Emoji serve different linguistic functions on different occasions:

- **Capabilities that ignore emoji, or bucket them as punctuation, ignore key aspects of computer-mediated communication.**
- **Analysis** that looks only at frequency or distribution ignores the distinctive communicative potentials of non-lexical characters.
- **Identifying where emoji is replacing text** can help us understand how emoji are used and where they are most effective. Recognizing the importance of content words in emoji communication helps us establish the types of questions we can ask about emoji.

### Collect tweets with emojis

- **Collect tweets with emojis** and annotate them with linguistics variables.
- **Woe to those from the public!** Twitter users use emoji in their tweets for a wide range of purposes. Some emoji are clearly polysemous; any, if any, may be polysemous.
- **Every email contains a facilitating range of pragmatic and emotional effects that may need refinement.**

### Feature engineering

#### Context helps; Unicode blocks can be a proxy for semantics; POS tagging is a nice hint

<table>
<thead>
<tr>
<th>Feature extracted for training</th>
<th>Metadata:</th>
<th>POS, part-of-speech tagging assigned by annotator</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>The tokens begin</strong></td>
<td></td>
<td>column 35</td>
</tr>
<tr>
<td><strong>The tokens are in use</strong></td>
<td></td>
<td>column 36</td>
</tr>
<tr>
<td><strong>The tokens are not in use</strong></td>
<td></td>
<td>column 37</td>
</tr>
<tr>
<td><strong>The tokens are in use</strong></td>
<td></td>
<td>column 38</td>
</tr>
<tr>
<td><strong>The tokens are not in use</strong></td>
<td></td>
<td>column 39</td>
</tr>
<tr>
<td><strong>The tokens are in use</strong></td>
<td></td>
<td>column 40</td>
</tr>
<tr>
<td><strong>The tokens are not in use</strong></td>
<td></td>
<td>column 41</td>
</tr>
<tr>
<td><strong>The tokens are in use</strong></td>
<td></td>
<td>column 42</td>
</tr>
<tr>
<td><strong>The tokens are not in use</strong></td>
<td></td>
<td>column 43</td>
</tr>
</tbody>
</table>

#### Inter-Annotator Agreement

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Original Tweets</th>
<th>Annotated Tweets</th>
<th>Inter-Annotator Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotators Agreement</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>99%</td>
<td>99%</td>
<td>99%</td>
<td></td>
</tr>
<tr>
<td>98%</td>
<td>98%</td>
<td>98%</td>
<td></td>
</tr>
<tr>
<td>97%</td>
<td>97%</td>
<td>97%</td>
<td></td>
</tr>
<tr>
<td>96%</td>
<td>96%</td>
<td>96%</td>
<td></td>
</tr>
<tr>
<td>95%</td>
<td>95%</td>
<td>95%</td>
<td></td>
</tr>
<tr>
<td>94%</td>
<td>94%</td>
<td>94%</td>
<td></td>
</tr>
<tr>
<td>93%</td>
<td>93%</td>
<td>93%</td>
<td></td>
</tr>
<tr>
<td>92%</td>
<td>92%</td>
<td>92%</td>
<td></td>
</tr>
<tr>
<td>91%</td>
<td>91%</td>
<td>91%</td>
<td></td>
</tr>
<tr>
<td>90%</td>
<td>90%</td>
<td>90%</td>
<td></td>
</tr>
</tbody>
</table>

#### Content words are easy to label; non-content words require an empirical inventory of emoji senses.

- **Mood/emotion** senses may need refinement, but the essential range of mood/emotion emoji is clearly identifiable.
- **Several small samples show a number of characteristics that are used both as content words and as emotional signals.**

### Metrics on CRF tagging

- **Table:** 
  - | Dataset | Original Tweets | Annotated Tweets | Inter-Annotator Agreement |
  - |---------------------------------|------------------|---------------------|
  - | Annotators Agreement | 100% | 100% | 100% |
  - | 99% | 99% | 99% |
  - | 98% | 98% | 98% |
  - | 97% | 97% | 97% |
  - | 96% | 96% | 96% |
  - | 95% | 95% | 95% |
  - | 94% | 94% | 94% |
  - | 93% | 93% | 93% |
  - | 92% | 92% | 92% |
  - | 91% | 91% | 91% |
  - | 90% | 90% | 90% |

- **Table:** 
  - | Dataset | Original Tweets | Annotated Tweets | Inter-Annotator Agreement |
  - |---------------------------------|------------------|---------------------|
  - | Annotators Agreement | 100% | 100% | 100% |
  - | 99% | 99% | 99% |
  - | 98% | 98% | 98% |
  - | 97% | 97% | 97% |
  - | 96% | 96% | 96% |
  - | 95% | 95% | 95% |
  - | 94% | 94% | 94% |
  - | 93% | 93% | 93% |
  - | 92% | 92% | 92% |
  - | 91% | 91% | 91% |
  - | 90% | 90% | 90% |

- **Table:** 
  - | Dataset | Original Tweets | Annotated Tweets | Inter-Annotator Agreement |
  - |---------------------------------|------------------|---------------------|
  - | Annotators Agreement | 100% | 100% | 100% |
  - | 99% | 99% | 99% |
  - | 98% | 98% | 98% |
  - | 97% | 97% | 97% |
  - | 96% | 96% | 96% |
  - | 95% | 95% | 95% |
  - | 94% | 94% | 94% |
  - | 93% | 93% | 93% |
  - | 92% | 92% | 92% |
  - | 91% | 91% | 91% |
  - | 90% | 90% | 90% |

### References