EMOJIS[1]

"An emoji is a pictogram, logogram, ideogram or smiley embedded in text and used in electronic messages and web pages. The primary function of emoji is to fill in emotional cues otherwise missing from typed conversation" (Source: Wikipedia)

😊 😎 🤣 😁 😆 🤪 😉

THE COMPLEX NATURE OF EMOJIS

→**Visually,** emojis are **images**   Human perspective

→**Technically**, emojis are ([unicode](#)) **characters**   Computer perspective

→**Analytically**, emojis are treated as **words**   Computer perspective

→**Semantically,** emojis can be **more than words**   Human perspective

CAIS CENTER FOR ADVANCED INTERNET STUDIES

3

EMOJIS & SENTIMENT

→ Since **emojis** are often used to confer emotions, they quite likely **provide valuable information for sentiment analysis**

→ However, while their inclusion should increase the accuracy of sentiment analysis, they are often ignored

→ The are two types reasons why emojis are **challenging** to deal with and, hence, often neglected:

  → **Technical** aspects

  → Questions related to **understanding & interpretation**

HUMAN UNDERSTANDING OF EMOJIS

→Emojis can have different meanings that depend on…

→context 🍆

→culture 🍙

→ also, subcultures 🤘

→inter-individual differences 😬

Note: A helpful resource for understanding (the use of) emojis is Emojipedia

TECHNICAL ASPECTS OF EMOJI USE

→ **emojis are rendered differently on different platforms**, meaning that they can potentially elicit different emotions
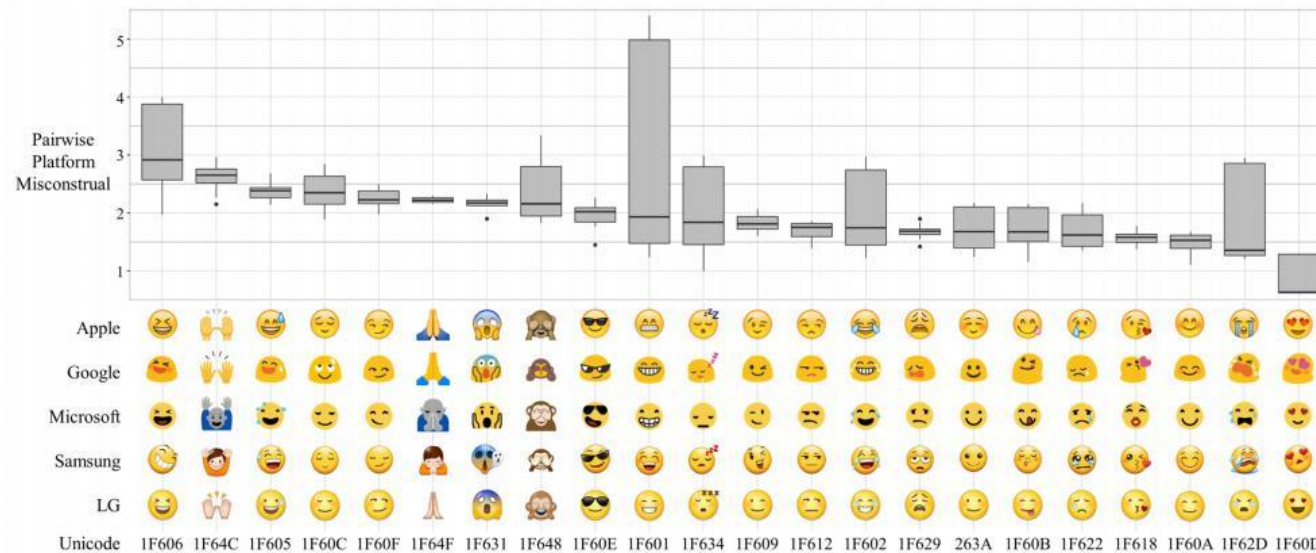

Figure 1. Across-platform sentiment misconstrual scores grouped by Unicode. Each boxplot shows the range of sentiment misconstrual scores across the five platforms. They are ordered by decreasing median platform-pair sentiment misconstrual, from left to right.

Source: Miller et al., 2016   Also see Emojipedia

TECHNICAL ASPECTS OF EMOJI USE

→emojis are difficult to deal with from a technical perspective due to the infamous character encoding hell 😈

- →emojis can come in one of multiple completely different encodings

- →your operating system (OS) has a default encoding that is used, e.g., when opening/writing files in a text editor

- →Python (as well as R and other programming languages) has a default character encoding (typically that of your OS)

TECHNICAL ASPECTS OF EMOJI USE

Unicode code point	character	UTF-8 (hex.)	name
U+1F600	😀	f0 9f 98 80	GRINNING FACE
U+1F601	😁	f0 9f 98 81	GRINNING FACE WITH SMILING EYES
U+1F602	😂	f0 9f 98 82	FACE WITH TEARS OF JOY
U+1F603	😃	f0 9f 98 83	SMILING FACE WITH OPEN MOUTH
U+1F604	😄	f0 9f 98 84	SMILING FACE WITH OPEN MOUTH AND SMILING EYES
U+1F605	😅	f0 9f 98 85	SMILING FACE WITH OPEN MOUTH AND COLD SWEAT
U+1F606	😆	f0 9f 98 86	SMILING FACE WITH OPEN MOUTH AND TIGHTLY-CLOSED EYES
U+1F607	😇	f0 9f 98 87	SMILING FACE WITH HALO
U+1F608	😈	f0 9f 98 88	SMILING FACE WITH HORNS
U+1F609	😉	f0 9f 98 89	WINKING FACE
U+1F60A	😊	f0 9f 98 8a	SMILING FACE WITH SMILING EYES

Source: https://www.utf8-chartable.de/unicode-utf8-table.pl?start=128512

DISCLAIMER

→ I mostly use R when I work with data

→ For the examples in the following, the data were collected, processed, and analyzed with R

→ The examples are based on the user comments for the Emoji Movie Trailer on *YouTube*

→ If you are interested in collecting, processing, and analyzing YouTube data with R (incl. sentiment analysis), you can check out the materials of the course "Automatic Sampling and Analysis of YouTube Comments" by Annika Deubel, M. Rohangis Mohseni, and me (note: we will quite likely offer the workshop again on February 14th & 15th, 2024; check the GESIS Training website for updates)

→ Side note: There also are multiple packages that you can use for sentiment analysis in R, such as:

  → syuzhet (dictionary-based)

  → sentimentr (sentence-based)

 → sentiment.ai (makes use of language embedding models)

EXAMPLE: SENTIMENT ANALYSIS WITH EMOJIS

→ For the YouTube data workshop, we implemented a dictionary-based bag-of-words/ -emojis approach for a sentiment analysis that includes emojis

→ Short summary of our approach:

  → data: N = 38,137 comments for the Emoji Movie Trailer collected via the YouTube API in February 2023

  → emoji sentiment dictionary from the lexicon package for R

    → The lexicon includes sentiment scores for the 734 most frequent emojis; since – similar to words – the distribution of emojis follows Zipf's Law, this should cover most of the used emojis

    → n = 4,193 of the comments contained at least one emoji from the dictionary

    → Sentiment scores range from -1 (negative) to +1 (positive)

  → For linking the emojis in the comments with the dictionary, we converted the emojis to textual descriptions: "Amazing movie 😀" becomes "Amazing movie EMOJI_GrinningFace"

  → For further details, see the sentiment analysis slides from that workshop

EXAMPLE: SENTIMENT ANALYSIS WITH EMOJIS

Distribution of Average Emoji Sentiment Scores by Comment

EXAMPLE: SENTIMENT ANALYSIS WITH EMOJIS
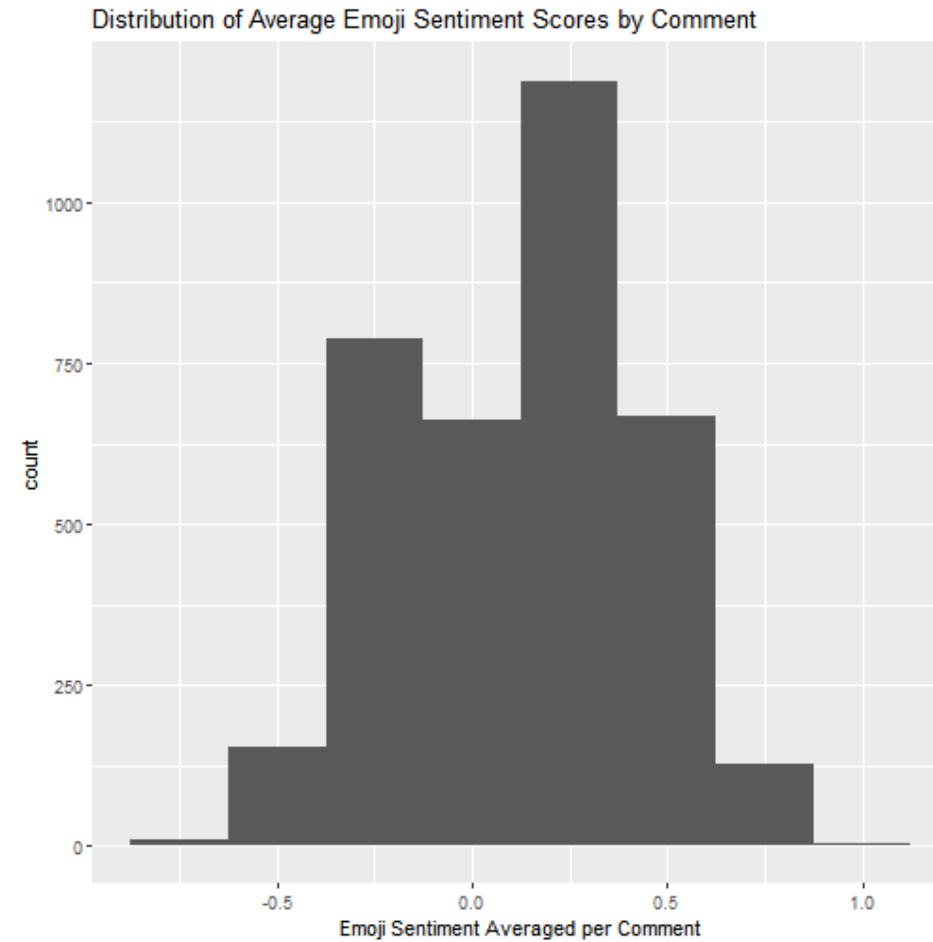
Averaged Sentiment Scores for Text and Emojis
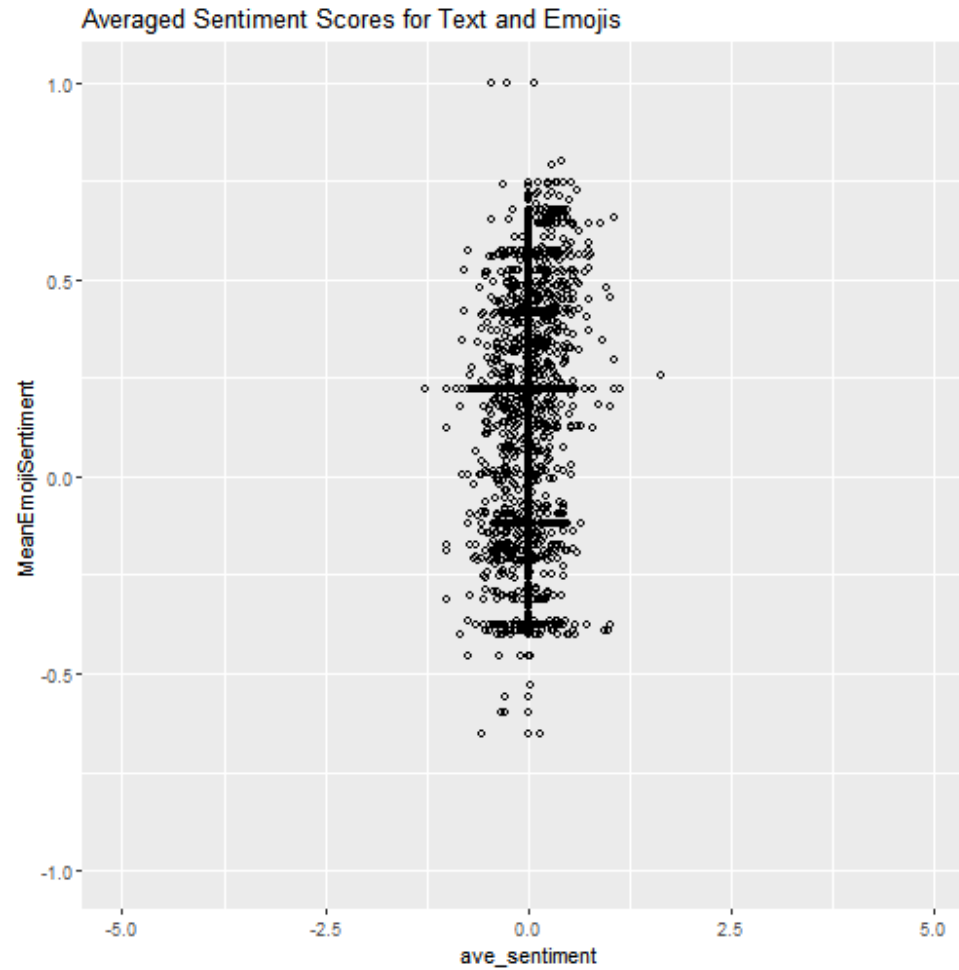
EXAMPLE: SENTIMENT ANALYSIS WITH EMOJIS

→ There seems to be no meaningful relationship between the sentiment scores of the text and the sentiment of the used emojis

→ Possible reasons for this include:

   → Comments that score very high (positive)/low (negative) on emoji sentiment often contain very little text

   → Most comment texts are scored as neutral

   → We only have sentiment scores for the most common emojis

   → Emojis are very much context-dependent, but we only consider a single sentiment score for each emoji

→ Limitations of dictionary-based bag-of-words/-emojis sentiment analysis

WHERE TO GO FROM HERE?

→ To properly include emojis in sentiment analyses, we probably need methods that consider context

→ Large language models (LLMs) are one option for this, and they have been proposed in several preprints as tool for automated text classification (see, e.g., Gilardi et al., 2023, Huang et al., 2023; Rathje et al., 2023; Yang & Menczer, 2023; Ziems et al., 2023)

    → However, it has also been noted that "Automated Annotation with Generative AI Requires Validation" (Pangakis et al., 2023)

    → *Google Colab* notebook to test *ChatGPT* for automated sentiment analysis

→ Of course, which models are used, how they are trained, and what prompts are used, can have a huge impact on the results

→ In the future, multi-modal LLMs (such as GPT-4) that accept images and text as input can further expand the options for sentiment analysis (e.g., of news articles or social media content)