Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 123 (2018) 360–365

Procedia Computer Science

www.elsevier.com/locate/procedia

8th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2017

Data-driven Social Mood Analysis through the Conceptualization of Emotional Fingerprints

Giovanni Pilato[1] and Ernesto D'Avanzo[1,2]

[1] ICAR-CNR, Palermo, Italy
giovanni.pilato@cnr.it
[2] University of Salerno, Salerno, Italy
edavanzo@unisa.it

Abstract

A body of knowledge shows the emerging of evidence according to a better account for the emotional spectrum is achievable by employing a complete selection of emotion keywords. Basic emotions, such as Ekman's ones, cannot be considered universal, but are related to with implicit thematic affairs within the corpus under analysis. The paper tracks some preliminary experiments obtained by employing a data-driven methodology that captures emotions, relying on domain data that you want to model. The experimentation consists of investigating the corresponding conceptual space based on a set of terms (i.e., keywords) that are representative of the domain and the determination. Furthermore, the conceptual space is exploited as a bridge between the textual content and its sub-symbolic mapping as an "emotional fingerprint" into a six-dimensional hyperspace.

Peer-review under responsibility of the scientific committee of the 8th Annual International Conference on Biologically Inspired Cognitive Architectures

Keywords: Emotion Detection from Text, Data Driven Conceptual Spaces, Social Sensing

1 Background

Cultural, political, social, and economic events have considerable, direct and peculiar effects on *public mood*, in all its dimensions. Microblogging, such as Twitter, can be considered as a kind of fine-grained evidence of *public mood* state. Tweets may express information about their author's mood status, no matter of their content and conscious use. Systematic and large-scale mood analysis could suggest collective emotional tendencies for existing social and economic indicators.

Collecting tweets in a given time window can expose changes happening in public mood state at a macroscopic scale and over the time. For instance, [5] discuss the *forecasting* power of real-world phenomena through social media analysis, showing how a Twitter sentiment analysis prototype can measure both the sentiment and the emotions expressed in posts broadcasted on the Twitter. The platform tested on different *hot trend* stories (i.e., *consumer electronics*, *public health* and *politics*) shows the plausibility of adopting computational intelligence to infer

Peer-review under responsibility of the scientific committee of the 8th Annual International Conference on Biologically Inspired Cognitive Architectures
10.1016/j.procs.2018.01.056

real-world phenomena through social media sentiments and emotions analysis.
However, if even a more-and-more increasing number of attempts take place, daily, to design and implement sentiment analysis platforms, able to capture the *collective mood* from social media, *emotion modelling* seems to gets methodological and technical cotributions from a variety of approaches to analyse *personal* and *social mood* which, not always, are useful for practical purposes. For instance, in their work, [3] extract six dimensions of mood (i.e.,*tension*, *depression*, *anger*, *vigor*, *fatigue*, *confusion*) using a version of the Profile of Mood States (POMS), then comparing the results obtained to fluctuations recorded by stock market and crude oil price indices and other events, such as the U.S. Presidential Election of November 4, 2008 and Thanksgiving Day. Findings show that the monitored events do have a relevant effect on the various dimensions of *public mood*. [2], on the other hand, in their work, propose an approach according to people transmitting their understanding of emotions through the language they employ, which surrounds mentioned emotion keywords. The methodology considers emotions as *conceptualised acts*. In other words, they can be interpreted in the same way as color, that, as known, despite it is a spectrum of visible light, they are categorized and communicated by human beings as discrete colors, employing the schemes offered by language, with all its expressiveness but also with its limits. Following this conceptual framework, the authors have been able to discriminate a set of eight basic emotion keywords, rather than Ekman's six (i.e., *anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise*), that performed better in all semantic tests than all of the basic emotion models analysed. It seems to emerge an evidence according to a better account for the emotional spectrum is achievable by employing a detailed selection of emotion keywords, just as it does in the case for the color range mentioned above.
From what has been discussed above, a body of knowledge, albeit briefly treated, claims that basic emotions, such as Ekman's ones, cannot be considered universal, but are related to with implicit thematic affairs within the corpus under analysis. Furthermore, it also emerges that emotions should be regarded as domain-dependent. Then, if these issue are read, all together, in the light of different domains, where emotions seem to play a key role to capture social trends, as briefly mentioned before, it can be drawn the need of a more fine-grained *emotion modelling*, depending on the domain data available; in other words it is claimed a data-driven *emotion modelling*.
The purpose of the investigation, which this paper only tracks some preliminary experiments, is precisely the design and implementation of a methodology that captures emotions, relying on domain data (i.e. data-driven) that you want to model. To this end, a viable route, which is explored below, and based on what has been discussed so far, consist of investigating the corresponding conceptual space based on a set of terms (i.e., keywords) that are representative of the domain. As known, *conceptual spaces* consist of "quality dimensions", often are derived from perceptual mechanisms [7], that can represent various kind of information and how they can be used to describe concept learning, such as the *emotion modeling* task.
In this sense, in the following it is adopted a *constructive* approach, because the aim is to build an *artifact* that can get to a precise cognitive task (i.e. *emotion modelling*).

2 Description of the system

The proposed system starts from the words contained in a well-known lexical dataset and uses them to retrieve tweets. The retrieved tweets, as well as other words and document artifacts, are used to build a data-driven conceptual space. The procedure allows mapping tweets into a six-dimensional emotional space, which constitutes the "emotional fingerprint" of the tweet. The whole procedure is illustrated in detail in the following subsections.

2.1 Emotional Lexicon

For the detection of the emotions have considered the six Ekman basic emotions: *anger*, *disgust*, *fear*, *joy*, *sadness* and *surprise*, exploiting an emotions lexicon obtained from the Word-Net Affect Lexicon, as described in [12] [6][11]. The lexicon associates a *synset* of WordNet to a set of affective labels in order to mark the specific synsets as representatives of an affective concept. The dataset we have used contains 1542 terms, each of which is associated to one of the six above mentioned emotions: 355 words for anger, 70 for disgust, 195 for fear, 553 for joy, 274 for sadness and 95 for surprise.

2.2 Twitter retrieval

The dataset object of analysis is retrieved by using the Twitter APIs with the default access level. The default access level gives a random sample of the streaming of publicly available tweets. For our approach, we use only the tweet text content, which is preprocessed before being exploited to build a data-driven conceptual space. Stop-words are filtered out, and links are removed before processing the text since they often hide off-topic posts or even spam. Abnormal sequences of characters were discarded. Each word of the emotional lexicon described in the previous section has been used to retrieve a set of tweets written in the English language.

2.3 LSA Space induction

The Latent Semantic Analysis (LSA) technique is a well-known methodology that is capable of giving a rough sub-symbolic encoding of word semantics [9] and of simulating several human cognitive phenomena [8]. The LSA procedure is based on a term-document occurrence matrix $\mathbf{A}$, whose generic element represents the number of times a term is present in a document. Let K be the rank of $\mathbf{A}$. The factorization named Singular Value Decomposition (SVD) holds for the matrix $\mathbf{A}$:

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T \tag{1}$$

Let R be an integer > 0 with $R < N$, and let $\mathbf{U}_R$ be the $M \times R$ matrix obtained from $\mathbf{U}$ by suppressing the last $N - R$ columns, let $\mathbf{\Sigma}_R$ be the matrix obtained from $\mathbf{\Sigma}$ by suppressing the last $N - R$ rows and the last $N - R$ columns; let $\mathbf{V}_R$ be the $N \times R$ matrix obtained from $\mathbf{V}$ by suppressing the last $N - R$ columns. Then:

$$\mathbf{A}_R = \mathbf{U}_R\Sigma_R\mathbf{V}_R^T \tag{2}$$

$\mathbf{A}_R$ is a $M \times N$ matrix of rank R, and it is the best rank R approximation of $\mathbf{A}$ (among the $M \times N$ matrices) with respect to the Frobenius metric. The i-th row of the matrix $\mathbf{U}_R$ may be considered as representative of the i-th word. The columns of the $\mathbf{U}_R$ matrix represent the R independent dimensions of the $\Re^{\Re}$ space S. Each j-th dimension is weighted by the corresponding value σ_j of $\mathbf{\Sigma}_R$. Furthermore, each j-th dimension can be tagged by considering the words having the highest module values of u_{ij}. This makes it possible to interpret the space S as a "conceptual" space, according to the procedure illustrated in [1][10].

2.4 Emotional fingerprint of a tweet

Any text d can be mapped into the Data Driven "conceptual" space by computing a vector $\mathbf{d}$ whose i-th component is the number of times the i-th word of the vocabulary, corresponding to the i-th row of $\mathbf{U}_R$, appears in d. This leads to the mapping of the text as:

$$\mathbf{d}_R = \mathbf{d}^T\mathbf{U}_R\Sigma_R^{-1} \tag{3}$$

The emotional lexicon has been split into six lists, each one associated to one of the basic Ekman emotions {*anger, disgust, fear, joy, sadness, surprise*}. Fixed an emotion e, a set of 300 artificial sentences has been built by using five randomly selected words belonging to the list related to e. This procedure has been done for each list associated with a fundamental Ekman emotion, leading to a set of 1800 artificial sentences. Furthermore, all the 1542 words of the lexicon have been considered. Each one of the 3342 (i.e. 1542+1800) b texts associated with an emotion e has been mapped into the data driven "conceptual space" induced by TSVD according to the transformation in eq.(2). The above procedure leads to a cloud of 3342 (i.e. 1542+1800) vectors that will be used to map a tweet from the conceptual space to the emotional space. In particular, we have six sets $E_{anger}, E_{disgust}, \cdots, E_{surprise}$ of vectors constituting the sub-symbolic coding of the words belonging to the lexicon for a particular emotion together with their artifact sentences. The generic vector belonging to one of the sets will be denoted in the following as $\mathbf{b}_i^{(e)}$ where $e \in \{$"*anger*", "*disgust*", "*fear*", "*joy*", "*sadness*", "*surprise*"$\}$ and i is the index that identifies the i-th $\mathbf{b}_i^{(e)}$ in the e set. Specifically, $\mathbf{b}_i^{(e)}$ is computed as:

$$\mathbf{b}_i^{(e)} = \mathbf{b}^T \mathbf{U}_R \Sigma_R^{-1} \tag{4}$$

where $\mathbf{b}$ is, time by time, the vector computed starting from one of the 3342 textual artifacts b according to the procedure illustrated at the beginning of this section.
Analogously, any textual content t of a tweet can be mapped into the Data Driven "conceptual" space by computing a vector $\mathbf{t}$ whose i-th component in the number of times the i-th word of the vocabulary, corresponding to the i-th row of $\mathbf{U}_R$, appears in t. This leads to the mapping of the tweet as:

$$\mathbf{t}_R = \mathbf{t}^T \mathbf{U}_R \Sigma_R^{-1} \tag{5}$$

Once the tweet t is mapped into the "conceptual" space as a vector $\mathbf{t}_R$, it is possible to compute its emotional fingerprint by exploiting the vectors $\mathbf{b}_i^{(e)}$, which act as "beacons" for the vector $\mathbf{t}_R$, helping in finding its position inside the conceptual space.
In particular, fixed $\mathbf{t}_R$, for each set E_e it is computed the weight:

$$w_e = \max cos(\mathbf{t}_R, \mathbf{b}_i^{(e)}) \tag{6}$$

once all the six w_e weights are computed, the vector $\mathbf{f}_t$, associated to the vector $\mathbf{t}_R$, and by consequence to the tweet t, is calculated as:

$$\mathbf{f}_t = \left[\frac{w_{(anger)}}{\sqrt{\sum_e w_e^2}}, \frac{w_{(fear)}}{\sqrt{\sum_e w_e^2}}, \cdots, \frac{w_{(surprise)}}{\sqrt{\sum_e w_e^2}} \right] \tag{7}$$

The vector $\mathbf{f}_t$ finally constitutes the *emotional fingerprint* of the tweet t in the emotional space. The emotional space is therefore a six-dimensional hypersphere where all tweets can be mapped and grouped. We call the fingerprint $\mathbf{f}_t$ "emoxel", analogously as the *knoxel* in the conceptual space paradigm [4].

3 Preliminary Experiments

Starting from the emotional lexicon, we have retrieved a set of 75078 unique tweets that contained at least one of the words included in the emotional lexicon. Furthermore, we have subdivided the lexicon into six subsets of words, each one corresponding to a given primary emotion, and we have created, for each emotion, 200 artificial sequences of ten words belonging

to the same subset. This leads to 1200 artificial documents that are oriented at artificially creating a bond among words belonging to the same subset of words carrying an emotional content. It is worthwhile to point out that adding these artifacts influences the Twitter dataset only for 1,68%. Moreover, we have included into the Twitter dataset also the dataset of 1250 news headlines used either as a trial or a test set in Strapparava et al. [12] for the task of classification of emotions in news headlines. We have chosen to add these texts to increase the number of short contexts that can involve emotional words in a more formal context concerning the language used in Tweets.
This led to a set of 77528 short documents from which inducing the "emotional" space. For the choice of the truncation parameter, we have experimentally chosen the threshold of $R = 100$.
Below we report the words describing the first ten dimensions of the space induced by the 77528 short texts used to build the data driven conceptual space. For each axis, we also report the weight of the singular value σ_j for that axis.

1. $\sigma_j = 13.85$ like, don't, love, people, good, out, feel, day, really, time
2. $\sigma_j = 11.13$ love, don't, like, people, brotherly, puppy, feel, much, out, god
3. $\sigma_j = 10.78$ pope, francis, high earned penchant pleas praise pictures photos crowd
4. $\sigma_j = 10.61$ best, ever, live skit comedy watch seen don't hilary checkout
5. $\sigma_j = 10.30$ don't, good, like, day, love, people, ever, happy, skit, live
6. $\sigma_j = 10.09$ like, don't, feel, people, good, care, look, trump, worry, health
7. $\sigma_j = 9.93$ people, good, don't, day, like, today, feel, most, morning, look
8. $\sigma_j = 9.79$ thoughts, leo, confident, between, tender, caught, good, don't, people, day
9. $\sigma_j = 9.53$ good, out, people, don't, new, trump, time, care, donald, ideas
10. $\sigma_j = 9.43$ trump, care, health, ideas, donald, republican, bewilder, experts, good, out

In the following, we report the preliminary results on emotional fingerprints of some tweets regarding two news stories at the time of the writing of this paper. In particular we have queried Twitter for "st.marteen" and the hashtag "#charliegard".

1) #boeing747 amazing landing at princess juliana airport, st. maarten. (klm asia) -youtube-'
$\mathbf{f_1} = [anger = 0.33, disgust = 0.08, fear = 0.26, Joy = 0.24, Sadness = 0.29, Surprise = 0.82]$
As it can be seen, in this case the main feeling is "surprise", maybe due to the "amazing landing";

2) @abc: tourist dies after being blown away by jet blast at a popular tourist spot near the end of a runway in st. maarten. ...'
$\mathbf{f_2} = [anger = 0.24, disgust = 0.04, fear = 0.85, Joy = 0.28, Sadness = 0.37, Surprise = 0.08]$
For this tweet the main emotion is "fear",

3) jet blast at st maarten's seaside airport leads to death of new zealand tourist video - the guardian
$\mathbf{f_3} = [anger = 0.53, disgust = 0.05, fear = 0.46, Joy = 0.36, Sadness = 0.60, Surprise = 0.09]$
In this case the main feeling is "sadness" followed by "anger" and "fear";

4) @fight4charlie: thank you to all of charlie's army for your unwavering support! keep it going! #charliegard"
$\mathbf{f_4} = [anger = 0.34, disgust = 0.08, fear = 0.33, Joy = 0.71, Sadness = 0.44, Surprise = 0.26]$
For this tweet the main emotion of the fingerprint is "Joy".

4 Conclusions and future works

We have illustrated a procedure that exploits a data-driven conceptual space as a bridge for mapping the textual content of a tweet as a point into a six-dimensional emotional hypersphere. The six dimensions are those of Ekman. The approach is promising, and it will be extended and improved by both fine tuning the parameters and modifying the mapping procedure, as well as making comparisons with other methods of emotion detection from texts.

References

[1] Francesco Agostaro, Agnese Augello, Giovanni Pilato, Giorgio Vassallo, and Salvatore Gaglio. A conversational agent based on a conceptual interpretation of a data driven semantic space. In *AI* IA*, volume 3673, pages 381–392. Springer, 2005.

[2] Eugene Y Bann and Joanna J Bryson. The conceptualisation of emotion qualia: Semantic clustering of emotional tweets. In *Proceedings of the 13th Neural Computation and Psychology Workshop on Computational Models of Cognitive Processes*, pages 249–263, 2013.

[3] Johan Bollen, Huina Mao, and Alberto Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *ICWSM*, 11:450–453, 2011.

[4] Antonio Chella, Marcello Frixione, and Salvatore Gaglio. A cognitive architecture for robot self-consciousness. *Artificial intelligence in medicine*, 44(2):147–154, 2008.

[5] E. DAvanzo, G. Pilato, and M. D.. Lytras. Using twitter sentiment and emotions analysis of google trends for decisions making. In *Program*, pages Vol. 51, Issue 3. 2017.

[6] Ernesto DAvanzo, AM Gliozzo, and Carlo Strapparava. Automatic acquisition of domain information for lexical concepts. In *Proceedings of the second MEANING workshop, Trento, Italy*, volume 150, 2005.

[7] Peter Gardenfors. Conceptual spaces as a framework for knowledge representation. *Mind and Matter*, 2(2):9–27, 2004.

[8] Thomas K Landauer and Susan T Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.

[9] Thomas K Landauer, Peter W Foltz, and Darrell Laham. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284, 1998.

[10] Sara Santilli, Laura Nota, and Giovanni Pilato. The use of latent semantic analysis in the positive psychology: A comparison with twitter posts. In *Semantic Computing (ICSC), 2017 IEEE 11th International Conference on*, pages 494–498. IEEE, 2017.

[11] C. Strapparava and R Mihalcea. Learning to identify emotions in text. In *SAC '08 Proceedings of the 2008 ACM symposium on Applied computing*. 2008.

[12] Carlo Strapparava and Rada Mihalcea. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 70–74. Association for Computational Linguistics, 2007.