Computers in Human Behavior xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Computers in Human Behavior

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

Sentiment detection in social networks and in collaborative learning environments

Francesco Colace [a], Luca Casaburi [a], Massimo De Santo [a], Luca Greco [b,*]

[a] DIEM, University of Salerno, Fisciano, Italy
[b] DIIN, University of Salerno, Fisciano, Italy

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Information extraction
Sentiment analysis
Latent Dirichlet allocation

ABSTRACT

Daily millions of messages appear on the web, which is becoming a rich source of data for opinion mining and sentiment analysis. The computational study of opinions, feelings and emotions expressed in a text often relates to the identification of agreement or disagreement with statements, contained in comments or reviews, that convey positive or negative feelings. The detection and analysis of sentiment in textual communication is a topic attracting attention also in the context of collaborative learning in social networks, being learners actively engaged in presenting and defending ideas and opinions, as well as exchanging moods about courses with peers. In this paper, we investigate the adoption of a probabilistic approach based on the Latent Dirichlet Allocation (LDA) as Sentiment Grabber. Through this approach, for a set of documents belonging to a same knowledge domain, a graph, the Mixed Graph of Terms, can be automatically extracted. The paper shows how this graph contains a set of weighted word pairs, which are discriminative for sentiment classification. The proposed method has been tested in different context: a standard dataset containing movie reviews; a real-time analysis of social networks posts; a collaborative learning scenario. The experimental evaluation shows how the proposed approach is effective and satisfactory.

1. Introduction

Thanks to blogs, microblogs, social networks or reviews collector sites, millions of messages appear daily on the web. In general, this textual information can be divided in two main categories: facts and opinions. Facts are objective statements while opinions reflect people's sentiments about products, other people and events; the latter appear to be extremely important, being able to influence the decisional process (Dascalu, Bodea, Lytras, de Pablos, & Burlacu, 2014; Lytras & de Pablos, 2008, 2011; Lytras, Sicilia, Naeve, de Pablos, & Lytras, 2008; Sebastiani, 2002). The interest, that potential customers show in opinions and reviews about products, is something that vendors are gradually paying more and more attention to. Companies are interested in what customers say about their products as politicians are interested in how different news media are portraying them. The detection and analysis of sentiment in textual communication is a topic attracting attention also in the context of collaborative learning in social networks, being learners actively engaged in presenting and defending ideas and opinions, as well as exchanging diverse beliefs with peers. Many institutions are adopting e-learning and collaborative learning both to improve their traditional courses and increase the potential audience since it allows more flexibility and quality in general. Anyway, e-classrooms are often composed by students inattentive or appearing bored and wondered. So the main question for the teacher becomes: why am I not able to reach these students and catch their attention? why are they not excited about the material although my efforts to present it in an organized and coherent manner? This sense of frustration increases when he faces students' poor performance on tests. Recent studies showed that emotions can affect the e-learning experience. What are emotions? A general definition for emotions is the following: emotions are complex psychophysical processes that evoke positive or negative psychological responses (or both) and physical expressions, often involuntary. Emotions are often related to feelings, perceptions or beliefs about elements, objects or relations between them, in reality or in the imagination. They typically arise spontaneously, rather than through conscious effort. An emotion (reaction or state) is often differentiated from a feeling (sensation or impression), although the word "feeling" is used as a synonym for "emotion" in some contexts. In fact, emotion has to do with how one feels. This feeling, if positive is believed to have a productive effect on the individual; otherwise it seems to impact

* Corresponding author.
E-mail address: lgreco@unisa.it (L. Greco).

http://dx.doi.org/10.1016/j.chb.2014.11.090
0747-5632/

negatively on the individual's learning experience. Obviously, the topic of emotions goes far beyond this simple definition and it is especially hard to detect in an e-learning environment. In a face-to-face class instructors can detect facial expressions of students but, in an online environment, students need to establish an online presence and the instructors need to be able to pick up on this. In this scenario, a promising approach is the sentiment analysis: the computational study of opinions, sentiments and emotions expressed in a text (Liu, 2010; Wang et al., 2013). In literature, there are many approaches to the sentiment analysis. A very broad overview of the existing work was presented in Pang and Lee (2008). The authors describe in detail the main techniques and approaches for an opinion oriented information retrieval. Early work in this area was focused on determining the semantic orientation of documents. In particular some approaches attempt to learn a positive–negative classifier at a document level. Turney (2002) introduces the results of review classification by considering the algebraic sum of the orientation of terms as respective of the orientation of the documents. Starting from this approach other techniques have been developed by focusing on some specific tasks as finding the sentiment of words (Wilson, Wiebe, & Hwa, 2004). Baroni and Vegnaduzzo (2004) proposed to rank a large list of adjectives according to a subjectivity score by employing a small set of manually selected subjective adjectives and computing the mutual information of pairs of adjectives using frequency and co-occurrence frequency counts on the web. The work of Turney and Littman (2002) proposes an approach to measure the semantic orientation of a given word based on the strength of its association with a set of context insensitive positive words minus the strength of its association with a set of negative words. By this approach sentiment lexicons can be built and a sentiment polarity score can be assigned to each word (Gamon & Aue, 2005; Neviarouskaya, Prendinger, & Ishizuka, 2011). Sentiment polarity score means the strength or degree of sentiment in a defined sentence pattern. Artificial Intelligence and probabilistic approaches have also been adopted for sentiment mining. In Pang, Lee, and Vaithyanathan (2002) three machine learning approaches (Naïve Bayes, Maximum Entropy and Support Vector Machines) has been adopted to label the polarity of a movie reviews datasets. A promising approach is presented in Prabowo and Thelwall (2009) where a novel methodology has been obtained by the combination of rule based classification, supervised learning and machine learning. In Shein (2009a) a SVM based technique has been introduced for classifying the sentiment in a collection of documents. Other approaches are inferring the sentiment orientation of social media content and estimate sentiment orientations of a collection of documents as a text classification problem (Colbaugh & Glass, 2010). More in general, sentiment related information can be encoded within the actual words of the sentence through changes in attitudinal shades of word meaning using suffixes as discussed in Esuli and Sebastiani (2006). This has been investigated in Neviarouskaya, Prendinger, and Ishizuka (2011) where a lexicon for sentiment analysis has been obtained. In Yu, Liu, Huang, and An (2012) a probabilistic approach to sentiment mining is adopted. In particular this paper uses a probabilistic model called Sentiment Probabilistic Latent Semantic Analysis (S-PLSA) in which a review, and more in general a document, can be considered as generated under the influence of a number of hidden sentiment factors (Hofmann, 1999). The S-PLSA is an extension of the PLSA where it is assumed that there are a set of hidden semantic factors or aspects in the documents related to documents and words under a probabilistic framework. In this paper, we investigate the adoption of a similar approach based on the Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). In LDA, each document may be viewed as composed by a mixture of various topics. This is similar to probabilistic latent semantic analysis, except that in LDA the topic distribution is assumed to have a Dirichlet prior. By the use of the LDA approach on a set of documents belonging to a same knowledge domain, a Mixed Graph of Terms can be automatically extracted (Clarizia, Colace, De Santo, Greco, & Napoletano, 2011; Clarizia, Colace, Greco, De Santo, & Napoletano, 2011; Colace, De Santo, Greco, & Napoletano, 2014; Napoletano, Colace, De Santo, & Greco, 2012). Such a graph contains a set of weighted word pairs (Colace, De Santo, Greco, & Napoletano, 2015), which we demonstrate to be discriminative for sentiment classification.

The main reason of such discriminative power is that LDA-based topic modeling is essentially an effective conceptual clustering process and it helps discover semantically rich concepts describing the respective "sentimental" relationships. By means of applying these semantically rich concepts, that contain more useful relationship indicators to identify the sentiment from messages and by using a terminological Ontology Builder which allows to identify the kind of semantic relationship between word pairs in mGT, the proposed system can accurately discover more latent relationships and make less errors in its predictions.

The rationale of this paper is the following: Section 2 discusses the extraction of a Mixed Graphs of Terms from a document corpus; Section 3 introduces the proposed approach for the sentiment extraction. The Section 4 discusses the experimental results. Finally, conclusions and further works are discussed.

2. Extracting a Mixed Graph of Terms

In this paper we explain how a complex structure, that we call a Mixed Graph of Terms (mGT), allows to capture and represent the information contained in a set of documents that belong to a well-defined knowledge domain. Such a graph can be automatically extracted from a document corpus and can be effectively used as a filter to employ in document classification as well as in sentiment extraction problems. Formally, a Mixed Graph of Terms can be defined as a graph $\mathbf{g} = \langle N, E \rangle$ where:

- $N = \{R, W\}$ is a finite set of nodes, covered by the set $R = \{r_1, \ldots, r_H\}$ whose elements are the *aggregate roots* and by the set $W = \{w_1, \ldots, w_M\}$ containing the *aggregates*. Aggregate roots can be defined as the words whose occurrence is most implied from the occurrence of all other words in the training corpus. Aggregates are defined as the words most related to aggregate roots from a probabilistic point of view.
- $E = \{E_{RR}, E_{RW}\}$ is a set of edges, covered by the set $E_{RR} = \{e_{r_1 r_2}, \ldots, e_{r_{H-1} r_H}\}$ whose elements are links between aggregate roots and by the set $E_{RW} = \{e_{r_1 w_1}, \ldots, e_{r_H w_M}\}$ whose elements are links between aggregate roots and aggregates.
  As better explained further, two aggregate roots are linked if strongly correlated (in a probabilistic sense):

$$e_{r_i r_j} = \begin{cases} 1 & \text{if } \psi_{ij} \geqslant \tau \\ 0 & \text{otherwise} \end{cases}$$

  Aggregate roots can be also linked to aggregates if a relevant probabilistic correlation is present:

$$e_{r_i w_s} = \begin{cases} 1 & \text{if } \rho_{is} \geqslant \mu_i \\ 0 & \text{otherwise} \end{cases}$$

Details about mGT building and thresholds τ and μ_i will be now discussed. The Feature Extraction module (FE) is represented in Fig. 1. The input of the system is the set of documents:

$$\Omega_r = (\mathbf{d}_1, \ldots, \mathbf{d}_M)$$

Please cite this article in press as: Colace, F., et al. Sentiment detection in social networks and in collaborative learning environments. *Computers in Human Behavior* (2014), http://dx.doi.org/10.1016/j.chb.2014.11.090