Information Processing and Management 50 (2014) 899–908

Contents lists available at ScienceDirect

Information Processing and Management

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

Computational approaches for mining user's opinions on the Web 2.0

Gerald Petz [a,*], Michał Karpowicz [a], Harald Fürschuß [a], Andreas Auinger [a], Václav Stříteský [b], Andreas Holzinger [c]

[a] University of Applied Sciences Upper Austria, Campus Steyr, Wehrgrabengasse 1-3, 4400 Steyr, Austria
[b] University of Economics, Prague, W. Churchill Sq. 4, 13067 Prague, Czech Republic
[c] Medical University, Graz, Medical Informatics, Statistics and Documentation, Auenbruggerplatz 2/V, 8036 Graz, Austria

ARTICLE INFO

Article history:
Received 14 August 2013
Received in revised form 14 February 2014
Accepted 24 July 2014
Available online 24 August 2014

Keywords:
Opinion mining
Noisy text
Text preprocessing
User generated content
Data mining

ABSTRACT

The emerging research area of opinion mining deals with computational methods in order to find, extract and systematically analyze people's opinions, attitudes and emotions towards certain topics. While providing interesting market research information, the user generated content existing on the Web 2.0 presents numerous challenges regarding systematic analysis, the differences and unique characteristics of the various social media channels being one of them. This article reports on the determination of such particularities, and deduces their impact on text preprocessing and opinion mining algorithms. The effectiveness of different algorithms is evaluated in order to determine their applicability to the various social media channels. Our research shows that text preprocessing algorithms are mandatory for mining opinions on the Web 2.0 and that part of these algorithms are sensitive to errors and mistakes contained in the user generated content.

1. Introduction

Opinion mining deals with analyzing people's opinions, attitudes and emotions towards different brands, companies, products and even individuals (Balahur, 2013; Liu, 2012; Pang & Lee, 2008). Although related research areas to opinion mining such as natural language processing (NLP), information extraction and information retrieval have quite a considerable history, the research on mining people's opinions has become quite popular in the last couple of years with the rise of the Web 2.0. User generated content on the Social Web can contain a variety of relevant market research information and deeply analyzing and exploiting it leads to more targeted business decisions (Guozheng, Faming, Fang, & Jian, 2008; Liu, 2008).

Analyzing opinions on the Social Web is met with a variety of challenges: (i) the "usual" challenges known from natural language processing (such as word sense disambiguation, topic recognition and co-reference resolutions) and (ii) challenges arising from user generated content:

* Corresponding author. Tel.: +43 (0)50804 33410.
E-mail addresses: gerald.petz@fh-steyr.at (G. Petz), michal.karpowicz@fh-steyr.at (M. Karpowicz), harald.fuerschuss@fh-steyr.at (H. Fürschuß), andreas.auinger@fh-steyr.at (A. Auinger), stritesv@vse.cz (V. Stříteský), andreas.holzinger@medunigraz.at (A. Holzinger).

http://dx.doi.org/10.1016/j.ipm.2014.07.005
0306-4573/

- *Noisy texts, language variations*: User generated texts tend to be less grammatically correct and often use specific characters to express emotions (emoticons), abbreviations and unorthodox capitalization. (Abbasi, Chen, & Salem, 2008; Dey & Haque, 2009). Moreover, social media texts typically assume a higher level of knowledge about the context by the reader than more formal texts (Maynard, Bontcheva, & Rout, 2012).
- *Relevance and boilerplate*: When web texts and social media texts are gathered using a web crawler, the gained texts usually contain irrelevant content like advertisements, navigational elements or previews of other articles (Maynard et al., 2012; Petz et al., 2012; Yi & Liu, 2003).
- *Target identification*: Search-based approaches have to deal with the problem, that topics of retrieved documents do not necessarily match the mentioned sentiment object (Maynard et al., 2012).
- *Big data challenges*: That can be broken into several contexts such as temporal, spatial and spatio-temporal contexts (Derczynski, Yang, et al., 2013; Maynard, Dupplaw, & Hare, 2013).

Due to these challenges, research papers usually deal with assumptions and constraints: Many of the approaches to analyze opinions assume linguistically correct texts (Dey & Haque, 2009), others focus on specific social media resources (e.g. *Twitter* as a basis for opinion mining Bollen, Mao, & Zeng, 2011; Davidov, Tsur, & Rappoport, 2010; Pak & Paroubek, 2010; or newswire text Balahur, Steinberger, van der Goot, Pouliquen, & Kabadjov, 2009; Sayeed, 2011; or Blogs Leshed & Kaye, 2006; Mishne & Glance, 2006; Zhang, Yu, & Meng, 2007). The utilization of text preprocessing steps prior to sentiment analysis approaches is quite important in order to achieve good results.

The objectives of this paper are (i) to investigate the differences between social media channels regarding opinion mining and (ii) to evaluate the effectiveness of various text preprocessing algorithms as a subtask of opinion mining in these social media channels. To attain these objectives, we set up the research methodology as follows:

(1) Identification of popular approaches and algorithms to carry out text preprocessing as a prior step to sentiment analysis.
(2) Identification of differences between social media channels and deduction of impacts on opinion mining and text preprocessing.
(3) Evaluation of the effectiveness and properness of several algorithms in order to determine their applicability.

The rest of the paper is organized as follows: in the next section we discuss some related work in the field of opinion mining. We then report in Section 3 on the characteristics of user generated content in different social media channels. Section 4 discusses the impacts of these characteristics on some frequently used algorithms and evaluates their performance regarding noisy text.

2. Related work, background

2.1. Sentiment analysis and opinion mining

Pang and Lee (2008) and Liu (2012) present a detailed review of opinion mining. Liu defines an opinion as a quintuple (e_i, a_{ij}, s_{ijkl}, h_k, t_l), where e_i is the name of an entity, a_{ij} is an aspect of e_i, s_{ijkl} is the sentiment on aspect a_{ij} of entity e_i, h_k is the opinion holder and t_l is the time when the opinion is expressed. An entity is the target object of an opinion; it is a product, service, topic, person, or event. The aspects represent parts or attributes of an entity (part-of-relation). The sentiment is positive, negative or neutral or can be expressed with numeric scores (such as star-ratings). The indices i, j, k, l indicate that the items in the definition must correspond to one another. (Liu, 2012; Wilson, Wiebe, & Hoffmann, 2009)

There are several main research directions (Kaiser, 2009; Pang & Lee, 2008): (1) *Sentiment classification*: The main focus of this research direction is the classification of content according to its sentiment about opinion targets; (2) *feature-based opinion mining* (or *aspect-based opinion mining* Hu & Liu, 2004b; Liu, Hu, & Cheng, 2005) is about analysis of sentiment regarding certain properties of objects (e.g. Hu & Liu, 2004a); (3) *comparison-based opinion mining* deals with texts in which comparisons of similar objects are made (e.g. Jindal & Liu, 2006a, 2006b). Other research directions focus on multilingual opinion mining (e.g. Banea, Mihalcea, & Wiebe, 2010; Steinberger, Lenkova, Kabadjov, Steinberger, & Goot van der, 2011) and on cross-domain sentiment analysis (e.g. Bollegala, Weir, & Carroll, 2011; Pan, Ni, Sun, Yang, & Chen, 2010).

The classification of texts regarding sentiment polarity can be done at three different levels: (1) document level, (2) sentence level and (3) entity and aspect-level. There are several approaches to analyze opinions: (1) corpus-based approaches (e.g. Hatzivassiloglou & Wiebe, 2000; Turney, 2002; Wiebe & Mihalcea, 2006) and dictionary-based/lexicon-based approaches (e.g. Ding, Liu, & Yu, 2008; Hu & Liu, 2004a; Kim & Hovy, 2004; Popescu & Etzioni, 2005; Steinberger et al., 2012), (2) machine learning approaches. These approaches can be categorized as follows:

(1) *Supervised learning*: Supervised learning ("classification") is a machine learning task of inferring a function from labeled training data, where statistical methods are applied to construct prediction rules. This type of learning is widely used in real-world applications. Typical supervised learning algorithms are Naïve bayes classifiers, maximum entropy, support vector machines (SVM) and K-Nearest neighbor learning, amongst others (Liu, 2008; Zhang, 2010).