Web Semantics: Science, Services and Agents on the World Wide Web 34 (2015) 27–39

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

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

SACI: Sentiment analysis by collective inspection on social media content

Leonardo Rocha [a,*], Fernando Mourão [a], Thiago Silveira [a], Rodrigo Chaves [a], Giovanni Sá [a], Felipe Teixeira [a], Ramon Vieira [a], Renato Ferreira [b]

[a] Universidade Federal de São João Del Rei, Department of Computer Science, São João Del Rei, MG, Brazil
[b] Universidade Federal de Minas Gerais, Department of Computer Science, Belo Horizonte, MG, Brazil

ARTICLE INFO

Article history:
Received 5 September 2014
Received in revised form
9 April 2015
Accepted 25 May 2015
Available online 15 June 2015

Keywords:
Sentiment analysis
Classification
Transition graph

ABSTRACT

Collective opinions observed in Social Media represent valuable information for a range of applications. On the pursuit of such information, current methods require a prior knowledge of each individual opinion to determine the collective one in a post collection. Differently, we assume that collective analysis could be better performed when exploiting overlaps among distinct posts of the collection. Thus, we propose *SACI* (Sentiment Analysis by Collective Inspection), a lexicon-based unsupervised method that extracts collective sentiments without concerning with individual classifications. *SACI* is based on a directed transition graph among terms of a post set and on a prior classification of these terms regarding their roles in consolidating opinions. Paths represent subsets of posts on this graph and the collective opinion is defined by traversing all paths. Besides demonstrating that collective analysis outperforms individual one w.r.t. approximating collection opinions, assessments on *SACI* show that good individual classifications do not guarantee good collective analysis and vice-versa. Further, *SACI* fulfills simultaneously requirements of efficacy, efficiency and handle of dynamicity posed by high demanding scenarios. Indeed, the consolidation of a *SACI*-based Web tool for real-time analysis of tweets evinces the usefulness of this work.

1. Introduction

Social media has emerged as an important environment wherein people publish their opinions about distinct topics on the Web, providing an unprecedented source of information for modeling and understanding user behavior. Hence, a growing number of efforts aim to adapt traditional computational analysis to this new scenario. Given the practical value of such subjective content, particular attention has been given to **Sentiment Analysis** (SA) on Social Media, that is, the automatic extraction and identification of subjective information from textual data [1–4]. However, SA on Social Media is even more challenging than on other scenarios, such as product reviews, since its content constitutes a fast-evolving stream of short, unstructured and domain-specific textual data [5].

This work addresses the challenge of identifying the collective sentiment about a target entity (e.g., product, person or service) mentioned in a stream of textual documents. By collective sentiment we mean the predominant sentiment observed in the whole set of documents. Traditionally, each document is classified individually w.r.t. its sentiment and the resulting classifications are combined in order to compose an aggregated analysis, since individual perspectives are in general not representative enough. However, classifying short documents (e.g., posts) is difficult due to the absence of contextual information. A typical domain that depicts this scenario is the SA of political candidates or parties in social networks. In this case, usually, we are not concerned about the opinion of a given individual, but about the predominant opinion of a population. Classifying each post published by each individual and, then, deriving a collective opinion might be challenging and unreliable. Differently, we intend to determine the collective sentiment by classifying bunches of documents simultaneously, which we named collective analysis, instead of each one individually. In this sense, we propose *SACI* (Sentiment Analysis by Collective Inspection), a new lexicon-based unsupervised method that handles efficiently a huge volume of documents extracting the collective sentiment without concerning with individual classifications.

SACI uses a directed probabilistic graph of transitions among terms that occur in a document set D about a target entity. Each node represents a term (i.e., a single word) and edges express the

* Corresponding author.
E-mail addresses: lcrocha@ufsj.edu.br (L. Rocha), fhmourao@ufsj.edu.br (F. Mourão), tssilveira@ufsj.edu.br (T. Silveira), rachaves@ufsj.edu.br (R. Chaves), giovannisa@ufsj.edu.br (G. Sá), fcteixeira@ufsj.edu.br (F. Teixeira), ramonv@ufsj.edu.br (R. Vieira), renato@dcc.ufmg.br (R. Ferreira).

http://dx.doi.org/10.1016/j.websem.2015.05.006
1570-8268/

probability of a term being adjacent to another in at least one document of *D*. Further, each node has an attribute that represents its lexicon-semantic class, which defines specific transformations performed by each single term on the sentiment of a sentence. Thus, each path in our graph represents a distinct subset of sentences observed within documents of *D*. We assign a single lexicon-semantic class to each path by traversing it and applying successive semantic transformations, according to the class of each reached term. As each path also has a probability of occurrence, the collective sentiment of the whole set *D* is given by summing up the probability of occurrence associated with positive, negative and neutral paths. The main hypothesis exploited by *SACI* is that the overlap among distinct documents may emphasize consensual or most frequent opinions, whereas rare individual viewpoints become less relevant for the collective sentiment.

Aiming to evaluate SACI, we compare its collective analysis with four aggregated strategies derived from two unsupervised [1,3] and two supervised [5,4] methods well-established in the literature in two real domains. The first domain comprises 20 USA TV series and the second one refers to the 2012 USA presidential election. Despite not being the most effective method to determine individual opinions, *SACI* outperformed aggregated analysis when identifying the collective opinion. These results evince that good individual sentiment classifications do not guarantee good collective analysis and vice-versa. Further, the proposed collective analysis allows us to fulfill simultaneously three main requirements posed by Social Media content, which are usually not aligned:

1. **Accuracy:** Collective analyses, such as performed by *SACI*, might provide statistical robustness against noises and complex behaviors (e.g., irony) by exploiting consensual opinions. Further, inductive and inferential procedures become more reliable, since more information is available in a set of opinions than in a single one, mainly within short textual data.
2. **Efficiency and scalability:** Besides not requiring a labeled training set, *SACI* relies only on a fast construction and analysis of transition graphs among terms, allowing us to classify quickly even data streams.
3. **Handle of dynamicity:** The probabilistic graphs defined by *SACI* can be instantaneously updated as soon as new data arrives, reflecting the most recent textual information while keeping the most common interactions among terms.

Indeed, *SACI*'s execution time grows linearly with the number of analyzed tweets. Further, the approximation error between the actual sentiment distribution and SACI's distribution was up to 86% smaller than the errors obtained by the aggregated strategies, even those based on supervised methods. In summary, we point out *SACI* as a promising method of collective SA, mainly, for high demanding scenarios, such as Social Media. As proof of concept, we also present a *SACI*-based Web tool able to conduct real-time analysis of tweets.

2. Related work

In the past few years, we have observed a growing interest in Sentiment Analysis (SA), whose goal is to extract subjective information from textual data [4,2,5,3]. Such interest stems from the consolidation of social media as a new, rich and huge source of subjective information about users. However, several studies have found that traditional methods of sentiment analysis, employed in scenarios such as product reviews, are not suitable for this new scenario [6,7]. Hence, some works have attempted to use additional information in order to enhance SA in this case. In [8], for instance, the authors used emoticons to train a Naïve Bayes classifier in order to perform the sentiment classification on tweets. Further, [6] presented a mathematical formulation for exploiting some sociological theories, such as sentiment consistency and emotional contagion. Besides facing difficulties to classify short textual data, due to the absence of contextual information, most of these methods cannot handle efficiently a huge volume of data. Several supervised machine learning algorithms, even when considering additional features extracted by natural language processing methods (e.g., syntactic classes of terms), struggle to perform properly SA on this scenario [9]. Thus, unsupervised methods are assuming an important role in the pursuit of effective and efficient approaches for sentiment analysis in social media content [10,1,11].

Most of the unsupervised sentiment analysis methods are based on sentiment lexicons and have usually two distinct steps. In the first one, the consolidation of a lexicon is performed. Based on such lexicon, the second step focuses on identifying the sentiment of each distinct document. Concerning about the consolidation of lexicons, most of the efforts can be divided into three main categories [6]. The first one refers to methods based on human annotations, in which terms are classified manually [12]. Despite providing good lexicons, these methods are labor intensive and, therefore, unfeasible for social media applications. The second category comprises dictionary-based methods that identify the sentiment of a term from semantically related terms in a specific dictionary (e.g., WordNet) [13,14]. These methods are, in general, computationally efficient, however, they do not take into account the application domain. Thus, terms always exhibit the same sentiment, regardless its domain of occurrence. Finally, the third category is known as corpus-based methods [15,1] on which a context-dependent sentiment is constructed, defining the sentiment of terms according to relations between terms observed in a corpus. The lexicons built by these methods might be as good as the manual construction with a reduced computational cost.

Considering the step of identifying the sentiment of each document, some works perform a straightforward use of the lexicons. For instance, [10] defined the sentiment of each tweet by verifying whether it contains positive or negative words, according to a lexicon from OpinionFinder [16]. Similarly, [11] used OpinionFinder and GPOMS (Google-Profile of Mood States) to determine the mood of each individual tweet. The goal in this case was to evaluate the correlation between mood states in Twitter and the Dow Jones Industrial Average over time. In turn, [17] investigated Twitter as a forum for political deliberation on the context of the German federal election. The authors used LIWC [18] to extract the sentiment of a set of tweets. LIWC is a text analysis tool that counts words in psychologically meaningful categories in order to assess emotional, cognitive, and structural components of text samples. Similarly, [19] developed a measure of positive/negative influence for popular users on Twitter using a term-based matching technique based on LIWC's output. Aiming to handle the lack of contextual information in the short text data from Social Media, recent works are taking advantage of all textual features available on posts [20–22]. For instance, Xia Hu et al. propose a new framework that considers emotional signals, such as emoticons or product ratings, present in each post to determine its sentiment [20]. In turn, the length of each word is exploited as another evidence of sentiment by [21].

Differently from previous efforts, *SACI* did not intend to classify documents individually on the second step. To the best of our knowledge, the only proposal in the literature that evaluates a collection of documents, instead of an individual approach, is [17]. The authors count word occurrence in a collection of documents, according to psychological categories defined by LIWC, and determine the collection sentiment as the categories' distribution. By modeling documents as a bag-of-words, this approach fails to capture useful information, such as co-occurrences and order of occurrence of words into each sentence. On the other hand, *SACI* uses a