Information Processing and Management 52 (2016) 5–19

Contents lists available at ScienceDirect

Information Processing and Management

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

Contextual semantics for sentiment analysis of Twitter

Hassan Saif[a,*], Yulan He[b], Miriam Fernandez[a], Harith Alani[a]

[a] Knowledge Media Institute, The Open University, United Kingdom
[b] School of Engineering and Applied Science, Aston University, United Kingdom

ARTICLE INFO

Article history:
Received 30 April 2014
Revised 24 November 2014
Accepted 28 January 2015
Available online 7 March 2015

Keywords:
Sentiment analysis
Contextual semantics
Twitter

ABSTRACT

Sentiment analysis on Twitter has attracted much attention recently due to its wide applications in both, commercial and public sectors. In this paper we present SentiCircles, a lexicon-based approach for sentiment analysis on Twitter. Different from typical lexicon-based approaches, which offer a fixed and static prior sentiment polarities of words regardless of their context, SentiCircles takes into account the co-occurrence patterns of words in different contexts in tweets to capture their semantics and update their pre-assigned strength and polarity in sentiment lexicons accordingly. Our approach allows for the detection of sentiment at both entity-level and tweet-level. We evaluate our proposed approach on three Twitter datasets using three different sentiment lexicons to derive word prior sentiments. Results show that our approach significantly outperforms the baselines in accuracy and *F*-measure for entity-level subjectivity (neutral vs. polar) and polarity (positive vs. negative) detections. For tweet-level sentiment detection, our approach performs better than the state-of-the-art SentiStrength by 4–5% in accuracy in two datasets, but falls marginally behind by 1% in *F*-measure in the third dataset.

1. Introduction

Twitter sentiment analysis has attracted much attention due to the rapid growth in Twitter's popularity as a platform for people to express their opinions and attitudes towards a great variety of topics. Approaches to Twitter sentiment analysis tend to focus on the identification of sentiment of individual tweets (*tweet-level sentiment detection*). Broadly speaking, existing work on tweet-level sentiment detection follows two main types of approaches, supervised learning or lexicon-based.

Supervised learning approaches require training data for sentiment classifier learning. In Twitter, training data are typically obtained by either assuming that tweets' polarities (positive, negative, neutral) can be inferred using emoticons (Go, Bhayani, & Huang, 2009; Kouloumpis, Wilson, & Moore, 2011; Pak & Paroubek, 2010; Saif, He, & Alani, 2012b) or by taking consensus from the results returned by the sentiment detection websites (Barbosa & Feng, 2010). Moreover, supervised approaches are domain-dependent and require re-training with the arrival of new data (Aue & Gamon, 2005). Given the great variety of topics that constantly emerge from Twitter, these limitations affect the applicability of such approaches.

On the other hand, lexicon-based approaches do not require training data. Instead, they use lexicons of words weighted with their sentiment orientations to determine the overall sentiment of a given text. These approaches have shown to work effectively on conventional text (Liu, 2010). However, traditional lexicons tend to be ill-suited for Twitter data, which often

* Corresponding author.
E-mail addresses: h.saif@open.ac.uk (H. Saif), y.he@cantab.net (Y. He), m.fernandez@open.ac.uk (M. Fernandez), h.alani@open.ac.uk (H. Alani).

http://dx.doi.org/10.1016/j.ipm.2015.01.005
0306-4573/

contains a large number of malformed words and colloquial expressions (e.g., "looov", "luv", "gr8"). Moreover, many lexicon-based approaches also make use of the lexical structure of a sentence to determine its sentiment, which becomes problematic in Twitter, where ungrammatical sentences are very common due to the 140-character length limit. Aiming to overcome these limitations, Thelwall, Buckley, Paltoglou, Cai, and Kappas (2010) and Thelwall, Buckley, and Paltoglou (2012) introduced a human-coded lexicon of words and phrases specifically built to work with social data. They proposed an algorithm called *SentiStrength* that utilises the lexicon to identify the sentiment strength of informal text (e.g., tweets, status updates). We refer to this lexicon as Thelwall-Lexicon hereafter.

SentiStrength has received much attention in recent years due to its relatively good and consistent performance on social media data. Nevertheless, similarly to other lexicon-based approaches, SentiStrength and its underlying Thelwall-Lexicon face two main limitations. Firstly, SentiStrength is confined with the fixed set of words that appear in the Thelwall-Lexicon. Words that do not appear in the lexicon are often not considered when analysing sentiment (Liu, 2010; Xu, Peng, & Cheng, 2012), which may create a problem when dealing with Twitter data, where new expressions and jargons constantly emerge. Secondly and more importantly, SentiStregnth and the like offer fixed, context-independent, word-sentiment orientations and strengths. For example, SentiStrength assigns the same sentiment strength to the word "`good`" in "`It is a very good phone indeed!`" and in "`I will leave you for good this time!`". Although a training algorithm has been proposed to optimise the terms' sentiment scores in Thelwall-Lexicon (Thelwall et al., 2010), it requires frequent retraining from human-coded data, which is labour-intensive and domain dependent.

In this paper we introduce an approach called SentiCircles (Saif, Fernandez, He, & Alani, 2014b), which builds a dynamic representation of words that captures their contextual semantics (i.e., semantics inferred from the co-occurrence patterns of words in text) in order to tune their pre-assigned sentiment strength and polarity in a given sentiment lexicon.

Contextual semantics (aka statistical semantics) (Wittgenstein, 1953) has been traditionally used in diverse areas of computer science, including Natural Language Processing and Information Retrieval (Turney & Pantel, 2010). The main principle behind the notion of contextual semantics comes from the dictum – "You shall know a word by the company it keeps!" (Firth, 1930–1955). This suggests that words that co-occur in a given context tend to have certain relation or semantic influence, which we try to capture with our SentiCircle approach.

We assess the performance of our proposed SentiCircle approach in two different sentiment analysis tasks: (i) *entity-level sentiment* detection, which detects sentiment towards a particular entity or topic (e.g., Obama, Microsoft, iPad) and (ii) *tweet-level sentiment* detection, which identifies the overall sentiment of *individual* tweets. To this end, we propose three different methods, which utilise several trigonometric identities on the SentiCircle representation to perform both sentiment analysis tasks.

We evaluate and test our approach under different settings (three different sentiment lexicons and three different datasets) and compare its performance against various lexicon baseline methods. We also compare our approach against SentiStrength, which, to our knowledge, is considered one of the best lexicon-based sentiment detection approaches for social media. For entity-level sentiment detection, our experimental results show that our proposed approach, based on SentiCircles, outperforms all the other methods by nearly 20% in accuracy and 30–40% in *F*-measure for subjectivity detection (neutral vs. polar). For tweet-level sentiment detection, our approach outperforms SentiStrength by 4–5% in accuracy in two datasets, but falls marginally behind by 1% in *F*-measure on the third dataset.

The main contributions of this paper can be summarised as follows:

- Introduce a novel lexicon-based approach using a contextual representation of words, called SentiCircles, which is able to capture the latent semantics of words from their co-occurrence patterns and update their sentiment orientations accordingly.
- Propose three different methods of employing SentiCircles for tweet-level sentiment detection.
- Conduct a series of experiments and test the effectiveness of our proposed approach for both entity- and tweet-level sentiment detection against several baselines, including SentiStrength.
- Perform a runtime analysis of our approach to demonstrate its scalability.
- Build and release the STS-Gold (Saif, Fernandez, He, & Alani, 2013), a new gold-standard dataset that allows for evaluating both, tweet- and entity-level sentiment analysis approaches.

The remainder of this paper is structured as follows. Related work on tweet-level and entity-level sentiment analysis is discussed in Section 2. The proposed SentiCircle representation of words is presented in Section 3. How to apply SentiCircles for sentiment analysis is described in Section 4. Experimental setup and results are presented in Sections 5 and 6 respectively. Discussion and future work are covered in Section 7. Finally, we conclude our work in Section 8.

2. Related work

Most existing approaches to *Twitter sentiment analysis* focus on classifying the individual tweets as positive or negative. They can be categorised as *supervised methods* (those which need training data) and *lexicon-based methods* (those based on dictionaries of terms with associated sentiment orientations).