Information Sciences 280 (2014) 275–288

Contents lists available at ScienceDirect

Information Sciences

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

An empirical study of sentence features for subjectivity and polarity classification

Jose M. Chenlo *, David E. Losada

Centro de Investigación en Tecnoloxías da Información (CITIUS), Universidad de Santiago de Compostela, Rúa de Jenaro de la Fuente Domínguez, Santiago de Compostela, Corunna, Spain

ARTICLE INFO

Article history:
Received 19 December 2013
Received in revised form 2 April 2014
Accepted 8 May 2014
Available online 16 May 2014

Keywords:
Sentiment Analysis
Opinion Mining
Sentence-level analysis
Subjectivity classification
Polarity classification

ABSTRACT

While a number of isolated studies have analysed how different sentence features are beneficial in Sentiment Analysis, a complete picture of their effectiveness is still lacking. In this paper we extend and combine the body of empirical evidence regarding sentence subjectivity classification and sentence polarity classification, and provide a comprehensive analysis of the relative importance of each set of features using data from multiple benchmarks. To the best of our knowledge, this is the first study that evaluates a highly diversified set of sentence features for the two main sentiment classification tasks.

1. Introduction

Sentiment Analysis (SA) – also known as Opinion Mining – is an active and influential research area concerned with automatically extracting subjectivity from natural language text [30,23,7,14]. Sentence-level analysis plays a major role because it permits a fine-grained view of different opinions expressed in text. Moving closer to opinion targets and sentiments on targets facilitates opinion extraction from text that may only contain a few sentences that discuss the topic of interest [23].

A wide range of studies have demonstrated that information provided by some textual features is valuable for sentiment classification. Many types of sentence features have been proposed and tested in the literature: *n*-grams, part-of-speech features, location-based features, lexicon-based features, syntactic features, structural or discourse features, just to name a few. However, there is a lack of substantive empirical evidence of their relative effectiveness with different types of texts. Some features have only been tested against product or movie reviews. Other features have only been tried with news datasets. Furthermore, some experiments were performed in a supervised setting while others were performed in an unsupervised setting. The specific evaluation tasks performed are also diverse, e.g., subjectivity classification, polarity classification, opinion summarisation, or polarity ranking. This heterogeneous array of experimental results makes it difficult to understand what features are effective, and when and how they are best used. For instance, some linguistic cues might be beneficial in news articles (due to the nature of the language utilised by journalists) and harmful in product reviews (where customers tend to use a more direct style).

To shed light on these issues, in this paper we aim to provide a comprehensive body of evidence regarding the effectiveness and limitations of different sentence features. Our main contribution is of an empirical nature: the focus of this paper is

* Corresponding author. Tel.: +34 881816396.
E-mail addresses: josemanuel.gonzalez@usc.es (J.M. Chenlo), david.losada@usc.es (D.E. Losada).

http://dx.doi.org/10.1016/j.ins.2014.05.009
0020-0255/

to report an extensive set of experiments planned to evaluate the relative effectiveness of different sentence features for subjectivity classification and polarity classification. Specifically, we designed a general experimental setting – with two SA tasks and different sources of data – to jointly evaluate features that have otherwise been tested independently. To the best of our knowledge, this is the first study of this kind for SA at sentence level.

Our analysis of results suggests possible use cases of the features considered and gives insights into the potential applicability of every feature set. The experimental outcome also reveals that some features claimed as effective in the literature have little value once the models incorporate other elemental features.

The remainder of this paper is organised as follows. In Section 2 we discuss related work on existing SA approaches and review how different aspects of content are typically involved in such methods. Sections 3 and 4 describe the method and experiments, respectively. In Section 5 we discuss the results of the evaluation and the main lessons learnt about the relative performance of each feature set. Last, in Section 6, we present our conclusions and suggest directions for future research.

2. Related work

The state-of-the-art in automated SA has been reviewed extensively. Existing surveys [30,23,7,14], which cover all important topics and recent advances, often pay special attention to sentence-level analysis. For instance, Chapter 4 of Liu's book [23] is devoted to sentence subjectivity and sentiment classification. Liu describes different types of sentence features and learning algorithms – supervised and unsupervised – that have been applied for sentiment classification.

Pang and Lee's survey [30] includes a comprehensive discussion of features that have been explored for SA and outlines how opinion extraction problems are often cast as sentence or passage-level classification problems. The reviews done by Cambria et al. [7] and Feldman [14] briefly discuss the main research problems related to sentence-level Sentiment Analysis and some of the techniques used to solve them.

Most of the methods to extract opinions at sentence level are supervised. Pang and Lee [29] made one of the first attempts to apply supervised learning. They showed that it is viable to build effective polarity classifiers for movie reviews. In the last decade, many other researchers have applied classification in SA [23,44]. As usual in Machine Learning, the selection of features is of paramount importance. Some typical features are *n*-grams, part-of-speech (POS) features, sentiment words features, syntactic patterns, location features, concept-level features, and discourse features. The next paragraphs briefly review how these features have been employed by different researchers.

2.1. N-grams

Pang et al. [31] demonstrated the usefulness of these features for polarity classification of film reviews. Unigram presence features turned out to be the most effective. Other features were considered, including bigrams, POS and location evidence, but none of these provided consistently better performance once unigrams were incorporated. Paltoglou and Thelwall [28] studied and compared different unigram weighting schemes and concluded that some variants of tf/idf are well suited for SA. Their study was done against movie reviews, product reviews and a blog dataset. A new unigram weighting scheme called Delta TFIDF was proposed for SA by Martineau and Finin [27].

2.2. POS

Turney showed that adjectives are important indicators of opinions [41], and Yu and Hatzivassiloglou [46] classified subjective sentences based on subsets of adjectives that were manually tagged as positive or negative.

2.3. Sentiment words

Kim and Hovy determined the sentiment orientation of a sentence by multiplying the scores of the sentiment words in the sentence [22]. Qiu et al. presented a self-supervised method that utilises sentiment lexicons to bootstrap training data from a collection of reviews [32]. Taboada et al. [37] exploited a dictionary of sentiment words and phrases with their associated orientations and strength, and incorporated intensification and negation to compute a sentiment score for each document.

2.4. Syntactic patterns

Turney defined in [41] five syntactic patterns to extract opinions from reviews. These patterns turned out to be very effective for sentiment classification in an unsupervised manner and have become reference rules for discovering opinions [23]. Wiebe and Riloff [42] discovered patterns to feed a rule-based method that generates training data for subjectivity classification. The rule-based subjectivity classifier classifies a sentence as subjective if it contains two or more strong subjective clues. Liu and Seneff proposed an approach for extracting adverb–adjective–noun phrases (e.g., "*very nice car*") based on the clause structure obtained by parsing sentences into a hierarchical representation [24]. Berardi et al. [3] combined POS, Syntactic Patterns and Sentiment Words to extract opinions from sentences that contain an hyperlink. The extracted