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

Mohammad M. Ghassemi, MIT
Roger G. Mark, MIT
Shamim Nemati, Emory University
• Motivation
• Hypothesis

INTRODUCTION
Clinical Notes are Shaped by Clinicians’ Sentiment

Structured Data
1. Waveforms
2. Lab Measures
3. Demographics

Unstructured Data
1. Physical Attributes
2. Psychosocial conditions

Sentiment

Database

Notes
Hypothesis

The judgment of care providers is driven by comprehensive observations of the patient, and this judgment may be reflected in the structural complexity and sentiment of their written patient notes.

We investigate our hypothesis by analyzing the evolution of the sentiment and language use over time and patient category.
The judgment of care providers is driven by comprehensive observations of the patient, and this judgment may be reflected in the structural complexity and sentiment of their written patient notes. We investigate our hypothesis by analyzing the evolution of sentiment and language use over time and patient category.
 METHODS

- Methods & Analysis
- Word2Vec Tool
- Illustration
The Word2vec Tool

- Word2Vec describes a class of neural network models that, given an unlabeled training corpus, produce a vector for each word in the corpus that encodes its semantic information.

- Semantic similarity is measured by cosine distance.
Individual Words in the Vector Space
Similar Words Cluster Together

Cardiac
Similar Words Cluster Together

Aorta
Similar Words Cluster Together

Valve
Similar Words Cluster Together
Clusters Form Meaningful Groups
Clusters Form Meaningful Groups
Clusters Form Meaningful Groups
Clusters Form Meaningful Groups

Heart

Thalamus
Clusters Form Meaningful Groups

Heart

Grey matter
Clusters Form Meaningful Groups
Ungrouped Words

Heart

Electrode

Brain
‘Electrode’ is Related to both Groups

Heart

Brain

Electrode
Centrality Indicates Importance for the Category
What about the Sentiment Terms?
Positive: Slightly More Related to the Heart
Negative: Strongly Related to the Brain
Conclusions Drawn from Illustrative Example

- There is greater positive sentiment for the **patient** category than negative sentiment

- There is greater negative sentiment for the **brain** group than the **heart** group
**Sentiment Score**

\[ s_s = \left( \frac{S_p}{S_n} - 1 \right) \times 100 \]

- Where \( S_p \) and \( S_n \) are the average cosine similarity between the ‘positive’ and ‘negative’ terms, and all other terms in the space.

\[ \cos(x, y) = \frac{x \cdot y}{\|x\| \|y\|} \]

- We can compare this against another metric that uses word counts:

\[ 100 \times \left( \frac{n_{positive}}{n_{negative}} - 1 \right) \]
Complexity Score

- Track the evolution of an optimal k, in the k-means algorithm
- Where optimality is determined by the Silhouette value
Visualizations

- Distributed Stochastic Neighbor Embedding (tSNE)
  - visualize distinctive word clusters
  - the evolution of language structure

- Principal Component Analysis (PCA)
  - characterize evolution over time
STUDY RESULTS

- Findings
- Conclusions
DAY 1
Center Words:
- anticoag
- pharmacy
- order
- subsequent
- potential

- compliant illness procedure
- pons thalamus cortex clear synch sounds
- platlets coagulopathic products
- blistered scabbed ulcerations
- endo endocrine
gi soft distended aline
positional radial
- thick secretions blood
- aroused awakes stimuli
- notes narrative nursing
- husband wife law children son
DAY 1
Center Words:
- anticoag
- pharmacy
- order
- subsequent
- potential

- cckgd
- starter
- cckday
- hepcc

- compliant
- illness
- procedure

- pons
- thalamus
- cortex

- clear
- synch
- sounds

- platlets
- coagulopathic
- products

- blistered
- scabbed
- ulcerations

- endo
- endocrine

- gi
- soft
- distended

- aline
- positional
- radial

- thick
- secretions
- blood
DAY 1
Center Words:
anticoag
pharmacy
order
subsequent
potential

- cckgd
- starter
- cckday
- hepcc
- compliant
- illness
- procedure
- clear
- synch
- sounds
- platlets
- coagulopathic
- products
- blistered
- scabbed
- ulcerations
- endo
- endocrine
- gi
- soft
- distended
- aline
- positional
- radial
- thick
- secretions
- blood

- Husband
- wife
- law
- children
- son
- notes
- narrative
- nursing

- aroused
- awakes
- stimuli
DAY 1
Center Words:
- anticoag
- pharmacy
- order
- subsequent
- potential

Husband
wife
law
children
son
notes
narrative
nursing
aroused
awakes
stimuli
platemlets
coagulopathic
products
blistered
scabbed
ulcerations
endo
endocrine
gi
soft
distended
aline
positional
radial
thick
secretions
blood
compliant
illness
procedure
pons
thalamus
cortex
clear
synch
sounds
Day 3
Center Words:
exertion
tolerated
alarms
specimens
penicillins
Day 5
Center Words:
prolonged
cardiac
hemorrhage
photo

intermit
fluctuating
scanner
night

nipple
vest
express
reinforce

alert
breathing
mmhg
iv
Sentiment Score also Evolves over Time and Outcome
Sentiment over Categories

- Sentiment differences across categories
- The vector-based score is more aligned with expected results than the score that uses ratio of word counts (RWC)

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

• Two main findings:
  – the sentiment of clinical notes evolve over time, patient condition, and patient background
  – The structure / complexity of clinical notes also evolves

• Results are preliminary, and will require further investigation to reach firm conclusions
Thank you

- Contact me for questions, or to collaborate!

ghassemi@mit.edu
Example: Similar Patients, Different Treatment Decisions

<table>
<thead>
<tr>
<th></th>
<th>Patient 1</th>
<th>Patient 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Structured Data</strong></td>
<td>Age = 35</td>
<td>Age = 35</td>
</tr>
<tr>
<td><strong>Treatment Decision</strong></td>
<td>Intubate</td>
<td>Don’t Intubate</td>
</tr>
</tbody>
</table>

*Source: Multiparameter Intelligent Monitoring in Critical Care Database*
# Example: Similar Patients, Different Treatment Decisions

<table>
<thead>
<tr>
<th></th>
<th>Patient 1</th>
<th>Patient 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Structured Data</strong></td>
<td>Age = 35</td>
<td>Age = 35</td>
</tr>
<tr>
<td><strong>Treatment Decision</strong></td>
<td>Intubate</td>
<td>Don’t Intubate</td>
</tr>
<tr>
<td><strong>Unstructured Data</strong></td>
<td>“blah blah blah”</td>
<td>“blah blah blah?”</td>
</tr>
</tbody>
</table>

*Source: Multiparameter Intelligent Monitoring in Critical Care Database*
CBoW

- C – word window size
- V – vocabulary size
- Y – output word
- X – input context words
**tf-idf**  (term frequency * inverse document frequency)

- Find “the red fox”
  - Remove all documents without the words
  - Count the number of times the words show up in each document (term frequency)
  - Because ‘the’ is common, we may over emphasize it and need to eliminate it’s effects
  - We diminish the weight of such terms using the inverse of their frequency in the set of documents (inverse document frequency).
Conclusions

- There is decreasing complexity of the language for patients who do not survive
  - This is not simply an artifact of the number of words
Silhouette analysis for KMeans clustering on sample data with n_clusters = 2

The silhouette plot for the various clusters.

The visualization of the clustered data.
Silhouette analysis for KMeans clustering on sample data with n_clusters = 3

The silhouette plot for the various clusters.

The visualization of the clustered data.

Cluster label

Feature space for the 1st feature

Feature space for the 2nd feature

The silhouette coefficient values

Silhouette
Silhouette analysis for KMeans clustering on sample data with n_clusters = 4

The silhouette plot for the various clusters.

The visualization of the clustered data.
Silhouette analysis for KMeans clustering on sample data with $n_{\text{clusters}} = 5$

The silhouette plot for the various clusters.

The visualization of the clustered data.