Procedia Computer Science

Volume 88, 2016, Pages 107–114

7th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2016

A Framework based on Semantic Spaces and Glyphs for Social Sensing on Twitter

Giovanni Pilato[1] and Umberto Maniscalco[1]

ICAR-CNR, Italian National Research Council Palermo, Italy
{giovanni.pilato,umberto.maniscalco}@cnr.it

Abstract

In this paper we present a framework aimed at detecting emotions and sentiments in a Twitter stream. The approach uses the well-founded Latent Semantic Analysis technique, which can be seen as a bio-insipred cognitive architecture, to induce a semantic space where tweets are mapped and analysed by soft sensors. The measurements of the soft sensors are then used by a visualisation module which exploits glyphs to graphically present them. The result is an interactive map which makes easy the exploration of reactions and opinions in the whole globe regarding tweets retrieved from specific queries.

Keywords: Social Sensing, LSA, Soft Sensors, Emotion analysis, Sentiment Analysis

1 Introduction

Microblogging has shown a massive growth in popularity in recent years. One of the most representative examples of microblogging services is Twitter [19]. On the other hand, the interest in social sensing, named also participatory sensing [1],[6], together with sentiment analysis methods and techniques for recognition of emotions and opinions in social networks have become a field of great interest. Since sentiments and emotions characterize many aspects of human living many researchers tried to identify if a text can be considered as being subjective or objective [2] and also if an opinion expressed has a positive or negative polarity [14][13]. Moreover, understanding the emotional content expressed in social networks can help in inferring and describing the emotional status of a community, a group of people, a city, or even a country with regard to specific topics[16][9].

In this paper we illustrate a preliminary framework which tries to implement social sensing on Twitter by exploiting a semantic space that can be automatically induced from data. The semantic space is built by means of a Latent Semantic Analysis (LSA) [11] approach applied to a set of tweets downloaded from the Twitter streaming. The LSA approach has been chosen since it has been successfully used to simulate many psycholinguistic phenomena [12] and it has also a statistical foundation[15]. Once the semantic space is built, it can be exploited to map in it the tweets that are sensed from the Twitter stream according to given queries. The system

Selection and peer-review under responsibility of the Scientific Programme Committee of BICA 2016

doi:10.1016/j.procs.2016.07.413

uses three soft sensors on the "perceived" tweets trying to infer both sentiments and emotions expressed in them, according to the belief that "it is possible to infer emotion properties from the emotion words"[5]. The first soft sensor is aimed at recognizing the sentiment orientation (positive, neutral or negative) in a tweet; the second one is aimed at sensing the emotions in a tweet. A third soft sensor extracts the geospatial coordinates, if they are available, of the tweets under analysis. The general emotional content regarding the topics described by the keywords is modeled and specified on Ekman's emotional scale [7], which assumes that there is a finite number of basic and discrete emotions and specifies the following six human emotions: anger, disgust, fear, happiness, sadness and surprise [3][4]. Once the sensing is completed, the information coming from the two sensors is sent to a visualization module which graphically illustrates in a world map, by using glyphs, the main sentiment and the emotions arising from the people for a specific topic. This approach can help for deep understanding people's behavior and for providing at the same time a number of indicative factors regarding specific problems, people, ideas, items, etc.

2 Latent Semantic Analysis

Latent Semantic Analysis (LSA) [11][15] is a paradigm to extract and characterize the meaning of words by statistical computations applied to a large corpus of texts. LSA is based on the vector space method: a text corpus is represented as a matrix $\mathbf{M}$ where rows are related to words, and columns are associated to documents or other contexts. The LSA paradigm defines a mapping between words and documents inducing from data a continuous vector space S, where the $i-th$ word w_i of a dictionary is associated to a vector $\mathbf{u}_i$ in S, and the $j-th$ document d_j of a text corpus is associated a vector $\mathbf{v}_j$ in S [20].

The S vector space is a "semantic space", since semantics conveyed by the presence of the i-th word in the j-th document can be measured by taking the dot product between the vector representing the word and the vector representing the document. The LSA approach learns automatically the similarity between the meanings of words, and bridges the gap between the information available in a set of text chunks and the knowledge people acquires after a large amount of experience. LSA has besides presented as a theory of learning, memory and knowledge; furthermore it is roughly equivalent to a neural network model, which is a typical bio-inspired cognitive architecture. As reported by Landauer et al, [12] it is supposed that the mind-brain stores and reprocesses its inputs in some manner that has the same effect of the Singular Value Decomposition operation.

3 The proposed approach

The proposed framework makes use of two kind of lexicons: *a)* a lexicon S containing a list of positive and negative subjective words from the Janyce Wiebe's subjectivity lexicon [17]; *b)* a lexicon E derived by Strapparava et al. in [18] containing words that are related to the six basic emotions of Ekman: *anger*, *disgust*, *fear*, *joy*, *sadness* and *surprise*. The whole approach consists of two phases: a *training* phase and an *production* phase, which are illustrated below.

During the *training* phase the two lexicons are merged together in order to obtain a unique list of words. This list is then used by a module that, exploiting the Twitter APIs, retrieves from the Twitter stream those tweets that are written in english language and that contain a given word. The set of retrieved tweets constitutes a corpus, which is used to automatically induce the semantic space by means of LSA. Once the LSA space is built, it can be used to map both