

# Investigating Redundancy in Emoji Use: Study on a Twitter Based Corpus

**Giulia Donato**

University of Copenhagen  
giulia.dnt@gmail.com

**Patrizia Paggio**

University of Copenhagen  
University of Malta  
paggio@hum.ku.dk  
patrizia.paggio@um.edu.mt

## Abstract

In this paper we present an annotated corpus created with the aim of analyzing the informative behaviour of emoji – an issue of importance for sentiment analysis and natural language processing. The corpus consists of 2475 tweets all containing at least one emoji, which has been annotated using one of the three possible classes: *Redundant*, *Non Redundant*, and *Non Redundant + POS*. We explain how the corpus was collected, describe the annotation procedure and the interface developed for the task. We provide an analysis of the corpus, considering also possible predictive features, discuss the problematic aspects of the annotation, and suggest future improvements.

## 1 Introduction

Nowadays emoji are widespread throughout mobile and web communication both in private conversations and public contexts such as blog entries or comments. In 2015, the Oxford Dictionary declared the emoji *Face with tears of joy* “Word of the year”, and since then the academic interest towards the topic, as well as the development of relevant resources, have grown substantially. Emoji are best known to be markers for emotions, and in this sense they can be considered an evolution of emoticons. However, these pictographs can be used to represent a much wider range of concepts than emoticons, including objects, ideas and actions in addition to emotions, and thus they interact with the content expressed in the surrounding text in more complex ways. Furthermore, emoji are used not only at the end of a message, e.g. a tweet, but can occur anywhere and possibly in sequences. Therefore, understanding the seman-

tic relation they have with the surrounding text, in particular whether emoji add independent meaning, is an important step in any approach attempting to process their contribution to the overall content of a given message, both for the purposes of sentiment analysis and natural language processing.

We are interested in investigating to what extent it is possible for a human annotator, and subsequently for an automatic classifier, to determine if emoji in tweets are used to emphasize or add information, which may well be emotional information, but could also have a different semantic flavour. If emoji do add meaning, we also ask how easy it is to understand if they are being used as syntactic substitutes for words. In this paper, we focus on the corpus of English tweets that was collected and annotated to provide training data for a number of classifiers aiming at predicting whether emoji in microblogs are used in a redundant or a non-redundant way.

The classification experiments achieved promising results (F-score of 0.7) for the best performing model, which combined LSA with handcrafted features and employed a linear SVM in a One vs. All fashion. The process and results of the experiments will be described in a future paper (in preparation).

In Section (2) we review related research, then in Section (3) we describe how the tweets were extracted and collected to create the corpus, and give counts of the various represented categories. In Section (4) the annotation process is described, Section (5) presents and discusses the results, and finally in Section (6) we provide a conclusion.

## 2 Related research

Several studies trace parallels between emoticons and emoji, sometimes using both terms inter-

changeably, with the purpose of dealing with emotion expression or automatic emotion detection, and thus only considering those pictographs that resemble facial features. [Boia et al. \(2013\)](#) focus on emoticons and their use in tweets. The authors attempted to determine the reliability of emoticon labels in sentiment classification by means of a user study and generated a sentiment lexicon from a corpus of 2.1 million tweets. They found that agreement between the sentiment expressed by emoticons and the sentiment expressed by the surrounding words is only slightly higher than random, showing that emoticons are likely to be used as a means to add emotion to an otherwise neutral text. The experiment based on the sentiment lexicon proved that emoticons are good indicators of sentiment in the tweet, but are less effective in retrieving related sentiment words, thus confirming that emoticons complement the text rather than stressing what is already expressed by the words.

The paper by [Hallsmar and Palm \(2016\)](#) is instead focused on the effectiveness of using emoji to automatically annotate training data for multiclass emotion classification. The researchers employed a training corpus of 400,000 tweets, 100,000 for each of four classes (sadness, anger, fear and happiness), then tested against 80 instances, manually collected and labeled according to their textual content. The results show that emoji can be effectively used to automatically annotate the emotion class in large sets of tweets, thus suggesting that emoji, in contrast with emoticons, may co-occur with semantically related words.

Other works have analyzed the semantics of emoji, mostly by means of distributional semantics. In [Barbieri et al. \(2016\)](#), the authors used the skip-gram model paired with different dataset sizes and different filtering methods to generate emoji embeddings. These were evaluated against a set of 50 emoji pairs manually annotated for *similarity* and *relatedness* scores. The similarity scores obtained by the models were strongly correlated with those in the gold standard, particularly if stop words and punctuation are removed from the dataset. This indicates that surrounding words and other emoji are useful for inferring the meaning of a given emoji, possibly indicating that the emoji is being used in a redundant way.

In [Eisner et al. \(2016\)](#), emoji embeddings were learnt from their description in the Unicode emoji

standard, and representations are thus obtained for all represented emoji including those that appear infrequently in online text. In spite of the model being trained on much less data, the authors claim to outperform [Barbieri et al. \(2016\)](#) on the task of Twitter sentiment analysis. These results point to the fact that the emoji descriptions in the Unicode standard are a valid source from which to model their semantics.

The issue whether emoji add content to the text they occur in, particularly in tweets, or whether they are largely redundant, as well as how their specific use in this respect can be predicted, is not investigated directly in any of the studies mentioned so far.

The paper by [Zanzotto et al. \(2011\)](#) addresses the problem of linguistic redundancy within the realm of microblogs. Although this study does not specifically target emoji, it is of particular interest for our work given the formal definitions provided for both redundancy and non redundancy as well as the methodology employed. The authors performed a classification experiment on 1242 pairs of tweets related to news, previously annotated considering four possible relations, i.e. entailment (redundant), paraphrase (redundant), related/unrelated (non-redundant), and contradiction (non-redundant). They used the annotated corpus to test different models in a classification experiment, and obtained the best results with a combination of syntactic and similarity features computed across the word vectors of each pair.

The methodology adopted in our work builds on the [Zanzotto et al. \(2011\)](#) study, both as concerns the fundamental question we ask, and the way we have collected and annotated our training corpus. A crucial difference is, however, that our analysis focuses on the use of emoji.

### 3 Corpus Preparation

To answer our research questions we set up a corpus of English tweets automatically extracted from Twitter with the aid of specific emoji keywords. The corpus was then annotated by four human coders to be further used in a machine learning experiment. The annotated corpus consists of tweets containing emoji paired with their counterparts where the emoji has been removed, for a total of 2475 pairs.

The purpose of the corpus collection and annotation was twofold. Our primary goal was to pro-




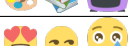






Category	Emoji	Names
Traveling/Commuting		<i>car, airplane, sailboat</i>
Events		<i>party popper, jack-o-lantern, graduation cap</i>
Places		<i>school, european castle, home + garden</i>
Other Activities		<i>artist palette, books, television</i>
Feelings		<i>smiling face with heart eyes, unamused face, crying face</i>
People		<i>man and woman holding hands, person walking, person raising one hand</i>
Eating & Drinking		<i>pizza, doughnut, hot beverage</i>
Nature & Animals		<i>dog, snowflake, maple leaf</i>
Music		<i>microphone, guitar, musical notes</i>
Sport		<i>trophy, swimmer, basketball and hoop</i>

Table 1: List of the emoji used to extract tweets for the corpus collection

vide training data to develop classifiers that could predict the relation of emoji in unseen tweets. A secondary goal was to investigate how easy it is for human coders to distinguish different uses of emoji with respect to their semantic contribution. In order to clarify this aspect, we run an inter-annotator agreement test on part of the annotated material.

### 3.1 Emoji Selection

To select a set of meaningful emoji to use for the data extraction, we start by defining a categorization of the whole emoji set. The Unicode consortium website provides the full emoji dataset, in which every emoji is annotated with a code, fourteen different graphic renderings, the emoji name, the date of addition to the Unicode standard, and a set of keywords that identify the content of each pictograph. Unicode separates groups of emoji according to similar renderings and, possibly, semantic relatedness, but does not provide an official ontology.

Previous studies interested in emoji semantics use different categorizations for their purposes. Cappallo et al. (2015) relied on the categories listed in the MSCOCO (Lin et al., 2014) dataset: *Person & Accessory, Animal, Vehicle, Outdoor Object, Indoor Object, Sport, Kitchenware, Food, Furniture, Appliance, Electronics*. These categories partially overlap the ones in Emojipedia (Burge, 2013): *Smileys & People, Animals & Nature, Food & Drink, Activity, Travel & Places,*

*Objects, Symbols, Flags*. Emojipedia categorizes the pictographs considering their graphical properties, while the MSCOCO categories are modeled for object recognition, thus they discriminate more precisely among inanimate objects.

Barbieri et al. (2016) used word embeddings, dimensionality reduction and clustering, to identify 11 clusters labeled as: *Sports & Animals, Nature, Body gestures & Positive, Free Time, Unclear, Love & Parties, Letters, Barber & Symbols, Eating & Drinking, Music, Sad & Tears*. These labels reflect the graphical and conceptual similarity of the data points included in a specific cluster. Nevertheless, some of the labels are claimed to be inconsistent since the relevant clusters include few and apparently unrelated pictographs.

Vidal et al. (2016) used a categorization based on Emojipedia which includes six categories: *Food & Drinks, Non-food objects, Celebrations, Activity, Travel & Places, Nature*.

After having considered the categorizations mentioned above we developed our own including the following labels: *Nature & Animals, Places, Traveling/Commuting, Sport, Events, Other Activities, Music, Eating & Drinking, People, Feelings*. Our intent was to select a small number of relatively broad and easily recognizable categories. Furthermore, we chose to keep events and activities separate from entities, as is done in many linguistically-oriented ontologies.

From each category in our list we have selected three emoji; in order to get clearly distin-

guishable pictographs we have considered both their frequency of use given by the Emojitracker, thus favouring the most frequent tokens, and their graphical features. The full list of emoji is shown in table 1.

### 3.2 Data Collection

All the data were collected between the 1st and 2nd of November 2016 by means of the Twitter Streaming API and the Python Tweepy wrapper.

To extract the data we added to the script a filter for the English language and passed the list of the selected emoji as the keywords parameter. Both the possibilities of filtering data by language and keywords are provided as features by the API.

The raw data included 501,342 tweets, subsequently reduced to 196,434 after removing all duplicate entries. A series of common preprocessing steps were applied before the annotation: in particular all the mentions of other users and all the links were replaced with placeholders.

The accepted character length on Twitter is 140; in the cleaned corpus the average length of the tweets was of 50 characters, 555 tweets were longer than 140 characters with a maximum length of 196 characters. Thus, as an additional step, all the tweets below a threshold length of 10 characters and above a threshold length of 140 characters were discarded. We checked again for the presence of duplicates after replacing mentions and links, since tweets may have the same content and differ only for these elements; this led to a resulting collection of 180,958 instances. In this cleaned version of the corpus the average tweet length is of 52 characters with a standard deviation of 32.

The best represented category is, unsurprisingly, *Feelings* with a total of 99,050 instances. Within *Feelings* the most frequent emoji is *Smiling face with heart shaped eyes* with 60,479 extracted tweets. The least represented category is *Places* with a total of 900 instances. Within this category the least frequent emoji is *School* with 47 extracted tweets.

From these data we created a balanced corpus by sampling 900 instances from each category, since this is the size of the least populated one. From the resulting corpus of 9000 instances we further removed all the tweets containing only the emoji used for the data extraction since this would have resulted in pairs containing one empty tweet and one tweet consisting in an emoji key-

word repeated multiple times. The final collection contained 8985 pairs; from this corpus we randomly sampled 4100 pairs for the annotation. The size was chosen considering the corpus size in the [Zanzotto et al. \(2011\)](#) paper, which we used as a methodological model for our work.

## 4 Annotation

The annotation of the 4100 tweet pairs took place remotely between the 21st and the 31st of December 2016 and was performed by four annotators, three located in Greece and one in the Netherlands. All the annotators were fluent English speakers. For the annotation we developed an ad-hoc user interface.

We chose a multiclass setup with three classes of interest: *Redundant*, *Non-redundant*, and *Non-redundant + POS*; we will further define these classes and explain them with examples shortly below. The annotators were asked to assign a class to each pair in the corpus.

### 4.1 Classes Definition

The general definition of redundancy is *repetition of already expressed information*; to describe the classes for the annotation we relied on [Zanzotto et al. \(2011\)](#), who define as redundant tweet pairs which are in a relation of paraphrase or entailment, while pairs in a relation of contradiction or relatedness are considered non-redundant. We expect an emoji to be considered redundant if it represents an object or an action also expressed by words in the text (the emoji is a synonym of another word) or if it represents an object or action whose presence is directly implied by the text (the emoji is entailed by the words).

The final set of classes includes three labels: *Redundant*, *Non-Redundant*, *Non-Redundant + POS*. The Redundant class indicates that the emoji of interest repeats the information present in the text or that its meaning is implied by the text.

On the contrary, we expect the Non-Redundant class to be assigned when the emoji adds information not already present or implied in the text.

Lastly the Non-Redundant + POS class, which can be considered as a subset of the Non Redundant class, indicates the case where the emoji is used with a syntactic function (and can be labeled with its POS), thus replacing a word. We provided a set of examples to the annotators and clarified possible edge cases. An extract from the examples

is listed here:

1. *Redundant*

- "We'll always have Beer. I'll see to it. I got your back on that one. 🍺"
- "@USER I need u in Paris girls 🇫🇷"

2. *Non-Redundant*

- "I wish you were here ✈️"
- "Hopin for the best 🎓"

3. *Non-Redundant + POS*

- "Thank you so so so so much ily Here's a 🍕 as a thank you gift x"
- "Good morning 🌍"

An edge case could be represented by:

- "Reading is always a good idea 📚. Thank you for your sincere support @USER. Happy reading."

In this case the emoji represents books which are related to the verb "reading", however the act of reading does not necessarily imply the presence of books (it is not an entailment) since it is possible to read newspapers, blogs, comments, emails; the emoji is narrowing down the meaning of the verb, therefore it is adding information and we should consider it non-redundant.

Emotions also represent a challenge since we need to rely on symbols or simplifications to depict complex expressions. While a case such as:

- "i'm so proud of myself 🥳 \*pats my back\*"

is clearly non-redundant (here the emoji is used ironically), a tweet like:

- "My forever love 🥰 @URL"

represents redundant use.

## 4.2 Interface

To annotate the tweets we set up a dynamic interface accessible online and hosted - until the completion of the task - on a server at the Demokritos Institute of Research in Athens (<http://www.demokritos.gr/>); we provided detailed guidelines explaining how to access and use the interface and describing the annotation criteria and the classes with the aid of examples.

Before the annotation started, we tested the interface on the latest versions of Mozilla Firefox and Google Chrome. Since browsers do not always render emoji automatically we provided our interface with a link to the Symbola Font, one of the richest in emoji renderings.

On the first page of the interface each annotator had a welcoming message and a briefer version of the instructions already provided in the guidelines. After the instructions and three examples of tweet pairs with the correspondent class checked, the annotators could move on to the annotation page which presented the pairs, a forced choice form to select the class and a submit button. The pairs were updated dynamically after each submission and the checked value was stored together with the index of the pair and the annotator id. The default value of the form was set to blank; we gave the annotators the possibility to submit a blank value whenever they were undecided about the class to pick; the blank submissions were recorded as *undefined*.

The screenshot shows a web interface for tweet annotation. It displays two examples of tweet pairs, labeled A and B. Example A is "minute to win it 🎰" and Example B is "minute to win it". Below the examples is a form with three radio buttons: "Redundant", "Non Redundant", and "Non Redundant + POS". The "Non Redundant + POS" option is selected. There are two buttons at the bottom: "Submit" and "End Session".

Figure 1: Screen capture of the annotation interface

The first 100 pairs were annotated by all the annotators to measure the inter-annotator agreement; after this set of common pairs the annotators had random access to further 1,000 pairs each among the remaining 4000. On completion of the task the annotators were redirected to a thanksgiving page. Furthermore, we gave them the option to interrupt and restart the annotation process in order to complete the task in multiple sessions. Their work was automatically saved to a csv (comma separated value) file after each session's interruption.

Due to the random access, after the first 100 pairs, some of the 1,000 pairs left were presented and annotated more than once, hence they were



discarded from the final corpus. Additionally, one of the annotators reported problems with the interface when saving the last part of her work. Therefore, and also considering the fact that we excluded the 100 pairs used to calculate the agreement, our final corpus consists of 2475 annotated pairs in total.

### 4.3 Annotation Reliability

To assess the inter-annotator agreement we adopt Cohen’s  $\kappa$  coefficient (Cohen, 1960).

Coehn’s  $\kappa$  is used to assess agreement between two annotators and it is considered more robust than simple percentage agreement since it corrects for chance agreement. Moreover, this choice allows us to compare our results with those obtained by Zanzotto et al. (2011) for a similar, although more complex, task.

Considering the agreement results described in Zanzotto et al. (2011) we expected to get a  $\kappa$  of 0.6, which is generally considered moderate agreement (Landis and Koch, 1977).

	A1	A2	A3	A4
A1	-	0.76	0.78	0.7
A2	0.76	-	0.81	0.8
A3	0.78	0.81	-	0.71
A4	0.7	0.8	0.71	-

Table 2: Observed agreement

	A1	A2	A3	A4
A1	-	0.57	0.62	0.48
A2	0.57	-	0.66	0.64
A3	0.62	0.66	-	0.5
A4	0.48	0.64	0.5	-

Table 3: Cohen’s  $\kappa$  agreement

In tables 2 and 3 we report the results for the percentage and Cohen’s  $\kappa$  agreement between each pair of annotators. The average percentage agreement is 76%, while the average Cohen’s  $\kappa$  is 0.576, a value only slightly lower than what we were aiming for. A discussion of the difficulties encountered by the annotators is provided in the next section. To comply with the suggestion given by one of the anonymous reviewers, we also calculated agreement using Fleiss’ kappa and Krippendorff’s alpha. The values we obtained, however, are very similar (0.575 and 0.576, respectively.)

## 5 Analysis and Discussion

### 5.1 Corpus Analysis

Our gold standard contains a total of 2475 annotated pairs, as stated in the previous section.

	<i>End</i>	<i>Not End</i>	<i>Total</i>
R	452 <b>(0.357)</b>	382 (0.316)	834 (0.337)
Non-R	768 <b>(0.607)</b>	660 (0.546)	1428 (0.577)
Non-R+POS	37 (0.029)	139 <b>(0.115)</b>	176 (0.071)
Undefined	9 (0.007)	28 <b>(0.023)</b>	37 (0.015)
Total	1266 (1)	1155 (1)	2475 (1)

Table 4: Conditional frequency of the emoji class given the emoji position: absolute counts and proportions. The largest proportion for each class in each condition is in boldface.

	<i>CD</i>	<i>NN</i>	<i>Other</i>	<i>Total</i>
R	362 <b>(0.348)</b>	328 (0.336)	144 (0.314)	834 (0.337)
Non-R	583 (0.560)	565 <b>(0.580)</b>	280 (0.610)	1428 (0.577)
Non-R+POS	73 (0.070)	79 <b>(0.081)</b>	24 (0.052)	176 (0.071)
Undefined	23 (0.022)	3 (0.003)	11 <b>(0.024)</b>	37 (0.015)
Total	1041 (1)	975 (1)	459 (1)	2475 (1)

Table 5: Conditional frequency of the emoji class given the emoji POS tag: counts and proportions. The largest proportion for each class in each condition is in boldface.

The distribution of the classes is as follows: the *Redundant* class has 834 instances (33.7%), the *Non-Redundant* class has 1428 instances (57.7%), the *Non-Redundant + POS* class has 176 instances (7.1%). Additionally, 37 instances are annotated as *undefined* (1.5%).

Table 4 details how the classes are distributed given the position of the emoji as either close to the end of the tweet or not<sup>1</sup>: 35.7% of the instances are annotated as Redundant (R in the tables), 60.7% as Non-Redundant (Non-R), 2.9% as

<sup>1</sup>The emoji position was computed by dividing the index of the emoji in the tokenized tweet by the number of tokens in the tweet. We considered close to the end those emoji with a value equal or above 0.7

Non-Redundant + POS (Non-R+POS), and 0.7% are undefined. In the opposite condition (when the position of the emoji is not close to the end of the tweet) 31.6% instances are Redundant, 54.6% are Non-Redundant, 11.5% are Non-Redundant + POS, and 2.3% are undefined. Interestingly, although not surprisingly, the Non-Redundant + POS class is the only one (leaving the undefined instances out) to show a higher probability of occurrence in the "not close to the end" than the "close to the end" condition.

From the distribution we can see that, at least in a corpus the size of ours, the distinction between close or non close to the end is not a strong indicator of whether the emoji is used to repeat or add information, with the exception of the case in which the emoji not only adds information but also replaces a word. The differences in the distribution are significant, as demonstrated by a  $\chi$ -squared test of independence ( $\chi$ -squared = 81.644,  $df = 3$ ,  $p$ -value  $< 0.001$ ). An analysis of the residuals confirmed that the effect of position is highest in the case of the Non-Redundant + POS class.

We had an intuition that the part-of-speech category of the emoji might be an interesting feature to look for the purposes of training classifiers to predict the relation of the emoji with the content of the rest of the text. Therefore, the corpus was run through the Stanford Tagger. We decided to use the standard Stanford POS Tagger from the Python NLTK wrapper since traditional POS taggers have been reported to achieve satisfactory results when compared with domain specific taggers (Derczynski et al., 2013), and also since Twitter-specific POS taggers do not seem to provide tags for emoji.

In table 5 we report the frequencies for the most frequent tags, which are *CD* or *NN* (cardinal number and noun, respectively). The column *Other* sums the frequencies of the remaining categories. The Stanford POS Tagger considers several features prior to assigning a tag to unknown words. This set of features includes capitalization, context (n-grams), hyphens, numbers, and allcaps (Toutanova et al., 2003). Tokens containing allcaps, a slash or a dash as well as numbers are tagged with *NN* (since they might be company names). Thus the POS-tag assigned to an emoji may either be the result of these specific features or may be based on the n-gram sequence in which the emoji is embedded.

From the numbers in the table, and again leaving out the undefined instances, it would appear that *NN* might be used as a predictor of the two Non Redundant classes, while *CD* seems more predictive of Redundant use. The differences are significant on a  $\chi$ -squared test of independence ( $\chi$ -squared = 21.385,  $df = 6$ ,  $p$ -value  $< 0.01$ ). An analysis of the residuals showed that, if we ignored the undefined instances, the largest contributions to the differences are found in the negative effect of *CD* on the Non-Redundant class, the negative effect of *Other* on Non-Redundant + POS and the positive effect of *NN* on that same class.

To sum up, the analysis shows that position (close to the end or not) and part-of-speech class might be useful features to consider when training a classifier to predict whether emoji in tweets are being used in a redundant or additive way.

## 5.2 Annotation difficulties

We saw earlier that the inter-annotator results are slightly lower than expected. Some annotators reported difficulty in assigning a class when the tweet content was not meaningful, thus a possible way to improve the annotation design and increase the agreement would be to filter out all the spam and advertisement tweets that contain little or non-informative text and keep only tweets from individual (possibly verified) users, avoiding corporate accounts and bots.

To gain a better understanding of the difficulties of the annotation process we considered a small sample of pairs where two annotators assigned the Non-Redundant class and the other two assigned the Non-Redundant + POS class, some examples are listed here:

- "🇵🇱 vs 🇵🇱 Legia Warsaw"
- "🍕 - like who comments 'ifb'"

From these examples it can be seen that disagreement emerges when the tweet content is very short and unstructured and the function of the emoji is ambiguous, given also the lack of syntactic cues in the text. Such cases include occurrences where the text of the tweet consists of hashtags only.

We also noted disagreement (mostly among a single annotator and the three others) in cases where the emoji is strongly related to other words in the text. E.g.:

- "I wish I was a pet so I could just stay home,

lounge all day and have no responsibilities 🐶”

- “@USER mom, my birthday is coming 🎉”

In such cases it is possible that one or more annotator identified a relation of synonymy or entailment (thus, label the instance as “Redundant”) while the others consider it as relatedness or similarity (thus, label the instance as “Non Redundant”). This suggests that identifying entailment at token level instead than from pairs of sentences, especially in unstructured and short text, is a hard task. We must also note that even though we balanced the amount of tweets per category in our corpus, we did not further balance the tweets in each category according to the emoji used to retrieve them. Therefore, we cannot exclude a possible effect derived from the most common of these emoji and we should consider to improve this aspect in future research.

Lastly, we cannot exclude that difficulties may have arisen due to renderings of other co-occurring emoji that were missing from the Symbola font we adopted.

## 6 Conclusion

We have presented an annotated corpus of tweets that was developed with the purpose of training models to classify the informative behaviour of emoji in tweets.

We have described the entire process of retrieving, cleaning, and presenting the data to the annotators through the graphical user interface specifically developed for the task. The interface source code is available at [https://github.com/giuliadnt/Annotation\\_gui](https://github.com/giuliadnt/Annotation_gui); the corpus can be provided by the main author on request.

The reliability of the annotation was measured and, although the average  $\kappa$  score was slightly lower than expected, it still showed close to moderate agreement among the annotators, an acceptable result given the difficulty of the task.

We have also provided an analysis of the corpus in terms of the distribution of three classes of emoji behaviour (Redundant, Non-Redundant, and Non-Redundant + POS) given the position of the emoji in the tweet, as well as their part-of-speech category. Both dimensions seem to provide at least some predictive power, and have in fact been used as features to develop classifiers of emoji informative behaviour in tweets (paper in preparation),

There are several aspects we have discussed in this work that may constitute a limitation and are, therefore, open to improvements and changes. The most important is perhaps the fact that the three classes of interest are far from being equally represented. Thus, more data should be collected. Doing so could also reduce the effect of noisy examples, such as those of tweets only consisting of emoji.

Regarding this aspect we could also consider the possibility of using a binary setup, thus merging Non Redundant and Non Redundant + POS into the same class and balancing the amount of instances related to each case within it. Improvements to the annotation interface should also be considered if more data is annotated.

Considering the confusion sometimes made by the annotators between similarity and entailment, more examples should be provided to train them more extensively to categorize such cases correctly.

Furthermore, the agreement can be improved by including additional annotators and removing from the corpus those tweets that result to be particularly problematic

As future work it will be interesting to evaluate emoji’s behaviour in the context of specific NLP tasks such as threads summarization. Moreover, it would be important to verify if the redundancy between emoji and words is equivalent or differs from the redundancy among the words alone in the context of the same tweet.

## Acknowledgments

We would like to express our gratitude to George Giannakopoulos, the [Institute of Informatics and Telecommunications](#) at NSRF Demokritos (Athens, GR) and everybody at [Sci.FY](#) for the help and support throughout the whole data collection and annotation process. We would also like to thank Daniela Schneevogt and Michael Schlichtkrull for all their suggestions and the reviewers for the feedback provided.

## References

- Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. 2016. What does this emoji mean? a vector space skip-gram model for twitter emojis. In *Language Resources and Evaluation conference, LREC*. Portoroz, Slovenia.



- Marina Boia, Boi Faltings, Claudiu-Cristian Musat, and Pearl Pu. 2013. A:) is worth a thousand words: How people attach sentiment to emoticons and words in tweets. In *Social Computing (SocialCom), 2013 International Conference on*. IEEE, pages 345–350.
- Jeremy Burge. 2013. Emojipedia. <https://emojipedia.org/>.
- Spencer Cappallo, Thomas Mensink, and Cees GM Snoek. 2015. Image2emoji: Zero-shot emoji prediction for visual media. In *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, pages 1311–1314.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1):37–46. <https://doi.org/10.1177/001316446002000104>.
- Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. 2013. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *RANLP*. pages 198–206.
- Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. 2016. emoji2vec: Learning emoji representations from their description. *arXiv preprint arXiv:1609.08359*.
- Frederik Hallsmar and Jonas Palm. 2016. Multi-class sentiment classification on twitter using an emoji training heuristic.
- J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* pages 159–174.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*. Springer, pages 740–755.
- Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, pages 173–180.
- Leticia Vidal, Gastón Ares, and Sara R Jaeger. 2016. Use of emoticon and emoji in tweets for food-related emotional expression. *Food Quality and Preference* 49:119–128.
- Fabio Massimo Zanzotto, Marco Pennacchiotti, and Kostas Tsioutsoulis. 2011. Linguistic redundancy in twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 659–669.