Identifying Language Biomarkers for Depression and Anxiety Among Trauma Patients

Nicole Mathews

MPS Candidate, Biomedical and Health Informatics

November 11, 2020
# Table of Contents

I. Internship Overview and Learning Objectives  
II. Background and AURORA Study  
III. My Role, Results, and Data Analysis  
IV. Twitter Study  
V. Challenges, Significance, and Lessons Learned  
VI. References
Internship Overview and Learning Objectives
Dr. Lisa Vizer, Assistant Research Professor in the School of Medicine at UNC-Chapel Hill

AURORA Study
Learning Objectives

- **PROCESS AND COMPARE LARGE DATA SETS**
  - Python
  - Excel

- **INTEGRATION OF INFORMATICS AND RESEARCH**
  - Data
  - Meaningful Knowledge

- **RESEARCH GRANT PROCESS**
  - Privacy restrictions

- **OVERCOME CHALLENGES**
  - Calculations
  - Writing code
Background & AURORA Study
Adverse posttraumatic neuropsychiatric sequelae (APNS) include posttraumatic stress, postconcussion syndrome, depression, and regional or widespread pain.

Trauma experiences include car accidents, physical assault, sexual assault, and other experiences.
AURORA Study

“Advancing Understanding of Recovery After Trauma”

Aims to analyze language use data from 5,000 participants recruited from emergency departments after a trauma experience.
Develop and evaluate **predictive and diagnostic biomarkers** for APNS (which includes depression and anxiety) as well as clinical interventions for trauma patients.
Patient encounters in Emergency Department
Veterans and trauma patients
- Received care following trauma incident
- Language use collected from mobile phone keyboard entries over specified time intervals
LIWC
(Linguistic Inquiry and Word Count)

“LIWC reads written or transcribed verbal texts which have been stored in a digital, computer-readable form (such as text files). The text analysis module then compares each word in the text against a user-defined dictionary... the dictionary identifies which words are associated with which psychologically-relevant categories” (LIWC 2015: How it Works, 2020).

<table>
<thead>
<tr>
<th>LIWC Dimension</th>
<th>Output Label</th>
<th>LIWC2015 Mean</th>
<th>LIWC2007 Mean</th>
<th>LIWC 2015/2007 Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word Count</td>
<td>WC</td>
<td>11,921.82</td>
<td>11,852.99</td>
<td>1.00</td>
</tr>
<tr>
<td>Summary Variable</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analytical Thinking</td>
<td>Analytic</td>
<td>56.34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clout</td>
<td>Clout</td>
<td>57.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authentic</td>
<td>Authentic</td>
<td>49.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emotional Tone</td>
<td>Tone</td>
<td>54.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Language Metrics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Words per sentence</td>
<td>WPS</td>
<td>17.60</td>
<td>25.07</td>
<td>0.74</td>
</tr>
<tr>
<td>Words+6 letters</td>
<td>Sixtir</td>
<td>15.60</td>
<td>15.89</td>
<td>0.98</td>
</tr>
</tbody>
</table>
Participants reporting symptoms that characterize PTSD will produce higher rates of the following word types:

Pronouns: first-person, first-person singular, plural, expressions of anxiety, religious words, expressions of anger

Participants reporting higher severity of symptoms that characterize PTSD will produce higher rates of the following word types:

Death and dying words, words focused on past, unique words, singular pronouns (reexperiencing symptoms)

Participants reporting symptoms that characterize PTSD will produce lower rates of the following word types:

Pronouns: singular, first-person plural, expressions of anxiety (hyperarousal symptoms)
Participants reporting higher severity of symptoms that characterize PTSD will produce lower word volume.

Participants reporting symptoms that characterize PTSD will produce higher rates of the following word types:

- Sensation/perception words
- Negative affect words
- Cognitive process words

Participants reporting higher severity of symptoms that characterize PTSD will produce lower rates of the following word types:

- Positive and negative affect words
- Cognitive process words (reexperiencing symptoms)
- Death and dying words (avoidance symptoms)
Word use will accurately predict severity of depression.

Word use will accurately predict severity of anxiety.

Participants reporting higher severity of depression, anxiety, and suicidal ideation will produce higher rates of words expressing absolutist thinking.
My Role, Results, and Data Analysis
Data Analytics Training

Python
Excel
Natural Language Processing (NLP)

Big Data Analytics
Literature Review

- **AUORA Study Website**
  - "Towards Assessing Changes in Degree of Depression Through Facebook"
  - "What Twitter Profile and Posted Images Reveal About Depression and Anxiety"
  - "Detecting Depression and Mental Illness on Social Media: An Integrative Review"
  - "In an Absolute State: Elevated Use of Absolutist Words is a Marker Specific to Anxiety, Depression, and Suicidal Ideation"
  - "Suicide Risk Assessment with Multi-level Dual-Context Language and BERT"
  - "Linguistic Predictors of Trauma Pathology and Physical Health"
  - "Psychological Aspects of Natural Language Use: Our Words, Our Selves"

"Towards Assessing Changes in Degree of Depression Through Facebook"

"Linguistic Predictors of Trauma Pathology and Physical Health"
Excel Tasks

- Sorting for unique users
- Writing formulas for:
  - First day of participation
  - Last day of participation
  - Determining each participant’s date range
  - Determining the number of missing days for each participant
  - Sorting for Week 1 and Week 2 Mean and StDev with at least 10/14 days of participation

LIWC Data for 3,227 Participants
~320,000 files
Count for ANDROID/IOS users

os_type
- ANDROID: 1698
- IOS: 1649
Percentage of days within first 8 weeks that have no Word Cloud data

Number of participants vs. Percentage of days within first 8 weeks that have no Word Cloud data for different operating systems (os_type)
Linear Regression Model

Of variables most correlated with raw depression assessment scores at 2 weeks after the trauma incident, the following are the most predictive:

- Week 1 Mean of cognitive process words
- Week 1 Mean of assent words
- Week 2 Mean of negative emotion words

Binary Logistic Model

Of the variables most correlated with the categorical depression indicator at two weeks after the trauma incident, the following are the most predictive:

- Week 1 Mean of sadness words
- Week 1 Mean of past focus words
- Week 1 Mean of assent words
- Week 2 Mean of function words
- Week 2 Mean of body words
- Weeks Mean of present focus words
"Freeze 2 Data Basis Document"

3. DEMOGRAPHICS

a) Sex

Description
Sex at Birth
Available at two time points
Emergency Department (ED)
Variable name for each time point
ED_GenderBirthCert
Variable type and acceptable range of the variable
Variable Type: Integer
Acceptable Range: 1: Male, 2: Female, -1: Present not Answered, -3: Present not Answered, -7: Refused
Algorithm details
Question asked in the survey: What was your sex at birth, as it appears on your birth certificate?
Response options: 1: Male, 2: Female, -1: Present not Answered, -3: Present not Answered, -7: Refused

b) Gender Identity

Description
Current Gender Identity
Available at two time points
Emergency Department (ED)
Variable name for each time point
ED_GenderNow
Variable type and acceptable range of the variable
Variable Type: Integer
Acceptable Range: 1: Female, 2: Male, -1: Present not Answered, -3: Present not Answered, -7: Refused
Importing packages – pandas, os
Reading in a csv
Error code
Creating a Dataframe
Creating reference names
Coding for specific rows in the file
Mean and Standard Deviation
Structuring a table as a csv
Writing a csv
## Mean & StDev Spreadsheet

![Spreadsheet Image]

### Table: LIWCStDevMean Excel Sheet

<table>
<thead>
<tr>
<th></th>
<th>Segment</th>
<th>Mean All</th>
<th>Mean Day 0-14</th>
<th>Mean Day 0-60</th>
<th>Mean Day 0-90</th>
<th>Mean Day 0-180</th>
<th>Mean Day 15-60</th>
<th>Mean Day 60-90</th>
<th>Mean Day 90-180</th>
<th>Mean Day 180-End</th>
<th>StDev All</th>
<th>StDev Day 0-14</th>
<th>StDev Day 0-60</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>WC</td>
<td>76.4954</td>
<td>179.91666777</td>
<td>183.777778</td>
<td>183.777778</td>
<td>102.6007736</td>
<td>185.708333</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>Analytic</td>
<td>75.20113</td>
<td>59.4225</td>
<td>56.27972222</td>
<td>56.27972222</td>
<td>32.745283</td>
<td>54.703333</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>Clout</td>
<td>71.05634</td>
<td>86.2875</td>
<td>83.56222222</td>
<td>83.56222222</td>
<td>75.6031205</td>
<td>82.195833</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>Authentic</td>
<td>31.28649</td>
<td>40.57416667</td>
<td>38.35083333</td>
<td>38.35083333</td>
<td>32.9870547</td>
<td>37.2391667</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>Tone</td>
<td>57.2200</td>
<td>80.5</td>
<td>79.30583333</td>
<td>79.30583333</td>
<td>62.8795238</td>
<td>78.77625</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>WPS</td>
<td>76.4854</td>
<td>179.91666777</td>
<td>183.777778</td>
<td>183.777778</td>
<td>102.6007736</td>
<td>185.708333</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>Dic</td>
<td>67.14835</td>
<td>78.13666667</td>
<td>77.89194444</td>
<td>77.89194444</td>
<td>69.0839622</td>
<td>77.769533</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>function</td>
<td>62.61026</td>
<td>39.41333333</td>
<td>36.45044444</td>
<td>36.45044444</td>
<td>36.45044444</td>
<td>36.45044444</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>prpron</td>
<td>4.388196</td>
<td>7.47</td>
<td>7.68027778</td>
<td>7.68027778</td>
<td>7.8255416</td>
<td>7.8571438</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>1</td>
<td>1.183505</td>
<td>1.003333333</td>
<td>0.88666667</td>
<td>0.88666667</td>
<td>0.95405604</td>
<td>0.82333333</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>2</td>
<td>0.61287</td>
<td>1.55833333</td>
<td>1.42</td>
<td>1.42</td>
<td>0.67813208</td>
<td>1.35033333</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>3</td>
<td>1.716495</td>
<td>3.09333333</td>
<td>3.14388889</td>
<td>3.14388889</td>
<td>2.06090434</td>
<td>3.29166667</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>4</td>
<td>0.06886</td>
<td>1.13</td>
<td>1.73888889</td>
<td>1.73888889</td>
<td>0.80917681</td>
<td>1.95333333</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>5</td>
<td>0.26654</td>
<td>0.595</td>
<td>0.49777778</td>
<td>0.49777778</td>
<td>0.32094437</td>
<td>0.44166667</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>6</td>
<td>2.052716</td>
<td>3.31916667</td>
<td>3.08722222</td>
<td>3.08722222</td>
<td>2.34566077</td>
<td>2.97125</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>7</td>
<td>0.97886</td>
<td>2.43166667</td>
<td>1.64944444</td>
<td>1.64944444</td>
<td>1.3245283</td>
<td>1.25833333</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>8</td>
<td>0.69686</td>
<td>10.07083333</td>
<td>8.69611111</td>
<td>8.69611111</td>
<td>9.13537758</td>
<td>8.00875</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Twitter Study
Comparing self-identifying PTSD patients to the general population

Tweepy, Python & Twitter API
Challenges, Significance, and Lessons Learned
Challenges

Data Delivery

Writing Code Under Multiple Conditions

Twitter Data Access
Significance

Understanding the link between trauma, mental health, and language use

Improving the understanding, prevention, and recovery of mental health conditions in trauma patients
Lessons Learned

- Process and compare large data sets using Python and Excel
- Overcome challenges in calculating formulas and writing code
- Understand research grant processes and methodologies
- Integrate informatics into a research study
- Work with privacy restrictions
- Turn data into meaningful knowledge
Thank you!

Nicole Mathews
MPS Candidate, Biomedical and Health Informatics
November 11, 2020


