Interpreting Emoji with Emoji:
🏖️ ⇒ 🌞😎🌊

Jens Helge Reelfs[1], Timon Mohaupt*[2], Sandipan Sikdar[3], Markus Strohmaier[4], Oliver Hohlfeld[1]
[1] Brandenburg University of Technology, [2] Talentschmiede Unternehmensberatung AG
[3] RWTH Aachen University, [4] University of Mannheim, GESIS, CSH Vienna
reelfs@b-tu.de, timon.mohaupt@me.com, sandipan.sikdar@cssh.rwth-aachen.de,
markus.strohmaier@uni-mannheim.de, hohlfeld@b-tu.de

Abstract

We study the extent to which emoji can be used to add interpretability to embeddings of text and emoji. To do so, we extend the POLAR-framework that transforms word embeddings to interpretable counterparts and apply it to word-emoji embeddings trained on four years of messaging data from the Jodel social network. We devise a crowdsourced human judgement experiment to study six use-cases, evaluating against words only, what role emoji can play in adding interpretability to word embeddings. That is, we use a revised POLAR approach interpreting words and emoji with words, emoji or both according to human judgement. We find statistically significant trends demonstrating that emoji can be used to interpret other emoji very well.

1 Introduction

Word embeddings create a vector-space representation in which words with a similar meaning are in close proximity. Existing approaches to make embeddings interpretable, e.g., via contextual (Subramanian et al., 2018), sparse embeddings (Panigrahi et al., 2019), or learned (Senel et al., 2018) transformations (Mathew et al., 2020)—all focus on text only. Yet, emoji are widely used in casual communication, e.g., Online Social Networks (OSN), and are known to extend textual expressiveness, demonstrated to benefit, e.g., sentiment analysis (Novak et al., 2015; Hu et al., 2017).

Goal. We raise the question if we can leverage the expressiveness of emoji to make word embeddings—and thus also emoji—interpretable. I.e., can we adopt word embedding interpretability via leveraging semantic polar opposites (e.g., cold / hot) to emoji (e.g., ❄️ / 🌞, or 😡 / 😊) for interpreting words or emoji w.r.t. human judgement.

*Timon Mohaupt performed this work during his master thesis at Brandenburg University of Technology and RWTH Aachen University.

Approach. Motivated and based upon POLAR (Mathew et al., 2020), we deploy a revised variant POLAR$^{\rho}$ that transforms arbitrary word embeddings into interpretable counterparts. The key idea is to leverage semantic differentials as a psychometric tool to align embedded *terms* on a scale between two polar opposites. Employing a projection-based transformation in POLAR$^{\rho}$, we provide embedding dimensions with semantic information. I.e., the resulting interpretable embedding space values directly estimate a *term*'s position on a-priori provided polar opposite scales, while approximately preserving in-embedding structures (§ 2).

The main contribution of this work is the large-scale application of this approach to a social media corpus and especially its evaluation in a crowdsourced human judgement experiment. For studying the role of emoji in interpretability, we create a word-emoji input embedding from on a large social media corpus. The dataset comprises four years of complete data in a single country from the online social network provider Jodel (48M posts of which 11M contain emoji). For subsequent main evaluation, we make this embedding interpretable with word and emoji opposites by deploying our adopted tool POLAR$^{\rho}$ (§ 3).

Given different expressiveness of emoji, we ask **RQ1)** How does adding emoji to POLAR$^{\rho}$ impact interpretability w.r.t. to human judgement? I.e., do humans agree on best interpretable dimensions for describing words or emoji with word or emoji opposites? And **RQ2)** How well do POLAR$^{\rho}$-semantic dimensions reflect a *term*'s position on a scale between word or emoji polar opposites?

Human judgement. We design a crowdsourced human judgement experiment (§ 4) to study if adding emoji to word embeddings and POLAR$^{\rho}$ in particular increases the interpretability—while also answering how to describe emoji best. Our human judgement experiment involves six campaigns explaining *Words* (W/*) or *Emoji* (E/*) with *Words*,

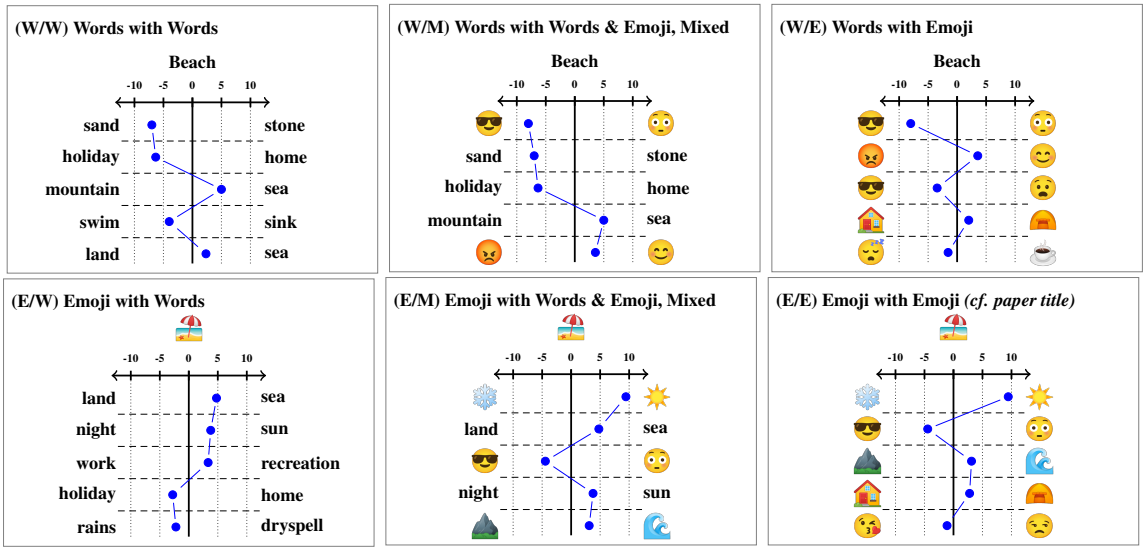


Figure 1: The POLAR-framework (Mathew et al., 2020) makes word embeddings interpretable leveraging polar opposites. It provides a new interpretable embedding subspace with systematic polar opposite scales: Along six use-cases, we evaluate which role emoji expressiveness plays in adding interpretability to word embeddings. I.e., how well can our adopted POLAR$^\rho$ interpret (W/*) words or (E/*) emoji with words, emoji or both (*/M), *Mixed*. We test POLAR$^\rho$ alignment with human judgement as represented in shown semantic profiles above.

Emoji, or both *Mixed*. We evaluate two test conditions to answer both research questions: *(RQ1)* a *selection* test studies if human subjects agree to the POLAR$^\rho$ identified differentials (e.g., how do emoji affect POLAR$^\rho$ interpretability?), and *(RQ2)* a *preference* test that studies if the direction on a given differential scale is in line with human judgement (e.g., how well does POLAR$^\rho$ interpret scales).

Results. POLAR$^\rho$ identifies the best interpretable opposites for describing emoji with emoji, yet generally aligning well with human judgement. Except interpreting words with emoji only probably due to lack of emoji expressiveness indicated by coder agreement. Further, POLAR$^\rho$ estimates an embedded *terms*' position on a scale between opposites successfully, especially for interpreting emoji.

Broader application. Not all emoji have a universally agreed on meaning. Prior work showed that differences in the meaning of emoji exist between cultures (Guntuku et al., 2019; Gupta et al., 2021). Even within the same culture, ambiguity and double meanings of emoji exist (Reelfs et al., 2020). Currently, no data-driven approach exists to infer the meaning of emoji—to make them interpretable. Our proposed approach can be used to tackle this challenge since it makes emoji interpretable.

2 Creating Interpretable Embeddings

We explain next our deployed tool for creating interpretable word-emoji embeddings: POLAR (Mathew et al., 2020); and provide detail on a revised POLAR extension via projection.

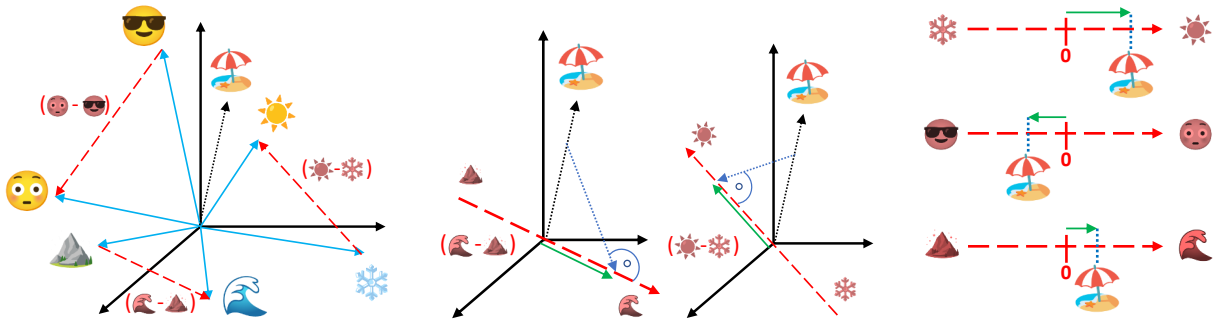
2.1 POLAR Approach

Semantic Differentials. Based upon the idea of semantic differentials as a psychometric tool to align a word on a scale between two polar opposites (Fig. 1), POLAR (Mathew et al., 2020) takes a word embedding as input and creates a new interpretable embedding on a polar subspace. This subspace, i.e., the opposites used for the interpretable embedding are defined by an external source.

That is, starting with a corpus and its vocabulary $\mathcal{V}$, a word embedding created by an algorithm a (e.g., Word2Vec or GloVe) assigns vectors $\overrightarrow{\mathbb{W}_v^a} \in \mathbb{R}^d$ on d dimensions to all words $v \in \mathcal{V}$ according to an optimization function (usually word co-occurrence). This pretraining results in an embedding $\mathbb{D} = \left[\overrightarrow{\mathbb{W}_v^a}, v \in \mathcal{V}\right] \in \mathbb{R}^{|\mathcal{V}| \times d}$.

Such embedding spaces carry a semantic structure between embedded words, whereas the dimensions do not have any specific meaning. However, we can leverage the semantic structure between words to transform the embedding space to carrying over meaning into the dimensions: POLAR uses N semantic differentials/opposites that are itself items within the embedding, i.e., $\mathbb{P} = \left\{(p_z^i, p_{-z}^i), i \in [1..N], (p_z^i, p_{-z}^i) \subseteq \mathcal{V}^2\right\}$.

As shown in Fig. 2a, given two anchor points for each polar opposite, a line between them represents a differential—which we name POLAR direction



(a) Polar Directions/Differentials. (b) Excerpt: Per Direction Orthogonal Projection. (c) Interpretable Scales for 🏖️.

Figure 2: POLAR (Mathew et al., 2020) with Projection in a nutshell: We showcase POLAR$^\rho$ interpreting emoji with emoji (E/E) (cf. paper title). *(a)* We leverage polar opposites (*here:* ❄️/☀️, 😎/😳, 🏔️/🌊) to provide embedding dimensions with semantic information. By using opposite differential directions (red dashed vectors), we create a new interpretable subspace. *(b)* Orthogonal projection (blue dotted vectors) of an embedded term (*here:* 🏖️) onto this subspace (e.g., *left:* 🏔️/🌊, *right:* ❄️/☀️) yields a direct scale measure between both opposites in the adjacent leg (green vectors, directed alike the differential). *(c)* The resulting interpretable embedding now contains a tangible position estimation along employed polar dimensions for each embedded term (*here:* 🏖️).

(red dashed vectors):

$$\overrightarrow{dir_i} = \overrightarrow{\mathbb{W}^a_{p^i_z}} - \overrightarrow{\mathbb{W}^a_{p^i_{-z}}} \in \mathbb{R}^d$$

Base Change. Naturally, we can use these differentials as a new basis for the interpretable embedding $\mathbb{E}$. Gathering all directions in a matrix $dir \in R^{N \times d}$, we obtain for all embedded *terms* $v \in \mathcal{V}$: $dir^T \overrightarrow{\mathbb{E}_v} = \overrightarrow{\mathbb{W}^a_v}$, and ultimately apply a base change transformation $\overrightarrow{\mathbb{E}_v} = (dir^T)^{-1} \overrightarrow{\mathbb{W}^a_v}$ yielding an interpretable subspace along the differentials $\overrightarrow{dir_i}$ that carries over specific geometric semantics from the input embedding. I.e., for each word $v \in \mathcal{V}$ within the resulting interpretable embedding $\mathbb{E}$, its embedding vector $\overrightarrow{\mathbb{E}_v}$ now carries a measure along each polar dimension's semantics.

Limitations. Polar opposites being very close in the original embedding space might tear apart. From a technical perspective, the used pseudo inverse for the base change becomes numerically ill-conditioned if $d \approx N$ (Mathew et al., 2020).

2.2 POLAR$^\rho$ Extension: Projection

While the base change approach seems natural, its given limitations lead us to propose a variant that comes with several benefits. Instead of creating a new interpretable vector space, we take measurements on the differentials dir defined as before (Fig. 2a, red dashed vectors). However, we now project each embedding vector $\overrightarrow{\mathbb{W}_v}$ for v orthogonally onto the differentials as shown in Fig. 2b (blue dotted vectors). This leads to a smallest distance between both lines w.r.t. the differential, yet simultaneously allows for a direct scale measure on the differential vector as shown in Fig. 2b & Fig. 2c (green vectors). Thereby, we also *decouple* the transformation matrix, which eases later add-ins to the interpretable embedding.

Orthogonal projection (blue dotted vectors) of each input embedding vector $\overrightarrow{\mathbb{W}^a_v}$ onto a differential i provides us the adjacent leg vector as follows:

$$\text{oproj}_{dir_i}(\overrightarrow{\mathbb{W}^a_v}) = \underbrace{\frac{\overrightarrow{\mathbb{W}^a_v} \cdot \overrightarrow{dir_i}}{|\overrightarrow{dir_i}|}}_{scalar} \cdot \underbrace{\frac{\overrightarrow{dir_i}}{|\overrightarrow{dir_i}|}}_{direction}$$

As this adjacent leg (green vectors)'s direction naturally equals the differential, we focus only on the scalar part representing a direct scale measure. By normalizing the differential vector lengths $\hat{dir} = dir \cdot |dir|^{-1}$, the projected scale value conveniently results in: $\text{oproj}^{scalar}_{dir_i}(\overrightarrow{\mathbb{W}^a_v}) = \overrightarrow{\mathbb{W}^a_v} \cdot \overrightarrow{\hat{dir}_i}$.

This transformation allows to create a new interpretable embedding $\mathbb{E} \in \mathbb{R}^{|\mathcal{V}| \times N}$ for each embedding vector $\overrightarrow{\mathbb{W}^a_v}$ (exemplified in Fig. 1) as follows:

$$\overrightarrow{\mathbb{E}_v} = \text{oproj}^{scalar}_{dir}(\overrightarrow{\mathbb{W}^a_v}) = \hat{dir}^T \overrightarrow{\mathbb{W}^a_v} \in \mathbb{R}^N$$

Computationally it requires an inital matrix multiplication for each embedded *term*; Dimension increments require a dot product on each *term*.

Downstream Tasks. Other experiments indicate POLAR$^\rho$ downstream task performance being on par with the input embedding, and an edge over base change POLAR if $d \approx N$ (not shown).

2.3 Measuring Dimension Importance

There can be many possible POLAR dimensions, which requires to select the most suitable ones.

That is, we want to define a limited set of opposites that best describes words or emoji w.r.t. interpretability across the whole embedding.

Extremal Word Score (EWSO). We propose a new metric to measure the quality of polar dimensions complementing heuristics from (Mathew et al., 2020). It measures the embedding confidence and consistency along available differentials. The idea of POLAR$^\rho$ is that directions represent semantics within the input embedding. We determine embedded *terms* shortest distance to these axes via orthogonal projection; we use resulting intersections as the position w.r.t. the directions.

That is, as a new heuristic, for each of our differentials dir_i, we look out for $k = 10$ embedded words at the extremes (having the highest scores in each direction) and take their average cosine distance within the original embedding $\mathbb{D}$ to the differential as a measure. This results in the average similarity of existing *extremal* words on our scale—a heuristic that represents the skew-whiffiness within the extremes on a differential scale.

3 Approach: Embedding & Polarization

We next propose an approach to improve the interpretability of word embeddings by adding emoji. It uses our extended version POLAR$^\rho$ and adds emoji to the POLAR space by creating word embeddings that include emoji.

3.1 Data Set

We create a word embedding out of a social media text corpus, since emoji are prominent in communication within Online Social Networks. We decided to use a corpus from the Jodel network, where about one out of four sentences contain emoji (see (Reelfs et al., 2020)).

The Jodel Network. We base our study on a country-wide complete dataset of posts in the online social network Jodel, a mobile-only messaging application. It is location-based and establishes local communities relative to the users' location. Within these communities, users can *anonymously* post photos from the camera app or content of up to 250 characters length, i.e., microblogging, and reply to posts forming discussion threads.

Corpus. The network operators provided us with data of content created in Germany from 2014 to 2017. It contains 48M sentences, of which 11M contain emoji (1.76 emoji per sentence on average).

Ethics. The dataset contains no personal information and cannot be used to personally identify users except for data that they willingly have posted on the platform. We synchronize with the Jodel operator on analyses we perform on their data.

3.2 Semantic Differential Sources

POLAR$^\rho$ can create interpretable embeddings w.r.t. a-priori provided opposites. We next describe how we select these opposites to make POLAR$^\rho$ applicable to our data. Most importantly, the approach requires being part of or locating desired opposites within the original embedding space.

Words. As we extend the word embedding space with emoji, we still want to use words. We find common sources of polar opposites in antonym wordlists (Shwartz et al., 2017) as used in the original POLAR work. To fit our German dataset, we translated and manually checked all pairs keeping 1275 items. From GermaNet (Hamp and Feldweg, 1997), we extracted 1732 word pairs via antonym relations leading to $|\mathbb{P}_{\text{words}}| = 1832$ word pairs.

Emoji. Being not ideal, but due to lack of better alternatives, we ended up heuristically creating semantic opposites from emoji through qualitative surveys across friends and colleagues resulting in $|\mathbb{P}_{\text{emoji}}| = 44$ emoji pairs, cf. Tab. 3. While we could use far more opposites especially of facial emoji, due to emoji clustering in the input embedding, spanned expressive space would arguably become redundant at similar EWSO scores for many directions. Effectively it may bias interpretability over proportionally towards facial emoji.

3.3 Polarization

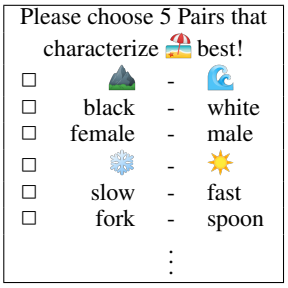
Preprocessing. We tokenize sentences with `spaCy` and remove stopwords. To increase amounts of available data, we remove all emoji modifiers (skin tone and gender): {🤦🏼‍♂️, 🤦‍♀️, 🤦🏿‍♂️} → 🤦‍♂️. Due to German language, we keep capitalization.

Original Embedding. We use `gensim` implementation of `Word2Vec` (W2V). A qualitative investigation suggests that skip-gram works better than CBOW (better word analogy). We kept training parameters largely at defaults including negative sampling, opting for $d = 300$ dimensions.

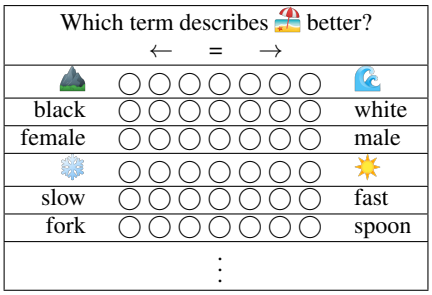
Interpretable Embedding. The actual application of embedding transformation is simple. We create the matrix of differentials dir, the POLAR subspace, according to our antonym-set $\mathbb{P}_{\text{words}} \cup \mathbb{P}_{\text{emoji}}$ (§ 3.2). After normalizing the subspace vectors, we create all embedding vectors via projec-

interpret	**W**	*with* **Mixed**	**E**
Words	(W/W)	(W/M)	(W/E)
Emoji	(E/W)	(E/M)	(E/E)

(a) Campaigns Overview. We interpret Words & Emoji with likewise Words, Emoji, and Mixed (both).



(b) Selection Task for Emoji/Mixed (E/M).



(c) Preference Task for Emoji/Mixed (E/M).

Figure 3: *(a)* We conduct six campaigns measuring human interpretability for including emoji to the POLAR$^\rho$ embedding space. Exemplified with the Emoji Mixed campaign (E/M): interpreting emoji with emoji and words. *(b)* In the Selection test, coders choose suitable differentials for describing a given term. *(c)* In the Preference test, coders provide their interpretation of a given term to a differential scale.

tion $\overrightarrow{\mathbb{E}_v} = \hat{dir}^T \overrightarrow{\mathbb{W}_v}, \forall v \in \mathcal{V}$. Though normalization requires careful later additions to the POLAR space, we opted for standard normalization, $\mathbb{E}_{\text{stdnrm}} = [\mathbb{E} - \text{mean}(\mathbb{E})] \cdot \text{std}(\mathbb{E})^{-1}$, to ensure that the whole embedding space aligns properly around the center of gravity on each differential scale. We select the best suited opposites for a given embedding space by using the Extremal Word Score (§ 2.3) for d=500+44 dimensions (words + emoji).

4 Human Evaluation Approach

While we have now created a supposedly interpretable embedding, it remains to be seen how well it is *perceived* by humans. That is, we next evaluate our two key RQs, discuss significance, and provide further details: *RQ1)* How well does POLAR$^\rho$ with EWSO perform in selecting most interpretable dimensions at varying expressiveness of words and emoji? *RQ2)* How well do POLAR$^\rho$ scalar values reflect directions on the differential scales? *i)* Do humans prefer emoji to words? *ii)* How well do human raters align w.r.t. interpretability? *iii)* What impact do demographic factors play in interpretability with or without emoji?

4.1 Evaluation design

To gather human judgement, we employ crowdsourcing on the Microworkers platform.

4.1.1 Questions & Evaluation Metrics

Our evaluation of the POLAR$^\rho$ approach including emoji to the differentials bases on two main questions next to demographics.

Selection test. Analogous to the original work, we want to find out whether humans agree on best interpretability of POLAR$^\rho$ selected differentials with a word intrusion task. The question asks our coders to select five out of ten differentials that describe a given word best as shown in Fig. 3b. We select half of these dimensions according to the highest absolute projection scale values (most extreme). The other half consists of a random selection from the bottom half of available differentials. I.e., if the projection approach determines interpretable dimensions well, humans would choose all five out of five POLAR$^\rho$ chosen differentials.

As any user might choose differently, we count how often coders choose certain differentials. The resulting frequencies immediately translate in a ranking that we leverage for calculating the fraction of Top 1..5 being POLAR$^\rho$ chosen differentials.

Preference test. Additionally, we introduce the preference test evaluating whether the direction on a given differential scale is in line with human judgement. That is, for the same words from the selection test, we display the same ten dimensions (5 top-POLAR$^\rho$, 5 random bottom) where coders select their interpretation of the given word on scales as shown in Fig. 3c. Typical for semantic differential scales (Tullis and Albert, 2008; Osgood et al., 1957), we deliberately use a seven point scale representing -3 to 3, allowing more freedom than 3 or 5 points (Simms et al., 2019). Further, we specifically allow a center point—being equal—as it might indicate both being *equally well* or *not good at all*.

Due to scale usage heterogeneity (Rossi et al., 2001), we normalize coder chosen directions (shift+scale according to mean) prohibiting disproportional influence of single coders. We evaluate the coder agreement by counting direction (sign) non-/alignment with the POLAR$^\rho$ projection scale.

Demographics. There is a multitude of other external factors that might have impact on coders' choices. To better understand participant back-

ground, we ask for their education, emoji usage (familiarity), smartphone platform (different emoji pictograms), and if they had used Jodel before.

4.1.2 Evaluation Setup

Crowdworker Campaigns We run a campaign for each of the cross product between words only, emoji only, and mixed Tab. 3a and Fig. 2. *(W/W)* word/word sets a baseline comparison to results from the original POLAR work, albeit now using the projection approach. *(W/M)*: word/mixed uses not only words, but includes emoji to the POLAR subspace. *(W/E)*: word/emoji uses only emoji to describe words. *(E/W)*: emoji/word provides another baseline as to how well emoji may be interpreted with words only. *(E/M)*: emoji/mixed uses both, emoji and words to interpret emoji. *(E/E)*: emoji/emoji may be the most interesting as we only use the expressiveness of emoji to describe emoji.

For mixed cases (emoji and words within the POLAR subspace), we create rankings from absolute scale values on both types (words/emoji) separately and then select them equally often to achieve similar amounts of word and emoji differentials.

Used Words & Emoji. We selected 50 words and emoji to be described in each campaign. To ensure that *i)* we only use common words that are very likely known to our coders, and *ii)* these words are captured well within the underlying embedding, we pick them out of the upper 25% quantile by occurrences in the corpus ($n \geq$1.6k). I.e., we chose emoji and words that appear frequently and should therefore be well-known. For words, we ensured that they are part of the German dictionary *Duden*.

Tasks Setup. Within our six campagins, we now have each 50 emoji or 50 words to be interpreted. We bundled this into 5 tasks each consisting of 10 emoji/words—resulting in 30 different tasks. Each of these tasks contains the Selection test, Preference test, and demographics.

Subjects. Human judgement and crowdsourced evaluations are noisy by nature. While it is usually sufficient to employ few trusted expert coders, it is suggested to use more in the non-expert case (Snow et al., 2008). Thus, we assign 5 different annotators to each of the 30 tasks. At estimated 10-15min duration, we provide 3$ compensation for answering a single task, above minimum wage in our country.

Quality Assurance. Any crowdsourcing task offers an incentive to rush tasks for the money, which requires us to employ means of quality assurance (QA). As we have an uncontrolled environment and thus untrusted coders, we handcraft test questions for the selection and preference test. This task is non-trivial as we require unambiguity in *correct* answers (we ensured this with multiple qualitative tests among friends and colleagues), while simultaneously not being too obvious. We place one test question for selection and one for preference randomly into each task (ending up in 11 words or emoji per task). This also means that each coder can only participate in up to 5 different tasks within a single campaign before re-seeing a test question.

We define acceptance thresholds of four out of five correct answers for both, the selection test and the correct direction for the preference test.

4.2 Results

Within the crowdsourcing process, we rejected about 10% of all tasks according to our QA measures, which then had to be re-taken. We ended up with 6 campaigns each having 50 words/emoji answered by 5 coders; summing up to completed 150 tasks. In total, 16 different coders accomplished this series of which 4 completed $\Sigma \geq 100$ tasks.

4.2.1 Interpreting Emoji

First we focus on the describing emoji campaigns (E/*). We present our main evaluation results in Tab. 1. Within columns, we show results for random, original POLAR, and our six campaigns. We split the rows into results from the selection test across Top1..5 entries and the preference test.

Selection Test. We find very good results along all emoji campaigns (E/*) being consistently better than any campaign describing words (W/*). The best performance was achieved for explaining emoji with emoji (E/E); others are on par.

We want to note however, that the small size of used emoji-differential set may ease selection. E.g., facial expression emoji regularly achieve higher embedding scores than others, which thus may bias the bottom control half (§ 4.1.1). However, interpreting emoji or words with words only, (E/W) and (W/W), achieve comparable performance.

Preference Test. Here, we make the same observation; The projected scales on the differentials are mostly well in line with human judgement.

4.2.2 Interpreting Words

Again, we refer to Tab. 1, but now change our focus to describing words, campaigns (W/*).

Selection Test. Albeit not being directly comparable, using POLAR$^{\rho}$ in compaings: describing

Task		Random	*POLAR*	(W/W)	(W/M)	(W/E)	(E/W)	(E/M)	(E/E)
Selection	Top 1	0.500	*0.876*	0.79	**0.83**	0.60	0.81	0.79	**0.88**
	Top 2	0.222	*0.667*	**0.62**	0.61	0.35	0.67	0.68	**0.77**
	Top 3	0.083	*0.420*	**0.45**	0.42	0.15	0.54	0.57	**0.67**
	Top 4	0.024	*0.222*	**0.30**	0.18	0.07	0.37	0.37	**0.59**
	Top 5	0.004	*0.086*	**0.14**	0.08	0.01	0.22	0.19	**0.38**
Preference		0.500	-	**0.740**	0.672	0.576	0.800	**0.848**	0.832

Table 1: Campaign results. Random & original POLAR baseline. Selection and Preference results across campaigns. Words are better described by word dimensions, and emoji are better described by emoji dimensions.

words with words (W/W), or describing words with words and emoji (W/M) achieved performance well on par with POLAR. Noteworthy, describing words with emoji (W/E) yielded the worst results. The projection scale values for the emoji dimensions were mostly lower compared to words. I.e., according to POLAR$^\rho$, for words only few emoji differentials would be among the top 5 opposites.

Preference Test. As for the preference test, describing words yield the best results using word opposites only (W/W). Explaining words with emoji (W/E) performs particularly worse.

4.2.3 Result Confidence

Significance. To test for differences within the coder alignment with POLAR$^\rho$, we model both, the selection and preference test. With our primary goal to understand the impact of including emoji to a POLAR$^\rho$ interpretable word embedding, we anchor to the (W/W) campaign as a baseline.

For the selection test, we count if coders aligned with POLAR$^\rho$ or chose any of the random alternatives across the Top 1..5 selection. For the preference test, we count whether coders aligned with POLAR$^\rho$'s scale direction. We apply double-sided chi-square tests χ^2 with $p < 0.05$ between the interpreting words with words (W/W) baseline and the remaining five campaigns.

We identify significant differences in coder-POLAR$^\rho$ alignment to the (W/W) baseline when describing words with emoji (W/E) over Top1..5 selection and preference. Counts from explaining emoji with emoji (E/E) signal significance for preference and selection Top3..5. Coder-POLAR$^\rho$ alignment in preferences is also significant for describing emoji with emoji and words (E/M).

4.2.4 Observations

Emoji. As a byproduct, we also show if emoji opposites are preferred over words. That is, we focus on the mixed campaigns describing words and emoji with words and emoji (*/M).

α	(W/W)	(W/M)	(W/E)	(E/W)	(E/M)	(E/E)
Selection	0.44	0.35	0.24	0.46	0.39	**0.55**
Preference	0.57	0.41	0.34	**0.61**	0.54	0.60
Preference *POLAR$^\rho$ only*	0.65	0.52	0.40	**0.70**	0.64	0.68
Preference *random only*	0.31	0.17	0.25	**0.31**	0.22	0.22

Table 2: Inter-rater agreement Krippendorff's α across campaigns. Coders achieve the best agreement in selection test of emoji-based campaigns (E/*) and generally within the preference test measuring differential scales.

We establish a baseline by filtering the counts for all non-POLAR$^\rho$ randomly chosen dimensions being word or emoji representing a Bernoulli experiment. I.e., along the random dimensions, our coders chose 228 vs. 221 and 167 vs. 187 words over emoji. Applying chi-squared statistics indicates, that both types (words and emoji) are chosen equally often at least cannot be rejected.

We next analyze the POLAR$^\rho$ chosen dimensions in the mixed campaigns. Here, coders chose words over emoji as follows: 465 vs. 336 in the (W/M), and 414 vs. 482 in the (E/M) campaign. We find statistically significant favors for words to interpret words and emoji to describe emoji.

Scale Usage. We find no evidence for any directional biases within our preference test (cf. 3c).

Coder Agreement. While the aggregate results are compelling, we use the Krippendorff-alpha metric to measure coder agreement along all six campaigns as shown Tab. 2; higher scores depict better agreement. We split the overall results by test first (Selection & Preference), but also show additional agreement results for preferences along POLAR$^\rho$ chosen dimensions and their random counterpart.

Most agreement is within the *moderate* regime. This observation does not come unexpected from our five non-expert classifiers per task. Overall, we find that coders agree better for well-performing campaigns. We identify the best agreement scores for interpreting emoji with emoji (E/E); coders agree least in the worst performing explaining words with emoji campaign (W/E).

For the preference test, we subdivide our results into POLAR$^{\rho}$ chosen differentials and compare them to the randomly chosen ones. While the agreement on the random opposites is only *fair*, the agreement on POLAR$^{\rho}$ chosen opposites is consistently better: Estimating differential scale directions via POLAR$^{\rho}$ for words yields *moderate* agreement, whereas coders consistently align *substantially* in interpreting emoji. We presume emoji may convey limited ideas, but are easier to grasp, have better readability; the campaings interpreting emoji (E/*) were generally accomplished faster.

4.2.5 Demographics

Though we are confident in applied QA measures, none of the demographics can be confirmed. The annotator sample-size is small and thus most likely not representative. Further, we find most workers providing contrasting answers across multiple tasks they participated in, rendering collected demographic information unusable.

5 Related Work

No universal meaning of emoji. Prior work showed that the interpretation of emoji varies (Miller et al., 2016; Kimura-Thollander and Kumar, 2019), also between cultures (Guntuku et al., 2019; Gupta et al., 2021). Even within the same culture, ambiguity and double meanings of emoji exist (Reelfs et al., 2020) and differences exists on the basis of an individual usage (Wiseman and Gould, 2018). These observations motivate the need to better understand the meaning of emoji. Currently, no data-driven approach exists to make emoji interpretable—a gap that we aim to close.

Interpretable *word* embeddings. Word embeddings are a common approach to capture meaning; they are a learned vector space representation of text that carries semantic relationships as distances between the embedded words. A rich body of work aims at making word embeddings interpretable, e.g., via contextual (Subramanian et al., 2018), sparse embeddings (Panigrahi et al., 2019), or learned (Senel et al., 2018) transformations (Mathew et al., 2020)—all focus on text only. Recently, (Mathew et al., 2020) proposed the POLAR that takes a word embedding as input and creates a new interpretable embedding on a polar subspace. The POLAR approach is similar to SEMCAT (Senel et al., 2018), but is based on the concept of semantic differentials (Osgood et al., 1957) for creating a polar subspace. It measures the meaning of abstract concepts by relying on opposing dimensions associated (good vs. bad, hot vs. cold, conservative vs. liberal). In this work, we extend and use POLAR.

Emoji embeddings. Few works focused on using word embeddings for creating emoji representations, e.g., (Eisner et al., 2016) or (Reelfs et al., 2020). (Barbieri et al., 2016) used a vector space skip-gram model to infer the meaning of emoji in Twitter data (Barbieri et al., 2016). Yet, the general question if the interpretability of word embeddings can be improved by adding emoji and if different meaning of emoji can be captured remains still open. In this work, we adapt the POLAR interpretability approach to emoji and study in a human subject experiment if word embeddings can be made interpretable by adding emoji and how emoji can be interpretated by emoji.

6 Conclusion

We raise the question whether we can leverage the expressiveness of emoji to make word embeddings interpretable. Thus, we use the POLAR framework (Mathew et al., 2020) that creates interpretable word embeddings through semantic differentials, polar opposites. We employ a revised POLAR$^{\rho}$ method that transforms arbitrary word embeddings to interpretable counterparts to which we added emoji. We base our evaluation on an off the shelf word-emoji embedding from a large social media corpus, resulting in an interpretable embedding based on semantic differentials, i.e., antonym lists and polar emoji opposites.

Via crowdsourced campaigns, we investigate the interpretable word-emoji embedding quality along six use-cases (cf. Fig. 1): Using word- & emoji-polar opposites (or both *Mixed*), to interpret words (W/W, W/E, W/M) and emoji (E/W, E/E, E/M), w.r.t. human interpretability. Overall, we find POLAR$^{\rho}$'s interpretations w/wo emoji being well in line with human judgement. We show that explaining emoji with emoji (E/E) works statistically significantly best, whereas describing words with emoji (W/E) systematically yields the worst performance. We also find good alignment to human judgement estimating a *term*'s position on differential scales, using the POLAR$^{\rho}$-projection.

That is, emoji can improve POLAR$^{\rho}$'s capability in identifying most interpretable semantic differentials. We have demonstrated how emoji can be used to interpret other emoji using POLAR$^{\rho}$.

Acknowledgements

We thank Felix Dommes, who was instrumental for this work by developing and implementing the POLAR$^{\rho}$ projection approach and the Extremal Word Score in his Master Thesis.

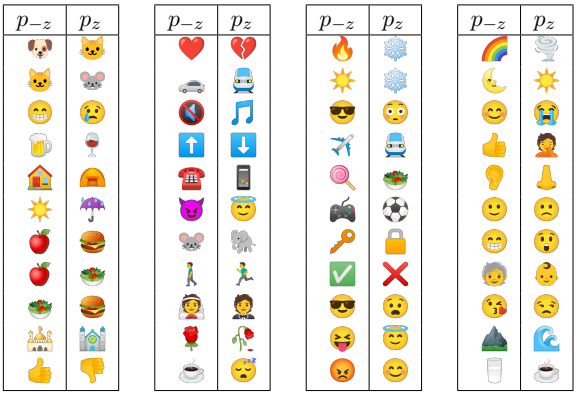


Table 3: Used heuristically identified polar emoji opposites $(p_{-z}, p_z) \in \mathbb{P}_{\text{emoji}}$. We opted for a diverse set of opposites selecting only few facial emoji differentials.

References

F. Barbieri, F. Ronzano, and H. Saggion. 2016. What does this Emoji Mean? A Vector Space Skip-Gram Model for Twitter Emojis. In *LREC*.

Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, 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*, Austin, TX, USA. Association for Computational Linguistics.

Sharath Chandra Guntuku, Mingyang Li, Louis Tay, and Lyle H Ungar. 2019. Studying cultural differences in emoji usage across the east and the west. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 226–235.

Mitali Gupta, Damir D Torrico, Graham Hepworth, Sally L Gras, Lydia Ong, Jeremy J Cottrell, and Frank R Dunshea. 2021. Differences in hedonic responses, facial expressions and self-reported emotions of consumers using commercial yogurts: A cross-cultural study. *Foods*, 10(6):1237.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a lexical-semantic net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

T. Hu, H. Guo, H. Sun, T. T. Nguyen, and J. Luo. 2017. Spice up your chat: the intentions and sentiment effects of using emojis. In *ICWSM*.

Philippe Kimura-Thollander and Neha Kumar. 2019. Examining the "global" language of emojis: Designing for cultural representation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–14.

Binny Mathew, Sandipan Sikdar, Florian Lemmerich, and Markus Strohmaier. 2020. The polar framework: Polar opposites enable interpretability of pre-trained word embeddings. In *WWW*, pages 1548–1558.

Hannah Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht. 2016. "blissfully happy" or "ready to fight": Varying interpretations of emoji.

P. . Novak, J. Smailović, B. Sluban, and I. Mozetič. 2015. Sentiment of emojis. *PloS one*.

C.E. Osgood, G.J. Suci, and P.H. Tannenbaum. 1957. *The Measurement of Meaning*. Illini Books, IB47. University of Illinois Press.

Abhishek Panigrahi, Harsha Vardhan Simhadri, and Chiranjib Bhattacharyya. 2019. Word2sense: sparse interpretable word embeddings. In *ACL*.

Jens Helge Reelfs, Oliver Hohlfeld, Markus Strohmaier, and Niklas Henckell. 2020. Word-emoji embeddings from large scale messaging data reflect real-world semantic associations of expressive icons. *Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media*.

Peter E Rossi, Zvi Gilula, and Greg M Allenby. 2001. Overcoming scale usage heterogeneity. *Journal of the American Statistical Association*, 96(453).

Lutfi Kerem Senel, Ihsan Utlu, Veysel Yucesoy, Aykut Koc, and Tolga Cukur. 2018. Semantic structure and interpretability of word embeddings. In *IEEE/ACM TASLP*, 10.

Vered Shwartz, Enrico Santus, and Dominik Schlechtweg. 2017. Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection. In *EACL*, pages 65–75.

Leonard Simms, Kerry Zelazny, Trevor Williams, and Lee Bernstein. 2019. Does the number of response options matter? psychometric perspectives using personality questionnaire data. *Psychological Assessment*, 31.

Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *EMNLP*.

Anant Subramanian, Danish Pruthi, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Eduard Hovy. 2018. Spine: Sparse interpretable neural embeddings. In *AAAI*, volume 32.

Thomas Tullis and William Albert. 2008. *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Sarah Wiseman and Sandy JJ Gould. 2018. Repurposing emoji for personalised communication: Why means "i love you". In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–10.