New methodologies to evaluate the consistency of emoji sentiment lexica and alternatives to generate them in a fully automatic unsupervised way

mfgavilanes@gti.uvigo.es, jonijm@gti.uvigo.es, sgarcia@gti.uvigo.es, kike@gti.uvigo.es, javier@det.uvigo.es
GTI Research Group
Telematic Engineering Dept., School of Telecommunication Engineering, University of Vigo, Vigo 36310 Spain

Abstract

Sentiment analysis aims at detecting sentiment polarities in unstructured Internet information. A relevant part of this information for that purpose, emojis, whose use in Twitter has grown considerably in these years, deserves attention. However, every time a new version of Unicode is released, finding out the sentiment users wish to express with a new emoji is challenging. In [KNSSM15], an Emoji Sentiment Ranking lexicon from manual annotations of messages in different languages was presented. The quality of these annotations affects directly the quality of possible generated emoji sentiment lexica (high quality corresponds to high self-agreement and inter-agreement). In many cases, the creators of the datasets do not provide any quality metrics, so it is necessary to use another strategy to detect this issue. Therefore, we propose an automatic approach to identify and manage inconsistent manual sentiment annotations. Then, relying on a new approach to generate emoji sentiment lexica of good quality, we compare two such lexica with lexica created from manually annotated datasets with poor and high qualities.

1 Introduction

Following a trend in the last years, emojis are being increasingly used in social applications. For example, 1% of the messages in a random sample of 22.14 billion tweets taken between July 2013 and March 2018 contained at least one emoji.

Emojis allow users to express feelings and emotions. Thus, it is interesting to try to extract from them useful knowledge on user opinions [HTGL13]. Natural Language Processing (NLP) allows us to analyze opinions, feelings, assessments, etc. on products, services or organizations [Liu12]. Until very recent times, researchers in the field of sentiment analysis (SA) only considered the information contributed by emoticons [BFMP13, DTR10, HBF+15]. Nevertheless, nowadays emojis are attracting considerable attention [GOB16, HGS+17]. For this reason, some recent studies have tried to obtain the sentiment expressed by emojis in the form of a lexicon [KNSSM15, LAL+16, KK17]. In many cases, however, the expected meaning of an emoji (in terms of positivity, neutrality or negativity), which is assumed to be universal, may change among languages and cultures [BKRS16].

Following this line, in [KNSSM15] the authors presented an Emoji Sentiment Ranking (ESR)², resulting from texts in 15 different languages containing emojis, whose sentiments were labeled manually by different human annotators over three months. However, the quality of manual labeling, measured in terms of self-agreement and inter-agreement as explained in [MGS16], may be poor.

Copyright © 2018 held by the author(s). Copying permitted for private and academic purposes.


1http://www.emojiTracker.com/api/stats
2Available at https://goo.gl/XEkJhZ
We can suppose that, if an emoji sentiment lexicon is generated from one of these single-language datasets, the most popular emojis should be highly correlated with those obtained from the overall SRS when the quality of manual labeling, measured in terms of self-agreement and inter-agreement, is acceptable (differences would be mainly due to emojis with different interpretations among languages). On the contrary, if at least one of these metrics is low, inconsistencies in manual sentiment annotations should be suspected, and the correlation would be seriously affected (the differences in emoji interpretations would be much greater). When these measurements are not provided by the dataset creators or they are unknown, an alternative should be sought to identify the inconsistencies. The final objective should be to create an emoji sentiment lexicon with the highest possible quality.

In this paper, we propose an approach to detect low-quality dataset annotations. In case of inconsistent annotations, we also present a fully automated approach to obtain emoji lexica with good quality.

The rest of the paper is organized as follows: Section 2 reviews related work on emoji sentiment analysis. Section 3 discusses the issue of labeling quality. Section 4 describes the proposed method. Section 5 presents experimental results. Finally, Section 6 summarizes the main contributions and conclusions.

2 Related work

Even though emoji sentiment interpretation (where sentiment is expressed as a positive, neutral or negative polarity) has already been studied in the field of NLP, a common practice in the case of Twitter was to filter Unicode symbols during message preprocessing, so that emojis’ information was lost [TK16]. But, for example, in the message “Today I have to go to the supermarket 😢”, the obvious negativity is given by the emoji.

Focusing on methods to guess the real sentiment of emojis, they can be classified in three types: manual, semi-automatic and automatic. Regarding manual methods, in [MTSC+16] the most popular Unicode emoji characters were manually labeled by multiple annotators, taking into account sentiment (positivity, neutrality and negativity) variance as well as semantics (meaning). In [KNSSM15], 83 native speakers of different languages labeled by hand the sentiment (positive, neutral or negative) of texts containing 751 different emojis. The authors calculated their sentiment based on their occurrences and the manual labels of the tweets containing them, by applying a discrete probability distribution. Finally, in [ELW+16] 78 strongly and 34 weakly subjective emojis were extracted from the list [KNSSM15] and given polarity values of +2, +1, -1 and -2 (strongly positive, weakly positive, weakly negative and strongly negative, respectively).

Currently, few approaches assign polarities to emojis with semi-automatic or automatic methods. In [HTAAAK16] the most used emojis in a dataset of Arabic tweets were classified into four categories: anger, disgust, joy and sadness. Subsequently, they were weighted with scores between -5 and +5 (most negative and most positive, respectively), according to those categories. The weights were obtained from the AFINN lexicon [Nie11], in which some entries are emojis. The Unicode short Common Locale Data Repository (CLDR) names of the missing emojis were obtained and the words composing them were searched in AFINN (one by one, independently). Finally, weights were also manually assigned according to the category of each emoji.

Regarding the approaches that obtain emoji sentiment lexica in a fully unsupervised way, we are only aware of the following examples. In [LAL+16], the authors analyzed emoji usage in text messages by country. In total, the sentiment of 199 emojis was obtained from their short CLDR names processed with the LIWC tool (which counts words that express positive, neutral or negative sentiment). This analysis did not exploit their real descriptions or their usage contexts. In [KK17], the authors extracted, for each word of a tweet that co-occurred with a target emoji, the set of synonyms or synsets available in WordNet. Then they recovered the most frequent affective label from WordNet-Affect. Five sentiment categories were differentiated: happiness, disgust, sadness, anger and fear, following a hierarchical structure. Finally they calculated a sentiment score vector for 236 emojis based on the mentioned co-occurrences. Again, this analysis also ignored the real descriptions or the usage contexts of the emojis. Finally, in [FGJMGM+18], a lexicon of 840 emojis was created using an unsupervised SA system, taking only into account emoji definitions in Emojipedia. This lexicon was then improved in different variants that took advantage of the sentiment distribution of informal texts including emojis.

3 Description of the problem

In general, a given emoji should have the same emotional meaning in different datasets written in the same language. This implies that emoji sentiment interpretation for each of them should be very close to

---

3http://unicode.org/emoji/charts/emoji-list.html
4https://liwc.wpengine.com/
5https://wordnet.princeton.edu/
6http://wndomains.fbk.eu/wnaffect.html
7https://emojipedia.org/
the interpretation for all datasets together. The problem arises when an *emoji* sentiment lexicon is created from multilingual datasets with manual sentiment annotations that are inconsistent for a language or some languages.

On the other hand, it seems logical to think that an *emoji* should have different emotional meanings across different languages and cultures. Nevertheless, according to [BKRS16], the semantics of the most popular *emojis* are strongly correlated most of the time in most languages in that regard. This was an interest finding, because both the vocabularies of the languages and the context words modeled by the semantic spaces are different. The authors stated that English and Spanish speakers interpret *emojis* in the most universal way, with a high correlation with all other languages, although strong differences may persist for some *emojis*. In this way, the sentiment of the most popular *emojis* in a particular language may differ from the “universal sentiment”, but they should be close in most cases.

Our main contributions are a method to detect anomalies in *emoji* sentiment lexicon due to inconsistent annotations and an alternative automatic approach to predict *emoji* sentiments with applications in *emoji* sentiment lexica generation.

4 Proposed methods

We first present a method for constructing automatically two *emoji* sentiment lexica [FGJMGM*18] (Figure 1). Summing up, (1) *emojis* are extracted from a set of informal texts and their descriptions are acquired from the *Emojipedia* repository. Then, (2) NLP techniques capture their linguistic peculiarities from both the descriptions and the informal texts, which are exploited independently by an *unsupervised* SA system with sentiment propagation across dependencies (USS-PAD) described in [FGALJM*16]. Depending on the combination of the polarities obtained from the SA, (3) two *emoji* sentiment lexica variants are created. In this regards, we remark that our aim is not a novel SA approach.

In Figure 1, the dotted arrow in the upper left corner represents the actions to gather a set of informal texts with *emojis*. The solid arrows represent the processes carried out on these texts to obtain the first *emoji* sentiment lexicon from an *emoji* sentiment ranking, from automatically labeled texts where they occur. The dashed arrows refer to the case in which a similar process is previously applied on each individual *emoji* description (extracted from *Emojipedia*), to obtain an initial *emoji* sentiment lexicon from the universal definitions by *emoji* creators. This lexicon, *unsupervisedDef*, is later applied as extra information into each particular informal text, to assign sentiment labels automatically and then obtain the second lexicon through the same *emoji* sentiment ranking. Next, we explain the method in more detail.

4.1 Acquiring *emoji* definitions

In order to extract *emoji* definitions, messages must be converted to a Unicode representation and regular expressions must be used for the extraction. Then, each *emoji* Unicode codepoint in hexadecimal notation is converted to UTF-8 hex bytes and submitted via a get request to the *Emojipedia* resource to retrieve its

---

8 This process was carried out using the *Emoji-java* library, available at [https://github.com/vdurmont/emoji-java](https://github.com/vdurmont/emoji-java).

9 [http://emojipedia.org/search/?q=](http://emojipedia.org/search/?q=)
English description, which is parsed through jsoup\textsuperscript{10}.

4.2 SA on texts and emoji definitions

At this point, the method performs SA on both the informal texts containing the emojis and their definitions. This consists of two main tasks: preliminary data treatment with lexical and syntactic analysis; and capturing linguistic peculiarities and applying USSPAD SA [FGALJM\textsuperscript{+16}]. In it, the final sentiment results from the propagation of sentiment term values (included in a sentiment lexicon) from the leaves to the parent nodes of each dependencies tree. Once these steps are completed, a polarity score is assigned to each informal text and emoji description, and emoji sentiment lexica can be created.

4.3 Creation of emoji sentiment lexica

Once all previous steps have been performed on informal texts and definitions, we are in a position to apply two different approaches to exploit polarity scores of texts and definitions, and create two emoji sentiment lexica. In the first variant (\(E1\)), the lexicon is created considering the ranking of polarity scores assigned to texts with emojis, applying the estimations in [KNSSM\textsuperscript{15}]. That is, following the solid arrows in Figure 1 we obtain \(R_{\text{unsupervised}}\). The second variant (\(E2\)) considers extra information. Lexicon \(R_{\text{unsupervised + unsupervised definition}}\) is created from sentiment scores obtained through automatic sentiment propagation on emoji definitions. These values are then included in the sentiment lexicon used in Section 4.2 to improve the SA of informal texts and obtain new polarity scores for them. Finally, the same estimations in [KNSSM\textsuperscript{15}] are applied to the resulting unsupervised sets. That is, following the dashed arrows in the figure, to obtain \(R_{\text{unsupervised + unsupervised definition}}\).

4.4 Detecting inconsistent annotations

Given the hypothesis that the sentiments of the most popular emojis are preserved across different languages, and that only a small percentage of them show language-specific usage patterns [BKRS\textsuperscript{16}], we assume that the correlation between the entries of an emoji sentiment lexicon created for a particular language and the entries of a multilingual emoji sentiment lexicon (ideally a universal lexicon) should be high. This is the base for the experiments in Section 5.2.

5 Evaluation and experimental results

5.1 Dataset

We used the annotated datasets in [KNSSM\textsuperscript{15}] in 15 different languages including Albanian, English, Polish and Spanish, among others. These datasets are available at the public CLARIN\textsuperscript{11} language resource repository. The entry for each labeled tweet consists of a tweet ID, a sentiment label (negative, neutral or positive) and an anonymized annotator ID. We focused on the four datasets in Table 1, discarding tweets without emojis and tweets with ambiguities among annotators. The authors reported good self-agreement (\(\text{Alpha}_\text{s}\)) and inter-agreement (\(\text{Alpha}_\text{i}\)) values for English and Polish and worse values for Albanian and Spanish.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#emojis</th>
<th>Label</th>
<th>#Tweets</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albanian</td>
<td>0.447</td>
<td>Negative</td>
<td>17</td>
<td>14.53%</td>
</tr>
<tr>
<td></td>
<td>0.126</td>
<td>Neutral</td>
<td>40</td>
<td>34.19%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Positive</td>
<td>60</td>
<td>51.28%</td>
</tr>
<tr>
<td>English</td>
<td>0.739</td>
<td>Negative</td>
<td>2,935</td>
<td>27.59%</td>
</tr>
<tr>
<td></td>
<td>0.613</td>
<td>Neutral</td>
<td>2,677</td>
<td>25.16%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Positive</td>
<td>5,027</td>
<td>47.25%</td>
</tr>
<tr>
<td>Polish</td>
<td>0.757</td>
<td>Negative</td>
<td>658</td>
<td>27.59%</td>
</tr>
<tr>
<td></td>
<td>0.571</td>
<td>Neutral</td>
<td>919</td>
<td>24.27%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Positive</td>
<td>2,229</td>
<td>58.87%</td>
</tr>
<tr>
<td>Spanish</td>
<td>0.245</td>
<td>Negative</td>
<td>1,022</td>
<td>16.85%</td>
</tr>
<tr>
<td></td>
<td>0.121</td>
<td>Neutral</td>
<td>3,431</td>
<td>26.89%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Positive</td>
<td>8,306</td>
<td>65.10%</td>
</tr>
</tbody>
</table>

Table 1: Distribution of negative, positive, and neutral tweets containing emojis for the datasets in the experiments

5.2 Practical case for detecting anomalies in annotations

Table 2 shows the correlations for positive, negative and neutral labels between the conventional ESF lexicon \(R_{\text{annotated definition}}\) created using the method in [KNSSM\textsuperscript{15}] from messages in 15 languages annotated by hand) and each emoji sentiment lexicon, which was created in the same way for a single language \(R_{\text{annotated definition}}\) for English, for instance). For a fair analysis, given the detection criterion, to calculate the correlation we considered the top 100 occurring emojis in each language lexicon as the most popular.

Looking at Table 2, score and ranking level correlations are high for English and Polish \(R_{\text{annotated definition}}\). Moreover, looking at Figures 2a and 2b, the associated linear regressions (represented with solid lines) have slightly less slope than the regression for the overall case that serves as gold-standard (represented with a dotted line). This suggests that the English and Polish datasets have consistent annotations, as

\textsuperscript{10}Available at https://jsoup.org/

\textsuperscript{11}http://hdl.handle.net/11356/1054.
Figure 2: Top 100 emoji sentiment scores comparing the general emoji lexicon with the lexicon of a particular language

Table 2: Score and rank correlations considering top 100 emojis ranked by score and occurrence evidenced by their good $\text{Alpha}_s$ and $\text{Alpha}_i$ values in [KNSSM15, MGS16].

However, when we compared the overall $R_{annotatedall}$ lexicon with the Spanish and Albanian lexica ($R_{annotatedes}$ and $R_{annotatedal}$), score and ranking correlations were worse. Indeed, in Figures 2c and 2d, the linear regression slopes are very flat, and therefore they move far from the overall case. This suggests that the Spanish and Albanian datasets have inconsistent manual annotations (as shown by $\text{Alpha}_i=0.121$ and $\text{Alpha}_i=0.245$ for Spanish and $\text{Alpha}_i=0.126$ for Albanian) [KNSSM15, MGS16]. In addition, if we focus on Figure 2c, a vast majority of emoji dots have positive polarity in the Spanish lexicon (X axis) while, for the overall case, polarities vary between positive and negative.

5.3 Alternative solution for lexica generation

Once we are able to detect annotation anomalies, we also have a methodology to validate an alternative solution to generate lexica automatically. We verified it on English and Spanish datasets as representative cases of which we have good and bad manual annotations, respectively. Two sentiment emoji lexica were created per language, corresponding to variants $E_1$, which only considers the automatic USSPAD annotation ($E_{1es}$ and $E_{1en}$), and $E_2$, which also considers Emojipedia definitions ($E_{2es}$ and $E_{2en}$). Subindex’s $es$ and $en$ denote Spanish and English, respectively.

Table 3: Score and rank correlations considering top 100 occurring emojis in English and Spanish

In Table 3, if we compare the English variants, we...
observe that the lexica are highly correlated. Introducing the effect of emoji definitions, correlation increases from $E_1$ to $E_2$ compared both with $R_{\text{annotated all}}$ and $R_{\text{annotated en}}$. This is clear in Figures 3a and 3c, where the line that serves as gold-standard and the regressions intersect at neutral emoji sentiments. However, in Figures 3b and 3d these lines intersect respectively at positive and neutral emoji sentiments. This shows that the definitions balance sentiments in the second variant.

On the other hand, given the fact that the emoji sentiment lexicon obtained from a manually annotated Spanish dataset $R_{\text{annotated es}}$ has poor quality due to annotation inconsistencies [MGS16], as confirmed by their authors and by Table 2 and Figure 2c in Section 5.2, its correlation with the automatic variants should also be low. This is verified in Table 3 for $E_1$ and $E_2$. The better behavior of $E_1$ in this case is not relevant, due to the anomalies in $R_{\text{annotated es}}$. However, in the comparisons with $R_{\text{annotated all}}$, the correlation with $E_2$ is higher both for ranking and score, as shown in Figures 4a and 4b, which is coherent with the observations for English.
5.4 Checking with SA the new approaches

\( R_{\text{annotated-all}} \) is biased by typical emoji usage worldwide and, to a lesser extent, by the vision of the annotator, who writes in a particular language. For this reason, we might worry about the influence of particular language subsets in the overall lexicon. Therefore, an independent evaluation of the generated emoji sentiment lexica is necessary.

Our objective here is to determine if our lexica variants for Spanish and English are good enough in a real-world scenario, by evaluating their impact with SA metrics (precision \( (P_{\text{macro}}) \), recall \( (R_{\text{macro}}) \) and \( F_{\text{macro}} \) macroaverages on the positive and negative classes).

In principle, in the Spanish subset this is impeded by bad labeling. We assumed that only a small percentage of the most popular emojis had significant sentiment differences between languages. For most properly annotated messages containing the top popular emojis, we could thus assume that any lexica should provide similar results. Therefore, we decided to restrict the SA test to the 100 most popular emojis in Spanish and English. This new dataset had a distribution with 3552 positive, 1998 negative and 1601 neutral messages. Table 4 shows the results.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Lexicon</th>
<th>( P_{\text{macro}} )</th>
<th>( R_{\text{macro}} )</th>
<th>( F_{\text{macro}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>English B</td>
<td>( R_{\text{annotated-en}} )</td>
<td>76.16%</td>
<td>69.45%</td>
<td>72.65%</td>
</tr>
<tr>
<td></td>
<td>( E2_{en} )</td>
<td>75.49%</td>
<td>69.20%</td>
<td>72.21%</td>
</tr>
<tr>
<td></td>
<td>( E1_{en} )</td>
<td>67.95%</td>
<td>67.74%</td>
<td>67.85%</td>
</tr>
<tr>
<td></td>
<td>( E2_{es} )</td>
<td>73.01%</td>
<td>67.84%</td>
<td>70.33%</td>
</tr>
<tr>
<td></td>
<td>( E1_{es} )</td>
<td>66.98%</td>
<td>67.89%</td>
<td>67.43%</td>
</tr>
<tr>
<td></td>
<td>( R_{\text{annotated-es}} )</td>
<td>56.42%</td>
<td>62.04%</td>
<td>59.10%</td>
</tr>
</tbody>
</table>

Table 4: Macroaveraging SA metrics of English dataset for the most popular emojis in English and Spanish

Our assumptions are validated by these results, sorted by \( P_{\text{macro}} \). The ordering is coherent with our expectations. \( R_{\text{annotated-en}} \) was created from consistent manual annotations, but \( E2_{en} \) only performs a bit worse. If we compare \( E1_{en} \) with \( E1_{es} \), on the one hand, and \( E2_{en} \) with \( E2_{es} \), on the other, their performances are comparable, bit for small percentages that can be explained by the small percentage of "top" emojis whose sentiment is not preserved across languages. An important finding is that our automatic approach performs satisfactorily compared to a lexicon produced from a well-annotated dataset.

6 Conclusions

A poorly labeled dataset (yielding low self-agreement and inter-agreement) may affect directly the quality of emoji lexica. In many cases the annotators do not publish any quality metrics, so it is difficult to determine beforehand if bad SA performance is due to the supporting lexicon or to the SA technique itself. In this paper we have proposed a method to detect low-quality annotations of tweet datasets written in particular languages containing emojis. We have also proposed a fully automated unsupervised approach to generate lexica with good quality. They have been validated on different datasets taken from [KNSSM15].

Acknowledgements

This work was partially supported by Mineco grant TEC2016-76465-C2-2-R and Xunta de Galicia grants ED341D R2016/012 and GRC2014/046, Spain.

References


