Knowledge-Based Systems 108 (2016) 65–78

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Building a Twitter opinion lexicon from automatically-annotated tweets

Felipe Bravo-Marquez*, Eibe Frank, Bernhard Pfahringer

Department of Computer Science, The University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand

ARTICLE INFO

Article history:
Received 12 November 2015
Revised 9 May 2016
Accepted 9 May 2016
Available online 10 May 2016

Keywords:
Lexicon expansion
Sentiment analysis
Twitter

ABSTRACT

Opinion lexicons, which are lists of terms labeled by sentiment, are widely used resources to support automatic sentiment analysis of textual passages. However, existing resources of this type exhibit some limitations when applied to social media messages such as tweets (posts in Twitter), because they are unable to capture the diversity of informal expressions commonly found in this type of media.

In this article, we present a method that combines information from automatically annotated tweets and existing hand-made opinion lexicons to expand an opinion lexicon in a supervised fashion. The expanded lexicon contains part-of-speech (POS) disambiguated entries with a probability distribution for positive, negative, and neutral polarity classes, similarly to SentiWordNet.

To obtain this distribution using machine learning, we propose word-level attributes based on (a) the morphological information conveyed by POS tags and (b) associations between words and the sentiment expressed in the tweets that contain them. We consider tweets with both hard and soft sentiment labels. The sentiment associations are modeled in two different ways: using point-wise-mutual-information semantic orientation (PMI-SO), and using stochastic gradient descent semantic orientation (SGD-SO), which learns a linear relationship between words and sentiment. The training dataset is labeled by a seed lexicon formed by combining multiple hand-annotated lexicons.

Our experimental results show that our method outperforms the three-dimensional word-level polarity classification performance obtained by using PMI-SO alone. This is significant because PMI-SO is a state-of-the-art measure for establishing world-level sentiment. Additionally, we show that lexicons created with our method achieve significant improvements over SentiWordNet for classifying tweets into polarity classes, and also outperform SentiStrength in the majority of the experiments.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Many sentiment analysis methods rely on opinion lexicons as resources for evaluating the sentiment of a text passage. An opinion or sentiment lexicon is a dictionary of opinion words with their corresponding sentiment categories or semantic orientations. A semantic orientation is a numerical measure for representing the polarity and strength of words or expressions. Lexicons can be used to compute the polarity of a message by aggregating the orientation values of the opinion words it contains [17,35]. They have also proven to be useful when used to extract features in supervised classification schemes [8,19,22,23,47].

Social media platforms, particularly microblogging services such as **Twitter**[1], are increasingly being adopted by people to access and publish information about a great variety of topics. The language used in **Twitter** provides substantial challenges for sentiment analysis. The words used in this platform include many abbreviations, acronyms, and misspelled words that are not observed in traditional media or covered by popular lexicons, e.g., omg, loove, #screwthis. The diversity and sparseness of these informal words make the manual creation of a Twitter-oriented opinion lexicon a time-consuming task.

In this article, we propose a method for opinion lexicon expansion for the language used in Twitter[2]. Taking SentiWordNet [2] as inspiration, each word in our expanded lexicon has a probability

* Corresponding author.
E-mail addresses: felipebravom@gmail.com, fjb11@students.waikato.ac.nz (F. Bravo-Marquez), eibe@waikato.ac.nz (E. Frank), bernhard@waikato.ac.nz (B. Pfahringer).

[1] http://www.twitter.com

[2] This article extends a previous conference paper [7] and provides a more thorough and detailed report.

http://dx.doi.org/10.1016/j.knosys.2016.05.018
0950-7051/© 2016 Elsevier B.V. All rights reserved.

distribution, describing how positive, negative, and neutral it is. Additionally, all the entries of the lexicon are associated with a corresponding part-of-speech tag. Estimating the sentiment distribution of POS-tagged words is useful for the following reasons:

1. A word can present certain levels of intensity [37] for a specific sentiment category, e.g., the word *awesome* is more positive than the word *adequate*. The estimated probabilities can be used to represent these levels of intensity. These probabilities provide a probabilistic interpretation of the underlying sentiment intensities conveyed by a word and can be used as prior knowledge in Bayesian models for sentiment inference [24]. In contrast, scores obtained by unsupervised methods such as point-wise-mutual information semantic orientation (PMI-SO) [39] lack a probabilistic interpretation.
2. The neutral score provided by the lexicon is useful for discarding non-opinion words in text-level polarity classification tasks. This can easily be done by discarding words classified as neutral. Note that unsupervised lexicon expansion techniques such as PMI-SO [39] provide a single numerical score for each word, and it is unclear how to impose thresholds on this score for neutrality detection.
3. Homographs, which are words that share the same spelling but have different meanings, should have different lexicon entries for each different meaning. By using POS-tagged words, homographs with different POS-tags will be disambiguated [42]. For instance, the word *fine* will receive different sentiment scores when used as an adjective (e.g., *I'm**fine**thanks*) and as a common noun (e.g., *I got a parking**fine**because I displayed the ticket upside down*).

This is not the first work exploring these properties for lexicon expansion. Sentiment intensities were described with probabilities in [2], and the disambiguation of the sentiment of words based on POS tags was studied in [35]. However, this is the first time that these properties are explored for the informal language used in Twitter.

Our expanded lexicon is built by training a word-level sentiment classifier for the words occurring in a corpus of positive and negative polarity-annotated tweets. The training words are labeled using a seed lexicon of positive, negative, and neutral words. This lexicon is taken from the union of four different hand-made lexicons after discarding all polarity clashes from the intersection. The expanded words are obtained after deploying the trained classifier on the remaining unlabeled words from the corpus of tweets that are not included in the seed lexicon.

All the words from the polarity-annotated corpus of tweets are represented by features that capture morphological and sentiment information of the word in its context. The morphological information is captured by including the POS tag of the word as a nominal attribute, and the sentiment information is captured by calculating association values between the word and the polarity labels of the tweets in which it occurs.

We calculate two types of word-level sentiment associations: PMI-SO [39], which is based on the point-wise mutual information (PMI) between a word and tweet-level polarity classes, and stochastic gradient descent semantic orientation (SGD-SO), which is based on incrementally learning a linear association between words and the sentiment of the tweets in which they occur.

To avoid the high costs of manually annotating tweets into polarity classes for calculating the word-level sentiment associations, we rely on two heuristics for automatically obtaining polarity-annotated tweets: **emoticon-based annotation** and **model transfer**. In the first approach, only tweets with positive or negative emoticons are considered and labeled according to the polarity indicated by the emoticon. This idea, which has been widely used before to train message-level sentiment classifiers [5,16] is affected by two main limitations:

1. The removal of tweets without emoticons may cause a loss of valuable words that do not co-occur with emoticons.
2. There are many domains, such as politics, in which emoticons are not frequently used to express positive and negative opinions. Thus, it is very difficult to obtain emoticon-annotated data from these domains.

To overcome these limitations, we pursue a model transfer approach by training a probabilistic message-level classifier from a corpus of emoticon-annotated tweets and using it to label a target corpus of unlabeled tweets with a probability distribution of positive and negative sentiment. Note that the model transfer produces soft sentiment labels, in contrast to the hard labels provided by the emoticons. We study how to compute our word-level sentiment association attributes from tweets annotated with both hard and soft labels.

We test our word-level sentiment classification approach on words obtained from different collections of automatically labeled tweets. The results indicate that our supervised framework outperforms using PMI-SO by itself when the detection of neutral words is considered. We also evaluate the usefulness of the expanded lexicon for classifying entire tweets to polarity classes, showing significant improvement in performance compared to the original lexicon.

This article is organized as follows. In Section 2, we provide a review of existing work on opinion lexicon expansion. In Section 3, we describe the proposed method in detail. In Section 4, we present the experiments we conducted to evaluate the proposed approach and discuss results. The main findings and conclusions are discussed in Section 5.

2. Related work on lexicon expansion

There are two types of resources that can be exploited for automatically building or expanding opinion lexicons: semantic networks, and document collections. Previous work on opinion lexicon expansion from these two types of resources is presented in the following two subsections.

2.1. Semantic networks

A semantic network is a network that represents semantic relations between concepts. The simplest approach, based on a semantic network of words such as WordNet[3], is to expand a seed lexicon of labeled opinion words using synonyms and antonyms from the lexical relations [18,21]. The hypothesis behind this approach is that synonyms have the same polarity and antonyms have the opposite. This process is normally iterated several times. In [20], a graph is created using WordNet adjectives as vertices and the synonym relations as edges. The orientation of a term is determined by its relative distance from the two seed terms *good* and *bad*. In [12], a supervised classifier is trained using a seed of labeled words that is obtained through expansion based on synonyms and antonyms. For each word, a vector space model is created from the definition or *gloss* provided by the WordNet dictionary. This representation is used to train a word-level classifier that is used for lexicon expansion. An equivalent approach was applied later to create SentiWordNet[4] [2,13]. In SentiWordNet, each WordNet *synset* or group of synonyms is assigned into classes *positive, negative* and *neutral*, with soft labels in the range [0, 1].

[3] http://wordnet.princeton.edu/

[4] http://sentiwordnet.isti.cnr.it/

Download English Version:

https://daneshyari.com/en/article/571788

Download Persian Version:

https://daneshyari.com/article/571788

Daneshyari.com