Expert Systems With Applications 103 (2018) 74–91

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Creating *emoji* lexica from unsupervised sentiment analysis of their descriptions

Milagros Fernández-Gavilanes*, Jonathan Juncal-Martínez, Silvia García-Méndez,
Enrique Costa-Montenegro, Francisco Javier González-Castaño

GTI Research Group, Telematic Engineering Department, School of Telecommunication Engineering, University of Vigo, Vigo, Pontevedra 36310, Spain

ARTICLE INFO

Article history:
Received 28 September 2017
Revised 25 February 2018
Accepted 26 February 2018
Available online 8 March 2018

MSC:
68Q55
68T50

Keywords:
Emoji analysis
Sentiment analysis
Opinion mining
NLP
Artificial intelligence

ABSTRACT

Online media, such as blogs and social networking sites, generate massive volumes of unstructured data of great interest to analyze the opinions and sentiments of individuals and organizations. Novel approaches beyond *Natural Language Processing* are necessary to quantify these opinions with polarity metrics. So far, the sentiment expressed by *emojis* has received little attention. The use of symbols, however, has boomed in the past four years. About twenty billion are typed in Twitter nowadays, and new *emojis* keep appearing in each new Unicode version, making them increasingly relevant to sentiment analysis tasks. This has motivated us to propose a novel approach to predict the sentiments expressed by *emojis* in online textual messages, such as tweets, that does not require human effort to manually annotate data and saves valuable time for other analysis tasks. For this purpose, we automatically constructed a novel *emoji sentiment lexicon* using an unsupervised sentiment analysis system based on the definitions given by *emoji* creators in `Emojipedia`. Additionally, we automatically created lexicon variants by also considering the sentiment distribution of the informal texts accompanying *emojis*. All these lexica are evaluated and compared regarding the improvement obtained by including them in sentiment analysis of the annotated datasets provided by Kralj Novak, Smailovic, Sluban and Mozetic (2015). The results confirm the competitiveness of our approach.

1. Introduction

Emojis are commonly used in smartphone texting, social media sharing, advertising, and more. For example, in 2015 nearly half of all texts posted on Instagram contained them (Dimson, 2015). Similarly, at the time of this research, in a 1% random sample of tweets published from July 2013 to August 2017, 19.88 billion tweets contained *emojis* according to Emojitracker.com [1]. *Emojis* differ from emoticons in that the former are represented by pictographs with a designated textual description, while the latter are typographic facial representations.

Even though *emojis* seem a recent alternative to emoticons, they have been around for 30 years. They were first used in Japan (*emoji* literally means "*image*" and "*character*") and originally could only be used on Japanese phones (D'Aleo, Perticone, Rizzo, & Tabacchi, 2015). They gained popularity when the Unicode standard incorporated them and Apple included them in its operating systems in 2011.

Since then, their number has continuously grown with the introduction of new characters in each new Unicode version, including not only faces but also pictographs representing concepts and ideas such as weather, vehicles and buildings, food and drinks, animals and plants, and emotions, feelings or activities, like running and dancing (Pavalanathan & Eisenstein, 2015).

Moreover, in 2007, Google completed the conversion of "*enhanced emotions*" to Unicode private-use codes, and in 2009 a set of 722 Unicode characters was defined collecting all Japanese *emoji* characters. More pictographs were added in 2010, 2012 and 2014 (Davis & Edberg, 2017). In November 2013, a study indicated that 74% of the United States population used these graphic symbols[2] In China, the percentage of population that used them in nonverbal communications was even higher, reaching 82% (Statista, 2013; Sternbergh, 2014).

This suggests the capability of *emojis* to express feelings or emotions in absence or other elements such as words, facial ex-

* Corresponding author:
E-mail addresses: milagros.fernandez@gti.uvigo.es, mfgavilanes@gti.uvigo.es (M. Fernández-Gavilanes), jonijm@gti.uvigo.es (J. Juncal-Martínez), sgarcia@gti.uvigo.es (S. García-Méndez), kike@gti.uvigo.es (E. Costa-Montenegro), javier@det.uvigo.es (F.J. González-Castaño).

[1] http://www.emojitracker.com/api/stats.

[2] https://blog.swiftkey.com/the-united-states-of-emoji-which-state-does-your-emoji-use-most-resemble/.

https://doi.org/10.1016/j.eswa.2018.02.043
0957-4174/© 2018 Elsevier Ltd. All rights reserved.

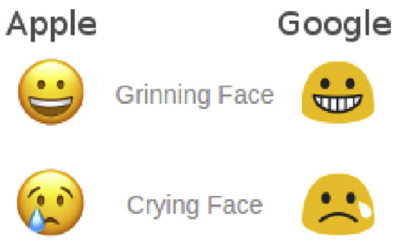

Fig. 1. Example of positive and negative emojis.

pressions or voice cues (Wallbott & Scherer, 1986), across different cultures (D'Aleo et al., 2015), which means that they can be exploited as *a priori* knowledge about opinions in user comments (Hu, Tang, Gao, & Liu, 2013). Consequently, they are appealing to *Sentiment Analysis* (SA), a subfield of *Natural Language Processing* (NLP). The latter combines computational science methods (such as artificial intelligence, automatic learning, or statistical inference) with applied linguistics to achieve computer-aided comprehension and processing of information expressed in human language. In this scenario, SA, also called *opinion mining*, is the field that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, events and topics, and their attributes (Liu, 2012). Although linguistics and NLP have a long history in common, little research about people's opinions and sentiments was conducted before 2000. Since then, it has become a very active research area, especially in the analysis of informal texts such as tweets. Only in recent years emoticons have been considered to play a role (Boia, Faltings, Musat, & Pu, 2013; Davidov, Tsur, & Rappoport, 2010; Hogenboom et al., 2015; Solakidis, Vavliakis, & Mitkas, 2014; Yamamoto, Kumamoto, & Nadamoto, 2014), albeit nowadays *emojis* are more popular. Even so there is still little research work devoted in both of them (Guibon, Ochs, & Bellot, 2016).

One of the most evident issues is the disparity of appearance of an *emoji* from one platform to another. Fig. 1 shows the *emojis* corresponding to *grinning* 😀 and *crying* 😢 for two different platforms, Apple and Google. It is considered that any symbolic representation based on a given name is totally valid, although the meaning associated to each symbol is unique: for example, the *emoji grinning face* 😀 refers to a positive emotion, whereas the *emoji crying face* 😢 clearly has a negative meaning.

Due to the variability of *emoji* representations in different platforms, and given the continuous introduction of new *emojis* in each new Unicode version, it is very difficult to understand meanings beyond affective stances in terms of positivity, neutrality or negativity, and those can vary with social context and author identity (Derks, Bos, & von Grumbkow, 2007; Park, Barash, Fink, & Cha, 2013; Schnoebelen, 2012). That is, in some cases the original meaning has nothing in common with that attributed by people in a particular context, and could be quite different to the initial intention of the creator.

Accordingly, some authors have constructed *emoji* sentiment lexica by manually annotating on informal texts (with the consequent arduous work that this entails) (Kralj Novak, Smailović, Sluban, & Mozetič, 2015). Therefore, it is interesting to adopt an approach that initially considers *emoji* information that does not require human annotation, such as the real meaning of an *emoji*, which is given by its definition, which, in turn, is still strongly linked to the emotional meaning conceived by its creators.

In this paper we present our research to *automatically* construct sentiment lexica with 840 *emojis* using an *Unsupervised System with Sentiment Propagation Across Dependencies* (USSPAD) approach, based on the analysis of the sentiment of informal texts in English and Spanish. The initial sentiment of each *emoji* is derived from a sentiment score obtained after applying the meaning assigned by its creator. Then this value is improved taking into account sentiment scores obtained from informal texts in which that *emoji* appears. So, the results reflect not only the actual use of *emojis* in a context, by applying SA to informal texts such as Twitter, but also the sentiments in the definitions describing such *emojis* in `Emojipedia`[3]. To the best of our knowledge, this is the first time that *emoji* definitions are considered in automatic *emoji sentiment lexicon* creation, where textual information is analyzed with USSPAD, and later combined with textual contexts.

Different experiments and results are presented. In this regard, comparing different approaches is extremely difficult due to the lack of a goldstandard *emoji sentiment lexicon*. Consequently, we compare our strategies with the few in which *emojis* were subject to SA, providing support for our main hypothesis. As a testbed, we employed the available annotated datasets provided by Kralj Novak et al. (2015). Only considering the "initial" sentiment of the *emojis* (i.e. by only taking their short names into account), our approach was competitive with that of Kralj Novak et al. (2015) (based on annotated data), and significantly better when also considering their definitions and usage contexts (messages contexts then are included in). Note that, unlike that approach, ours is fully unsupervised. At the same time these results confirm that *emoji* descriptions add discriminating information that could be exploited in more advanced social NLP systems, given the improvement in accuracy and macroaveraging metrics they achieve.

The paper is organized as follows. Section 2 reviews related work on *emoji* SA. Section 3 describes the proposed SA system. Section 4 discusses experimental results for Twitter dataset. Finally, Section 5 summarizes the main findings and conclusions.

2. Related work

In spite of the fact that *emojis* may be considered a language form, they have been little studied from an NLP perspective, in contrast to their predecessors the *emoticons*. The few exceptions include studies on *emojis* usage and semantics.

For example, Barbieri, Anke, and Saggion (2016a) constructed a vector space model aiming at providing a common semantic ground in which *emojis* are naturally distributed according to geolocation in metropolitan areas. In (Barbieri, Kruszewski, Ronzano, & Saggion, 2016b), the study was extended to countries with different languages. Finally, Ljubešić and Fišer (2016) investigated the global distribution of *emojis*, performing a cluster analysis over countries and a correlation analysis between *emoji* distributions and World Development Indicators.

Regarding semantics studies, Barbieri, Ronzano, and Saggion (2016c) generated, validated, and shared semantic vectorial models built over 10 million tweets posted by USA users by consistently mapping in the same vectorial space both words and *emojis*. They applied skip-gram word embedding models (effectiveness was validated by comparing the output of these models with human assessment using semantic similarity experiments). Their aim was to estimate the degree of similarity between two *emojis* in a situation where both can occur. Later, Eisner, Rocktäschel, Augenstein, Bosnjak, and Riedel (2016) used a similar distributional semantic models, but instead of running skip-gram models on large collections of *emojis* and their tweet contexts, *emoji* embeddings were directly trained on Unicode short *Common Locale Data Repository* (CLDR) names[4] (thus requiring much less training data).

[3] http://emojipedia.org/.

[4] These are annotations which provide names and keywords for Unicode *emojis*, which are available at http://unicode.org/emoji/charts/emoji-list.html.