Computer Standards & Interfaces 35 (2013) 442–453

Contents lists available at SciVerse ScienceDirect

Computer Standards & Interfaces

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

Sentiment analysis via dependency parsing

Luigi Di Caro [a,*], Matteo Grella [b]

[a] University of Turin, Turin, Italy
[b] Parsit, Turin, Italy

ARTICLE INFO

Available online 23 November 2012

Keywords:
Sentiment analysis
Data visualization
User interfaces

ABSTRACT

Nowadays, Sentiment Analysis (SA) is receiving huge attention because of the wide range of its direct applications like analyses of products, customer profiles, political trends, and so forth. Still, the availability of big amounts of data coming from the World Wide Web makes easier the study of both new techniques and evaluation methods. Current literature mainly focuses on two approaches which rely on sentiment lexicons (i.e., lists of words associated to scores of sentiment polarity) or on Natural Language Processing techniques (NLP). In this paper, on one hand, we introduce and evaluate a novel algorithm for SA that relies on a simple set of propagation rules applied at syntactic level within a dependency parse tree. On the other hand, we propose a context-based model where the users' sentiments (or opinions) are tuned according to some context of analysis. Finally, we present the system called SentiVis which implements these ideas through an orthogonal approach to SA that directly leans on Data Visualization. Extracted sentiments, with respect to some query of analysis, are ordered and represented graphically in a 2-dimensional space, conveying information about their strength and variability. This way, we avoid cumbersome rankings of objects and associated opinions by directly mapping such information on the screen. The user is then able to interact with the visualized data in order to discover interesting facts as well as removing false positive (or negative) opinions deriving by the used algorithm. We then evaluate the efficacy of the proposed system through several case studies.

1. Background

Sentiment Analysis can be generally defined as the extraction of users' opinions from texts. Given the large amount of data available on the Web, it is now possible to investigate high-level Information Retrieval tasks like users' intentions and feelings about facts or objects discussed [38] is maybe the most accurate and complete survey on this topic.

This task has an interesting range of applications. For example, a product seller may be interested in the users' opinions about its products. Still, a politician could want to understand what people think about her or him [35], and so forth.

There is a large literature on Social Sciences and Psychology talking about emotions. For instance, several studies demonstrated that the perception of the emotions changes with respect to the experience of a person and her personal issues [9], and to the gender [48]. Many studies have also tried to understand a sort of primary set of emotions although without a perfect agreement [15,41,53]. Still, it has also been demonstrated that people assume different and even independent emotions within the same assertion [20]. Other aspects are the roles of the emotions in general discourses [7,40,45], and in on-line communications [19,23,31].

In general, there are several granularities regarding this type of analysis. A common one is to understand whether a text represents an *objective* or a *subjective* thought [36,25]. A deeper approach is based on the analysis of the *polarity* (or *valence*) of the text, i.e., positive or negative. Most of the approaches concentrate on this task. Going further, another problem is to figure out the *strength* (also called *arousal*) of an emotional state underlying a text [52,49]. A more complex problem is finding the exact *emotions* covered by a textual expression, like "happy", "sad", "angry" and so on [58,13]. Finally, the most challenging problem is represented by the extraction of users' intentions, arguments and speculations [56].

From a procedural point of view, Sentiment Analysis can be done at different levels, i.e., at word level, sentences level, or document level. Moreover, the analysis has to be carefully adapted to the context and the source of the data. Texts could be news articles (substantially well-written and formal), tweets (short text messages coming from the Web Service Twitter) [1] [10,11], phone messages [22], bulletin boards, chatrooms, sites ([8,14]). In this sense, with the advent of the Web 2.0 and its related social services, other aspects of subjectivity came out like the concept of *mood* (i.e., an emotional state that the users can attach with the text). [29,30] are a few attempts on the

* Corresponding author.
E-mail addresses: dicaro@di.unito.it (L. Di Caro), matteo.grella@parsit.it (M. Grella).

[1] http://twitter.com/

0920-5489/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csi.2012.10.005

analysis of such data. Still, *emoticons* play an important role in the expressiveness of this kind of texts [18,54,32,43].

2. Motivation

As already mentioned in Section 1, there are several things to take into account when facing a Sentiment Analysis task. Generally speaking, there exist two main approaches (and hybrid solutions): the use of sentiment lexicons (eventually with Natural Language Processing (NLP) techniques), and the use of Machine Learning classification algorithms. A Sentiment lexicon is a list of words that are associated to polarity values (positive or negative). One of the most known is ANEW [12], tebradley1999affective, developed for linguistic studies even before the concept of Sentiment Analysis in Computer Science. Other resources are General Inquirer [47], Opinion Finder [55], Wordnet-Affect [50], Senti-Wordnet [6], Q-Wordnet [3], Senti-strength [52] and a few others [34,4]. NLP techniques represent a deeper level of analysis since they take into account the context in which the words appear (it is well known that words in natural languages can have different meaning with respect to the context). Some of the approaches presented in literature include the use of n-grams [2,33,51], lexical and syntactic patterns [44], rule-based systems [42,58] and others.

On the other hand, SA can be viewed as a Machine Learning classification problem. Generally speaking, the task becomes to classify the opinion as falling under one of two opposing sentiment polarities, or locate its position on the continuum between these two polarities, given an opinionated piece of text [38]. There exist many works that face SA under this view, like [39] which uses Support Vector Machines (SVM) over a dataset of movies. [2] In our opinion, relevant papers on this direction are [1,2,5,21,29,57,37]. Our intent, instead, is more related to NLP rather than Machine Learning, that is, we aim to develop an unsupervised system which does not learn from any specific training set. For this reason we do not extensively go over such Machine Learning approaches in this paper.

In this paper, we first propose a new algorithm for SA based on some propagation rules that work at syntactic level, then we present a Data Visualization approach which is independent from any algorithm used for SA, so it can be seen as complementary. In fact, we think that apart from the goodness and the effectiveness of one algorithm for the extraction of users' opinions, the intrinsic automatism of such methods would inevitably carry around some error, due to the algorithm itself or to subjective aspects of the users' comments [9]. For this reason, there is the need of *taking the human in the loop*, that means to show her at-a-glance views over the opinions and let her decide to take some and discard the rest according to her purposes. This could be done by Data Visualization techniques which map texts and extracted opinions into 2-dimensional displays.

3. SPR: sentiment analysis by syntactic-based propagation rules

Our algorithm for SA relies on the concept of *Sentiment Propagation*, which assumes that each linguistic element like a noun, a verb, etc. can have an intrinsic value of *sentiment* that is propagated through the syntactic structure of the parsed sentence.

In the next sections we present a set of syntactic-based rules that aim at covering a significant part of the sentiment salience expressed by a text. It is important to note that this approach is based on the result of a dependency parser, thus the effectiveness of the method directly depends on the quality of the parsing procedure.

3.1. Dependency parsing

Our approach for SA is totally based on a deep NLP analysis of the sentences, using a dependency parsing as pre-processing task. While we initially developed this technique for Italian, using the Italian parser TULE [28], we propose here an adaptation for English making use of the Stanford Dependency Parser [16] (abbreviated with SDP from now on). After selecting some of the dependencies defined for SPD [17], we selected five categories of propagation processes, one referring to a specific linguistic construct:

- **MODS** Modifiers are all constructs that modify a *verb*, a *noun*, or another *modifier*. They can be simple *adjectives*, or *nouns* (functioning as adjectives), or *adverbs*.
- **TUNS** The tuners are a kind of adverbial modifiers that strengthen (or weaken) the sentiment value of a word. For example, the adverb "*very*" tends to increase the strength of the governor (both positive or negative), while "textitalmost" is assumed to reduce it.
- **INVS** The inverters come from negation-like dependencies, and they are represented by words like "*not*" and "*never*".
- **PREPS** The prepositions "*to*", "*with*", "*in*" and so forth, are considered as channels that vehiculate sentiment values between the terms that they connect.
- **VERBS** The verbs can have sentiment connotations whose values are transmitted to their arguments.

The idea is that the sentiment values of such components influence each other at syntactic level, so their actual *sentiment* is the result of a propagation within the network represented by the dependency tree. Note that this approach formalizes and generalizes what recently assumed in [25] concerning the existing linguistic phenomena and their relative influence on polarity.

In order to define the propagation mechanism, we will use the following simple notation: given a term x, its sentiment value is defined as S_x.

3.1.1. INVS: the inverters

This first rule applied by SPR manages to invert the sign of the sentiment values in case of negative-like dependencies. For instance, in the sentence:

I didn't like the food of that restaurant.

here the word *like* has a negative connotation, even if it actually has a positive polarity if taken alone. Given a dependency neg(x, n), where

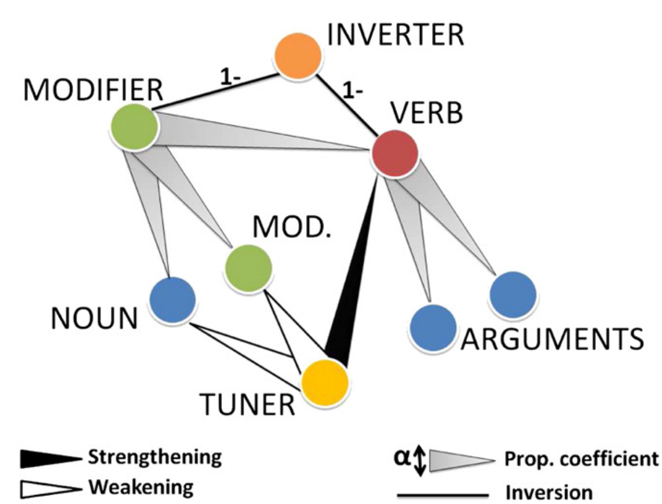

Fig. 1. The general architecture of the *sentiment propagation* between linguistic components. The *sentiment* is propagated from modifiers to verbs, nouns and modifiers, and from verbs to the arguments (that can be other noun phrases, recursively). The degree of propagation is defined by α, that is the sender's percentage of sentiment value to be transmitted to the recipient.

[2] http://www.cs.cornell.edu/people/pabo/movie-review-data/