How Gender and Skin Tone Modifiers Affect Emoji Semantics in Twitter

Francesco Barbieri
LASTUS Lab, TALN
Universitat Pompeu Fabra
Barcelona, Spain
name.surname@upf.edu

Jose Camacho-Collados
School of Computer Science and Informatics
Cardiff University
United Kingdom
camachocolladosj@cardiff.ac.uk

Abstract

In this paper we analyze the use of emojis in social media with respect to gender and skin tone. By gathering a dataset of over twenty two million tweets from United States some findings are clearly highlighted after performing a simple frequency-based analysis. Moreover, we carry out a semantic analysis on the usage of emojis and their modifiers (e.g. gender and skin tone) by embedding all words, emojis and modifiers into the same vector space. Our analyses reveal that some stereotypes related to the skin color and gender seem to be reflected on the use of these modifiers. For example, emojis representing hand gestures are more widely utilized with lighter skin tones, and the usage across skin tones differs significantly. At the same time, the vector corresponding to the male modifier tends to be semantically close to emojis related to business or technology, whereas their female counterparts appear closer to emojis about love or makeup.

1 Introduction

Gender and race stereotypes are still present in many places of our lives. These stereotype-based biases are directly reflected on the data that can be gathered from different sources such as visual or textual contents. In fact, it has been shown how these biases can lead to problematic behaviours such as an increase in discrimination (Podesta et al., 2014). These biases have already been studied in diverse text data sources (Zhao et al., 2017), and have been proved to propagate to supervised and unsupervised techniques learning from them, including word embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017; Zhao et al., 2018) and end-user applications like online ads (Sweeney, 2013).

In this paper we study the biases produced in a newer form of communication in social media from an analytical point of view. We focus on the use of emojis and their interaction with the textual content within a social network (i.e. Twitter). We study emojis as another part of the message, as it could be words. An interesting feature about emojis, apart from their increasing use in diverse social media platforms, is that they enable us to numerically measure some biases with respect to gender and race. Recently, emojis have introduced modifiers as part of their encoding. With these modifiers the same emoji can be used with different features: as male or female, or with different skin colors, for example.

We approach the problem from two methodological perspectives. First, we analyze the use of emojis and their modifiers from a numerical point of view, counting their occurrences in a corpus. This already gives us important hints of how these emojis are used. Then, we leverage the SW2V (*Senses and Words to Vectors*) embedding model (Mancini et al., 2017) to train a joint vector space in which emojis and their modifiers are encoded together, enabling us to analyze their semantic interpretation. While there have been approaches attempting to model emojis with distributional semantics (Aoki and Uchida, 2011; Barbieri et al., 2016; Eisner et al., 2016; Ljubešic and Fišer, 2016; Wijeratne et al., 2017), to the best of our knowledge this is the first work that semantically analyzes modifiers as well. In fact, even though the information provided by modifiers can be extremely useful (for the modeling of emojis in particular, and of messages in social media in general), this has been neglected by previous approaches modeling and predicting emojis (Barbieri et al., 2017; Felbo et al., 2017).

Following our two complementary methodological perspectives, we reached similar conclusions: many stereotypes related to gender and race are also present in this new form of communication.

*Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 101–106
New Orleans, June 5-6, 2018.

Figure 1: Recent tweets using the dark fist emoji (dark skin color in the first and medium-dark in the second).

Moreover, we encountered other interesting findings related to the usage of emojis with respect to gender and skin tones. For instance, our analysis revealed that light skin tones are more widely used than dark ones and their usage is different in many cases. However, incidentally the dark raised fist emoji (i.e. ✊🏿) is significantly more used, proportionally, than its lighter counterparts. This is mainly due to the protest black community started in favour of human rights, dating back from the Olympic Games of Mexico 1968. This sign was known as the *Black Power salute* (Osmond, 2010) and is still widely used nowadays, especially in social media symbolized by the above-mentioned dark-tone fist emoji. Figure 1 shows two recent tweets using this emoji as a response to some Donald Trump critics to NFL players protesting for black civil rights by kneeing during the national anthem before their games. As far as gender-based features are concerned, female modifiers appear much closer to emojis related to love and makeup, while the male ones are closer to business or technology items.

2 Methodology

For this paper we make use of the encoding of emoji modifiers (Section 2.1) and exploit an embedding model that enables us to learn all words, emojis and modifiers in the same vector space (Section 2.2).

2.1 Emoji Modifiers

Emoji modifiers are features that provide more precise information of a given emoji. For example, a hand-based emoji (e.g. 🖐) can have different skin colors: light, medium-light, medium, medium-dark, or dark. This information has been recently added in the official encoding of emojis[1]. At the same time, some emojis like a person rising a hand could be displayed as a woman (i.e. 💁) or a man (i.e. 💁‍♂️). We exploit this information provided by modifiers to study the role of gender and skin color in social media communication.

2.2 Joint Vector Space Model

We construct a vector space model in which words, emojis and their modifiers share the same space. To this end, we exploit SW2V[2] (Mancini et al., 2017), which is an extension of Word2Vec (Mikolov et al., 2013) and was originally designed for learning word and sense embeddings on the same vector space. Given an input corpus, SW2V trains words and its associated senses simultaneously, exploiting their intrinsic connections. In our work, however, we are not interested in learning embeddings for senses but for emojis and their modifiers.

Formally, we use the SW2V model by extending the input and output layers of the neural network with emoji modifiers. The main objective function of the CBOW architecture of Word2Vec aiming at predicting the target word in the middle does not change, except when the model has to predict an emoji with its modifier(s). In this case, instead of simply trying to classify the word in the middle, we also take into account the set of associated emojis. This is equivalent to minimizing the following loss function:

$$-\log(p(e_t|E^t, M^t)) - \sum_{m \in M_t} \log(p(m|E^t, M^t))$$

where M_t refers to the set of modifier(s) of the target emoji e_t. $E^t = w_{t-n}, ..., w_{t-1}, w_{t+1}, ..., w_{t+n}$ and $M^t = M_{t-n}, ..., M_{t-1}, M_{t+1}, ..., M_{t+n}$ both represent the context of the target emoji. While E^t includes surface words (w_i) as context, M^t includes the modifiers of the emojis (M_i) within the surrounding context, if any[3].

The resulting output is a shared space of word, emoji and modifier embeddings. In addition, we

[1] http://unicode.org/reports/tr51/#Emoji_Modifiers_Table

[2] http://lcl.uniroma1.it/sw2v/

[3] M^t may be empty if no modified emoji occurs in the context of the target emoji.

propose a second variant[4] of the SW2V architecture modeling words, non-modified emojis and emojis associated with their modifiers (e.g. 👌). For example, for the emoji black hand (🖐) this variant would learn the embedding for the hand without any modifier, and the same emoji with the modifier *dark* (i.e. 🏿) instead of the embedding for the modifier alone learned in the main configuration of our SW2V model.

The advantages of using this model with respect to a usual word embedding model are manifold: first, it enables us to separate modifiers from emojis so we can learn accurate representations for both types; second, with this model we can learn embeddings for words, emojis and their modifiers in the same vector space, a property that is exploited in our experiments; third, since an emoji with modifiers may occur quite infrequently, by using this approach we take into account the semantic of the emoji (e.g. 🖐) so the representation of the emoji with their modifiers (e.g. 🖐🏻, 🖐🏿...) is more accurate; finally, with this model we can associate a given emoji with one or more modifiers (e.g. skin color and gender on the same emoji).

3 Experiments

All our experiments are carried out on a corpus compiled from Twitter, including all tweets geolocalized in United States from October 2015 to January 2018. The corpus contains over 22M tweets and around 319M tokens overall. In the corpus we encode emojis and their modifiers as single joint instances. Taking this corpus as reference, we inspect the use of emojis with respect to skin tone and gender from two complementary methodological perspectives: frequency-based (Section 3.1) and semantics-based (Section 3.2).

3.1 Frequency

By exploring the frequency of emojis in Twitter we can obtain a clear overview of their diverse use regarding skin tone. To this end, we carried out a frequency analysis on hand-related emojis with different skin color modifiers: light, medium-light, medium, medium-dark, dark, and neutral (i.e. no modifier). Table 1 shows the frequency of the top twenty most frequent hand-related emojis according to skin tone. As can be clearly seen, the emojis without any particular skin tone modifier (yellow), which are displayed by default, are the most frequent. However, it is surprising to note the gap between the usage of the light-tone emojis (over 70K occurrences with over almost 17% overall) with respect to dark-tone emojis (less than 8K occurrences which corresponds to less than 2% overall). Nevertheless, this gap may be simply due to demographics, since many Twitter users employ modifiers as a form of self-representation (Robertson et al., 2018).

	No mod	🏻	🏼	🏽	🏾	🏿
Abs	121,343	70,139	102,397	61,865	50,871	7,621
Rel	29.3	16.9	24.7	14.9	12.2	1.8

Table 1: Absolute and relative (%) frequency of hand-related emojis. These frequency estimators indicate the number of tweets where an emoji occurs, without considering repetitions.

In addition to the raw frequencies of these emojis and their modifiers we analyze how these emojis were used proportionally for each skin tone. Table 2 displays the proportion of emojis used per skin color. Interestingly, the pattern followed by the darker emojis is clearly different from the distribution followed by lighter ones (Pearson correlation of 98% between light and medium-light tones in comparison to the relatively low 71% between light and dark tones). For example, the emoji corresponding to the raised fist (i.e. ✊) is significantly more used for the dark tone than the light ones (10.3% to 1.6%). The reason, as explained in the introduction, dates back from the Olympic Games of 1968 (Osmond, 2010). It represents the fight of the black community for human rights, which is still present nowadays, as highlighted in the recent tweets of Figure 1. Additionally, the hand emoji representing the middle finger raised (i.e. 🖕), which is often used as an insult, occurs proportionally significantly more often with the dark skin color (2.2% to 0.5%). In contrast, light skin tone emojis tend to be more used for emojis including some form of assertion: e.g. 👌 (12% vs 6.7%), and the 👍🏻 (7.8% vs 3.7%).

3.2 Semantics

For inspecting the semantics of each emoji and its modifiers we rely on the joint semantic vector space (SW2V) of words, emojis and modifiers described in Section 2.2. We ran SW2V in our Twitter corpus with the following hyperparame-

[4]We use this second variant in our last semantics-based experiment in Section 3.2.2.

🙌	13.0	15.7	14.9	16.1	17.5	12.2
💪	12.9	8.8	11.3	13.7	14.2	13.5
👌	12.5	12.0	10.9	9.0	7.8	6.7
🙏	12.0	9.9	9.0	15.0	18.5	20.0
✌	11.4	12.8	15.0	10.6	8.8	8.6
👍	10.3	7.8	5.3	3.5	2.9	3.7
🤘	6.0	14.1	14.3	11.0	7.8	8.3
👊	4.6	4.5	4.6	3.9	2.6	3.5
👏	4.1	3.9	3.3	3.0	2.5	2.2
👋	2.4	3.1	2.9	2.6	2.0	1.4
👉	2.3	1.1	1.0	1.4	1.5	1.7
✊	2.2	1.6	2.4	4.4	7.7	10.3
☝	2.0	1.3	1.7	1.7	1.7	2.1
👆	1.0	0.8	0.8	0.9	0.9	0.9
👐	0.8	0.4	0.6	0.7	0.6	0.7
✋	0.7	0.6	0.5	0.5	0.4	0.7
👇	0.6	0.2	0.3	0.4	0.5	0.4
🖕	0.5	0.8	0.7	1.2	1.6	2.2
👎	0.5	0.3	0.3	0.2	0.2	0.3
👈	0.5	0.3	0.2	0.4	0.5	0.7

Table 2: Relative frequency (%) of the top twenty hand-related emojis with respect to skin modifiers (from left to right: no modifier, light, medium-light, medium, medium-dark, dark).

					♂	♀
💕	💕	😏	💯	✊	👍	👼
👯	👯	💯	🗣	💯	💦	😉
🎩	💙	😈	💽	💰	👜	🤳
💗	👬	😛	😬	🔒	🎯	🤰
💝	💛	💋	🇭🇹	🏚	😎	💅
😊	🏐	😌	🔌	🔋	💼	🙈
✨	💗	👯	📶	🔌	🕴	😜
😌	💟	🔒	💸	💽	💭	👧
👭	😛	😎	😤	⬛	💨	😊
💕	💘	🔐	™	😤	👼	👠
🐻	💚	💕	🔋	💸	👌	😏
👭	👬	😜	💰	🏓	🍾	😌
💙	💞	🔑	🏧	🏧	🤝	👴
💖	😌	💞	🔒	💹	💪	💦
🌈	🎁	🎯	❌	🇭🇹	🏚	🎀

Table 3: Fifteen emoji nearest neighbours of the seven modifiers in our analysis.

ters: 100 dimensions and window size of 6 tokens. We performed two kinds of experiment: one relying on the nearest neighbours in the vector space to understand the main semantics of skin tone and gender modifiers (Section 3.2.1) and another experiment in which we analyze the main semantic divergences between opposing modifiers (Section 3.2.2), i.e. dark vs. light (skin tone) and male vs. female (gender).

3.2.1 Nearest Neighbours

For this experiment we analyze the nearest neighbours of skin tone and gender modifiers in the SW2V vector space using cosine similarity as comparison measure. Table 3 shows the fifteen nearest neighbours for the five skin tone and two gender modifiers. For the skin tone modifiers it is noteworthy the fact that while lighter tones contain love-related emojis as nearest neighbours, these do not appear on the list of darker tones. Instead, we can see some money-related (e.g. 💰, 💸 or 🏧) and electric-related emojis (e.g. a battery 🔋 or a plug 🔌) as nearest neighbours of dark tone emojis. These two electric emojis are often used in the context of music, sport or motivational tweets along with hashtags like *#energy* or *#chargedup* (e.g. *The GRIND begins!!! Refuse to settle for average!!* 🔋 ⚡ 🤵 *#chargedup*). As a possibly more worrying trend we found many versions of the (often derogatory) word *nigger* and *gang* as nearest neighbours of dark tone modifiers. A more focused analysis on this issue would be required in order to understand the possible racist implications.

As far as gender modifiers are concerned, business-related emojis (e.g. a briefcase 💼, a suit 🕴 or a handshake 🤝) are among the closest emojis to the man modifier in the SW2V vector space, while nail polishing (i.e. 💅) or the selfie emoji (i.e. 🤳), for example, are among the nearest neighbours of the female modifier.

3.2.2 Semantic Divergences

In addition to the nearest neighbours experiments, we analyze the highest semantic similarity gap between skin tone and gender modifiers. In Table 4 we display in each row the emojis with the highest similarity gap with respect to the opposite modifier (light vs. dark and male vs. female), being more similar to the corresponding modifier row. In this case we can see a similar pattern as in the nearest neighbours experiment. A money-related emoji appears again semantically close to the dark-skin modifier (💰) but far from the light skin modifier, and love-related emojis closer to the light skin modifier (e.g. 💙, 💕 and 💓). Likewise,

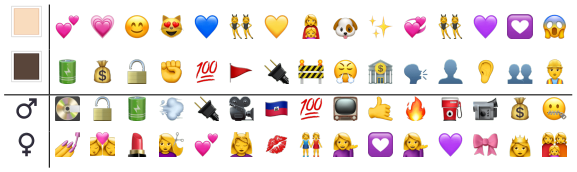

Table 4: Emojis with highest similarity gap between opposite modifiers (light *vs* dark, male *vs* female).

we can see how technology-related emojis (e.g. a CD, a video camera or a television) are close to the man modifier and far from the female one. In contrast, makeup-related emojis like nail polishing (i.e. 💅) or the lipstick emoji (i.e. 💄) are clearly female-based.

In order to complement this experiment, we also inspect the emojis whose similarity was lower when changing the modifier.[5] We compare the similarity between all emojis which can have a skin color or gender modifier. Table 5 shows the fifteen emojis whose semantic similarity, as measured by cosine similarity, was lower by switching to the corresponding opposite modifier. The first surprising finding that arises is the low similarity values (negative values lower than -0.6 in some cases), considering that the only change is the modifier, while the emoji does not change. The emojis that change most when switching the skin tone are in the main hand gestures. Conversely, the emojis that change most when switching the gender modifier are people in job roles such as detective (i.e. 🕵), judge (i.e. 💂), police officer (i.e. 👮) or teacher (i.e. 👩‍🏫). From these four items only the teacher emoji is closer to the female modifier, while the other three are closer to the male modifier. In contrast, emojis referring to other jobs like fireman (i.e. 👨‍🚒), artist (i.e. 👨‍🎨) or singer (i.e. 👩‍🎤) do not seem to considerably change their meaning when switching their gender.

4 Conclusion

In this paper we have studied the role of gender and skin tone in social media communication through emojis. Thank to the modifiers associated with different emojis and the usage of a joint semantic vector space of words, emojis and modifiers, we were able to model the semantics of emojis with respect to gender and skin tone features[6].

Skin Tone		Gender	
🙌	-0.621	👨	-0.422
👐	-0.601	👩	-0.346
👌	-0.590	👴	-0.331
🤙	-0.541	👱‍♀️	-0.289
✋	-0.535	💂‍♀️	-0.277
👶	-0.490	👷‍♂️	-0.222
👮‍♂️	-0.427	🧚‍♀️	-0.195
👍	-0.409	👱‍♂️	-0.191
💪	-0.388	👮‍♀️	-0.185
🤘	-0.375	👰	-0.174
🙆‍♀️	-0.374	👱‍♂️	-0.169
👂	-0.366	🧜‍♀️	-0.144
🎅	-0.349	🕵️‍♂️	-0.127
👈	-0.347	💂‍♂️	-0.117
👴	-0.344	👯‍♂️	-0.114

Table 5: Emojis with lowest similarity using opposite modifiers (light *vs* dark, male *vs* female).

Our analysis on a corpus of tweets geolocalized in United States reveals clear connotations associated with each gender. For example, male modifiers being much closer to business and technology while female ones are often associated with love and makeup. Other connotations are present with respect to the skin color, being dark tone hand emojis more associated with derogatory words and emojis[7]. In a more general perspective, these modifiers clearly increase the ambiguity of emojis, which were already shown highly ambiguous in many cases (Wijeratne et al., 2016; Miller et al., 2017). In fact, modifiers can render emoji meanings very far apart, as clearly showed in Table 5.

While in this work we have approached the problem from a purely analytical point of view, our work can also be viewed as a starting point for the development of accurate education guidelines that could contribute to a reduction of gender- and race-associated stereotypes in society. Additionally, the understanding of emoji semantics provided in our analysis paves the way for the development of debiasing techniques to be leveraged on supervised and unsupervised models which make use of social media data, in the lines of Bolukbasi et al. (2016) and Zhao et al. (2017).

[5]For this last experiment we used the SW2V variant in which emojis with their modifiers are included in the vector space (cf. Section 2.2).

[6]Code and SW2V embeddings are available at https://github.com/fvancesco/emoji_modifiers

[7]This goes in line with some previous findings about the use of modifiers in other platforms such as Apple: goo.gl/UalXoK

Acknowledgments

We thank Roberto Carlini for the initial idea of the paper, and Luis Espinox-Anke for fruitful discussions on this topic. Francesco B. acknowledges support from the TUNER project (TIN2015-65308-C5-5-R, MINECO/FEDER, UE) and the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502).

References

Sho Aoki and Osamu Uchida. 2011. A method for automatically generating the emotional vectors of emoticons using weblog articles. In *Proceedings of the 10th WSEAS International Conference on Applied Computer and Applied Computational Science, Stevens Point, Wisconsin, USA*, pages 132–136.

Francesco Barbieri, Miguel Ballesteros, and Horacio Saggion. 2017. Are emojis predictable? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 105–111.

Francesco Barbieri, German Kruszewski, Francesco Ronzano, and Horacio Saggion. 2016. How cosmopolitan are emojis?: Exploring emojis usage and meaning over different languages with distributional semantics. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 531–535. ACM.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bosnjak, and Sebastian Riedel. 2016. emoji2vec: Learning emoji representations from their description. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 48–54, Austin, TX, USA. Association for Computational Linguistics.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625.

Nikola Ljubešic and Darja Fišer. 2016. A global analysis of emoji usage. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 82–89, Berlin, Germany. Association for Computational Linguistics.

Massimiliano Mancini, Jose Camacho-Collados, Ignacio Iacobacci, and Roberto Navigli. 2017. Embedding words and senses together via joint knowledge-enhanced training. In *Proceedings of CoNLL*, Vancouver, Canada.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Hannah Miller, Daniel Kluver, Jacob Thebault-Spieker, Loren Terveen, and Brent Hecht. 2017. Understanding emoji ambiguity in context: The role of text in emoji-related miscommunication. In *11th International Conference on Web and Social Media, ICWSM 2017*. AAAI Press.

Gary Osmond. 2010. Photographs, materiality and sport history: Peter norman and the 1968 mexico city black power salute. *Journal of Sport History*, 37(1):119–137.

John Podesta, Penny Pritzker, Ernest J. Moniz, John Holdren, and Jefrey Zients. 2014. *Big data: Seizing opportunities, preserving values*. White House, Executive Office of the President.

Alexander Robertson, Walid Magdy, and Sharon Goldwater. 2018. Self-representation on twitter using emoji skin color modifiers. *arXiv preprint arXiv:1803.10738*.

Latanya Sweeney. 2013. Discrimination in online ad delivery. *Queue*, 11(3):10.

Sanjaya Wijeratne, Lakshika Balasuriya, Amit Sheth, and Derek Doran. 2016. Emojinet: Building a machine readable sense inventory for emoji. In *International Conference on Social Informatics*, pages 527–541. Springer.

Sanjaya Wijeratne, Lakshika Balasuriya, Amit Sheth, and Derek Doran. 2017. A semantics-based measure of emoji similarity. In *Proceedings of the International Conference on Web Intelligence*, pages 646–653. ACM.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of NAACL*, New Orleans, LA, United States.