Verbal Behavior Event Detection Using Textual and Acoustic Semantics

Asif Salekin
Sarah Masud Preum
Motivation: Detecting Verbal Agitation

• Agitation affects people with dementia, autism, Alzheimer
• 65% demented elderly patients are hospitalized due to agitation
• Manage cognitive disorder
• Solution:
  – continuous monitoring by caregivers
  – automatic detection
Motivation: Verbal Agitation Metrics

Cohen Mansfield inventory for verbal agitation
1. Crying
2. Laughing
3. Screaming
4. Negativism
5. Cursing
6. Saying repetitive sentence
7. Asking for help
8. Making sexual advance
Related Work: Detect Verbal Agitation Events Using *Acoustic* Features

- [W. Huang, 2010], [L.S. Kennedy, 2004], [S. Petridis, 2008], [K. P. Truong, 2007]
Related work: Detect Verbal Agitation Events Using **Textual** Features

- Detecting negativism is equivalent to sentiment analysis [M. Weigand, 2010] [B. Pang, 2008], [T. Wilson 2005]

- Detecting repetitive sentence is a sequence mining problem: detect recurring subsequences [J. Pei, 2004]
Goals

• Detect **cursing**
• Detect **asking for help**  *Patients in hospital*
• Detect **verbal sexual advances**  *Detect sexual harassment in office environment*

Less explored but important
Challenges

• Cursing: Ambiguity in word with multiple senses
  – “I rode my ass up the mountain” vs “Stop being an ass!”

• Asking for help: Textual content can often be misleading, need acoustic semantics (tone of the speaker)
  – In a urging tone: “Please help me!”
  – In a neutral tone : “The man asked your help”

• Verbal sexual advances
  – Similar challenges
Goals and Challenges

• Detect three events from the Cohen Mansfield agitation inventory for verbal agitation

Acoustic
- Crying
- Laughing
- Screaming

Textual
- Negativism
- Repetitive sentence
- Cursing

Acoustic + Textual
- Asking for help
- Sexual advances
Detecting Asking For Help & Verbal Sexual Advances
Overview of Approach

- Behavior event data collection

  - Acoustic preprocessing
    - Acoustic Features
  - Textual preprocessing
    - Textual Features

  - Binary Classification
Extracting Acoustic Features

• “human behaviors remain consistent with the specific emotion concepts” [Y. Zemack-Rugar, 2007]
  Verbal sexual advance → arousal

• Goal: Represent the emotional concepts reflected in the tone of speech.

<table>
<thead>
<tr>
<th>Acoustic features</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero crossing rate</td>
<td>Detect voice vs non voice</td>
</tr>
<tr>
<td>Harmonic-to-noise ratio</td>
<td>Anger vs non anger</td>
</tr>
<tr>
<td>Energy</td>
<td>Arousal vs non arousal</td>
</tr>
<tr>
<td>Pitch</td>
<td></td>
</tr>
<tr>
<td>F0 fundamental frequency</td>
<td>Joy-surprise vs disgust-anger</td>
</tr>
</tbody>
</table>
Processing Text Data

• Speech to text conversion:
  Dragon NaturallySpeaking 12 (95%-99% accuracy)

• Stop word list reduction:
  some stop words are important features in the problem domain (e.g., help, please)

• Stemming: Porter stemmer

• Normalization: punctuation, case conversion
Extracting Textual Features

Text document: converted text from audio clips

Document 1: please please help me

....

Bag of word representation: smoothed vector space of words

\[
\hat{t}f(w, d) = \begin{cases} 
1 + \log tf(w, d) & \text{if } tf(w, d) > 0 \\
0 & \text{otherwise}
\end{cases}
\]

\[
idf(w) = 1 + \log \left( \frac{N}{DF(w)} \right)
\]

<table>
<thead>
<tr>
<th>...</th>
<th>help</th>
<th>please</th>
<th>...</th>
<th>...</th>
<th>me</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4.2</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0.2</td>
</tr>
</tbody>
</table>
Combined Feature Space Formation

Acoustic Features

Textual Features

Combine Features

Binary Classification
Detecting Cursing
Overview of Approach

1. Behavior Event Data Collection
2. Speech to Text Conversion
3. Textual preprocessing
4. Extracting Textual Features
5. Binary Classification
Resolving Ambiguity of Curse Detection

• Generate curse dictionary of 165 words:
  – http://www.noswearing.com/dictionary/d

• 36 Ambiguous words: dog, ass, etc.

• Word sense disambiguation
  – WordNet knowledge base
  – Modified Lesk algorithm
Resolving Ambiguity of Curse Detection: An Example

- “I rode my ass up the mountain” vs “Stop being an ass!”

<table>
<thead>
<tr>
<th></th>
<th>Word sense</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non 1</td>
<td>Hardy and sure footed animal smaller and with longer ears that horse</td>
</tr>
<tr>
<td>curse 2</td>
<td>The fleshy part of the human body that you sit on</td>
</tr>
<tr>
<td>Curse 3</td>
<td>A pompous fool</td>
</tr>
<tr>
<td>4</td>
<td>Slang for sexual intercourse</td>
</tr>
</tbody>
</table>

- Adapted Lesk Algorithm -> modification
  - Multiclass problem
  - Binary class Problem
Experiment Design

• Data
  – Controlled experiments with 4 volunteers
  – Movie clip extracts

• Ground truth: manual labeling

• Performance metrics: accuracy, F-1

• Analysis: kappa statistics for confidence analysis
Result: Detecting “Asking for Help”

- Using only acoustic features

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>Kappa Statistics</th>
<th>F1-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>75.8</td>
<td>0.42</td>
<td>0.75</td>
</tr>
<tr>
<td>K-nearest neighbor</td>
<td>80.9</td>
<td>0.57</td>
<td>0.81</td>
</tr>
<tr>
<td>Random forest</td>
<td>80.2</td>
<td>0.50</td>
<td>0.79</td>
</tr>
</tbody>
</table>

- Using only textual features performs even worse

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>Kappa Statistics</th>
<th>F1-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>84.7</td>
<td>0.67</td>
<td>0.86</td>
</tr>
<tr>
<td>K-nearest neighbor</td>
<td>83.5</td>
<td>0.62</td>
<td>0.84</td>
</tr>
<tr>
<td>Random forest</td>
<td>89.6</td>
<td>0.75</td>
<td>0.89</td>
</tr>
</tbody>
</table>

~11% increase
~10% increase
Result: Detecting “Verbal Sexual Advances”

• Using only acoustic features

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>Kappa Statistics</th>
<th>F1-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>71.4</td>
<td>0.42</td>
<td>0.72</td>
</tr>
<tr>
<td>K-nearest neighbor</td>
<td>80.3</td>
<td>0.61</td>
<td>0.81</td>
</tr>
<tr>
<td>Random forest</td>
<td>79.7</td>
<td>0.61</td>
<td>0.80</td>
</tr>
</tbody>
</table>

• Using only textual features performs even worse

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>Kappa Statistics</th>
<th>F1-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>73.2</td>
<td>0.45</td>
<td>0.72</td>
</tr>
<tr>
<td>K-nearest neighbor</td>
<td>76.4</td>
<td>0.53</td>
<td>0.74</td>
</tr>
<tr>
<td>Random forest</td>
<td>86.6</td>
<td>0.72</td>
<td>0.87</td>
</tr>
</tbody>
</table>

~8% increase

~7% increase
Result: Cursing

• Baseline 1: Use all words in curse dictionary:
• Baseline 2: Use only single meaning words in dictionary as curse

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline 1</td>
<td>0.73</td>
<td>1</td>
<td>0.84</td>
</tr>
<tr>
<td>Baseline 2</td>
<td>1</td>
<td>0.74</td>
<td>0.85</td>
</tr>
<tr>
<td>Our approach</td>
<td>0.95</td>
<td>0.96</td>
<td>0.96</td>
</tr>
</tbody>
</table>
Analyzing Results: Identifying New Challenge

• Random forest better than kNN, Naïve Bayes
  – Probably data not linearly separable

• Curse of Dimensionality??
  – Small dataset but large feature space

• Potential Solution:
  Evaluate feature reduction: PCA
Principle Component Analysis (PCA)

Figure 1

Figure 2

Figure 4
Problem: PCA

Results:
**Asking for help:**
Reduced Feature number: 126
Accuracy: from 89.6% to 84%

**Verbal sexual advances:**
Reduced Feature number: 213
Accuracy: from 86.6% to 80%
Project to higher dimension with RBF kernel

\[ f_1 = \text{similarity}(x, l^{(1)}) = \exp \left( - \frac{\| x - l^{(1)} \|_2^2}{2\sigma^2} \right) = \exp \left( - \sum_{j=1}^{n} (x_j - l_{j}^{(1)})^2 \right) \]
Kernel PCA

- Increase the dimension up to number of training data
- Perform PCA on that higher dimension data
- Results:
  - Asking for help: Accuracy from 89.6% to 73%
  - Verbal sexual advances: Accuracy from 86.6% to 67.8%
- Limitation
  - We have limited number of training data so, cannot increase dimension!!!
  - Hence, we need more data to use KPCA.
Conclusion

• Combine text mining and signal processing
• First to detect cursing, asking for help, verbal sexual advances
• Textual features enhances classification performance
• Future works:
  – Evaluate on a larger dataset: validity of feature reduction
  – Include more contextual features
Thanks!

Question?