Knowledge-Based Systems 105 (2016) 236–247

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A multilingual semi-supervised approach in deriving Singlish sentic patterns for polarity detection

Siaw Ling Lo[a], Erik Cambria[b,*], Raymond Chiong[a], David Cornforth[a]

[a] School of Design, Communication and Information Technology, The University of Newcastle, Callaghan, NSW 2308, Australia
[b] School of Computer Science and Engineering, Nanyang Technological University, 639798 Singapore

ARTICLE INFO

Article history:
Received 16 November 2015
Revised 13 April 2016
Accepted 24 April 2016
Available online 26 April 2016

Keywords:
Sentic computing
Polarity detection
Semi-supervised
Singlish
Twitter

ABSTRACT

Due to the huge volume and linguistic variation of data shared online, accurate detection of the sentiment of a message (polarity detection) can no longer rely on human assessors or through simple lexicon keyword matching. This paper presents a semi-supervised approach in constructing essential toolkits for analysing the polarity of a localised scarce-resource language, Singlish (Singaporean English). Corpus-based bootstrapping using a multilingual, multifaceted lexicon was applied to construct an annotated testing dataset, while unsupervised methods such as lexicon polarity detection, frequent item extraction through association rules and latent semantic analysis were used to identify the polarity of Singlish n-grams before human assessment was done to isolate misleading terms and remove concept ambiguity. The findings suggest that this multilingual approach outshines polarity analysis using only the English language. In addition, a hybrid combination of the Support Vector Machine and a proposed Singlish Polarity Detection algorithm, which incorporates unigram and n-gram Singlish sentic patterns with other multilingual polarity sentic patterns such as negation and adversative, is able to outperform other approaches in comparison. The promising results of a pooled testing dataset generated from the vast amount of unannotated Singlish data clearly show that our multilingual Singlish sentic pattern approach has the potential to be adopted in real-world polarity detection.

1. Introduction

Sentiment analysis has been a popular research area over the past few years. It has gained even more attention with the prevalence of social media usage, where 'netizens' freely and openly express their views about anything; be it a product, a policy, or even a picture. Although the content shared on social media can be a potential gold mine for companies and organisations to analyse sentiment and gather feedback, it is challenging to detect polarity with high accuracy, as the content is known to mix with linguistic variations where localised expression is commonly used [1].

There are mainly two approaches in sentiment analysis – subjectivity and polarity detection. While subjectivity detection is about understanding if the content contains personal views and opinions as opposed to factual information, polarity detection focuses on subjectivity analysis with varying polarities, intensities or rankings [2]. Being one of the first studies on creating Singlish language digital resources, here we have chosen polarity detection as it is able to identify content that is emotional and convey true feeling of the netizens. Positive and negative sentiments can be used as a litmus test to the well-being of a company or an organisation.

Most polarity analysis studies in the literature are limited to the English language [3], but with the popularity of social media worldwide, it is no longer sufficient to extract just English language content for analysis purposes. In fact, only 28.6% of Internet users speak English[1]. It is thus essential to explore the construction of resources and tools in languages other than English. To fully understand sentiments on the ground, analysing informal scarce-resource languages commonly used on social media alongside other formal languages is highly necessary.

While increasing attention has been paid to creating resources on alternative formal languages, limited resources are available when it comes to languages that are not commonly used in official communication or formal news reporting, due to their informal and evolving nature. These languages often evolve from a main national language, such as English, and are broadly used by some local community in daily conversation, both in the physical and online world. In addition, it is not uncommon to mix a few languages and use a localised lingual range to form a unique language to express emotion, especially in a multi-cultural environment [4]. This

* Corresponding author.
E-mail address: cambria@ntu.edu.sg, erik@sentic.net (E. Cambria).

[1] http://www.internetworldstats.com/stats7.htm

http://dx.doi.org/10.1016/j.knosys.2016.04.024
0950-7051/

is evident as the native or localised vernacular is able to resonate with the community[2] better than a formal language. One such example is Singlish, which is essentially the colloquial Singaporean English that has incorporated elements of some Chinese dialects and the Malay language [5]. It is hence not surprising that Twitter, as an informal channel in spreading news and information, is packed with localised or multilingual idioms so that messages can be conveyed more personally or effectively. In this study, we focus on shared information of Twitter (tweets) to extract Singlish content with polarity.

Even though Singlish is mainly associated with Singapore, citizens from the neighbouring country, Malaysia, with a similar multi-cultural environment, are able to understand the language with ease. This is not the case for others from different cultural backgrounds. The understanding of localised expression (e.g., it is a widely known practice to append an English sentence with "lah" in Singlish) is not sufficient with merely the knowledge of the English language, as Singlish is often presented with a mixture of multiple dialects and languages including English. Clause-final discourse particles [5] such as "lah", "hah", "ah" usually play a role in exaggerating the expression and do not particularly carry any polarity, and hence understanding Singlish polarity should be treated as deciphering another 'new' language. Besides that, due to the mixture of a few languages and its ever evolving nature, relevant research studies mainly concentrate on the linguistics aspect [5,6] and construction of dictionaries[3,4]. To date, there is no known polarity resource or tool available for the language.

Sentiment analysis for a language is usually dependent on manually or semi-automatically constructed lexicons [7,8], found in a dictionary or corpus [9]. The availability of these resources enables the creation of rule-based sentiment analysis or construction of a training dataset for classification tasks. However, as creating lexical or corpus resources for a new language can be very time-consuming and resource intensive, various multilingual sentiment analyses [9,10] have been done by relying on some available English knowledge base, such as SentiWordNet [11]. While the lexicon-based approach is still important in sentiment analysis, an alternative concept-based approach, which incorporates commonsense reasoning [12,13], is fast developing and provides the potential to manage more subtle sentiments that are often not captured or handled in current sentiment analysis research. SenticNet [14] being the core resource available, contains 30,000 commonsense concepts. It can be used for different sentiment analysis tasks, including polarity detection. In addition to concept-based analysis, the dependency relation of concepts is taken into consideration in the form of sentic patterns [11]. It has been shown that a better understanding of the contextual role of each concept within a sentence can improve polarity detection markedly [15].

In this paper, we aim to leverage SenticNet's sentic patterns, which include handling of English negation and adversative terms, to derive a unique set of Singlish sentic patterns for polarity detection. We use a multilingual semi-supervised approach to extract Singlish unigrams, bigrams and trigrams with polarities before multilingual negation and adversative terms as well as Twitter's retweet structure are incorporated in a Singlish Polarity Detection (SinglishPD) algorithm to identify the sentiment of a given tweet.

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

- To the best of our knowledge, our work in this paper is the first study using a semi-supervised approach to extract the polarity of Singlish.
- We create Singlish resources including a Singlish-English dictionary with relevant Part-Of-Speech (POS) notations and a set of Singlish annotated testing data.
- A list of Singlish sentic patterns that play an important role in determining the polarity of a Singlish tweet has been extracted. It includes multilingual negation/adversative terms, Twitter's retweet structure and Singlish unigram, bigram and trigram sentic patterns.
- Singlish sentic patterns have been shown to outperform English sentic patterns in detecting polarity, as the English lexicon is unable to fully capture the sentiment of Singlish tweets.
- From the observation of our results, the SinglishPD algorithm incorporated with Singlish sentic patterns can be used for enhancing the accuracy of polarity assignment for Singlish tweets, especially when coupled with machine learning.

In the next section, we will discuss some related work in polarity detection with emphasis on multilingual sentiment analysis. Following which, we outline the resources needed and methods used in Sections 3 and 4, respectively. In Section 5, we describe our findings and results. We then discuss our observations of the findings and future plans in Section 6 before conclusions are drawn in Section 7.

2. Related work

There are different granularities of polarity analysis. Some researchers focused on polarity analysis where an opinion is regarded as highly positive, positive, negative or highly negative [16]. Others [14] worked on human emotions such as joy or anger so that appropriate actions can be taken through insights gained from the content analysed.

As our study is based on Twitter data, this review of related work concentrates on multilingual polarity detection on Twitter. Volkova et al. [17] proposed an approach for bootstrapping subjectivity clues from Twitter data and evaluated the approach on English, Spanish and Russian Twitter streams. They used the multiperspective question answering (MPQA) lexicon [18] to bootstrap sentiment lexicons from a large pool of unlabelled data using a small amount of labelled data to guide the process. Cui et al. [19] focused on building emotion tokens or SentiLexicon using emoticons, repeating punctuations and repeating letters. Their comparative evaluation with SentiWordNet [20] indicated that emotion tokens are helpful for both English and non-English Twitter sentiment analyses.

Although lexical resources are still used for detecting polarity in text, machine learning approaches are more commonly adopted for polarity analysis of larger scale. In the domain of English polarity detection on social media, Barbosa and Feng [21] and Davidov et al. [22] employed machine learning based approaches to work on datasets with different genres and/or in a target-independent way for Twitter sentiment analysis studies. Specifically, Barbosa and Feng [21] proposed a two-step approach to classify the sentiment of tweets using Support Vector Machine (SVM) classifiers with abstract features. Davidov et al. [22] used a supervised k-nearest neighbours-like classifier for classifying tweets into multiple sentiment types using hashtags and smileys as labels. In contrast, Pak and Paroubek [23] collected a corpus of 300,000 text posts from Twitter for objectivity and positive/negative-emotion analysis. They concluded that Twitter users tend to use syntactic structures to describe emotion or state facts, and that POS tags may be strong indicators of emotional text.

Singlish is considered a scarce-resource language where limited electronic resources are available and very minimal Natural Language Processing (NLP) tools can be found. The following studies concentrate on approaches for sentiment analysis on such lan-

[2] http://mypaper.sg/top-stories/officials-use-singlish-dialects-reach-out-20150211
[3] http://www.singlishdictionary.com/
[4] http://www.talkingcock.com/html/lexec.php