

Construct validity of five sentiment analysis methods in the text of encounter notes of patients with critical illness

Gary E. Weissman, MD, MSHP,^{1,2,3} Lyle H. Ungar, PhD,⁴ Michael O. Harhay, PhD,^{2,5}

Katherine R. Courtright, MD, MSHP,^{1,2,3} Scott D. Halpern, MD, PhD^{1,2,3,5}

February 1, 2017

1 Pulmonary, Allergy, and Critical Care Division, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

2 Palliative and Advanced Illness Research (PAIR) Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

3 Leonard Davis Institute of Health Economics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

4 Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA

5 Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Corresponding author: Gary E. Weissman, MD, MSHP, Palliative and Advanced Illness Research (PAIR) Center, Perelman School of Medicine, University of Pennsylvania, 306 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104. Email: gary.weissman@uphs.upenn.edu Telephone: 215-746-2887

Keywords: Critical Care; Forecasting; Natural Language Processing; Electronic Health Records; Attitude of Health Personnel

Manuscript word count: 1,956

Abstract

Sentiment analysis may offer insights into patient outcomes through the subjective expressions made by clinicians in the text of encounter notes. We analyzed the predictive, concurrent, convergent, and content validity of five sentiment methods in a sample of 791,216 multidisciplinary clinical notes among 40,602 hospitalizations associated with an intensive care unit stay. None of these approaches improved early prediction of in-hospital mortality. However, positive sentiment measured by Pattern (OR 0.09, 95% CI 0.04 – 0.17), sentimentr (OR 0.37, 95% CI 0.25 - 0.63), and Opinion (OR 0.25, 95% CI 0.07 - 0.89) were inversely associated with death on the concurrent day after adjustment for demographic characteristics and illness severity. Median daily lexical coverage ranged from 5.2% to 20.5%. While sentiment between all methods was positively correlated, their agreement was weak. Sentiment analysis holds promise for clinical applications, but will require a novel domain-specific method applicable to clinical text.

INTRODUCTION

In the era of widespread adoption of electronic health records (EHRs)[1] and learning health systems[2] there is growing interest in improving utilization of free-text data sources. Among patients with critical illness, the text of clinical notes has been used to identify diagnoses and interventions in the intensive care unit (ICU) and to improve predictions of future health states.[3–6] Clinical text contains important diagnostic information not found in structured data sources within the EHR.[7,8] But clinicians also make subjective assessments[9] and express attitudes about patient outcomes that may be purposefully or unwittingly inscribed in clinical notes. It is unknown if analysis of these subjective attitudes may augment existing yet imperfect mortality predictions,[10] improve communication by highlighting affective dynamics underlying patient-provider and patient-surrogate relationships,[11] or provide a feedback mechanism to clinicians regarding their implicit biases.[12]

The study of attitudes expressed in text is called “sentiment analysis” or “opinion mining.”[13] Dictionaries of terms (i.e. lexica) containing words with associated sentiment vary across different domains.[14] For example, “soft” may imply a different sentiment whether used with respect to sports or toys.[15] The analysis of sentiment in a medical context has been limited to patient opinions expressed in online social media[16,17] and in suicide notes,[18] the association of sentiment in hospital discharge documents with mortality and readmission,[19] and a descriptive comparison between nursing and radiology notes.[20]

Therefore, we sought to determine the construct validity of existing sentiment methods derived from other domains when used for analysis of clinical text among patients with critical illness. Specifically, we examined the predictive, concurrent, content, and convergent validity of these methods to assess different aspects of the sentiment construct.

METHODS

Population and data source

We analyzed the Medical Information Mart for Intensive Care (MIMIC) III database which comprises all hospital admissions requiring ICU care at the Beth Israel Deaconess Medical Center in Boston, MA, between 2001 and 2012.[21] Only hospital admissions with at least one clinical encounter note and a length of stay (LOS) ≤ 30 days were included.

Text sources and sentiment methods

We aggregated clinical encounter notes at the patient-day level for each hospital admission and included notes from physicians, nurses, respiratory therapists, and other clinical specialties. We calculated the proportion of positive sentiment in each collection of daily aggregated notes as

$$\text{proportion of positive sentiment} = \frac{\sum_w P_w}{\sum_w P_w + \sum_w |N_w|}$$

where P_w and N_w are the positive and negative sentiment scores, respectively, for each word w in the daily aggregated text. We calculated separate scores using the Opinion,[22] AFINN,[23] EmoLex,[24] Pattern,[25] and sentimentr[26] methods. All five methods use simple dictionary lookups, and the latter two also account for valence shifters (e.g. “very” and “not”).

Predictive validity

A sentiment measure with predictive validity should be strongly associated with some future outcome.[27] Therefore, for each method we trained a logistic regression model based on a random 75% sample of all hospital admissions to predict in-hospital mortality using data from the first day of the hospitalization. These were compared to a baseline model that did not include any sentiment measures. All performance measures were reported using the remaining 25% hold-out testing sample. The proportion of positive sentiment on the first hospital calendar day was the primary exposure, and the model was adjusted for age, gender, initial ICU type, modified Elixhauser score,[28,29] and initial sequential organ failure assessment (SOFA) score.[30] Model discrimination was assessed with the C-statistic and comparisons made with the DeLong test.[31] Calibration was assessed with the Brier score[32] and comparisons made using a bootstrapped t-test with 1,000 replicates.

Concurrent validity

A sentiment measure with concurrent validity should be strongly associated with an outcome in the same time period.[27] Therefore, we examined the relationship between daily sentiment and the risk of mortality on the same day. We constructed a multivariable mixed-effects logistic regression model using the daily proportion of positive sentiment as the primary, time-varying exposure and daily risk of in-hospital death as the dichotomous outcome. The model was adjusted for age, gender, initial ICU type, and modified Elixhauser score.[28,29] A random effect was included for each hospital admission. A SOFA score (30) ≥ 7 was included as a dichotomous, time-varying exposure to account for daily changes in clinical severity. While daily SOFA scores have not been studied with respect to the daily

risk of death, a time-varying score of ≥ 7 has been associated with an approximately 20% mortality rate in the ICU.[33]

Convergent validity

The object toward which sentiment is directed (e.g. the patient, the prognosis, the tumor) may vary significantly. Each lexicon may also vary by the content of their terms and associated sentiment depending on the domain in which the method was developed and its original purpose.[20] Therefore, each sentiment method may provide a measure of some different aspect of the complex tapestry of sentiment found in clinical encounter notes. To assess the degree to which these five sentiment methods described the same phenomena, i.e. their convergence,[34] we measured their agreement with Cronbach's alpha and calculated pair-wise Pearson correlations (r) at the patient-day level.

Content validity

A useful construct of sentiment in clinical encounter notes should rely on keywords commonly used in the medical domain. Thus, the content validity is the extent to which a sentiment approach is capable of accounting for words and phrases found in these texts.[27] We measured this lexical coverage as the proportion of words in each patient-day's aggregated text sample that was found in the lexicon.

Mixed-effects regression models were built using Stata (version 14.2, StataCorp, College Station, TX). Extraction of sentiment and training of other models were performed with the R language for statistical computing (version 3.3.2). The Pattern sentiment method was implemented using the Python programming language (version 2.7.13). We used a two-sided alpha = 0.05 as a threshold for significance and adjusted all tests for multiple

comparisons (Bonferroni correction). This study was considered exempt by the Institutional Review Board of the University of Pennsylvania.

RESULTS

We analyzed 40,602 unique hospital admissions comprising 176,541 patient-days. The median hospital LOS was 3 days (Interquartile range [IQR] 2 – 5), the median age at admission was 60.8 years (IQR 39.4 – 75.8), and 3,728 (9.2%) patients died in the hospital. Each hospital admission contained a median of 8 (IQR 4 – 22) clinical encounter notes with median 1,480 words (IQR 588 – 5,289). These totaled 791,216 encounter notes containing 228,472,074 words. The distribution of daily sentiment for each method is summarized in Table 1.

The unadjusted temporal trajectories of sentiment stratified by in-hospital mortality are presented in Figure 1. However, the baseline model and all models with the addition of sentiment had C-statistic 0.79 without clinically relevant differences in discrimination ($p = 0.06 - 0.76$ for all comparisons). There were no meaningful differences in calibration with the addition of sentiment to a baseline model despite some comparisons achieving statistical significance (all models had Brier score 0.075; $p = 0.005 - 0.481$).

Sentiment was strongly associated with death when measured on the concurrent day for three of the five sentiment methods (Table 1). Even when adjusting for baseline characteristics and daily severity of illness, the proportion of positive sentiment measured by the Pattern method was inversely associated with the daily risk of death (OR 0.09, 95% CI 0.04 – 0.17).

As a measure of convergence, the Cronbach's alpha for sentiment estimates for each patient-day was 0.65 (95% CI 0.64 – 0.65). All correlations between methods were positive and statistically significant, but most were of a modest magnitude ($p < 0.001$; Figure 2). The median proportion of daily lexical coverage by hospital admission (Figure 3) ranged from 5.2% to 20.5%.

The most common terms from the Opinion lexicon and representative samples of text are presented in Table 2. The associated polarity of these terms included instances with both concordant and discordant meanings in the medical domain.

Table 1: Adjusted odds ratio estimate for the proportion of daily positive sentiment for each sentiment method based on mixed-effects logistic regression model to assess concurrent validity; and distribution of daily sentiment.

Sentiment method	Odds Ratio (95% CI)	p value	Median (IQR)
Opinion	0.25 (0.07 - 0.89)	0.033	0.50 (0.39 - 0.62)
EmoLex	1.89 (0.41 - 8.69)	0.412	0.60 (0.53 - 0.69)
AFINN	0.65 (0.23 - 1.87)	0.428	0.56 (0.44 - 0.68)
Pattern	0.09 (0.04 - 0.17)	<0.001	0.63 (0.53 - 0.74)
sentimentr	0.37 (0.25 - 0.63)	<0.001	0.71 (0.35 - 0.96)

Table 2: The most common terms from the Opinion lexicon found in the clinical text sample and their polarity.

Term	Polarity	Appearances (n)	Representative context
pain	negative	658,808	'Pt reports back pain', 'Continue to monitor pain'
patient	positive	588,213	'Encouraged patient to take his medicine', 'I saw and examined the patient'
stable	positive	411,028	'stable frontal infarct', 'remains hemodynamically stable'
right	positive	383,482	'only moving right arm', 'elevation of the right hemidiaphragm'
clear	positive	368,261	'w/o clear evidence of infiltrates', 'Nutrition: clear liquids, advance diet'
well	positive	365,899	'get radiation as well as this decision', 'sating well, no resp distress'
support	positive	325,814	's/p arrest requiring ventilatory support', 'Emotional support given to patient & family'
soft	positive	290,426	'abdomen soft slightly distended', 'possibility of soft tissue pus collection'
failure	negative	268,838	'PNA with hypercarbic respiratory failure', 'R-sided heart failure leading to hepatopedal flow'
bs	negative	259,638	'PULM: decreased bs on left', 'soft distended with hypoactive bs'

DISCUSSION

In our assessment of multidisciplinary encounter notes of patients hospitalized with critical illness, existing sentiment approaches demonstrated little evidence of most types of validity and exhibited high variability between methods. These results argue against the use of available sentiment methods to inform bedside clinical decisions, but also highlight opportunities to make sentiment methods more clinically applicable.

Many of the covered terms in this analysis had discordant polarity when applied in the medical domain. For example, the term "right" in medical parlance most often expresses anatomic laterality (e.g. "right ventricle"), thus should carry a neutral rather than positive

sentiment with respect to prognosis. Similarly, the term “bs” is a shorthand abbreviation with multiple senses and may indicate “breath sounds”, “bowel sounds”, or “blood sugar” depending on the context. It should carry a neutral sense for all of these medical uses, but carried a negative polarity in the Opinion lexicon, where it may have been used originally to indicate a vulgar term in the online consumer reviews of electronics products.

The strong concurrent validity after adjustment for clinical and demographic characteristics suggests a temporal sensitivity of the sentiment to the patient’s clinical condition on the same the day. This finding was true even with adjustment for changes in severity of illness on each day, highlighting the presence of additional information encoded in free-text data not found in structured data sources such as laboratory values and vital signs. The models with the strongest effect sizes in this analysis (Pattern and sentimentr) were the only two that accounted for valence shifters. Nuances in expression of clinician sentiment are likely better captured by these approaches.

However, the addition of sentiment measures to a baseline prediction model resulted in no meaningful improvements to its discrimination or calibration. This lack of predictive validity suggests the temporal correlation of sentiment to mortality risk, while strongly concurrent, does not project forward in time.

Although all sentiment estimates were positively correlated with each other, their overall agreement was poor. The Opinion, AFINN, and EmoLex approaches were more highly correlated with each other ($r=0.58 - 0.68$), while the Pattern and sentimentr approaches were weakly correlated ($r=0.33$). These findings suggest a weak convergence towards two distinct constructs. More work is needed to distinguish between the sources, objects, and

aspects of sentiment in clinical text. Additionally, the sentiment associated with objective medical terms (e.g. “cardiac arrest”) is distinct from the expression of a private state[35] of a clinician (e.g. “Mr. Jones is an unpleasant and uncooperative 65 year old man”). Each of these has separate analytic and ethical implications for use in clinical predictive modeling that have yet to be explored.

Finally, the content of sentiment lexica demonstrated coverage of medical terms that was higher than in previous analyses of medical text, but low compared to sentiment use in other domains. For example, Denecke et al. found coverage of 5% – 11% in radiology reports, 6% – 11% in discharge summaries, and 8% – 12% in nursing notes, depending on the sentiment lexicon.[20] Coverage for the widely used SemEval Dataset range from 8% to 89% percent using commonly available sentiment lexica.[36]

The results of this study should be interpreted in the context of some limitations. First, the study analyzed data from a single academic center and may not generalize to the documentation style or patient population in other settings. Second, our analysis did not distinguish between the emotional valence of objective and subjective terms which conflates their practical use in clinical risk prediction.

In conclusion, this is the first study to examine sentiment in a set of multidisciplinary clinical encounter notes of critically ill patients and to assess the validity of these measures. Sentiment is strongly and concurrently associated with the risk of death even after adjustment for baseline characteristics and severity of illness. Our findings highlight the need for a domain-specific sentiment lexicon that has good coverage of medical terminology. Future work should seek to validate these findings in a broader population,

better distinguish sources and objects of sentiment, and address potential ethical challenges of using sentiment to guide clinical care.

FUNDING

GEW received support from NIH T32-HL098054.

REFERENCES

1. Henry J, Pylypchuk Y, Searcy T, et al. Adoption of electronic health record systems among us non-federal acute care hospitals: 2008-2015. *The Office of National Coordinator for Health Information Technology*. 2016.
2. Krumholz HM. Big data and new knowledge in medicine: The thinking, training, and tools needed for a learning health system. *Health Aff (Millwood)* 2014;33(7):1163–70.
3. Weissman GE, Harhay MO, Lugo RM, et al. Natural language processing to assess documentation of features of critical illness in discharge documents of acute respiratory distress syndrome survivors. *Ann Am Thorac Soc* 2016;13(9):1538–45.
4. Abhyankar S, Demner-Fushman D, Callaghan FM, et al. Combining structured and unstructured data to identify a cohort of icu patients who received dialysis. *J Am Med Inform Assoc* 2014;801–7.
5. Marafino BJ, Boscardin WJ, Dudley RA. Efficient and sparse feature selection for biomedical text classification via the elastic net: Application to ICU risk stratification from nursing notes. *J Biomed Inform* 2015;54:114–20.
6. Lehman LW, Saeed M, Long W, et al. Risk stratification of ICU patients using topic models inferred from unstructured progress notes. *AMIA Annu Symp Proc* 2012;2012:505–11.
7. Zhou L, Baughman AW, Lei VJ, et al. Identifying patients with depression using free-text clinical documents. *Stud Health Technol Inform* 2015;216:629–33.

8. Navathe AS, Zhong F, Lei VJ, et al. Hospital Readmission and Social Risk Factors Identified from Physician Notes. *Health Serv Res* 2017 Epub.
9. Detsky ME, Harhay MO, Bayard DF, et al. Discriminative Accuracy of Physician and Nurse Predictions for Survival and Functional Outcomes 6 Months After an ICU Admission. *JAMA* 2017;317:2187-95.
10. Sinuff T, Adhikari NK, Cook DJ, et al. Mortality predictions in the intensive care unit: Comparing physicians with scoring systems. *Crit Care Med* 2006;34:878-85.
11. Jacobowski NL, Girard TD, Mulder JA, et al. Communication in critical care: Family rounds in the intensive care unit. *Am J Crit Care* 2010;19:421-30.
12. Chapman EN, Kaatz A, Carnes M. Physicians and implicit bias: How doctors may unwittingly perpetuate health care disparities. *J Gen Intern Med*. 2013;28:1504-10.
13. Liu B. Sentiment analysis and subjectivity. *Handbook of Natural Language Processing*, Second Edition, Eds. N. Indurkha, F.J. Damerau. 2010;2:627-66.
14. Andreevskaia A, Bergler S. When Specialists and Generalists Work Together: Overcoming Domain Dependence in Sentiment Tagging. In: *Proceedings of ACL-08: HLT*. Columbus, Ohio: Association for Computational Linguistics; 2008;290-8.
15. Hamilton WL, Clark K, Leskovec J, Jurafsky D. Inducing domain-specific sentiment lexicons from unlabeled corpora. *Proc Conf Empir Methods Nat Lang Proess*. 2016;2016:595-605.

16. 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(Supplement C):148–58.
17. Ji X, Chun SA, Wei Z, et al. Twitter sentiment classification for measuring public health concerns. *Social Network Analysis and Mining* 2015;5(1):13.
18. Pestian JP, Matykiewicz P, Linn-Gust M, et al. Sentiment analysis of suicide notes: A shared task. *Biomed Inform Insights* 2012;5(Suppl 1):3–16.
19. 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 10:1–10.
20. Denecke K, Deng Y. Sentiment analysis in medical settings: New opportunities and challenges. *Artif Intell Med* 2015;64:17–27.
21. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016;3:160035.
22. Hu M, Liu B. Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, Seattle, WA. ACM, New York; 2004:168–77.
23. Nielsen FÅ. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In: *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages* 2011;93-8.

24. Mohammad SM, Turney PD. Crowdsourcing a word-emotion association lexicon.

Computational Intelligence 2013;29:436–65.

25. Smedt TD, Daelemans W. Pattern for Python. *Journal of Machine Learning Research*

2012;13:2063–7.

26. Rinker TW, Spinu V. sentimentr: Calculate text polarity sentiment.

doi:10.5281/zenodo.222103. Available from: <http://github.com/trinker/sentimentr>

Accessed December, 2017.

27. Cronbach L, Meehl P. Construct validity in psychological tests. *Psychol Bull* 1955;4:281–

302.

28. Elixhauser A, Steiner C, Harris DR, et al. Comorbidity measures for use with

administrative data. *Med Care* 1998;36:8–27.

29. Walraven C van, Austin PC, Jennings A, et al. A modification of the Elixhauser

comorbidity measures into a point system for hospital death using administrative data.

Med Care 2009;47:626–33.

30. 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 Jul;22:707–10.

31. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more

correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*

1988;837–45.

32. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: A framework for some traditional and novel measures. *Epidemiology* 2010;21:128.
33. Ferreira FL, Bota DP, Bross A, et al. Serial evaluation of the SOFA score to predict outcome in critically ill patients. *JAMA* 2001;10;286:1754–8.
34. Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull* 1959;56:81–105.
35. Wiebe J, Wilson T, Cardie C. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation* 2005;39:165–210.
36. Gatti L, Guerini M, Turchi M. SentiWords: Deriving a high precision and high coverage lexicon for sentiment analysis. *IEEE Transactions on Affective Computing* 2016;7:409–21.

FIGURE LEGENDS

Figure 1: Unadjusted trajectories of proportion of positive sentiment by sentiment method using a generalized additive model smoother with 95% confidence intervals

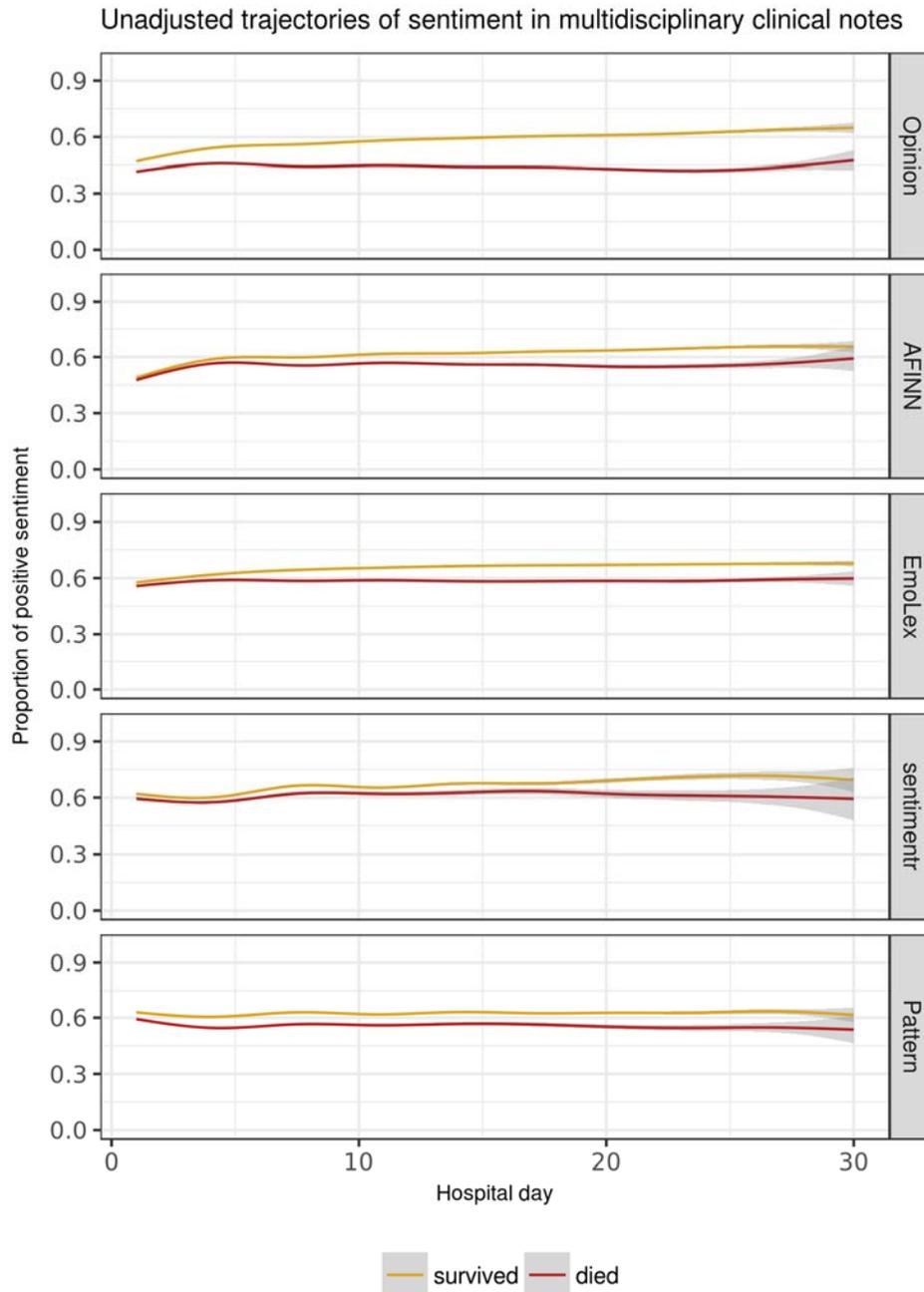


Figure 2: Pair-wise Pearson correlations between methods of calculated sentiment by patient-day. All estimates have $p < 0.001$ after adjustment for multiple comparisons.

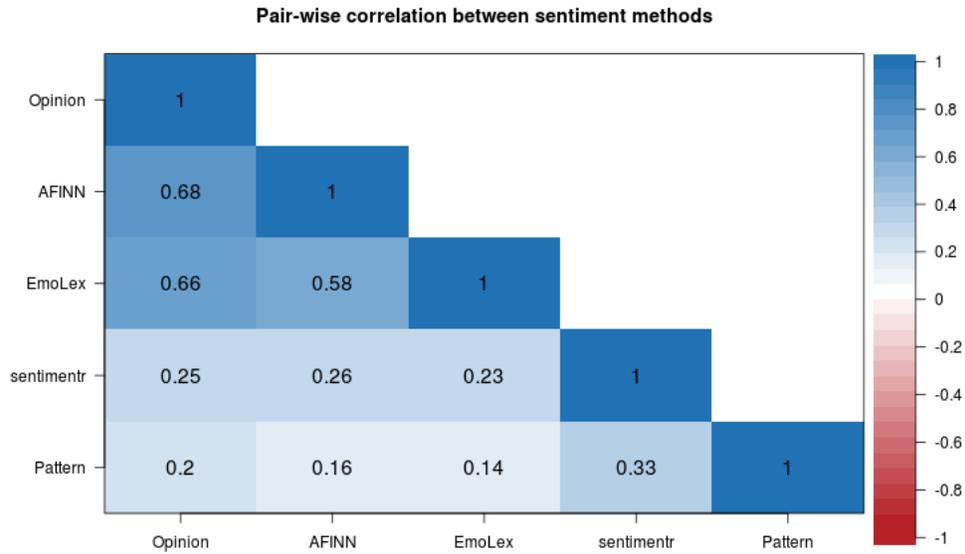


Figure 3: The distribution of median proportion of covered words for each hospital admission by sentiment method.

