

A Visualization of Evolving Clinical Sentiment Using Vector Representations of Clinical Notes

Mohammad M. Ghassemi¹, Roger G. Mark¹, Shamim Nemati²

Abstract

Our objective in this paper was to visualize the evolution of clinical language and sentiment with respect to several common population-level categories including: time in the hospital, age, mortality, gender and race. Our analysis utilized seven years of unstructured free text notes from the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) database. The text data was partitioned by category and used to generate several high dimensional vector space representations. We generated visualizations of the vector spaces using Distributed Stochastic Neighbor Embedding (tSNE) and Principal Component Analysis (PCA). We also investigated representative words from clusters in the vector space. Lastly, we inferred the general sentiment of the clinical notes toward each parameter by gauging the average distance between positive and negative keywords and all other terms in the space. We found intriguing differences in the sentiment of clinical notes over time, outcome, and demographic features. We noted a decrease in the homogeneity and complexity of clusters over time for patients with poor outcomes. We also found greater positive sentiment for females, unmarried patients, and patients of African ethnicity.

1. Introduction

Electronic Medical Record (EMR) systems are home to an increasingly large volume of structured and unstructured data. Investigation of these records reveal variance in care practice for patients with similar structured data profiles. Presumably, these differences in care practice arise because clinicians are considering features that are not captured by the structured data [1]. Indeed, the judgment of care providers is driven by more comprehensive observations of the patient, and this judgment may be reflected in

This work was supported in part by Salerno Foundation. The content of this article is solely the responsibility of the authors.

¹ M.M. Ghassemi and R.G. Mark are in the department of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA. Correspondence email: ghassemi@mit.edu

² S. Nemati is in the department of Biomedical Informatics at Emory University, Atlanta, GA 30322, USA. Correspondence email: shamim@seas.harvard.edu

the structure and sentiment of their written patient notes.

In this paper, We utilize the *word2vec* tool to investigate the evolution of clinical sentiment, language use, and complexity in a large clinical database. Several studies have already applied the *word2vec* tool to medical notes for a variety of purposes including: the disambiguation of clinical abbreviations [2], the identification of adverse drug-events [3] information retrieval, and relationship mining [4] [5]. To our knowledge, however, this is the first instance of application of *word2vec* to clinical sentiment analysis.

2. Methods

We extracted 1,237,977 medical notes from 38,390 unique Intensive Care Unit (ICU) stays from the publicly available Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) database [6]. The notes in MIMIC are of three types: radiology reports, discharge summaries and nursing notes.

We pre-processed the notes by removing all numbers, stop-words, punctuation and white-space characters (new line, tabs, etc.) from the extracted text. We also cast all words into lower-case and removed any single-character words from the text ('a' and 'I', for instance) or words that appeared less than five times. Lastly, we replaced all positive sentiment terms (such as 'good', 'happy', 'better', etc.) in the text with the single term 'POSITIVE', and all negative sentiment terms with the single term 'NEGATIVE'. The collection of negative and positive terms were based on a 2005 paper by Liu [7].

Following pre-processing of the notes, we separated the text into groups according to patient age, ethnicity, gender, marital status, outcome and the hospital stay day. In Table 1 we list the extracted note categories and the corresponding total word-count for each.

After grouping notes we applied the *word2vec* tool on the text collection to each of the categories in Table 1. The first five days of clinical notes were further separated by patient outcome and also analyzed. *Word2vec* is a tool that analyzes a corpus of text and generates vector representations of the words in the text using the skip-gram and continuous bag-of-words approaches [8]. The tool has several parameters that affect the nature of the embedding. For our analysis, we used a continuous bag-of-words approach

Word Count (Millions)	Note Category
4.44	Age <25
16.66	Age: 25 - 49
44.25	Age: 50 - 75
28.60	Age >75
21.36	Deceased
105.70	Survived
6.65	Day 1
11.32	Day 2
9.26	Day 3
8.69	Day 4
8.18	Day 5
45.08	Married
22.27	Single
32.69	Female
77.50	Male
3.95	Asian
92.87	White
13.14	African

Table 1. *Extracted note categories used for analysis (right) and corresponding word counts within those categories (left).*

with a neighborhood of five words, an embedding size of 100 dimensions and 15 training iterations of the algorithm.

For visualizations, we reduced the dimensionality of the word vector spaces using Principal Component Analysis (PCA) and Distributed Stochastic Neighbor Embedding (tSNE) [9]. PCA was used to visualize the evolution of the language, while tSNE was used for visualizing distinctive word clusters and the evolution of language complexity.

We applied k-means clustering on the reduced vector spaces produced by tSNE. The number of clusters, k , that corresponded to the optimal silhouette value [10] was selected. We tested values of k ranging from 2 through 25 and interpreted the identified value as an indicator of language complexity.

In addition to the generation of visuals, we performed analysis to understand the difference in sentiment across time periods, patient category, and outcome. We defined a simple sentiment score, s_s , as the ratio of positive to negative sentiments in the text:

$$s_s = \left(\frac{s_p}{s_n} - 1 \right) * 100, \quad (1)$$

where s_p and s_n are computed as the average cosine similarity between the 'POSITIVE' and 'NEGATIVE' word vectors respectively, and all other terms in the in the vector space. The cosine similarity between two vectors x and y is defined as:

$$\cos(x,y) = \frac{x \cdot y}{\|x\| \|y\|} \quad (2)$$

Equation (1) yields a sentiment score which is greater than zero if notes are optimistic, and less than zero if notes are

pessimistic. The sentiment score was computed and compared for all note categories. We compared our simple sentiment score against an alternative measure of sentiment which utilized a scaled ratio of word-counts to gauge sentiment: $100 * (n_{positive}/n_{negative} - 1)$. This simple score provided a baseline that allowed us to gauge if there was any additional value provided by using the vector representations.

3. Results

In Fig. 1 we show a PCA-based visualization of the word vector space partitioned over time, and patient outcomes. The temperature in the illustration represents the density of words in the area (with red corresponding to greater density).

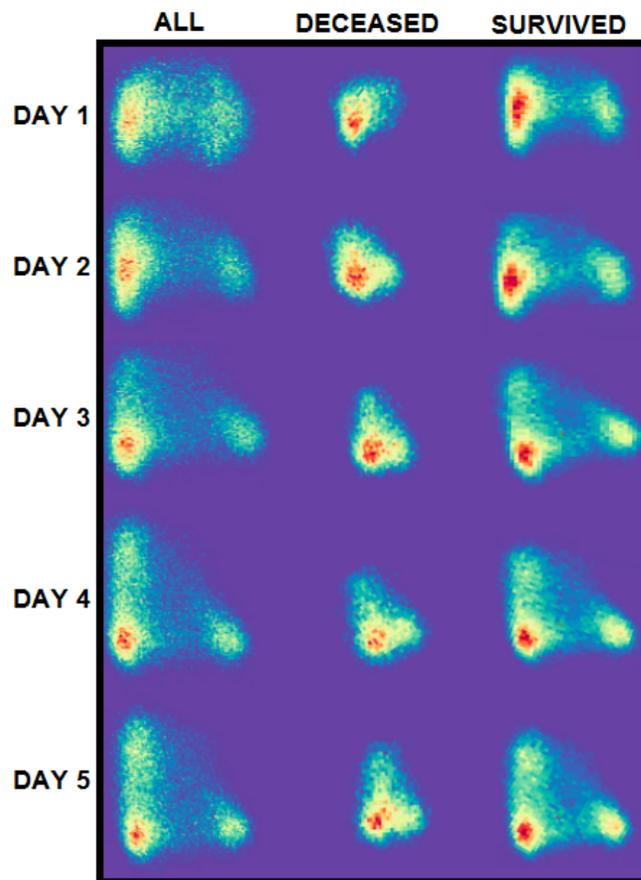


Figure 1. *The note data over the first five days of hospital stay for all patients and partitioned by outcome. We observe a evolution of the language structure over time.*

In the first column, which illustrates the vector space computed for all patients, we observe a gradual evolution of the word vector space from bi- to tri-modal. In the second column, which represents the space of patients with poor outcomes, we observe little evolution in the language

from day to day compared to those patients which survived (third column). We found that the first two principal components represented between 7-10% of the data variance across our note categories. This makes PCA sub-optimal from an overall representational standpoint. However, the simplicity of the PCA-depiction allows for a clear illustration of language evolution in the word vector space.

In Fig. 2 we illustrate the sentiment score (defined above) for patients partitioned by outcome and day of hospital stay. For both outcome classes, we observe a optimistic initial sentiment on the first day. Afterwards, the sentiment of the notes is consistently optimistic for patients who survived, and pessimistic for those who expired. This indicates that after only 24 hours of hospital stay, the outcome of the patients is evidenced by linguistic features in the notes. The results in Fig. 2 provide a level of confidence that our sentiment score is in fact sensible.

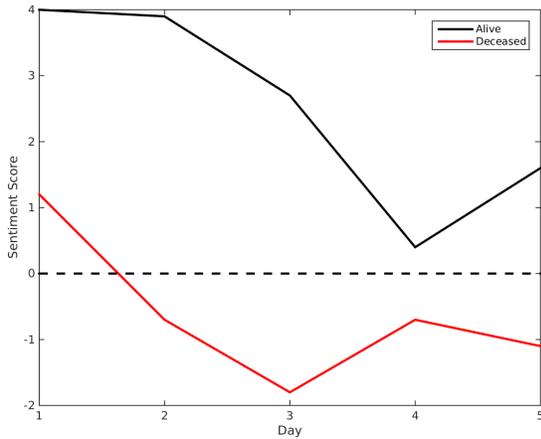


Figure 2. The sentiment score of the notes, partitioned by outcome class, and hospital stay day.

In Table 2, we show the sentiment analysis for the other note categories using both the vector-based sentiment measure we defined, and the key-word baseline. The negative trend between our note sentiment score and patient age follows intuition and provides additional confidence in the measure.

Interestingly, our approach reports greater positive sentiment for females, compared to males, and greater pessimism for married, as opposed to single patients. Furthermore, we observe pessimism for Asian patients as compared to optimism for patients of Caucasian and African ethnicity.

In Fig. 3 we illustrate the tSNE representation of the vector space for patients who did not survive their hospital stay. For each of the days we have listed a collection of words that were closest to the center of the vector space ("Center Words") and that held enhanced representative

Sentiment Score	RWC	Note Category
-1.98	-6.84	Deceased
0.57	63.3	Survived
0.63	72.16	Age <25
-0.31	-2.97	Age: 25 - 49
-1.36	-0.42	Age: 50 - 75
-1.82	1.65	Age >75
-0.28	2.16	Married
-0.08	5.25	Single
0.90	54.54	Female
0.39	47.59	Male
-1.07	115.51	Asian
0.14	41.06	White
0.45	62.99	African

Table 2. A comparison of our sentiment score and an alternative score using the ratio of positive to negative terms (RWC) for each of the note categories.

power for the notes from that day. In the first day's notes, we observe several sensible word clusters that emerge from the space discussing a range of topics from family members to brain areas, to skin conditions. Interestingly, the individual clusters on the fifth day are less homogeneous, containing words across a wider variety of topics. The k-means clustering of the vector spaces yielded an optimal value of k that decreased over successive hospital stay days from 6 on the first though third day, to 4 on the the fourth day, to 3 on the fifth day. We interpret this result as a simplification of note themes later in the stay.

4. Discussion

Our results highlight two interesting and intuitive facts about clinical notes: 1) the sentiment of clinical notes evolve over time, patient condition, and patient background. 2) The structural complexity of clinical notes for patients who ultimately do not survive, decreases over time.

While the sentiment differences for outcome and age are expected, other features such as gender, marital status and race are less clear and provide opportunities for further thought and investigation. The pessimistic sentiment towards the marital group may be the result of this group's increased age (and thus, additional complications). On the other hand, increased positive sentiment for females may reflect gender differences in disease types. The differences between racial groups are interesting, but will require further analysis with adjustment for potential confounding effects before any conclusions may be drawn.

The reduction in complexity of the tSNE projected vector space may be due to redundant or boilerplate text, but this will require further investigation. It is important to note that the evolution of the language is not simply an artifact of the number of words available to the *word2vec*

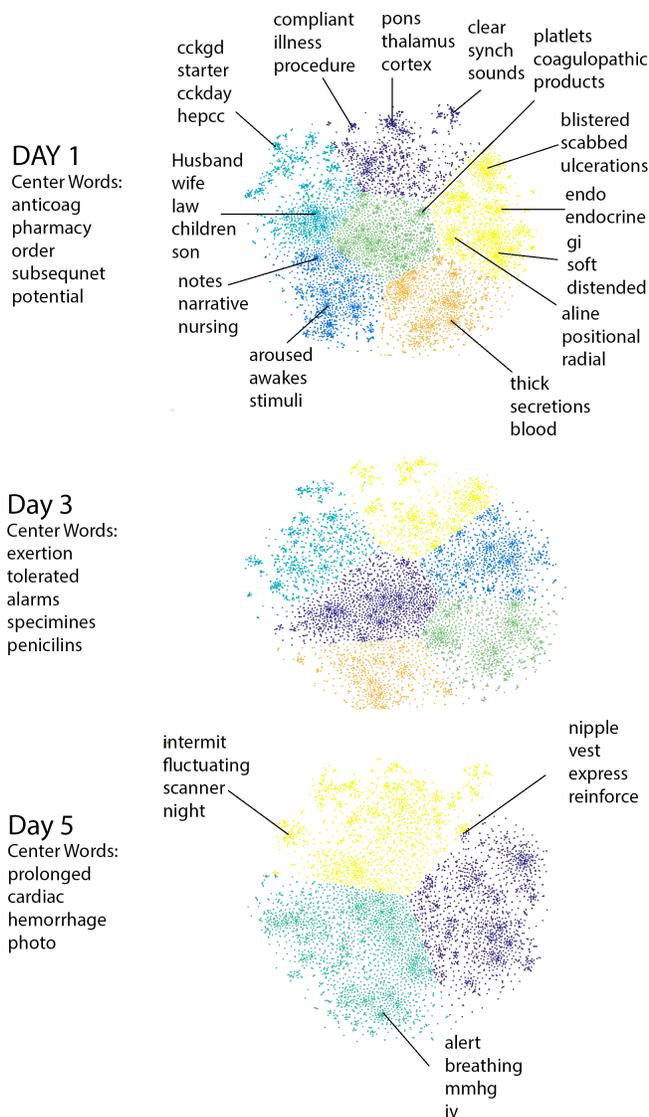


Figure 3. *tSNE visualizations for patients who did not survive their hospital stay. On the left we report both the day, and the words that were closest to the center of the vector space. Colors represent the boundaries of the k-means clusters identified using silhouette optimization.*

algorithm. In fact, in Table 1 we observe that the greatest volume of words is actually available on the second day of patient stay.

In Fig. 2 we observed a noteworthy sentiment inflection point on the fourth day of hospital stay, where notes were most pessimistic about the survivors, and least pessimistic about the non-survivors (excluding the first day). This result calls for further investigation across a longer time frame.

The results of this paper present a novel demonstration of how vector space representations may be used to in-

fer sentiment in the context of clinical notes. Follow-up work will benefit from a more comprehensive comparison of sentiment analysis techniques [11]. While the methods in this paper were applied to clinical notes from the ICU, it is important to emphasize their general applicability to other areas of clinical text analysis, such as monitoring of mental health status of patients [12].

References

- [1] M. M. Ghassemi, S. E. Richter, I. M. Eche, T. W. Chen, J. Danziger, and L. A. Celi, "A data-driven approach to optimized medication dosing: a focus on heparin," *Intensive care medicine*, vol. 40, no. 9, pp. 1332–1339, 2014.
- [2] Y. Wu, J. Xu, Y. Zhang, and H. Xu, "Clinical abbreviation disambiguation using neural word embeddings," *ACL-IJCNLP 2015*, p. 171, 2015.
- [3] A. Henriksson, M. Kvist, H. Dalianis, and M. Duneld, "Identifying adverse drug event information in clinical notes with distributional semantic representations of context," *Journal of biomedical informatics*, 2015.
- [4] J. A. Minarro-Giménez, O. Marín-Alonso, and M. Samwald, "Exploring the application of deep learning techniques on medical text corpora," *Studies in health technology and informatics*, vol. 205, pp. 584–588, 2013.
- [5] J. A. Miñarro-Giménez, O. Marín-Alonso, and M. Samwald, "Applying deep learning techniques on medical corpora from the world wide web: a prototypical system and evaluation," *arXiv preprint arXiv:1502.03682*, 2015.
- [6] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L. H. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark, "Multiparameter intelligent monitoring in intensive care (MIMIC II): a public-access intensive care unit database." *Crit Care Med*, vol. 39, no. 5, pp. 952–960, May 2011.
- [7] B. Liu, M. Hu, and J. Cheng, "Opinion observer: analyzing and comparing opinions on the web," in *Proceedings of the 14th international conference on World Wide Web*. ACM, 2005, pp. 342–351.
- [8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
- [9] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 2579–2605, p. 85, 2008.
- [10] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009, vol. 344.
- [11] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, vol. 1631. Citeseer, 2013, p. 1642.
- [12] G. C. M. D. C. Harman, "Quantifying mental health signals in twitter," *ACL 2014*, p. 51, 2014.