Multimodal Humor Detection

Lin Ai
COMS 6998
Spring 2022
Why Study Humor?

• To understand human interaction
• To detect when people are being humorous rather than serious to evaluate the content of what they say
• To learn the characteristics of humorous speech to be able to synthesize it (e.g. for robots, chatbots, games, advertisements)
• Because it’s interesting…
How Do We Define Humor?

1. Producer + Perceiver
2. Positive emotional reactions (laughter)
3. Highly individualistic & cultural specific

Lack of multimedia data annotated with humor
Humor Detection in Text

- 16k one-liners (Mihalcea and Strapparava, 2005)
  - Humor-Specific Stylistic Features: alliteration/rhyme, antonymy, adult slang
    - “A clean desk is a sign of a cluttered desk drawer”
- One-liners + 1k news article from “The Onion” (Mihalcea and Pulman, 2007)
  - Human-centeredness and negative polarity
    - “Take my advice; I don’t use it anyway”
- The New Yorker Cartoon Caption Contest (Radev et al, 2015)
  - Negative sentiment, human-centeredness
    - “If that’s Theseus, I’m not here.”
Humor Detection in Text

• Extract humor anchor in one-liners (Yang et al., 2015)
  – The subset of candidates that provides the maximum decrement of humor scores
    • “The one who invented the door knocker got a No-bell prize.”

• 1k tweets (Zhang and Liu, 2014)
  – Phonetic + morpho-syntactic + lexico-semantic + pragmatic + affective features
    • “I generally avoid temptation unless I can't resist it. - Mae West #quote #humor”

• TED talk transcripts (Chen and Lee, 2017)
  – Sentences containing or immediately followed by markup ‘(Laughter)’
    • “If you’re a dog and you spend your whole life doing nothing other than easy and fun things, you’re a huge success! (Laughter)”
Multimodal Humor Detection

• TV sitcoms
  – Use canned laughter to label humor
    • FRIENDS (Purandare and Litman, 2006)
    • The Big Bang Theory (Bertero and Fung, 2016)
    • Seinfeld (Bertero and Fung, 2016)
  – No study has shown that canned laughter actually represents the audience’s perception of humor.

Fig. 1: Example from The Big Bang Theory:
LEONARD: I did a bad thing.
SHELDON: Does it affect me?
LEONARD: No.
SHELDON: Then suffer in silence. LAUGH
Danmu/bullet curtain – *Time-aligned Comments*

https://www.bilibili.com/video/BV1nJ411h7ax?share_source=copy_web
https://www.nicovideo.jp/
Hypothesis

Audiences tend to respond to humor in videos with laughing
A high volume of laughing comments at a given time

- Laughing indicators
  - ‘233’ (internet meme)
  - ‘哈哈’ & ‘hh’ (onomatopoeia of laughter)
Data Collection

‘Papi酱’
• A Chinese influencer
• Famous for discussing trending topics in a humorous way
• 7 million subscribers, 660 million views on Bilibili.com
Data Collection

• We use early videos created by ‘Papi酱’
  – Filtered out videos containing dialects and advertisements
  – 100 videos, 93,593 time-aligned comments
    • 5,064 comments with ‘233’
    • 7,255 comments with ‘哈哈’
    • 730 with ‘hh’
• Segmentation
  – One-second unit level
  – Inter-pausal unit (IPU) level: 3 seconds on average
Constructing Unsupervised Labels

• Users typically do not pause to comment
• Response Time = reaction time + typing time
• Smooth number of laughing comments by response time distribution
• Set threshold to distinguish humor from non-humor segments
• One-second unit level
  – 6,508 humorous segments; 17,847 non-humorous segments
• Inter-pausal unit (IPU) level
  – 2,531 humorous segments; 5,394 non-humorous segments
Constructing Unsupervised Labels

Before smoothing  
After smoothing  
After labeling
Verification: Human Annotation

• We need a manually annotated test set to verify our unsupervised labeling method
• Three human annotators
  – Label each second with humor/non-humor
  – Average Cohen’s Kappa: 0.65
  – Fleiss’ Kappa: 0.65
• Gold labels on test set: majority vote
  – Unsupervised labels’ accuracy
    • One-second units: 0.78
    • Inter-pausal units: 0.76
Features — Acoustic-Prosodic

• Tools: Praat, openSMILE, Google ASR API
• Features:
  – Min, max, mean, range, std of pitch
  – Min, max, mean, range, std of intensity
  – Pitch existence: whether extractable pitch values exists in the segment
  – 384 features from openSMILE
    • More features, more functions
  – Speaking Rate: Number of characters per second (from ASR transcript)
- The existence of pitch is positively correlated with humor.
- Exclude segments with no pitch values in the analysis of other speech features.
Analysis - Speech Features

• Humorous speech has
  – Higher pitch value
  – Larger change in pitch
  – Higher intensity value
  – Smaller change in intensity
  – Slower speaking rate

• Humor techniques
  – Exaggeration and bombast

<table>
<thead>
<tr>
<th></th>
<th>One-second Unit</th>
<th>Inter-pausal Unit (IPU)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$t$</td>
<td>$p$</td>
</tr>
<tr>
<td>Pitch existence</td>
<td>8.71</td>
<td>$p&lt;0.001$</td>
</tr>
<tr>
<td>Pitch min</td>
<td>3.68</td>
<td>$p=0.403$</td>
</tr>
<tr>
<td>Pitch max</td>
<td>4.62</td>
<td>$p&lt;0.001$</td>
</tr>
<tr>
<td>Pitch mean</td>
<td>6.21</td>
<td>$p&lt;0.001$</td>
</tr>
<tr>
<td>Pitch range</td>
<td>2.40</td>
<td>$p=0.016$</td>
</tr>
<tr>
<td>Pitch stddev</td>
<td>0.93</td>
<td>$p=0.352$</td>
</tr>
<tr>
<td>Intensity min</td>
<td>6.91</td>
<td>$p&lt;0.001$</td>
</tr>
<tr>
<td>Intensity max</td>
<td>16.88</td>
<td>$p&lt;0.001$</td>
</tr>
<tr>
<td>Intensity mean</td>
<td>7.02</td>
<td>$p&lt;0.001$</td>
</tr>
<tr>
<td>Intensity range</td>
<td>-5.02</td>
<td>$p&lt;0.001$</td>
</tr>
<tr>
<td>Intensity stddev</td>
<td>-3.57</td>
<td>$p&lt;0.001$</td>
</tr>
<tr>
<td>Speaking rate</td>
<td>-10.12</td>
<td>$p&lt;0.001$</td>
</tr>
</tbody>
</table>
(Hamlet) In the end, surprisingly and also not surprisingly — everyone died!
Features — Transcript-based

- Tools: Google ASR API, Jieba, LIWC
- Audio preprocessing:
  - ‘Papi酱’ speeds her videos, so we slowed them down to 0.75 times the original speed for ASR
  - Normalized intensity and pitch
- Transcript preprocessing:
  - Word segmentation using ‘Jieba’
- LIWC (CLIWC): 91 word categories:
  - e.g. function words, affect words, social words, etc.
Analysis - Lexical Features

One-second unit level
• Positively correlated with humor:
  – Strategy: Anxiety, risk, netspeak, i
  – Content: Power, drive, religion
• Negatively correlated with humor:
  – Strategy: Cognitive process, insight
  – Content: Sexual, female, biological process

IPU level
• Positively correlated with humor:
  – Strategy: i
  – Content: religion
• Negatively correlated with humor:
  – Strategy: Cognitive process, cause, interrogatives, auxverb, they
  – Content: Female, biological process, body
Analysis - Lexical Features

• Humorous one-liners vs. non-humorous short sentence (Mihalcea and Pulman, 2007)
  – Negative polarity, Human-centeredness

• Negative polarity
  – Negation: not significant
  – Negative emotion: ‘anxiety’ significant on one-second unit level

• Human-centeredness
  – ‘i’ (first person pronouns): significant on both one-second unit and IPU level
  – Other personal pronouns: not significant
Features — Visual

• Frame similarity:
  – Assumption: difference between frames may capture visual patterns such as change of scenes and large body movements
  – Extracted 1 frame in each 10ms and compute similarity with neighbouring extracted frames
  – Measure: structural similarity index (SSIM)
  – Features: min, max, mean, range, std
Features — Visual

- Body poses
  - Extraction: AlphaPose
  - 17 keypoints of body junctions with confidence scores
  - Used binary features to indicate the appearance of hips and legs
  - Features: mean, std, mean of frame-level differences, std of differences
Features — Visual

- Facial landmarks:
  - Extraction: dlib library
  - 68 coordinates of facial landmarks
  - Preprocessing: rescaled, computed relative position, exclude keypoints for jawline
  - Features: mean, std, mean of frame-level differences, std of differences
Analysis - Visual Features

• SSIM - frame similarity
• Humor segments
  – Are unlikely to be motionless
  – But also have fewer complete scene changes

<table>
<thead>
<tr>
<th></th>
<th>One-second Unit</th>
<th>Inter-pausal Unit (IPU)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>t</td>
<td>p</td>
</tr>
<tr>
<td>SSIM min</td>
<td>0.75</td>
<td>p=0.452</td>
</tr>
<tr>
<td>SSIM max</td>
<td>-23.05</td>
<td>p&lt;0.001</td>
</tr>
<tr>
<td>SSIM mean</td>
<td>-19.83</td>
<td>p&lt;0.001</td>
</tr>
<tr>
<td>SSIM range</td>
<td>-6.57</td>
<td>p&lt;0.001</td>
</tr>
<tr>
<td>SSIM stddev</td>
<td>-6.51</td>
<td>p&lt;0.001</td>
</tr>
</tbody>
</table>
Analysis - Visual Features

Good news for those who are single! In 2016 — you will still be a single dog.
Analysis - Visual Features

• Body poses:
  – One-second unit: keypoints above hips are significant
  – IPU unit: keypoints above shoulder are significant
  – The movements of keypoints are correlated with humor, but the movement directions are not significant

• Facial landmarks:
  – Most significant keypoints: brows, nose (head-turning information)
Classification Experiments

• 70 videos (unsupervised labels) in training set, 30 videos (human labels) in test set

• Feature dimensions:
  – 396 speech features (11 from Praat, 384 from openSMILE, speaking rate)
  – 91 text features (CLIWC)
  – 522 visual features (5 from frame similarity, 408 from facial landmarks, 109 from body pose)

• Model: random forest classifier with 1000 estimators
Classification Experiments

- IPU segmentation outperforms one-second unit segmentation.
- Speech features are the most useful.

<table>
<thead>
<tr>
<th></th>
<th>One-second Unit</th>
<th>Inter-pausal Unit (IPU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech</td>
<td>0.71</td>
<td>0.76</td>
</tr>
<tr>
<td>Text</td>
<td>0.70</td>
<td>0.70</td>
</tr>
<tr>
<td>Visual</td>
<td>0.72</td>
<td>0.72</td>
</tr>
<tr>
<td>Speech + Text</td>
<td>0.72</td>
<td>0.76</td>
</tr>
<tr>
<td>Speech + Visual</td>
<td><strong>0.73</strong></td>
<td>0.75</td>
</tr>
<tr>
<td>Text + Visual</td>
<td>0.72</td>
<td>0.72</td>
</tr>
<tr>
<td>All Features</td>
<td><strong>0.73</strong></td>
<td>0.75</td>
</tr>
</tbody>
</table>
Future Directions

• Collect more videos from different types of humorous video creators
  – Current videos mainly include humor techniques like exaggeration and bombast
  – Explore larger variety of characteristics in humor
• Apply to different types of emotions and reactions
• Examine other platforms and create automatic labeling of video segments
  – Use videos collected from other sources such as YouTube live chats
Thanks233!
Next Week

Topic: Speech Analysis: Deception and Trust

Any questions?