UNIVERSITY OF OULU

Skin Tone Emoji and Sentiment on Twitter

Steven Coats

English Philology, University of Oulu
steven.coats@oulu.fi

3rd DHN Conference, Helsinki
March 7th, 2018

Outline

1. Emoticons and Emoji
2. Skin tone emoji
3. Global distribution of skin tone emoji
4. Skin tone emoji and sentiment
5. Skin tone emoji and word embeddings

Note: This is a .pdf version of html slides that include interactive elements. To access the interactive version, please go to http://cc.oulu.fi/~scoats/dhn18.html

Emoticons and Emoji

- Emoticons: Pictorial representations of (mainly) facial expressions, created with sequences of (mainly) ASCII or Latin-1 characters

  :) :D :-/ :^) (^_^) ʊ_ʊ ʕ •ᴥ•ʔ (͡° ͜ʖ ͡°)

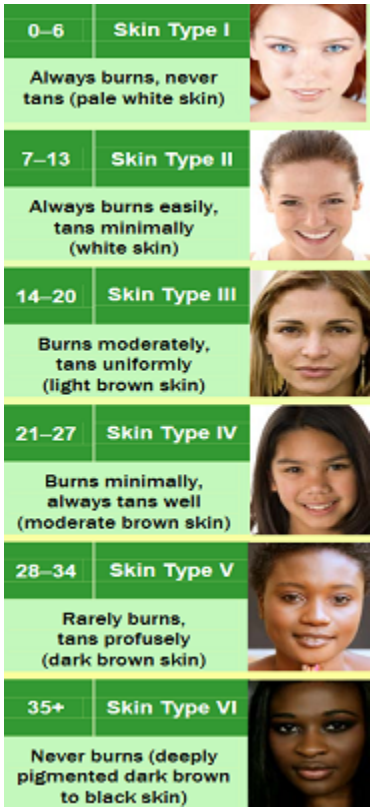
- Emoji: Pictorial representations in graphical form. Origins in Japan in 1990s, introduced into Unicode late 2000s as dedicated code points. Currently 2,789 unique emoji, more with every Unicode update.

Glyphs from twemoji

- Emoji are used in computer-mediated communication in most languages, making them interesting for NLP

Skin tone emoji

Since Unicode 8.0 (June 17, 2015), skin tone characters are part of Unicode



source

Emoji Modifier Fitzpatrick Type-1-2

Emoji Modifier Fitzpatrick Type-3

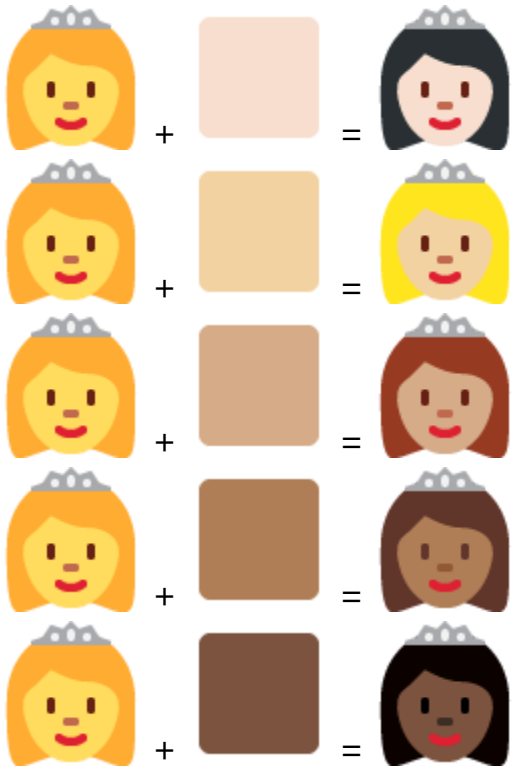
Emoji Modifier Fitzpatrick Type-4

Emoji Modifier Fitzpatrick Type-5

Emoji Modifier Fitzpatrick Type-6

Skin tone emoji use

Skin tone is shown using sequences of Unicode characters



\U0001f478\U0001f3fb

\U0001f478\U0001f3fc

\U0001f478\U0001f3fd

\U0001f478\U0001f3fe

\U0001f478\U0001f3ff

Emoji sequences

Since Unicode 9.0 (late 2016), emoji sequences can also be used to indicate activities, professions, groups, etc. These can usually be combined with skin tone as well.

Sequences can utilize additional **zero-width joiner** and **variation selector** code points to show that the sequence is to be parsed as one character

 = \U0001f9d6\U0001f3fb\U0000200d\U00002642\U0000fe0f

- Parsing and tokenization of emoji sequences can present difficulties

Research questions

How are these skin tone emoji being used globally?

- How often do users select a skin tone variant compared to a default version?
- Which skin tones are being used?

Emoji and sentiment

- What does emoji use tell us about sentiment? (Kralj-Novak et al. 2015)
- Is there a relationship between skin tone emoji and sentiment?

Meanings associated with individual emoji types

- Do skin tone emoji types exhibit similar semantic profiles?
- Word embeddings to explore (skin tone) emoji meanings

Data collection and methods

- 653,457,659 tweets with *place* attributes collected from Twitter's Streaming API from November 2016 – June 2017 (retweets excluded)
- Dictionary of **potential** skin tone emoji used to count occurrences of the types that can be modified with skin tone and their values
- Median and average skin tone values per country calculated (1 = Emoji Modifier Fitzpatrick Type-1-2, 5 = Emoji Modifier Fitzpatrick Type-6)
- Overview of skin tone use geographically, correlation of skin tone and sentiment (using Kralj-Novak et al. sentiment dictionary), word embeddings to investigate skin tone emoji meanings (tokenization issue)

Global skin tone emoji summary statistics

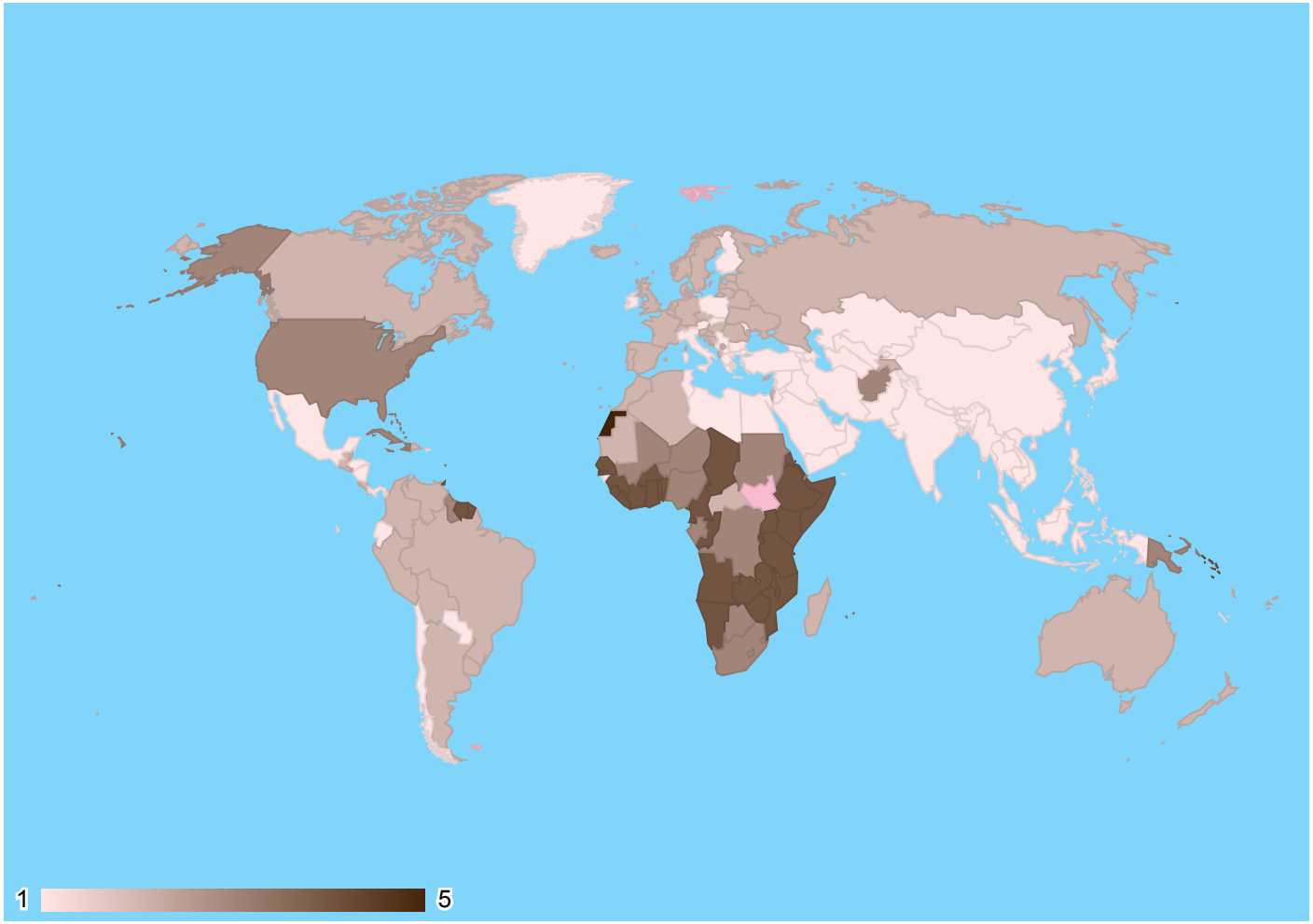
Show 247 entries Search:

Country	nTweets	Pot_skin_tone	nSkintone	propSkinton	avg	med
Andorra	26072	2771	909	0.33	1.48	1
United Arab Emirates	2768770	192935	121591	0.63	1.75	1
Afghanistan	25959	871	482	0.55	3.01	3
Antigua and Barbuda	87015	2628	1988	0.76	3.36	3
Anguilla	3238	247	94	0.38	2.88	3
Albania	31198	1185	477	0.4	1.91	2
Armenia	44311	2534	384	0.15	1.58	1
Angola	105193	12125	8462	0.7	3.47	4
Antarctica	19398	53	16	0.3	2.44	3
Argentina	20023675	1836721	290912	0.16	1.87	2
American Samoa	15274	1722	1317	0.76	2.15	2

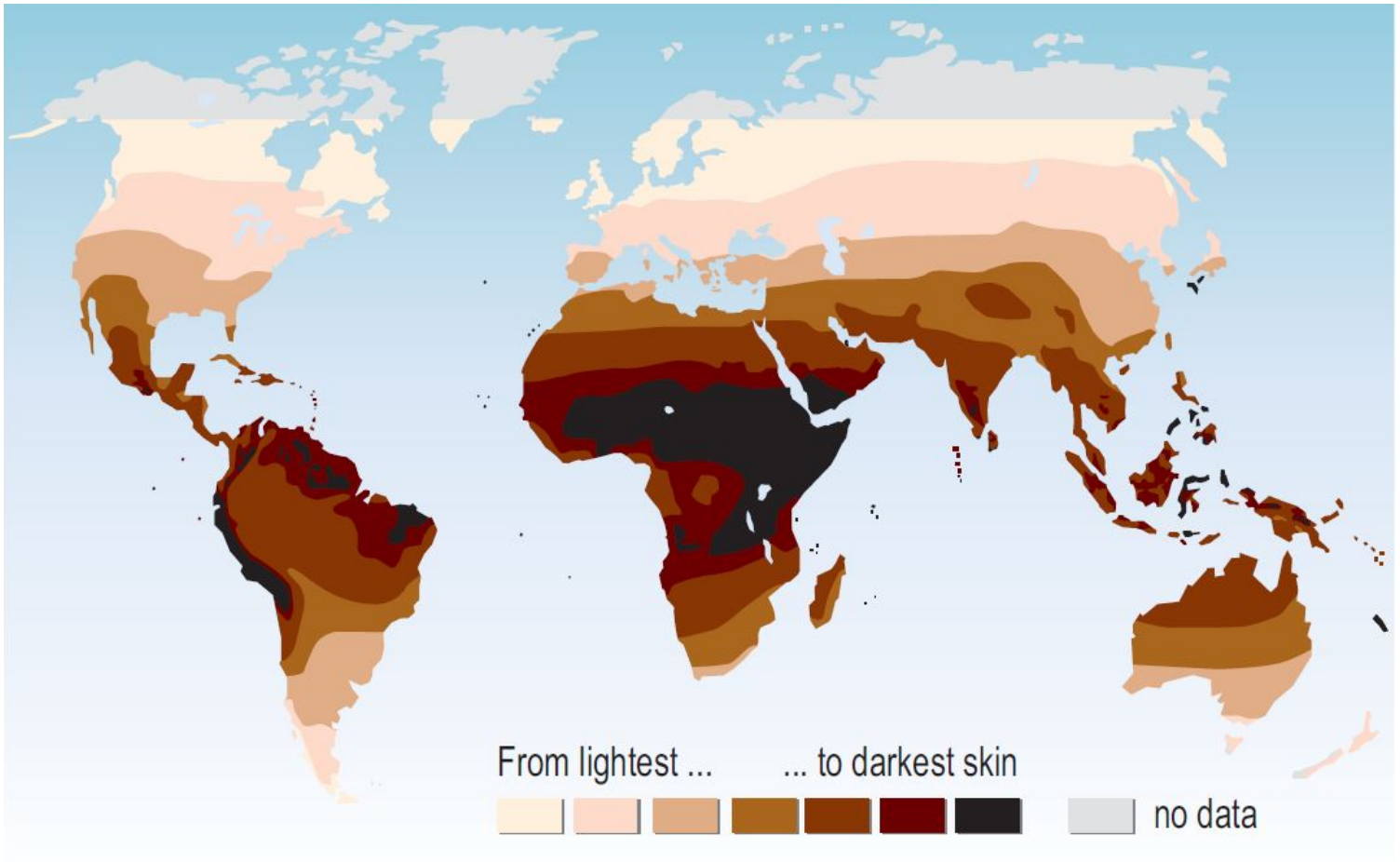
Showing 1 to 247 of 247 entries Previous 1 Next

Proportion of potential skin tone emoji assigned skin tone

Median skin tone values



Global distribution of skin colors



source

Emoji sentiment rankings (Kralj-Novak et al. 2015)

- L1 Annotators categorized tweets containing emoji in 13 European languages as "negative", "neutral", or "positive"
- Aggregate statistics were used to assign sentiment values to individual emoji
- This dictionary was applied to evaluate sentiment in my data

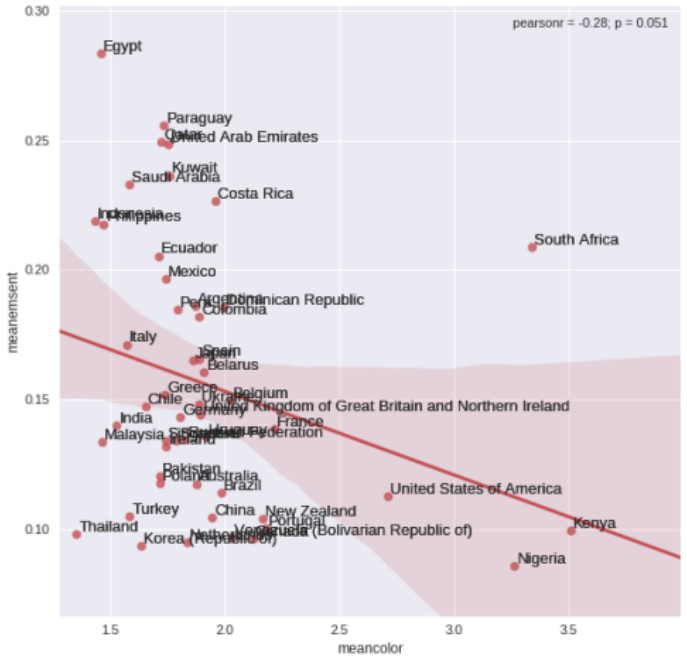
Emoji sentiment rankings (Kralj-Novak et al. 2015)

Calculation of mean sentiment by country/territory for tweets with potential skin tone emoji

- ~25m tweets with potential skin tone emoji
- Tweets stripped of usernames, URLs, and hashtags, then tokenized
- Mean values per country/territory

320582740	GB	en	@JoyboySj hi Steve 👰yeah busy but ok thanks, b...	0.491525	[hi, steve, 👰, yeah, busy, but, ok, thanks, ,,...
320582798	CM	und	🤞🏽😎	0.957515	[🤞🏽, 😎]
320582804	BR	pt	VIADOS parem de idolatrar #Lula e essa merda d...	1.229623	[viados, parem, de, idolatrar, e, essa, merda,...
320582810	EG	ar	🤞🏻ما يمكن خير https://t.co/RKhixSSvAN	0.464146	[🤞🏻, ما, يمكن, خير]
320582815	BR	pt	@Chellepvcs @bdlins12 não quero mt assunto n ..👆	0.522114	[não, quero, mt, assunto, n, .., 👆]
320582843	BR	pt	Que dor de cabeça 👎👎	-0.379845	[que, dor, de, cabeça, 👎, 👎]

Correlation of tweet sentiment with skin tone emoji



Word embeddings

- Distributional hypothesis (Harris 1968)
- Collocational information can be represented with vectors of co-occurrence probabilities
- Similarity of collocational context (and thus meaning) for any two types can be quantified using cosine similarity

- For types A and B, corresponding to vectors $\mathbf{A}$ and $\mathbf{B}$:

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

- Based on a span of 5 tokens to the left and right and at least 10 occurrences in the tweets with potential skin tone

Cosine similarity to skin emoji codepoints, 1,000 most frequent types in corpus[1]

[1]~25m tweets with potential skin tone emojis.

Tokenizing emoji sequences

- Python script incorporating elements from nltk.tokenize.casual, tinysegmenter for Japanese, jieba for Mandarin, and emojione data for sequences

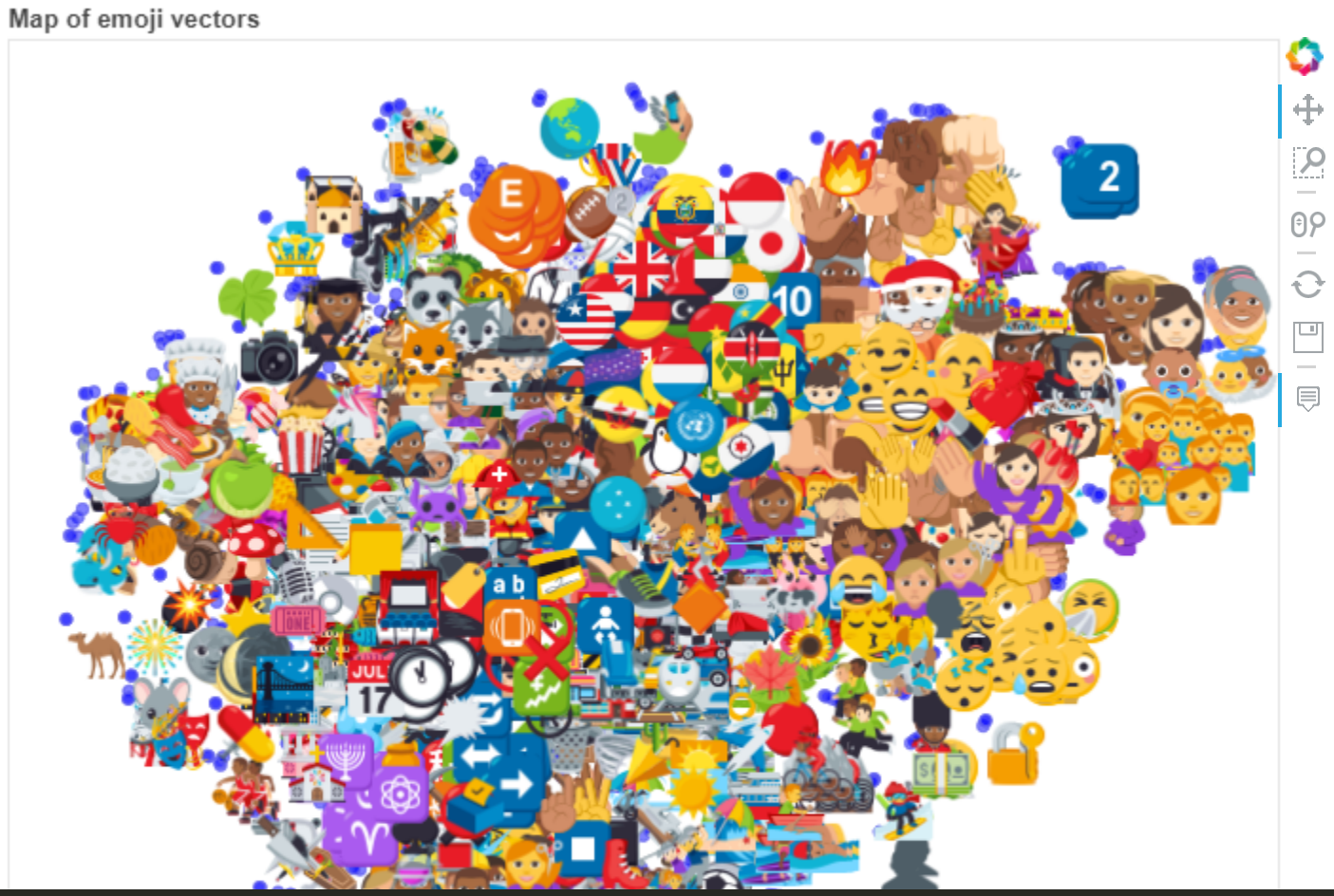
'This is a tokenizer test 👍🏿🇸🇪'

- ['This', 'is', 'a', 'tokenizer', 'test', '👍', '🏿', '🇸', '🇪'] = bad!
- ['This', 'is', 'a', 'tokenizer', 'test', '👍🏿', '🇸🇪'] = good!

'私は絵文字が大好き！🙋🏻'

- ['私は絵文字が大好き', '！', '🙋', '🏻'] = bad!
- ['私', 'は', '絵文字', 'が', '大好き', '！', '🙋🏻'] = good!

400 to 2 dimensions: t-SNE (van der Maaten and Hinton 2008) of emoji vectors



Summary

- About half of emojis that can take skin tone have skin tone – more popular in Anglophone countries
- Lighter skin tone emoji are favored in Asia, the Middle East, and parts of Latin America
- Negative correlation between sentiment and darker skin tone (Helliwell et al. 2017, Ljubešić and Fišer 2016)
- Light median skin tone values: prevailing cultural standards? (Peltzer et al. 2016; Swami et al. 2008, Li et al. 2008, Sahay & Piran 1997)
- Darker skin tone emoji are closer in meaning to informal (AAVE) English lexical items
- Kral-Novak et al. sentiment dictionary only up to Unicode 6.0 (2014), only European languages, low levels of inter-annotator agreement
- More detailed examination possible (USA, Europe, Nordics, E. Asia, e.g.)
- Analysis of evaluative use or correlation with affective language (swearing/profanity)

Thank you!

UNIVERSITY OF OULU

References

Davis, M., and Edberg, P. (2015). Unicode emoji (Unicode Technical Standard #51).

Harris, Z. (1968). *Mathematical structures of language*. New York: Interscience.

Helliwell, J. F., Huang, H., and Wang, S. (2017). The social foundations of world happiness. In: Helliwell, J. F., Layard, R., and Sachs, J. (eds.), *Word Happiness Report 2017*. New York: Columbia University Center for Sustainable Development.

Kralj-Novak, P., Smailovic, J., Sluban, B., and Mozetic, I. (2015). Sentiment of emojis. *PLoS ONE* 10(12).

Li, E., Min, H., Belk, R., Kimura, J., and Bahl, S. (2008). Skin lightening and beauty in four Asian cultures. In: Lee, A., and Soman, D. (eds.), *Advances in Consumer Research Volume 35*, pp. 444–449. Duluth, MN: Association for Consumer Research.

Ljubešić, N., and Fišer, D. (2016). A global analysis of emoji usage. In: *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pp. 82–89. Stroudsburg, PA: Association for Computational Linguistics.

van der Maaten, L., and Hinton, G. (2008). Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research* 9:2579–2605.

Peltzer, K., Pengpid, D., and James, C. (2016). The globalization of whitening: prevalence of skin lighteners (or bleachers) use and its social correlates among university students in 26 countries. *International Journal of Dermatology* 55(2), 165–172.

Sahay S., and Piran, N. (1997). Skin-color preferences and body satisfaction among South Asian-Canadian and European-Canadian female university students. *Journal of Social Psychology* 137(2), 161–171.

Swami, V., Furnham, A., and Joshi, K. (2008). The influence of skin tone, hair length, and hair colour on ratings of women's physical attractiveness, health and fertility. *Scandinavian Journal of Psychology* 49, 429–437.