provider behaviors, and further understanding of provider cognitive biases.

ACKNOWLEDGMENTS

We thank Multiparameter Intelligent Monitoring of Intensive Care and Massachusetts Institute of Technology Laboratory for Computational Physiology for developing and managing the deidentified dataset used for this study. We also thank Timothy A. Heintz for his contributions: conceiving the study design; sentiment conceptualization, labeling guide creation, and manual annotation; and serving as clinical subject matter expert from broad and diverse clinical domains.

1 *Department of Psychiatry, Harvard Medical School, Boston, MA.*

2 *Center for Precision Psychiatry, Massachusetts General Hospital, Boston, MA.*

3 *Department of Anesthesia and Perioperative Care, University of California San Francisco, San Francisco, CA.*

4 *Critical Care and Palliative Medicine, Department of Internal Medicine, University of California San Francisco, San Francisco, CA.*

5 *Department of Surgery, University of California San Francisco, San Francisco, CA.*

6 *Bakar Computational Health Sciences Institute, University of California San Francisco, San Francisco, CA.*

7 *Department of Anesthesiology, Perioperative and Pain Management, Stanford University, Stanford, CA.*

8 *Division of Pulmonary, Allergy, and Critical Care Medicine, Duke University, Durham, NC.*

9 *Division of Geriatrics, Department of Medicine, University of California San Francisco, San Francisco, CA.*

10 *Geriatrics, Palliative, and Extended Care, Veterans Affairs Medical Center, San Francisco, CA.*

11 *Department of Anesthesia, Anesthesia Service, San Francisco VA Health Care System, San Francisco, CA.*

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (http://journals.lww.com/ccejournal).

Drs. Cobert, Chiu, Gologorskaya, and Kennedy conceived the study design, contributed to data collection, and performed data analysis. Drs. Cobert, Chiu, Chapman, Farhan, Ashana, Lee, Smith, Pirracchio, Kennedy, Han, Hodgson, and Lazzereschi helped with the interpretation of the results. Drs. Han, Hodgson, Lazzereschi, and Cobert contributed to sentiment conceptualization, labeling guide creation, and manual annotation. Drs. Cobert, Chiu, Chapman, Farhan, Ashana, Lee, Smith, and Pirracchio also served as clinical subject matter experts from broad and

diverse clinical domains. Data visualizations were made by Drs. Gologorskaya and Kennedy. All authors contributed to the writing and review of the article. Dr. Cobert is the guarantor of the article.

Dr. Cobert was supported by the University of California, San Francisco (UCSF) Initiative for Digital Transformation in Computational Biology and Health, the Hellman Fellows Foundation, and supported by the UCSF Claude D. Pepper Older Americans Independence Center funded by the National Institute of Aging (NIA) (P30 AG044281). Dr. Lee was supported by the NIA K24AG066998 and R01AG057751. Dr. Smith was supported by grants from the NIA (R01AG057751 and K24AG068312). Dr. Pirracchio is supported by the Food and Drug Administration of the U.S. Department of Health and Human Services as part of a financial assistance award Center of Excellence in Regulatory Science and Innovation grant to UCSF and Stanford University, U01FD005978. The remaining authors have not disclosed any potential conflicts of interest.

For information regarding this article, E-mail: Julien.cobert@ucsf.edu

Deidentified data are available at the Multiparameter Intelligent Monitoring of Intensive Care Critical Care Database (physionet.org). University of California, San Francisco validation data are not sharable due to institutional requirements.

Replication code is available at https://github.com/ck37/mimic-clinical-sentiment and sentiment analysis software is available at https://github.com/ck37/clinsent.

REFERENCES

1. Korkontzelos I, Nikfarjam A, Shardlow M, et al: Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. *J Biomed Inform* 2016; 62:148–158
2. Vincent JL, Moreno R, Takala J, et al: The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intensive Care Med* 1996; 22:707–710
3. McCoy TH, Castro VM, Cagan A, et al: Sentiment measured in hospital discharge notes is associated with readmission and mortality risk: An Electronic Health Record Study. *PLoS One* 2015; 10:e0136341
4. Waudby-Smith IER, Tran N, Dubin JA, et al: Sentiment in nursing notes as an indicator of out-of-hospital mortality in intensive care patients. *PLoS One* 2018; 13:e0198687
5. Weissman GE, Ungar LH, Harhay MO, et al: Construct validity of six sentiment analysis methods in the text of encounter notes of patients with critical illness. *J Biomed Inform* 2019; 89:114–121
6. Johnson AE, Pollard TJ, Shen L, et al: MIMIC-III, a freely accessible critical care database. *Sci Data* 2016; 3:160035
7. He P, Gao J, Chen W. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. arXiv:2111.09543 Preprint posted online 2021.
8. Krippendorff K: Chapter 10: Analytical/Representational Techniques. In: *Content Analysis: An Introduction to Its Methodology*. Fourth Edition. Thousand Oaks, California, SAGE, 2018, pp. xiv, 451 pages
9. Krippendorff K: Systematic and random disagreement and the reliability of nominal data. *Commun Methods Meas* 2008; 2:323–338
10. Qi P, Yuhao Z, Zhang Y, et al: Stanza: Apython natural language processing toolkit for many human languages. Association for Computational Linguistics (ACL) System Demonstrations. Stanford. 2020. Available at: https://stanfordnlp.github.io/stanza/. Accessed December 15, 2022
11. Rinker T. *Sentimentr: Calculate Text Polarity Sentiment version 2.7.1. 2021*. Available at: http://github.com/trinker/sentimentr. Accessed December 15, 2022
12. Smedt T, Daelemans W: Pattern for python. *J Mach Learn Res* 2012; 13:2063–2067
13. Myers L, Sirois MJ. Spearman correlation coefficients, differences between. *Encyclopedia of Statistical Sciences*. 2006
14. Benoit DD, Jensen HI, Malmgren J, et al; DISPROPRICUS study group of the Ethics Section of the European Society of Intensive Care Medicine: Outcome in patients perceived as receiving excessive care across different ethical climates: A prospective study in 68 intensive care units in Europe and the USA. *Intensive Care Med* 2018; 44:1039–1049
15. McCoy TH, Castro VM, Roberson AM, et al: Improving prediction of suicide and accidental death after discharge from general hospitals with natural language processing. *JAMA Psychiatry* 2016; 73:1064–1071
16. Ghassemi MM, Al-Hanai T, Raffa JD, et al: How is the doctor feeling? ICU provider sentiment is associated with diagnostic imaging utilization. *Annu Int Conf IEEE Eng Med Biol Soc* 2018; 2018:4058–4064
17. Alvarez-Mellado E, Holderness E, Miller N, et al: Assessing the efficacy of clinical sentiment analysis and topic extraction in psychiatric readmission risk prediction. Proceedings of the 10th International Workshop on Health Text Mining and Information Analysis. 2019, pp. 81–86
18. Holderness E, Cawkwell P, Bolton K, et al: Distinguishing clinical sentiment: The importance of domain adaptation in psychiatric patient health records. Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019, pp. 117–123