Decision Support Systems 74 (2015) 46–56

Contents lists available at ScienceDirect

Decision Support Systems

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

Polarity classification using structure-based vector representations of text

Alexander Hogenboom [a], Flavius Frasincar [a,*], Franciska de Jong [a,b], Uzay Kaymak [c]

[a] Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, the Netherlands
[b] Universiteit Twente, P.O. Box 217, NL-7500 AE Enschede, the Netherlands
[c] Eindhoven University of Technology, P.O. Box 513, NL-5600 MB Eindhoven, the Netherlands

ARTICLE INFO

Article history:
Received 23 June 2014
Received in revised form 12 March 2015
Accepted 2 April 2015
Available online 12 April 2015

Keywords:
Sentiment analysis
Rhetorical structure
Machine learning
Support vector machines

ABSTRACT

The exploitation of structural aspects of content is becoming increasingly popular in rule-based polarity classification systems. Such systems typically weight the sentiment conveyed by text segments in accordance with these segments' roles in the structure of a text, as identified by deep linguistic processing. Conversely, state-of-the-art machine learning polarity classifiers typically aim to exploit patterns in vector representations of texts, mostly covering the occurrence of words or word groups in these texts. However, since structural aspects of content have been shown to contain valuable information as well, we propose to use structure-based features in vector representations of text. We evaluate the usefulness of our novel features on collections of English reviews in various domains. Our experimental results suggest that, even though word-based features are indispensable to good polarity classifiers, structure-based sentiment information provides valuable additional guidance that can help significantly improve the polarity classification performance of machine learning classifiers. The most informative features capture the sentiment conveyed by specific rhetorical elements that constitute a text's core or provide crucial contextual information.

1. Introduction

In the past decade, the Web has experienced an exponential growth into a network of more than 555 million Web sites, with over two billion users [1]. The Web has become an influential source of information with an increasing share of user-generated content, produced by many contributors [2]. This ubiquitous and ever-expanding user-generated content ranges from (micro)blog posts to reviews.

The abundance of user-generated content has the potential to act as a catalyst for well-informed decision making, as the data can be used to monitor the wants, the needs, and the opinions of large quantities of (potential) stakeholders, such as customers. Monitoring user-generated content enables decision makers to identify issues and patterns that matter, and to track and predict emerging events [3]. However, in this era of Big Data, potentially valuable data is often unstructured, scattered across the Web, and expanding at a fast rate, thus rendering manual analysis of all available data unfeasible [4]. Yet, automated tools for information monitoring and extraction can provide timely and effective support for decision making processes.

Today's information monitoring and extraction tools can process information from many heterogeneous sources in dynamic environments [5,6] in order to, e.g., detect trending topics in (on-line) conversations [7], or to identify discussed entities (e.g., products or brands) and the events in which these entities play a role [8]. The past decade has brought forth a surge of research interest in extracting one type of valuable information in particular — people's sentiment with respect to entities or topics of interest [9–12]. This development is driven by the significant electronic word-of-mouth effects of user-generated content [13] on, e.g., sales [14,15] and stock ratings [16].

Many automated sentiment analysis techniques focus on determining the polarity of natural language text, typically by making use of specific cues, e.g., words, parts of words, or other (latent) features of natural language text. This is often done in machine learning methods [17,18]. However, rule-based methods – often relying on sentiment lexicons that list words and their associated sentiment – are attractive alternatives, as the nature of typical rule-based sentiment analysis methods allows for intuitive ways of incorporating deep linguistic analysis into the sentiment analysis process [19].

Solely focusing on explicit cues for sentiment, e.g., words, has been shown not to yield a competitive polarity classification performance [20]. Therefore, successful rule-based approaches account for semantic [3] and structural [19,21–23] aspects of content in order to improve the classification performance. Such methods typically use a text's structure in order to distinguish important text segments from less important ones in terms of their contribution to the text's overall sentiment, and subsequently weight each segment's conveyed sentiment in accordance with its identified importance.

[*] Corresponding author. Tel.: +31 010 408 1340; fax: +31 010 408 9162.
E-mail addresses: hogenboom@ese.eur.nl (A. Hogenboom), frasincar@ese.eur.nl (F. Frasincar), f.m.g.dejong@utwente.nl (F. de Jong), u.kaymak@ieee.org (U. Kaymak).

http://dx.doi.org/10.1016/j.dss.2015.04.002
0167-9236/© 2015 Elsevier B.V. All rights reserved.

The performance of competitive rule-based approaches, albeit comparably robust across domains and texts, is typically inferior to the performance of machine learning polarity classification systems [24]. The latter systems typically exploit patterns in (large) vector representations of texts, mainly signaling the presence of specific words or word groups in these texts. However, as structural aspects of content have been proven useful in rule-based approaches [19,21–23], we propose to incorporate new, structure-based features in vector representations of text in order to further improve the polarity classification performance of machine learning approaches to sentiment analysis.

The main contribution of our work lies in our novel structure-based features, which facilitate a richer representation of natural language text that should enable a more accurate classification of its polarity. We evaluate the usefulness of our structure-based features in a machine learning sentiment analysis method. We thus aim to provide insight in the importance of accounting for structural aspects of text in a machine learning approach to sentiment analysis, such that automated sentiment analysis systems can be used more effectively for supporting decision making processes.

The remainder of this paper is structured as follows. First, in Section 2, we provide an introduction to the field of sentiment analysis, with a specific focus on typical features used to represent text, as well as on structure-based sentiment analysis. Then, in Section 3, we propose novel, structure-based features that can be used for sentiment analysis. We evaluate the usefulness of these features for machine learning polarity classification of text in Section 4 and we conclude in Section 5.

2. Related work

The significant electronic word-of-mouth effects of user-generated content [13] on, e.g., sales [14,15] and stock ratings [16] advocate a need for automated sentiment analysis methods in decision support systems [3]. With the help of such systems, organizations can pinpoint the effect of specific issues on customer perceptions, thus enabling them to respond with appropriate marketing and public relations strategies in a timely and effective manner [25]. Advances in automated sentiment analysis are hence of paramount importance for today's decision support systems.

The field of automated sentiment analysis is an upcoming field that has been attracting more and more research initiatives in the past decade [17,18]. This surge in research interest in automated sentiment analysis techniques is fueled by the potential of sentiment analysis for real-life decision support systems [10,26]. Several trends can be observed in existing sentiment analysis methods, as briefly addressed in Section 2.1. The vector representations of text, used by the (performance-wise) most competitive approaches are discussed in Section 2.2. In Section 2.3, we then elaborate on promising recent advances in sentiment analysis, where the analysis of the sentiment conveyed by a piece of natural language text is guided by the text's structure.

2.1. Sentiment analysis

Existing methods for sentiment analysis focus on various tasks. Some methods deal with distinguishing subjective text segments from objective ones [27], whereas other approaches have been designed to determine the polarity of words, sentences, text segments, or documents [17]. The latter task is commonly treated as a binary classification problem, which involves classifying the polarity of a piece of text as either positive or negative. More polarity classes – e.g., classes of neutral or mixed polarity, or star ratings ranging from one to five stars – may be considered as well, yet in this paper, we address the binary classification problem for the polarity of documents. Existing binary polarity classification approaches range from rule-based to machine learning methods [17,18].

Rule-based methods are rather intuitive methods that typically rely on sentiment lexicons, which list explicit sentiment cues like words [28] or emoticons [29], along with their sentiment scores. The scores of individual cues are typically combined in accordance with predefined rules and assumptions (e.g., by summing or averaging these scores) in order to obtain an overall sentiment score for a text, which is then used as a proxy for the text's polarity class. In this process, negation [30] or intensification [24] of sentiment may be accounted for. Moreover, rule-based sentiment analysis allows for intuitive ways of incorporating deep linguistic analysis into the process, for instance by weighting text segments in accordance with their importance, as identified based on their respective rhetorical roles [19]. The performance of rule-based methods tends to be comparably robust across domains and texts [24], and the nature of these methods allows for insight into the motivation for assigning a particular polarity class to a text.

Machine learning methods typically involve building Support Vector Machine (SVM) classifiers or the like, trained for specific corpora by means of supervised methods that aim to exploit patterns in vector representations of natural language text [24]. Such classifiers tend to yield comparably high polarity classification accuracy on the collections of texts that they have been optimized for [17,18,24,31], but they require a lot of (annotated) training data, as well as training time in order to reach this performance level. Nevertheless, their superior performance renders machine learning polarity classifiers particularly useful for specific, rather than generic, domain- or corpus-independent applications.

2.2. Common features for sentiment analysis

Various types of features are used by existing machine learning approaches to sentiment analysis in order to construct vector representations of text. The most common and most useful features signal the presence or frequencies of specific words (i.e., unigrams) or groups of words (i.e., n-grams) [17]. Such features constitute a so-called *bag-of-words* vector representation of a text, which in itself has been shown to be rather effective in polarity classification [32,33]. Binary features that indicate word presence have been shown to outperform frequency-based features [32], which may indicate that a text's sentiment, as opposed to its topic, is not necessarily highlighted through repeated use of the same terms [17]. Nevertheless, frequency-based features have been shown to be useful in later work [34].

Another type of information captured by features for sentiment analysis is part-of-speech (POS) information, enabling the distinction between (types of) nouns, verbs, adjectives, and adverbs. The correlation between the subjectivity of a piece of text and the presence of adjectives in this text [35] has been mistakenly taken as evidence of adjectives being good indicators for sentiment [17], thus resulting in a possibly misplaced focus on using adjectives as features in the sentiment analysis process [36–38]. Other POS types may contribute to sentiment expression too [17]. As such, a more fruitful approach is to differentiate words in the *bag-of-words* representation of a text by their POS [18].

As subjectivity is associated with word meanings rather than lexical representations of words [39–41], it is important to account for semantics when performing sentiment analysis [3]. POS information can be useful here to a limited extent [42], yet more advanced methods involve accounting for semantics by grouping words with similar meanings [38,43].

Opinion-conveying texts are significantly different from objective texts in terms of the presence of sentiment-carrying words [44]. Specific sentiment-carrying words have therefore been used as features in so-called *bag-of-sentiwords* vector representations of text, capturing the presence of sentiment-carrying words derived from a sentiment lexicon [20,45]. In other work, text has been represented as a *bag-of-opinions*, where features denote occurrences of unique combinations of opinion-conveying words, amplifiers, and negators [46]. Other features capture the length of a text segment, and the extent to which it conveys opinions [2,20].