Statistics of Internet Discussions with Illustrations on Emoticon Use

Juha Alho\(^1\)

University of Helsinki
Department of Social Research

Oct, 1, 2019

\(^1\)This research was supported by the Academy of Finland grant 292906, project Citizen Mindscapes.
Discussion Forum suomi24: Characteristics

Owner: Aller Publishers → research use!

Data from 2001-2015 (now 2001-2017), longest in the world?

Number of posts $48 \times 10^6$ (now $84 \times 10^6$)

Posts form threads. They differ in type:

<table>
<thead>
<tr>
<th></th>
<th>Endemic</th>
<th>Epidemic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Endogenous</td>
<td>&quot;bible&quot;</td>
<td>&quot;dispute about cars, authority&quot;</td>
</tr>
<tr>
<td>Exogenous</td>
<td>&quot;ski resort&quot;</td>
<td>immigration, cyber attacks,...</td>
</tr>
</tbody>
</table>
Emoticons Considered (16/23)

emot/ion + icon (≈ emoji pictographs, but in ASCII)

<3 :) :)) :))) ;) ;)) :-) ;-)

=) :( :-( :D :DD ;D :-D XD

= =E =f =O : o :P =P

Bottom 7 excluded: many were parts of internet addresses

Source:

Parser → lemmas → punctuation → list of potential emoticons
Total Counts (Dotted) and Share of Messages with at Least One Emoticon (Blue): Missing Data?
Weekly Shares of Posts by Hour of the Day
Generated by voluntary participation

Moderation by Owner of Data

Cyber Attacks

Data Processing Errors

No target population $\rightarrow$ no missing data!

Counts do not reflect the burstyness of the social world only $\rightarrow$ Observed changes of level are biased evidence of underlying moods/rhythms etc.

Total counts in posts are not related with emoticon use $\rightarrow$ No bias in relative shares, even when participation, moderation or processing errors vary.
Marginal Occurrence of Emoticons ($N \sim 48 \times 10^6$)

Messages with at least one emoticon are 8.2%

Of these messages, the percentages containing specific emoticons are

<table>
<thead>
<tr>
<th></th>
<th>:)</th>
<th>;)</th>
<th>:D</th>
<th>:=)</th>
<th>:-(</th>
<th>:-)</th>
<th>:)</th>
<th>;-)</th>
</tr>
</thead>
<tbody>
<tr>
<td>%</td>
<td>48.2</td>
<td>16.1</td>
<td>15.8</td>
<td>7.8</td>
<td>6.2</td>
<td>3.7</td>
<td>2.6</td>
<td>2.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>;D</th>
<th>;)</th>
<th>XD</th>
<th>:)))</th>
<th>:-D</th>
<th>:-)</th>
<th>:DD</th>
<th>&lt;3</th>
</tr>
</thead>
<tbody>
<tr>
<td>%</td>
<td>1.8</td>
<td>1.0</td>
<td>0.9</td>
<td>0.9</td>
<td>0.6</td>
<td>0.4</td>
<td>0.3</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Snoeleben's (2012) top ten in U.S. twitter data:

<table>
<thead>
<tr>
<th></th>
<th>:)</th>
<th>;)</th>
<th>:(</th>
<th>:D</th>
<th>:-)</th>
<th>XD</th>
<th>:=)</th>
<th>;-)</th>
<th>;D</th>
<th>:)</th>
</tr>
</thead>
<tbody>
<tr>
<td>%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Association Between Discussion Areas and Emoticons

460 discussion areas, select 410 with more than 500 messages → 410 × 16 matrix of counts \(Y_{ij}\)

**Correspondence Analysis:**

\[
Y_{+j} = \sum_{i=1}^{410} Y_{ij}, \quad Y_{i+} = \sum_{j=1}^{16} Y_{ij}, \quad Y_{++} = \sum_{i=1}^{410} \sum_{j=1}^{16} Y_{ij}
\]

\[
p_{ij} = Y_{ij} / Y_{++}, \quad r_i = Y_{i+} / Y_{++}, \quad c_j = Y_{+j} / Y_{++}
\]

\[
P = (p_{ij}), \quad r = (r_1, \ldots, r_{410})^T, \quad c = (c_1, \ldots, c_{16})^T
\]

**Least Squares** approximation via SVD:

\[
D(r)^{-1/2}(P - rc^T)D(c)^{-1/2} \approx UD(\alpha)V^T,
\]

where \(U^TU = V^TV = I_2\)
Motivation for least squares: e.g., under a Poisson model,

\[ \frac{(Y_{ij} - E[Y_{ij}])/E[Y_{ij}]^{1/2}}{\text{asymptotically}} = N(0, 1) \]

But, for display, \textbf{relative deviations from the expected} can be more interpretable

\[ D(r)^{-1/2}(P - rc^T)D(c)^{-1/2} \approx D(r)^{-1/2}UD(\alpha)V^TD(c)^{-1/2} \]

\textbf{Principal Coordinates:} \( D(r)^{-1/2}UD(\alpha) \)

\textbf{Standard Coordinates:} \( D(c)^{-1/2}V \)
Associations Between Emoticons from Correspondence Analysis (Standard Coordinates)
Shares of Main Types of Emoticons

Week

Relative Level

:) ;) =) :(

:-) ;-) :-( :-D

:D :DD ;D XD

:)) :))) ;))
Main Types of Emoticons by Hour of the Day

- :)) :))) ;))
- :D :DD ;D XD
- :) ;) =) :(
- :-) ;-) :-( :-D

Graph showing the share of different emoticons across different hours of the day.
### Topics ↔ Style ↔ Individuals

Top 5 Sports ordered by relative emoticon shares

<table>
<thead>
<tr>
<th></th>
<th>Golf</th>
<th>Extreme Skiing (CC+DH)</th>
<th>Ice swimming</th>
<th>Rollerskating</th>
<th>Skateboard Riding</th>
<th>Skating</th>
<th>Gymnastics</th>
<th>Icehockey</th>
<th>Weightlifting</th>
<th>Wrestling</th>
<th>Diving</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</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>2</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>3</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>4</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>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Emoticons Modify Meaning and Characterize User

Scott E. Fahlman’s proposal (Sept. 1982):

:) for joke, :( for not joke

suggests that emoticons are metatext. This is too narrow since

(1) by correspondence analysis, ostensibly similar emoticons (e.g. :) , :-) , or :D, :-D) can fall into different groups,

(2) emoticon use varies by time and by topic,

(3) excessive punctuation (!!!, !!!?, !!!!?? etc.): 8.7 % overall, but 11.8 % in messages with emoticons, and

(4) there are heavy-users of emoticons,

so emoticons are better characterized as enrichments whose meanings become determined in their context of use - suitable for the ASCIIII world.
Appendix: Data Use Toolbox Available for Research and Teaching Involving suomi24

- **Laptop** (R and other programs)
- **TAITO cluster computer** (UNIX, R, MySQL)
- **cPOUTA IaaS** (suomi24 as MariaDB, MySQL)

**Assumption:** Intended Users have little past experience in R, UNIX, MySQL

**Principle of Use:** Copy/Paste from a set of prototype examples

**Student Experience:** Logic Works, but Silly Technical Problems Arise: Character Sets, Where am I?