Expert Systems With Applications 58 (2016) 57–75

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Unsupervised method for sentiment analysis in online texts

Milagros Fernández-Gavilanes*, Tamara Álvarez-López, Jonathan Juncal-Martínez, Enrique Costa-Montenegro, Francisco Javier González-Castaño

AtlantTIC, University of Vigo, Campus, 36310 Vigo, Spain

ARTICLE INFO

Article history:
Received 26 October 2015
Revised 17 March 2016
Accepted 18 March 2016
Available online 1 April 2016

MSC:
68Q55
68T50

Keywords:
Sentiment analysis
Opinion mining
NLP
Artificial intelligence

ABSTRACT

In recent years, the explosive growth of online media, such as blogs and social networking sites, has enabled individuals and organizations to write about their personal experiences and express opinions. Classifying these documents using a polarity metric is an arduous task. We propose a novel approach to predicting sentiment in online textual messages such as tweets and reviews, based on an unsupervised dependency parsing-based text classification method that leverages a variety of natural language processing techniques and sentiment features primarily derived from sentiment lexicons. These lexicons were created by means of a semiautomatic polarity expansion algorithm in order to improve accuracy in specific application domains. The results obtained for the Cornell Movie Review, Obama-McCain Debate and SemEval-2015 datasets confirm the competitive performance and the robustness of the system.

1. Introduction

The field of *sentiment analysis* (SA) has received increasing attention in recent years (Liu, 2012), particularly due to the explosive growth of social media, blogs and forums, which has enabled individuals and organizations to write about experiences and express opinions using colloquial and compact language. This new form of expression is potentially a source of extremely valuable information. For example, Twitter, one of the most popular social media networks, grew from 5000 new tweets per day in 2007 to 500 million tweets per day in 2013[1] by over 240 million users[2] and today has over 500 million users. Consequently, an increasing number of companies are focusing their marketing campaigns on the analysis of online comments from these potential customers, for instance, to predict the acceptance level of certain products (Jansen, Zhang, Sobel, & Chowdury, 2009).

However, it is difficult and costly to manually extract relevant knowledge from such large volumes of data, which is why automated machine prediction is so attractive (Bothos, Apostolou, & Mentzas, 2010). SA represents an interdisciplinary challenge that leverages a variety of *natural language processing* (NLP) techniques in order to determine the sentiment expressed in texts and decide whether they are positive, negative or neutral.

Of the different approaches applied to polarity classification, we can basically distinguish between *supervised machine learning* (ML) and *unsupervised lexicon-based* approaches. Although ML has proven to be extremely useful in the field of SA, an obvious disadvantage is its limited applicability to subject domains other than the domain it was designed for. Moreover, training of the classifier requires labeled datasets (Moreno Ortiz & Pérez Hernández, 2013), which are often difficult or even impossible to obtain. This is because their generation require people labeling data which is too labor-intensive and time-consuming.

We describe an unsupervised method for SA in English that used dependency parsing to determine the polarity of tweets and a previously created sentiment lexicon that took into consideration the special structure and linguistic content of messages. We analyzed the linguistic peculiarities of the texts and used a new SA algorithm based on sentiment propagation for dependency parsing that does not need prior or specific training.

While some of these unsupervised studies tried to compare their results with "*ad-hoc*" supervised methods, developed directly by the same author, in our case we preferred to compare our results with existing ones. As a testbed, we evaluated the performance of our approach for the movie review domain,

* Corresponding author. Tel.: +34 986 814081.
E-mail addresses: milagros.fernandez@gti.uvigo.es (M. Fernández-Gavilanes), talvarez@gti.uvigo.es (T. Álvarez-López), jonijm@gti.uvigo.es (J. Juncal-Martínez), kike@gti.uvigo.es (E. Costa-Montenegro), javier@det.uvigo.es (F. Javier González-Castaño).

[1] http://www.telegraph.co.uk/technology/twitter/9945505/Twitter-in-numbers.html.

[2] http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/.

http://dx.doi.org/10.1016/j.eswa.2016.03.031
0957-4174/

represented by the *Cornell Movie Review*. Our system was able to correctly classify 74.80% of the test set, a result that improves on results obtained for this dataset by the methods in other works: (Annett & Kondrak, 2008; Zhou, Zhao, & Zeng, 2014) (unsupervised), (Li, Zhang, & Sindhwani, 2009) (semisupervised), (Annett & Kondrak, 2008) (supervised) and (Carrillo de Albornoz, Plaza, & Gervás, 2010) (hybrid), all without using training data.

We also evaluated our approach on the Obama-McCain Debate dataset (Shamma, Kennedy, & Churchill, 2009) in comparison with other unsupervised methods (Hu, Tang, Gao, & Liu, 2013; Zhou et al., 2014); as well as on the dataset provided for SemEval-2015 Task 10[3] on SA in Twitter (Rosenthal et al., 2015). For this message-level task (40 submissions by 40 teams), in which we participated for the first time, we achieved satisfactory results. The results for both datasets therefore confirm the competitive performance and robustness of our system.

The paper is organized as follows. Section 2 discusses related work on polarity classification. Section 3 describes the system proposed for this task and the semiautomatic domain adaptation of the sentiment lexicon. Section 4 describes experimental results for the Cornell Movie Review Data, Obama-McCain Debate and SemEval-2015 datasets. Finally, Section 5 summarizes the main findings and conclusions.

2. Related work

SA systems development follows two basic steps: identification and classification (Medhat, Hassan, & Korashy, 2014; Pang & Lee, 2008).

The first step identifies subjective features in texts. These can be selected using (which is not always trivial) methods that can use some paradigm words and word similarities in order to obtain words expressing similar opinions. According to the way the similarities are obtained, these methods may be divided into semantic thesaurus-based and domain corpus-based approaches.

Semantic thesaurus-based approaches rely on the existence of semantic thesaurus created by human annotators, like WordNet[4] or General Inquirer[5]. The approaches in this category depend on the kind of relationships, synonyms or antonyms, between sentiment terms and the glosses in the thesaurus, expanding the polarity lexicon from a small set of seed words with known polarity. In (Hu & Liu, 2004; Kim & Hovy, 2004) two positive and negative verb and adjective seed lists were bootstrapped using WordNet in order to produce a larger lexicon. Similarly, in (Kamps, Marx, Mokken, & de Rijke, 2004) a lexical network was built by linking synonyms provided by the thesaurus, and the sentiment polarity was defined by the distance from the seed words "*good*" and "*bad*" in the network. Furthermore, in (Esuli & Sebastiani, 2007) an inverse and bidirectional model of random walking algorithm was proposed. These methods commonly rely on the assumption that adjectives share the same polarities with their synonyms and opposite polarities with their antonyms. It could be argued that they rely on prior semantic thesaurus resources without considering the domain-dependent characteristic of the sentiment lexicon.

In recent years, domain corpus-based approaches have been more widely studied. They are built on the basic assumption that polar terms conveying the same polarities co-occur with each other in domain corpuses, with context-specific orientations, usually relying on syntactic or statistical techniques like co-occurrence of a word with another word of known polarity. For example, in (Hatzivassiloglou & McKeown, 1997), the orientation of adjectives was predicted using other adjectives linked to the first ones by "*and*" (the same orientation) and "*but*" (the opposite orientation). Another example can be found in (Turney, 2002) where semantic orientation was assigned by means of association relationships between an unknown word and a set of selected seeds (like "*excellent*" and "*poor*"). In other studies, such as (Read & Carroll, 2009), the polarity of a word was identified by studying its frequency in a large annotated corpus of texts. If the word occurred more frequently among positive (negative) texts, then polarity was assumed to be positive (negative). If neither positive nor negative texts were dominant, polarity was assumed to be neutral. Qiu, Liu, Bu, and Chen (2011) analyzed manually and summarized eight dependency rules between opinion words and opinionated targets, and proposed a double propagation algorithm to expand the opinionated targets and sentiment lexicon iteratively. Other studies treated the problem of detecting polarities of words by means of graph propagation algorithms, such as Rao and Ravichandran (2009) (with a label propagation algorithm) and Huang, Niu, and Shi (2014) (with a constrained label propagation one using chunk dependency information and prior generic lexicon).

More recently, many studies have also tried to exploit prior sentiment knowledge in source domains to assist sentiment lexicon construction in the target domain. This was the case in (Tan & Wu, 2011), where the lexicon construction was modeled as a random walking process over four types of relationships between documents and words from both the source and target domains. In (Liu˜K., 2015), a method for co-extracting opinion targets and opinion words by using a word alignment model is described. They detected opinion relations between them. In (Zhang & Singh, 2014), a semisupervised framework was proposed. Instead of using sentences, they used segments of them and their dependency relation pairs in order to capture the contextual sentiment words for sentiment lexicon construction. Other works, such as (Lu, Castellanos, Dayal, & Zhai, 2011), focus on the problem of learning a sentiment lexicon that is not only domain specific but also dependent on aspects in some context given an unlabeled opinionated text collection. Finally, in (Tang, Wei, Qin, Zhou, & Liu, 2014a), they have applied a seed expansion algorithm to enlarge a small list of sentiment seeds using prior web knowledge.

The second step is correct classification of the overall sentiment of a given text. As already commented, methods can be broadly divided into two categories: supervised ML and unsupervised lexicon-based approaches (Maynard & Funk, 2012). The former are often classifiers built from linguistic features that use two sets of documents: a labeled training set to learn the differentiating characteristics of texts and a test set to check classifier performance.

The most widely used supervised ML techniques applied for SA described in the literature include naive Bayes, maximum entropy and support vector machines (SVM) (Vohra & Teraiya, 2013). Once a supervised classification technique is selected, it is important to determine how the documents are represented, as the selection of adequate features is crucial for classification success. Many authors treat the documents as *bags of words* (Hu & Liu, 2004; Pak & Paroubek, 2010) comprising unigrams or *n*-grams with their frequencies, because of the resulting simplicity of classification (Pang, Lee, & Vaithyanathan, 2002). Other authors propose including linguistic information such as part-of-speech (POS) tagging in order to disambiguate text sense according to lexical category (Pang & Lee, 2008), for instance, identifying adjectives and adverbs used as sentiment indicators. Negation words reverse the sentiment (König & Brill, 2006; Pang & Lee, 2008) with the polarity of opinion words and phrases determined using WordNet (Hu & Liu, 2004). In this direction, most studies focus on selecting effective features to improve performance

[3] SemEval is an international forum for natural-language shared tasks and dataset is available at http://alt.qcri.org/semeval2015/.

[4] Available at http://wordnet.princeton.edu/

[5] Available at http://www.wjh.harvard.edu/˜inquirer/