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

1st International Workshop on Emoji Understanding and Applications in Social Media

---

Milagros Fernández-Gavilanes

GTI Research Group

Stanford (California), June 25th, 2018

atlanTTic research center for Telecommunication Technologies

UniversidadeVigo

Table of contents

1. Motivation
2. Dataset
3. Detecting inconsistent annotations
4. Alternative solution for lexica generation
5. Conclusions
6. References

MOTIVATION

Motivation

Sentiment Analysis (SA)

- Extract the opinion (P, N or NEU).
- Examples:
  - The Spanish simply have the best national anthem, P
  - The Spanish national anthem 😍, P
  - #ITAESP look at the bad weather, N
  - #ITAESP look at the weather 😥, N

Emojis are a relevant part:

- Adequate emoji sentiment lexicon is required.

Problem description

Existence of some emoji sentiment lexica:

- created from manual annotations [KNSSM15].
  - considered as gold-standard.
- created from automatic annotations [LAL16, KK17, FJGCG18].
  - evaluation performed comparing with a gold-standard.

Problems:

- each new emoji → new manual annotations (gold-standard).
- different emotional emoji meanings across languages → new manual annotations for each language (gold-standard).
- anomalies between annotators can be found for a language.

How can we solve these problems?

Dataset

Dataset

Use of the multilingual annotated dataset from [KNSSM15]:

- written in 15 different languages (EN, ES, PT, etc.).
- manually annotated over 3 months.
- self-agreement ($Alpha_s$) and inter-agreement ($Alpha_i$) values reported in [MGS16].

Emoji Sentiment Ranking lexicon proposed as **"universal"** (ESR)

- emoji sentiment lexicon can be created for each language.

Dataset (II)

Focusing on Albanian, English, Polish and Spanish subsets:

Dataset	*#emojis*	Label	#Tweets	%
Albanian		Negative	17	14.53%
$Alpha_s = 0.447$	48	Neutral	40	34.19%
$Alpha_i = 0.126$		Positive	60	51.28%
English		Negative	2,935	27.59%
$Alpha_s = 0.739$	624	Neutral	2,677	25.16%
$Alpha_i = 0.613$		Positive	5,027	47.25%
Polish		Negative	638	27.59%
$Alpha_s = 0.757$	369	Neutral	919	24.27%
$Alpha_i = 0.571$		Positive	2,229	58.87%
Spanish		Negative	1,022	16.85%
$Alpha_s = 0.245$	613	Neutral	3,431	26.89%
$Alpha_i = 0.121$		Positive	8,306	65.10%

$R_{annotated_{al}}$

$R_{annotated_{en}}$

$R_{annotated_{po}}$

$R_{annotated_{es}}$

Detecting inconsistent annotations

Previous assumptions

In general, **an emoji** should have:

- same emotional meaning in datasets written in a language.
- different emotional meanings across different languages.

However, for the **most popular emojis** [BKRS16]:

- their semantics are **strongly correlated in most languages**.
- people interpret them in an **universal way**:
  - high correlation between languages.
  - strong differences may persist for some of them.

Hypothesis, for **the most popular emojis**:

- their **sentiments in a language may differ from "universal" one**, but they are **close in most cases**.

Checking our hypothesis for detecting anomalies

So, correlations of **the most popular entries** between:

- ESR lexicon (universal), denoted by $R_{annotated_{all}}$; and
- ESL of each language.

should be:

- **high** $\Rightarrow$ **consistent annotations**.
- **low** $\Rightarrow$ **inconsistent annotations**.

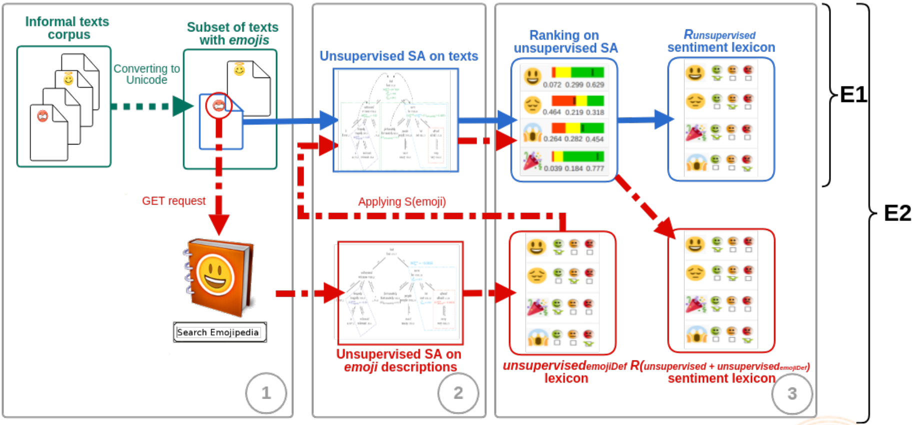
Correlations of top 100 *emojis* ranked by score and occurrence

Lexicon x	Lexicon y	$r_{score}(x,y)$	$r_{rank}(x,y)$
$R_{annotated_{all}}$	$R_{annotated_{en}}$	93.57%	89.46%
	$R_{annotated_{po}}$	88.74%	86.40%
	$R_{annotated_{es}}$	34.07%	37.35%
	$R_{annotated_{al}}$	36.37%	39.30%

Alternative solution for lexica generation

Alternative solution for lexica generation

Method for constructing ESL automatically using SA [FJGCG18]:



Applied on EN and ES datasets:

- $E1_{en}$ and $E1_{es}$: automatic USSPAD annotations.
- $E2_{en}$ and $E2_{es}$: also considers Emojipedia definitions.

Checking the alternative solution for lexica creation

Correlations of **the most popular entries** between variants and:

- a particular language ESL, or
- the ESR considered as "universal".

Lexicon x	Lexicon y	$r_{score}(x, y)$	$r_{rank}(x, y)$
$E1_{en}$	$R_{annotated_{en}}$	82.91%	76.20%
	$R_{annotated_{all}}$	79.70%	75.25%
$E2_{en}$	$R_{annotated_{en}}$	83.72%	79.37%
	$R_{annotated_{all}}$	86.90%	80.71%
$E1_{es}$	$R_{annotated_{es}}$	**47.19%**	**47.18%**
	$R_{annotated_{all}}$	74.93%	74.78%
$E2_{es}$	$R_{annotated_{es}}$	**30.06%**	**44.09%**
	$R_{annotated_{all}}$	81.32%	79.07%

Checking with SA the new alternative lexica

How these language subsets can influence the overall lexicon?

An **independent evaluation** of $E1_{en}$, $E1_{es}$, $E2_{en}$, $E2_{es}$ is needed.

- lexica variants checked in a real-world scenario with SA.
- SA measures applied on P and N classes.
  - precision (P_{macro}), recall (R_{macro}), F (F_{macro}).

Following our assumption, for **the most popular emojis**:

- most messages containing them → similar results with any lexica

Checking with SA the new alternative lexica (II)

So, to check our variants, we need:

- a subset of a consistent dataset with only popular emojis.
- to apply SA using USSPAD on this subset with the emoji lexica.

Dataset	Lexicon	P_{macro}	R_{macro}	F_{macro}
English B	$R_{annotated_{en}}$	76.16%	69.45%	72.65%
	$E2_{en}$	75.49%	69.20%	72.21%
	$E1_{en}$	67.95%	67.74%	67.85%
	$E2_{es}$	73.01%	67.84%	70.33%
	$E1_{es}$	66.98%	67.89%	67.43%
	$R_{annotated_{es}}$	56.42%	62.04%	59.10%

Conclusions

Conclusions

Assumptions:

- a poorly labeled dataset may affect emoji lexica quality.
- annotators do not always publish quality metrics.
  So, it is difficul to determine if:
  - bad SA performance is due to the supporting lexicon, or
  - the SA technique itself.

Contributions:

- a method to detect low-quality annotations of tweet datasets written in a particular language containing emojis.
- a fully automated unsupervised approach to generate lexica with good quality.
- a method to validate lexica created automatically.

References

References

KNSSM15: Kralj Novak, Petra & Smailović, Jasmina & Sluban, Borut & Mozetič, Igor (2015). Sentiment of Emojis.
PLOS ONE 10(12), 1 – 22.

LAL16: Lu, Xuan & Ai, Wei and Liu, Xuanzhe & Li, Qian & Wang, Ning & Huang, Gang & Mei, Qiaozhu (2016). Learning from the ubiquitous language: an empirical analysis of emoji usage of smartphone users.
Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, 770 – 780.

KK17: Kimura, Mayu & Katsurai, Marie (2017). Automatic Construction of an Emoji Sentiment Lexicon.
Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, 978-1-4503-4993-2 1033–1036.

FJGCG18: Milagros Fernández-Gavilanes & Jonathan Juncal-Martínez & Silvia García-Méndez & Enrique Costa-Montenegro & Fco. Javier González-Castaño (2018). Creating emoji lexica from unsupervised sentiment analysis of their descriptions.
Expert System with Application Journal 103, 74–91.

MGS16: Igor Mozetič & Miha Grčar & Jasmina Smailović (2016). Multilingual Twitter sentiment classification: The role of human annotators.
PloS one, 11(5)(5), 1–26.

BKRS16: Francesco Barbieri & Germán Kruszewski & Francesco Ronzano & Horacio Saggion (2016). How cosmopolitan are emojis?: Exploring emojis usage and meaning over different languages with distributional semantics.
In Proc. of the 2016 ACM Conf. on Multimedia Conference, MM 2016, 531–535.

Thank you for your attention

Milagros Fernández-Gavilanes
mfgavilanes@gti.uvigo.es

atlanTTic research center for Telecommunication Technologies

UniversidadeVigo