8-2016

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

Siaw Ling LO
Singapore Management University, sllo@smu.edu.sg

Erik CAMBRIA

Raymond CHIONG

David CORNFORTH

Citation

LO, Siaw Ling; CAMBRIA, Erik; CHIONG, Raymond; and CORNFORTH, David. A multilingual semi-supervised approach in deriving Singlish sentic patterns for polarity detection. (2016). *Knowledge-Based Systems*. 105, 236-247.
Available at: https://ink.library.smu.edu.sg/sis_research/4872

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

guages. Banea et al. [24] created a subjectivity lexicon for the Romanian language using a small set of seed words, a basic dictionary, and a small raw corpus. A bootstrapping approach was used to add new related words to a candidate list, and both Pointwise Mutual Information [25,26] and Latent Semantic Analysis (LSA) [27] were used to filter noise from the lexicon. They showed that unsupervised learning using a rule-based sentence level subjectivity classifier is able to achieve a subjectivity F-measure score of 66.2, which is an improvement compared to previously proposed semi-supervised methods. Bakliwal et al. [28] constructed a Hindi subjective lexicon for polarity classification of Hindi product reviews. Using WordNet [29] and a graph-based traversal method, a full (adjective and adverb) subjective lexicon was built. A small seed list with polarity was used to leverage the synonym and antonym relations of WordNet in order to expand on the initial lexicon. The subjectivity lexicon was then used in the review classification, with 79% accuracy achieved using unigram and polarity scores as features. Another approach by Chowdhury and Chowdhury [30] used both Bengali and English words to perform sentiment analysis on tweets. A semi-supervised bootstrapping method was used to create the training corpus for machine learning classification, and 93% accuracy was achieved through an SVM using unigrams with emoticons as features.

While lexicon-based and machine learning approaches or a combination of these approaches have been used for sentiment and polarity analysis, concept-based techniques are gaining popularity due to their ability to detect subtly expressed sentiments [31,32]. SenticNet [14] is a concept-based English resource, and recently it has been extended with a collection of concept disambiguation algorithms implemented to discover contexts in the Chinese language [33]. Google translate is used to do mapping of the English and Chinese languages. Various Chinese resources are utilised to discover language-dependent sentiment concepts through translation.

Recently, cross-lingual sentiment analysis has been explored extensively due to its ability to exploit existing labelled information from a resource-rich source language to build a sentiment classifier on a target language and minimise the needs to manually annotate the target language. Methods such as translation, co-training, and parallel corpus analysis have been adopted. However, as there is no translation machine available for Singlish, neither does any parallel corpus exist, it is hence not feasible to rely on methods and techniques developed for cross-lingual sentiment analysis in discovering Singlish polarity terms or sentic patterns.

In short, none of the above discussed studies is directly related to this work, of which a multilingual, multifaceted lexicon that includes English polarity terms, Malay polarity terms and emoticons for polarity detection has been compiled. For evaluation purposes, we constructed an annotated Singlish testing dataset through a corpus-based bootstrapping approach using the multilingual lexicon to obtain tweets with polarities before manual annotation was done by three human assessors. Machine learning with emoticons as features has also been implemented to assess the feasibility of using a semi-supervised approach to create Singlish polarity training datasets (through polarity detection based on emoticons). Sentic pattern, Singlish unigram, bigram and trigram analyses using LSA and association rules were carried out to extract terms and concepts with polarities. More details of these can be found in the next sections.

3. Details of resources needed

3.1. Construction of a Singlish dictionary

As there is no available de-facto Singlish dictionary, manual construction of a Singlish dictionary by combining several Internet resources has to be done. The list of resources used includes the Dictionary of Singlish and Singapore English[5], Coxford Singlish Dictionary[6] and Wikipedia Singlish vocabulary[7]. In order to standardise the English description of a Singlish term, a simple description is used instead of elaborated explanation, in the hope that the corresponding English term is able to replace the Singlish term in a sentence (if necessary) so as to derive the meaning of the sentence in English. Then, further analysis of POS of the Singlish term is done. Terms with multiple parts of speech are labelled with the corresponding POS types (where the types considered are noun, adverb, adjective and interjection), and the English description given is ensured to be consistent with the POS types assigned. The finalised list of this Singlish-English dictionary contains 1,024 terms with 978 unique Singlish expressions.

3.2. Construction of the multilingual (English, Malay) and multifaceted polarity lexicon

Singlish is a localised form of English that incorporates elements of different languages, in particular the Malay language and Chinese dialects. There is thus a need to consider the multilingual aspects of it in creating a useful polarity lexicon. In addition to the Singlish dictionary built in this study (which incorporates many of the Chinese dialects), polarity lexicons of two major formal languages, i.e., Malay and English, are used to help in identifying tweets and terms that may have polarities. While there are plenty of available polarity lexicons for English, the resources in Malay are limited. In this work, the Malay sentiment lexicon published in [34] has been used.

As for the English polarity lexicon to be used, different sources have been processed and analysed. The positive lexicon was extracted from the positive list of a Twitter sentiment analysis study[8] and a set of positive vocabulary word lists[9], while the negative lexicon was extracted from the negative list of the same Twitter sentiment analysis study[8] and a set of negative and adjective reference words[10]. Each list was checked to ensure the uniqueness of terms, and terms found with both positive and negative polarities were removed. The end result is an English resource with 2,640 positive terms and 5,127 negative terms.

In view of the fact that emoticons are commonly used in Twitter to express emotion, and related studies [19,22,30] have shown that emoticons can be used effectively to extract the polarity of content, the multilingual polarity lexicon constructed in this study is expanded to include emoticons (i.e., multifaceted) on top of the above two Malay and English lexicons. A list of positive and negative emoticons was extracted from Blake's IM emotions quick reference[11]. In total, 66 positive and 73 negative emoticons have been used in this work.

3.3. Statistics of the Twitter dataset used

In order to collect tweets shared by users from Singapore, Twitter's Search API was used to follow a list of Twitter users who had been tweeting topics relevant to Singapore. Besides verifying via location information stated on the platform, users who consistently shared information about Singapore were consolidated into

[5] http://www.singlishdictionary.com/
[6] http://www.talkingcock.com/html/lexec.php
[7] https://en.wikipedia.org/wiki/Singlish_vocabulary
[8] https://github.com/jeffreybreen/twitter-sentiment-analysis-tutorial-201107/tree/master/data/opinion-lexicon-English
[9] http://positivewordsresearch.com/positive-vocabulary-words-list/
[10] http://dreference.blogspot.sg/2010/05/negative-ve-words-adjectives-list-for.html
[11] http://computer-ease.com/emotposi.htm

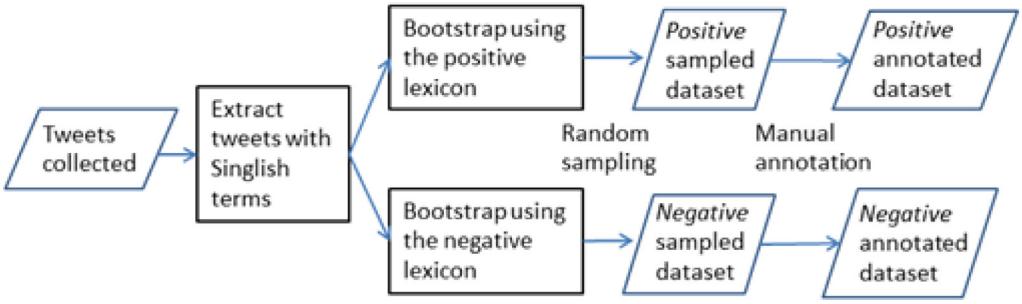

Fig. 1. Construction of Singlish annotated testing datasets.

1,498 users for data collection purposes. The period of data collection was from January 2013 to April 2014 and a total of 10,178,217 records were collected.

A detailed language analysis using the Language Detection Library for Java[12] shows that 23% of the data collected contains mixed languages, and 47% of it is detected as having English language content while 19% falls under the category of Indonesia/Malay languages. In order to leverage available English resources such as the sentic pattern [15] for polarity analysis, this study used a subset of the English content detected. The 978 unique Singlish terms mentioned in Section 3.1 were used to extract 2,745,822 tweets with Singlish content. Due to the large size of this Singlish dataset, the Apache Lucene[13] index was built and further analysis on matching hits of each Singlish term shows that out of the 978 unique terms, 466 of them are found in the Singlish tweet dataset. Terms that are not found are mainly Chinese dialects describing local food items such as "hae mee" (prawn noodle) and "chwee kueh" (steamed rice cake with salted radish).

4. Methods and setups

4.1. Pre-processing of tweets

Tweets are known to be noisy and often mixed with linguistic variations. The pre-processing tasks that are commonly used include lowercase conversion, handling of stop words, stemming, lemmatisation, as well as removal of URLs, Twitter's usernames found in the content (in the format of @username) and hashtags (with the # symbol). However, as the main focus of this study is to extract sentic patterns with polarity, it is important to consider the order and context of the words in tweets. Thus, the following pre-processing steps have been carried out:

- Handling of URLs, Twitter's usernames and hashtags: All the URLs and Twitter's usernames are normalised to two placeholders, *twitterurl* and *twitterusername*, respectively. Hashtags are preserved but the # symbol is removed.
- Handling of punctuations: As this study is about polarity detection, and other research [35] has shown that emoticons play an important role in differentiating different sentiments, punctuations used to form a term/emoticon such as ":-)" or ";-)" are kept so that emoticon analysis can be done.
- Handling of expressive lengthening terms: It is not unusual to find informal or expressive lengthening terms such as "goooood" and "hahahaha" being used to exaggerate the sentiment. Regular expression is used to detect such a repeating structure and the structure is reduced to two occurrences. For example, "gooood" is converted to "good" and "hahahaha" is changed to "haha". This process also applies to punctuations so "?????" is transformed into "??" for consistency's sake.

```
analysePolarity (content) :
if an item from the lexicon is found
    extract the polarity
if an item from the negation list is found
    if it is before any item from the lexicon
            reverse the polarity

for each tweet in the tweet list
    if an item from the adversative list is found
            analysePolarity(second section of the tweet)
    else
            analysePolarity(whole tweet)

if polarity is consistent
    assign the final polarity
else if both polarities found
    assign "both"
else
    assign "-"
```

Fig. 2. The algorithm used to assign polarity to tweets.

In an attempt to remove duplications after the above pre-processing, tweets have been lowercased and those having the same content would not be included in the subsequent analysis.

4.2. Construction of Singlish annotated testing datasets

While several annotated polarity corpora such as the Internet Movie Database archive [36] and MPQA opinion corpus [37] are available for sentiment analysis in English, a gold standard for Singlish is not readily available. It is therefore necessary to construct a Singlish annotated testing dataset in order to assess the performance of different approaches considered in this study.

Due to the sheer volume of tweets collected, we carried out random sampling with unsupervised class distribution based on the recommendation from a study by Wang et al. [38] that random sampling of each class separately can often improve the performance of a classifier when the sample reduction rate is high. As depicted in Fig. 1, the polarity lexicon developed in Section 3.2 is used as the first layer to filter the tweets containing Singlish terms into positive and negative classes.

Here, the list of tweets collected is 'bootstrapped' using the multilingual, multifaceted polarity lexicon. The pseudo-code of an algorithm used to assign polarity to these tweets is shown in Fig. 2. As can be seen, the algorithm relies on terms (i.e., "item" in the figure) from the lexicon. It is possible that one or more terms could

[12] http://code.google.com/p/language-detection/
[13] https://lucene.apache.org/

be found in a tweet. We use three polarity types, namely positive, negative or both. A tweet that is not assigned with any polarity type is considered as having no polarity. If the tweet contains one or more positive terms, taking into account of the negation, it is considered as positive. A similar process is applied in assigning negative tweets. However, if a tweet contains both positive and negative terms, it is considered to be the both type. In addition, the linguistic dependency relation between clauses is also considered. For example, tweets having adversatives such as "but" or similar are handled by the algorithm, as it was found that polarity of such a sentence/tweet tends to be determined by the second clause/section of the sentence or tweet [15].

The resulting datasets were then randomly sampled with 500 records each, before manual annotation was done on the sampled datasets to create the annotated datasets. Three Singlish native speakers annotated the tweets individually. These three assessors are graduates who are proficient in English, Malay, Mandarin as well as various Chinese dialects, and they are familiar with the culture of Singapore and its regions. Fleiss' kappa has been used to calculate the inter-rater agreement rate, as it is known to be able to assess the reliability of agreement among multiple 'raters' or assessors for categorical assignment [39]. The agreement of the three assessors based on Fleiss' kappa is 0.74 for the positive sampled dataset and 0.79 for the negative sampled dataset. To ensure that the annotated datasets have tweets that are labelled consistently across the two sampled datasets, tweets that have been consistently labelled as positive in the negative sampled dataset are also used in the positive annotated dataset and vice versa. In the end, the total numbers of positively labelled tweets are 215 and negatively labelled tweets are 459.

4.3. Use of supervised machine learning

4.3.1. The SVM

The SVM, a supervised machine learning method for two- or multi-class classification, is the chosen learning approach to be used in this work. It has been successfully applied to many application domains, including text categorisation [40] and social media analytics [41,42]. The SVM utilises an optimally separating hyperplane that separates labelled training data of positive and negative samples in the feature space. Consider a set of N distinct samples (x_i, y_i) with $x_i \in \Re^D$ and $y_i \in \Re^d$, an SVM can be modelled as

$$\sum_i a_i K(x, x_i) + b, i \in [1, N] \quad (1)$$

where $K(x, x_i)$ is the kernel function, and α and b are the parameter and threshold of the SVM, respectively. For more details, see [43].

We have used the sigmoid kernel with the LibSVM implementation of RapidMiner[14] in this study, since it produces higher precision prediction than other kernels such as the radial basis function or polynomial kernel. Given that the training input of an SVM is through a matrix of feature vectors, we created the feature vectors via term frequency-inverse document frequency analysis of the tweet content after data pre-processing (see Section 4.1 for details), stop-word removal and word stemming using Porter [44] were done.

As the SVM is a supervised learning approach, there is a need to prepare annotated datasets of different classes to train the SVM model. In view of the huge un-annotated tweet data, it is not practical to manually annotate each of the tweets. Fortunately, several research studies (e.g., see [19,22,35,45]) have shown that emotions can be used to label the polarity. This leads us to creating an SVM model with emoticons for our work.

[14] http://rapidminer.com/

4.3.2. An SVM model built using polarity emoticons

The list of positive and negative emoticons constructed, as described in Section 3.2, was used to extract tweets containing polarity emoticons in the Singlish dataset. After extracting the Singlish tweets with positive and negative emoticons, random sampling of some of the tweets indicated that only a portion of them have polarities while quite a few others were either ambiguous or came with elements of sarcasm, especially those among the tweets with positive emoticons. In order to create a 'cleaner' training dataset, Singlish terms with known polarity (identified through translation from the Singlish-English dictionary and the English, Malay multilingual polarity lexicon) have been applied as an additional layer of filtering. The resulting records (244 negative and 389 positive) were then manually annotated. 61% of the assigned negative tweets have been labelled as "negative" while only 32% of the assigned positive tweets are annotated as "positive". The high 'rejection rate' is partly due to the fact that tweets with ambiguous polarities or mixed emotions have been omitted. It is observed that emoticons seem to work well with negative sentiments but not so well with positive sentiments, as some of the tweets with positive emoticons can have mixed emotions and hence cannot be labelled as having positive sentiment. The two annotated training datasets were subsequently used to build an SVM model. Fig. 3 shows the detailed steps of creating a set of annotated training datasets for the SVM.

4.4. Construction of Singlish sentic patterns

While not much detailed research has been done on Singlish's sentence structure, a related study has shown that Singlish's grammar differs quite markedly from the standard English, with topic-prominent language features being most noticeable [5]. It is common for Singlish speakers to establish the topic first, e.g., at the beginning of a sentence, before referring to it subsequently. For example, "`Christmas -- we don't celebrate because we are not Christians`" [5]. Because of this, English POS was not adopted in this study to understand the sentence structure. Instead, three approaches, namely polarity lexicons, association rules and LSA [27], have been used to discover Singlish sentic patterns in an unsupervised manner. The English sentic patterns published in [15], which include negation and adversative patterns, have also been considered in this study together with Twitter's specific structure of re-tweet or RT. In addition, due to the limited resources available for Singlish, an extensive Singlish n-gram study was conducted to assess the effects of various Singlish unigrams, bigrams and trigrams on polarity detection. Other special considerations such as ambiguous and misleading terms were included in the analysis too. The following subsections describe each of these in detail.

4.4.1. English and Singlish sentic patterns

In our analysis of the English sentic patterns, a polarity reversing rule based on English negation terms such as "not", "couldn't" and "shldnt" was implemented. In short, if a polarity term was found after a negation, the polarity of the term was reversed. The English polarity lexicon constructed, as described in Section 3.2, was used to detect the polarity of the term. Besides negation handling, tweets containing adversative terms such as "but" were further processed to ensure the correct polarity was assigned. Specifically, if an adversative term was detected, the tweet was separated into two parts based on the adversative term. Only the polarity of the second part of the tweet was considered. For example, considering the following tweet:

"replying happy over and over again but not happy"

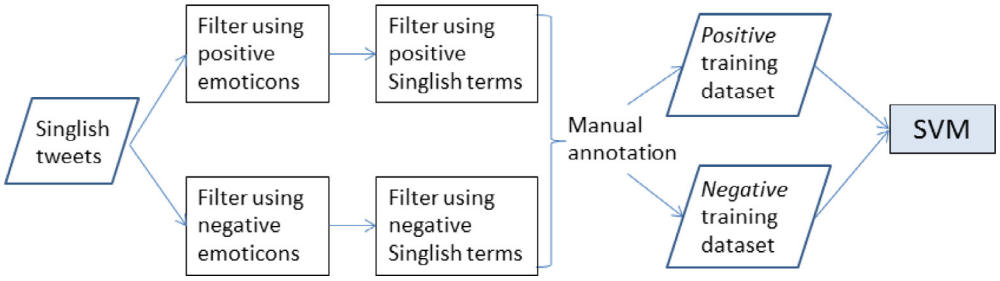

Fig. 3. The SVM model built with polarity emoticons.

In this example, the correct polarity is detected at the second part of the tweet, which is "not happy", and thus the tweet is assigned with negative polarity.

As for the Singlish tweet dataset, it contains a mixture of languages and online expressions. The multilingual and multifaceted polarity lexicon constructed using English and Malay polarity lexicons and emoticons (see Section 3.2) is therefore used in conjunction with negation and adversative terms for the Singlish sentic pattern analysis. Malay negation terms such as "tak" and "tidak" as well as Malay adversative terms like "tetapi" and "tapi" are also included to capture the correct sentiment expressed in a tweet.

4.4.2. Handling of RT

Re-tweet or RT is one of the most commonly used Twitter features to share content. It is similar to forwarding a tweet to other Twitter users by adding RT to the original content. It also allows users to add their own comments and expressions before the RT content. As a result, there is a need to take a tweet with RT into special consideration so that the right sentiment of the user who retweeted the content can be captured. Let us consider the two tweets containing RT below:

> Tweet1: It's okay to be lame. As long as you tak tweet cinta macam macai macai lol XD RT twitterusername sorry i'm so fucking lame :(

> Tweet2: Ye la mana nak jumpa. RT twitterusername guess who's botak now? hehehehehehehe. A few days left. twitterusername

The polarity of Tweet1 is positive even though the RT content is negative. This is because the reply of the RT content is of positive sentiment. On the other hand, as the user of Tweet2 is only asking a question "`Ye la mana nak jumpa`" (yes, where to meet), the polarity of Tweet2 is determined by the content of the original tweet or the RT content, which is also of positive sentiment. This tweet is likely about sharing the joy of being able to get out of the national service camp, which all males in Singapore need to attend and have their head shaved (the meaning of "botak" in the tweet), in a few days.

From the examples, it is clear that there is a need to cross check if the prepended content of the RT message has polarity before deciding on the final polarity. We therefore have taken the following steps to assess the polarity of a tweet containing RT:

(i) If the reply to the RT content has polarity, the polarity of the tweet is assigned by considering the polarities of both the reply and RT content. Sometimes, it is possible that even if the polarities of both the reply and RT content are negative, the polarity of the tweet could be positive (e.g., `It's not ok indeed. RT twitterusername I don't feel right about it man`).
(ii) If the reply to the RT content has no polarity, then the polarity will be based on the polarity of the original RT content.

4.4.3. Singlish unigram sentic patterns

As the aim of this study is to identify Singlish terms or patterns with polarity, we analysed and extracted the list of terms from the Singlish dictionary constructed (from Section 3.1) via two approaches. The first approach is based on polarity detection using the multilingual, multifaceted polarity lexicon detailed in Section 3.2 on tweets having the dictionary terms. For each dictionary term, the polarity is determined by the normalised occurrences of a positive or negative vocabulary from the list of tweets containing the term. The second approach is to identify frequent items associated with the dictionary term so as to infer the sentiment of the term. Frequent item analysis is done using the Frequent Pattern Growth association rule algorithm [46]. As the number of tweets per dictionary term is different, a minimum support parameter needs to be dynamically calculated to ensure that all frequent items retrieved have the minimum support of two records. These frequent items identified are then analysed using the lexicons mentioned above and a dictionary term found to have polarity frequent items associated with them will be assigned with the corresponding polarity.

4.4.4. Singlish bigram and trigram sentic patterns

In order to minimise the need to annotate a large amount of content, the unsupervised LSA method has been used to extract a list of bigrams and trigrams with polarity. It has been shown that LSA can provide similar results in detecting subjectivity candidates with a faster response time and less training data [24]. Hence, we adopted LSA to extract terms that are similar to a given bigram or trigram.

There are two steps in the extraction process. The first step is to identify a bigram or trigram with the possibility of collocation via hypothesis testing based on a t-test with significance level 0.005. The main purpose of the test is to remove a bigram or trigram that is occurring by chance. In short, the notion of collocation in this study is the same as previously defined by [47], where it is referred to as a sequence of two or more consecutive words having characteristics of a syntactic and semantic unit, and whose exact and unambiguous meaning cannot be derived directly from its individual components.

The second step is to implement LSA to extract the top 100 terms (deduced empirically) that are most similar to the bigram or trigram based on values from the dot product of individual terms. Details of the algorithm for extracting polarity bigrams of a given term are given in Fig. 4. The same algorithm is also used to extract all the trigrams.

Clause-final discourse particles [5] such as "lah", "hah", "ah", "lor", "leh", "one" and "wor", which represent a stereotype of Singlish, are omitted due to its sheer volume detected and its usage mainly as a marker in a sentence. However, "meh" has a negative notation compared to the rest, as it is sometimes used in a standalone way in a tweet. For example, "`No mood to study geography now. Thought got someone accompany me. But meh, goodnight.`" and "`Seriously got so nice to sleep meh. Don't even know why the fuck I fell`

for each Singlish term,
  extract all bigrams of a given Singlish term (include prefix and postfix of the term)
  for each bigram,
    calculate the t-test value of each term in the bigram
    if the t-test value is of significance
      add the bigram in the potential_bigram list
      for each potential_bigram,
        - extract all the tweets (*pre-processed according to Section 4.1 and all common stop words including twitterusername, twitterurl and rt are removed*)
        - calculate the dot product of the potential_bigram and the extracted representative words from tweets
        - select the top 100 words based on the value of the dot product
        - calculate polarity based on the occurrences of the positive and negative lexicon words
        - assign polarity

Fig. 4. The algorithm for extracting bigrams with polarity.

`back asleep`". Removing the "meh" in these sentences can potentially change the polarity of the tweets. In addition, some terms that are found in the Singlish dictionary with commonly known polarities, such as "fuck", and food items such as "satay" (roasted meat on a stick), "roti" (bread or similar) or related food cooking descriptions like "goreng" (fried) and "lemak" (oily), have been excluded in the bigram and trigram polarity analysis.

4.4.5. Special considerations

Ambiguous terms detected during the bigram and trigram analysis as well as manual annotation were specially extracted. For example, "fuck" is a common term used in the dataset and it is mostly negative but the following are some patterns that defile the trend. For example, "fucking nice sia" or "fucking love you" is actually having positive sentiment. As a result, these patterns have been added into the Singlish sentic pattern.

There are some Singlish terms in the dictionary that can be misleading if each individual term is analysed. For example, "mee siam" and "hor fun" are the names of local food items. However, "siam" can also be a negative expression that means "get out of the way" while "fun" is usually assigned as positive. Keeping such terms can be misleading in the analysis of Singlish polarity. These terms are hence considered as Singlish stop words and have been omitted in the detailed analysis.

4.5. The SinglishPD algorithm

In order to leverage the various resources constructed and sentic patterns derived in this study, a SinglishPD algorithm was implemented to integrate the resources and patterns for polarity assignment. Fig. 5 shows the details of the algorithm.

This SinglishPD algorithm is an enhancement or a more comprehensive version of the algorithm described in Section 4.2 (see Fig. 2), used to extract the polarity samples for creating the Singlish annotated testing datasets. Instead of using the polarity lexicon, a set of sentic patterns has been incorporated in the algorithm, including Singlish's polarity n-gram, Twitter's RT structure and misleading term handling.

4.6. A hybrid approach for polarity analysis using Singlish sentic patterns and machine learning

As the SinglishPD algorithm relies very much on the sentic patterns, and due to the limited resources available for Singlish, it is possible that a tweet would not be assigned any polarity. In such a situation, the SVM described in Section 4.3 will be used to complement the Singlish polarity assignment. This hybrid approach is able to leverage the strength of knowledge-based polarity assignment via sentic pattern detection and tap on the classification ability of a machine learning algorithm at the same time. The overall architecture is depicted in Fig. 6.

4.7. Performance evaluation

4.7.1. F-measure

Typical accuracy metrics used for statistical analysis of binary classification, which take into consideration the true positive (TP) and true negative (TN), have known issues in terms of reflecting the performance of a classifier [48]. We have therefore used F-measure as the metric when assessing the performance of the various approaches proposed in addition to the correct assignment or accuracy percentage of positive and negative datasets.

The formulas of F-measure are as follows:

$$precision = TP/(TP + FP) \tag{2}$$

$$recall = TP/(TP + FN) \tag{3}$$

$$F - measure = 2 \times \frac{precision \times recall}{precision + recall} \tag{4}$$

where TP, TN, FP and FN represent the true positive, true negative, false positive and false negative, respectively.

4.7.2 . A pooling strategy

While most classification studies would end with an accuracy performance analysis of the annotated testing dataset, we extended our assessment of the various sentic patterns proposed with the SinglishPD algorithm to the entire un-annotated dataset in order to evaluate if it is feasible to adopt the approaches in a real-world application. Due to the huge amount of un-annotated data, an adapted pooling strategy [49] has been used to measure the relative performance of the various sentic patterns. The following steps comprise the pseudo process of generating a pooled testing dataset for assessment purposes:

(i) Extract the tweets with polarities using the SinglishPD algorithm;
(ii) Identify tweets that have various sentic patterns;

```
for each tweet,
    if sentic_pattern_RT is found
        detectPolarity(tweet_portion) where the portion to be extracted is based
        on Section 4.4.2
    else
        detectPolarity(whole_tweet)

detectPolarity(content)
• handleAdversative()
• removeMisleadTerms()
• handlePunctuation()
• removeRepeatedChar()
• removeStopwords()
    for each item in content
        if an item is found in the lexicon
            store the location and polarity in a location_polarity map
            if the lexicon is an n-gram
                skip the item to be analysed accordingly
        if an item is found in the negation list
            check the location_polarity map
            if negation is found before a lexicon item
                reverse the polarity
• calculate polarity based on the number of processed lexicon items
• assign polarity
```

Fig. 5. The SinglishPD algorithm.

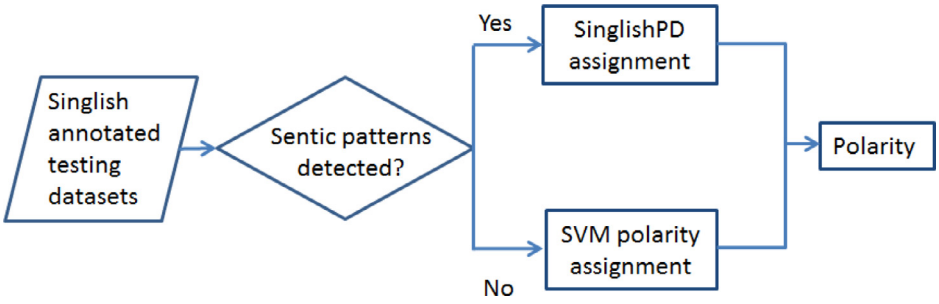

Fig. 6. Hybrid polarity analysis with Singlish sentic patterns and the SVM.

(iii) Randomly sample each of the different types (e.g., negation, adversative, RT);
(iv) Combine the various types to form a pool for human assessors to judge.

As a result, a total of 500 tweets were generated through this approach. The 500 consist of 250 positive and 250 negative tweets (determined by the SinglishPD algorithm). For each polarity type, i.e., positive and negative, 50 tweets were extracted for each of the three sentic pattern types, namely tweets with negation, adversative and RT. In other words, $50 \times 3 = 150$ tweets with selected sentic patterns were randomly sampled separately from the positive and negative tweet groups. Another 100 tweets with no specified sentic patterns were also extracted to further ascertain the performance of the SinglishPD algorithm. These culminated in a 500-tweet pooled testing dataset.

5. Results

5.1. Results of Singlish sentic patterns

5.1.1. Singlish unigram sentic patterns

As different terms were being identified by the two approaches mentioned in Section 4.4.3, namely the polarity lexicon and association rules, further analysis was done to finalise the list of unigrams with polarity. Singlish terms that were labelled as having "noun" POS have been omitted, since previous research [50] has shown that polarity terms are usually associated with verbs, adverbs, adjectives and not nouns. In addition, as a number of terms identified with positive sentiment were actually food items, these terms were manually verified to ensure the quality of the terms selected before creating the final list. See Table 1 for the details of Singlish unigram sentic patterns with polarity.

5.1.2. Singlish bigram and trigram sentic patterns

As shown in Table 2, bigrams and trigrams have been successfully filtered into two separate lists having polarity. While t-test collocation analysis was able to select the more possible co-occurring terms, it was not surprising that through random sampling, there were some n-grams having negation or the opposite sentiment found in the lists. For example, quite a few negation associated terms were found, such as "tak boleh" (cannot), "cant tahan" (cannot tolerate) and variants like "cannt", "cnnt" and "cnt". These have been removed to avoid affecting negation handling by the SinglishPD algorithm. Besides that, the final list was verified using the multilingual lexicon. For instance, "die liao" has been moved from positive to negative. Similarly, those with positive sentiment have been moved from negative to the positive list, e.g., "smoke clear". Those with numbers, like "level 50″ and "drop 5″, were all removed.

Table 1
Singlish unigrams with polarity.

	Polarity lexicon	Association rules	Final	Examples with English translation in brackets
Positive	184	190	60	heng (fortunate), lepak (relaxing), shiok (satisfaction), stylo (stylish), zai (smart)
Negative	253	99	73	Amacam (somehow), kacau (interfere), kiasu (afraid of losing), mati (die), toot (stupid person)

Table 2
Singlish bigrams and trigrams with polarity.

	Original	Collocation t-test	LSA – polarity lexicon – n-gram verifier	Final
Bigram	45,363	7,884	Positive – 1,524; Negative – 1,337	Positive – 1,327; Negative – 1,339
Trigram	44,695	11,593	Positive – 254; Negative – 300	Positive – 179; Negative – 277

Table 3
Results based on the Singlish annotated testing datasets.

Method	Positive	Negative	F-measure
SVM	62.8% (135/215)	76.0% (349/459)	0.587
English Sentic pattern	71.6% (154/215)	71.2% (327/459)	0.615
Sentic pattern	76.7% (165/215)	73.2% (336/459)	0.656
Singlish unigram Sentic pattern	77.2% (166/215)	73.6% (338/459)	0.661
Singlish bigram Sentic pattern	77.7% (167/215)	73.6% (338/459)	0.664
Singlish trigram Sentic pattern	77.7% (167/215)	73.6% (338/459)	0.664
Hybrid	***84.2% (181/215)***	***84.7% (389/459)***	***0.777***

Clause-final discourse particles were further analysed using bigrams and trigrams extracted. It was found that the particles are often associated with a lexicon with known sentiment, which is evidence that these discourse particles do not have polarity on their own but they do emphasise the expression of the user in some way. The only exception is “meh” – negative associations were detected in tweets found with the term and hence this term has been included in the polarity unigram. The rest of the discourse particles are left as having no polarity.

Due to the unsupervised nature of LSA and omission of named entity handling in this first study, some known local named entities such as “sheng siong” (which is the name of a popular supermarket in Singapore) had been retrieved. “siong” is a common Singlish word that means tough or labourious. As a result, manual assessment was done to remove possible concept ambiguity and known named entities. These terms have been considered in the special handling process mentioned in Section 4.4.5. The following are some examples of ambiguous Singlish trigrams using the term “drop” – “samsung share drop” should be negative but “drop in coe” and “jobless claims drop” are positive. Besides that, “drop dead gorgeous” is positive but “drop dead tired” is negative. These sentic patterns found have been included in the Singlish trigram sentic pattern list.

5.2. Performance of various polarity assignment approaches based on Singlish annotated testing datasets

The manually annotated Singlish testing datasets were used to evaluate the various polarity assignment approaches. These include the SVM (trained using emoticons), English sentic patterns where the English lexicon is used for detecting polarity, sentic patterns where the multilingual, multifaceted polarity lexicon is incorporated, three Singlish n-gram sentic patterns and the hybrid approach that combines Singlish trigram sentic patterns with the SVM. The polarity accuracy and F-measure results can be found in Table 3.

As expected, the SVM does not perform as well as the rest, since it is heavily dependent on the training datasets. While the emoticons used for training purposes do contain elements of Singlish terms and other data with polarity, it is highly likely that emoticons are the most important features learned during the training process. In fact, the features are well-learned as the F-measure score of 5-fold cross validation using the emoticon training datasets is 0.89. However, the Singlish annotated testing dataset was randomly sampled and some tweets may or may not have emoticons. It is therefore understandable that the performance of the SVM is not satisfactory. On the other hand, it is obvious from the results of Table 3 that the multilingual, multifaceted polarity lexicon or the sentic pattern performs better than merely using an English polarity lexicon. This is due to the fact that the tweets are of multilingual nature and thus having multifaceted lexicons improves the performance.

The performances of the various Singlish n-gram sentic patterns are competitive compared to that of the sentic pattern (using the multilingual and multifaceted polarity lexicon). Part of the reason is because Singlish sentic patterns have been used, since the algorithm relies on the detection of the patterns for polarity assignment. Indeed, a further analysis done on the annotated testing dataset found that there is very little content of Singlish unigrams: six occurrences (2.8%) of the Singlish polarity unigram are found in the positive dataset and 41 occurrences (8.9%) are found in the negative dataset. As for the other sentic patterns: negation 14%, adversatives 4.7% and RT 5.6% in the positive dataset, and negation 31.2%, adversatives 8.1% and RT 1.5% in the negative dataset. Even though the results indicate that sentic patterns indeed help in improving the accuracy, it is of interest to conduct a more in-depth analysis on another randomly selected dataset containing most of the Singlish polarity terms and RT structure to ascertain the effect of the sentic patterns.

Interestingly, the hybrid approach has achieved the highest F-measure value, indicating that hybridisation of sentic patterns and the SVM is able to leverage the strength of both approaches and complement each other to obtain better results than faring on their own.

5.3. Results from the pooling strategy

To wrap up, we also created a type-specific pooled testing dataset from the un-annotated dataset using the approach specified in Section 4.7.2 to ascertain the effect of the Singlish sentic patterns for polarity detection. The results from the pooling strategy are shown in Table 4.

Table 4
Results based on the pooled testing dataset with different types of sentic patterns (positive and negative datasets each having 250 records).

	Adversative	Negation	RT	Others
Positive	78%	62%	84.2%	81%
Negative	98%	96%	73.3%	85%

As can be seen from the table, the accuracy of negation is not as good as the other sentic patterns in the positive dataset. Further analysis done on this revealed that actually only five out of the 50 tweets have been assigned as negative (32 as positive and the rest no polarity). It is observed that quite a few of the tweets containing the negation sentic pattern are having either ambiguous meaning or sarcastic expression, which is a known challenge in polarity analysis research. One of the examples with sarcastic expression is shown below:

> "I got mention name meh? Don't have right? Just a random tweet, what makes you think I saying you? Unless you ownself guilty conscience.#LOL??"

It is interesting to observe that the SinglishPD algorithm is able to achieve accuracy above 80% for the "Others" type for both positive and negative polarities. This is clear evidence that the algorithm incorporated with Singlish sentic patterns can differentiate the two types of polarities with high accuracy.

In order to ascertain the effect of RT on Singlish polarity detection, the SinglishPD algorithm was amended and run twice on the RT tweets, once with RT handling integrated and once without the consideration of RT structure. The result strongly indicates that it is essential to include RT handling for polarity detection, as the accuracy increases from 60% to 73.3% for the negative dataset. As for the positive dataset, the accuracy without RT handling is 75.5% and with RT handling it increases to 84.2%.

As the pooled testing dataset is a specially curated Singlish dataset, it is of interest to know how the other approaches, such as the English sentic pattern fares in classifying the data. Table 5 shows the classification results of different approaches based on the pooled testing dataset. The results clearly show that classifiers having Singlish sentic patterns outperform the rest. It is thus important to include Singlish sentic patterns in analysing content with Singlish sources and localised expressions.

6. Discussions and future plans

Due to the limited resources available for Singlish polarity detection, this study focused on deriving a set of sentic patterns that is able to aid in differentiating whether a Singlish tweet has a positive or negative sentiment. Related studies have used dictionary-based bootstrapping to expand the subjectivity [24] or polarity lexicons [30] on other scarce-resource languages. However, this dictionary-based approach does not work well on Singlish, because Singlish is not a 'full-bloom' complete language but rather a variant of English. Most of the approaches in the literature (e.g., see [51,52]) rely on English resources such as WordNet [29] to extract related words for other languages such as Spanish and Hindi. Singlish, being a derived language with localised expression is often unique in describing an expression and/or feeling, which may or may not have the exact or single word translation in English. As a result, bootstrapping processes using candidate synonyms would not work well for the informal language compared to other complete languages, such as the Romanian language [24].

To overcome Singlish's scarce-resource constraints, unsupervised approaches such as corpus-based bootstrapping, classification based on lexicons and emoticons, as well as LSA have been used to construct polarity datasets in this study. However, as there is no gold standard annotated dataset or dictionary available, human assessors were asked to annotate the various datasets needed for this study so that more detailed analyses can be done. Through the annotation efforts, three datasets have been created. The first dataset (i.e., the Singlish annotated testing dataset) was created based on corpus-based bootstrapping using the multilingual, multifaceted lexicon as well as random sampling. The second dataset was created via corpus-based bootstrapping using emoticons, and the third dataset was generated by the SinglishPD algorithm based on the entire un-annotated dataset with random sampling of specific sentic patterns. The results of manual annotation are shown in Table 6. It is clear that the results of the proposed SinglishPD algorithm are promising, with 77.6% accuracy over the positive dataset and 85.6% accuracy over the negative dataset. These accuracies represent a marked difference compared to the other two datasets extracted through lexicons (for the construction of the Singlish annotated testing dataset) and emoticons (for the training dataset of SVM). In general, accuracies on the positive datasets are consistently lower than those of the negative datasets. The differences are even more contrasting on datasets generated through lexicons and emoticons. A detailed analysis was done and it was found that the majority of the tweets are having either no polarity or ambiguous polarities where the three human assessors had provided different annotations. Specifically, there are 43.6% of such tweets found in lexicon-classified tweets, while 60.6% are found in tweets extracted using emoticons. This is partly due to the fact that expression of true positive sentiment may not be straightforward. While positive lexicons or emoticons can be detected in the content, the tweets may carry sarcasm and hence pose a challenge if a polarity is to be recognised via keyword matching, of which the context of the whole text is ignored. It is thus important to include n-gram analysis and also concepts of specific terms in the lexicons to improve the accuracy.

As shown in Tables 3 and 5, the SVM does not perform well on the Singlish annotated and pooled testing datasets, respectively. These results are in contrast to other findings. Chowdhury and Chowdhury [30] used both Bengali and English words to perform sentiment analysis on tweets and achieved 93% accuracy for the SVM using unigrams with emoticons as features. Go et al. [35] used emoticons in texts and constructed a training corpus from tweets with the best result of 81% accuracy obtained through a Naïve Bayes classifier. The SVM's not-so-satisfactory results here may be because more features are required in Singlish polarity detection compared to other languages, due to its multilingual nature. In fact, a further analysis on the emoticon dataset shows that Singlish sentic patterns do help in improving the accuracy compared to using merely the English sentic pattern. See Table 7 for details. It is encouraging to see that incorporating Singlish sentic patterns in the SinglishPD algorithm can achieve highly competitive results.

While a Singlish dictionary has been developed in this study, it should not be considered as a complete reference, as analysis on the whole Singlish dataset indicates that there are still many unknown terms that cannot be recognised by standard English and Malay dictionaries. Moreover, the online jargons and expressions used do not follow any formal format and cannot be found in a dictionary; they evolve according to trending topics and cultural influences. Singlish being a localised language is thus highly evolving and there is a need to keep the dictionary up-to-date with regular analysis on expressions used by the netizens.

Our preliminary study on clause-final discourse particles like "lah" and "leh", done through empirical assessment by three human assessors on a small random bigram and trigram dataset,

Table 5
Classification results based on the pooled testing dataset.

Method	Positive	Negative	F-measure
SVM	62.8% (115/183)	69.5% (148/213)	0.634
English Sentic pattern	44.3% (81/183)	82.2% (175/213)	0.603
Sentic pattern	58.5% (107/183)	81.7% (174/213)	0.691
Singlish unigram Sentic pattern	84.7% (155/183)	86.4% (184/213)	0.855
Singlish bigram Sentic pattern	87.4% (160/183)	90.6% (193/213)	0.890
Singlish trigram Sentic pattern	87.4% (160/183)	91.1% (194/213)	0.892

Table 6
Accuracy results from manual annotations.

Method	Positive	Negative
Lexicon classification	27.6% (138/500)	65.4% (327/500)
Emoticon classification	32.4% (126/389)	61.1% (149/244)
SinglishPD classification	77.6% (194/250)	85.6% (214/250)

shows that such terms do not carry any sentiment meaning. Thus, the terms have not been added to the Singlish unigram list, but included in the bigram and trigram lists if they are found to have possibility of being co-localised. While the association of Singlish with these terms is well-known, they do not appear to play a direct role in Singlish polarity detection. Having said that, a more comprehensive study needs to be conducted to investigate the effect of such terms further.

Besides the polarity sentic patterns, there is also a need to recognise tweets that are neutral or having 'spam' content. Several research studies [53,54] have reported that by implementing a hierarchical analysis with subjectivity analysis filtering at the first level will help improve polarity analysis at the second level. Subjectivity detection is a study on understanding if the content contains personal views and opinions as opposed to factual information. Often, these subjective expressions are due to the culture or experience of the person or community and hence can be very localised and specific to a society. As a result, subjectivity is usually first studied before detailed sentiment analysis is done, as it is essential to filter out factual content to have a better understanding of issues that are shared among the netizens. It is worth investigating if the observed findings from subjectivity detection are also applicable to our Singlish dataset. On the other hand, several potential 'spamming' structures were detected during the various analyses and annotation processes in this study. These include request to follow back (e.g., please follow <username>, follow back), job advertisement, location sharing information (e.g., I am at <location>), and song sharing (e.g., nowplaying). The accuracy of polarity detection will definitely improve with the removal of such content.

The results from Tables 3 and 7 indicate that the sentic pattern is able to perform well in classifying data with some Singlish content. However, this observation is not found in Table 5 with data having mostly Singlish sources and localised expressions. In contrast, Singlish polarity n-gram sentic patterns consistently perform well with the latter. These Singlish polarity n-gram sentic patterns can be used as the core resources for further analysis on Singlish. Nevertheless, as observed in Section 4.4.5, special considerations are required to handle ambiguous terms or localised named entities. This can be achieved through concept-based analysis by SenticNet [14]. While bigram and trigram analysis has the ability to filter out these terms, it is not capable of discovering polar sentence structures with terms that are not adjacent to each other. Frequent itemset analysis can be adapted to extract out common occurrences of such terms. Future work incorporating concept disambiguation handling will enable detection of subtle expression, including sarcasms and hence improve the accuracy, especially on tweets with positive polarity. In addition, we plan to use the polarity sentic patterns discovered for Singlish concept-based knowledge base construction, as well as topic-based and domain specific polarity studies in the future.

7. Conclusion

This study was among the first research done on Singlish polarity detection through the construction of Singlish NLP resources and toolkits via a multilingual semi-supervised approach. Besides assembling a Singlish dictionary with relevant POS, Singlish polarity n-gram sentic patterns have been identified. Due to the huge amount of un-annotated data, unsupervised learning including corpus-based bootstrapping using a multilingual, multifaceted lexicon, lexicon polarity detection, frequent item extraction through association rules and LSA have been adopted to create annotated datasets and identify terms and n-grams with polarity before further verification by human assessors. The results have established the importance of having various polarity sentic patterns such as negation, adversative and RT structure in Singlish polarity detection. A proposed hybrid approach combining the SinglishPD algorithm (with n-gram sentic patterns) and SVM is able to show distinctive performance with F-measure of 0.78. The findings have clearly demonstrated that multilingual consideration is essential in analysing localised languages, with Singlish being an example, as English sentic patterns and emoticons are unable to distinguish the different polarities with good accuracy. In view of the many informal localised scarce-resource languages used on social media, the multilingual semi-supervised approach proposed in this paper is vital for polarity detection research, so that sentiment analysis

Table 7
Results based on the emoticon dataset.

Method	Positive	Negative	F-measure
SVM*	89.2% (116/130)	89.3% (133/149)	0.885
English Sentic pattern	52.3% (68/130)	53.7% (80/149)	0.509
Sentic pattern	92.3% (120/130)	79.2% (118/149)	0.854
Singlish unigram Sentic pattern	92.3% (120/130)	83.9% (125/149)	0.876
Singlish bigram Sentic pattern	93.8% (122/130)	84.6% (126/149)	0.887
Singlish trigram Sentic pattern	93.8% (122/130)	84.6% (126/149)	0.887

* based on 5-fold cross validation

can be done more comprehensively on all sorts of content shared rather than merely the English content.

References

[1] E. Cambria, H. Wang, B. White, Guest editorial: Big social data analysis, Knowl.-Based Syst. 69 (2014) 1–2.
[2] E. Cambria, B. Schuller, B. Liu, H. Wang, C. Havasi, Statistical approaches to concept-level sentiment analysis, IEEE Intell. Syst. 28 (3) (2013) 6–9.
[3] E. Cambria, Affective computing and sentiment analysis, IEEE Intell. Syst. 31 (2) (2016) 102–107.
[4] A. Zielinski, U. Bügel, L. Middleton, S. Middleton, L. Tokarchuk, K. Watson, F. Chaves, Multilingual analysis of twitter news in support of mass emergency events, in: Proceedings of European Geosciences Union General Assembly Conference, 2012, pp. 8085–8089.
[5] J.R. Leimgruber, Singapore English, Lang. Linguist. Compass 5 (1) (2011) 47–62.
[6] C. Tan, English or Singlish? The syntactic influences of Chinese and Malay on the learning of English in Singapore, J. Lang. Learn. 3 (1) (2005) 156–179.
[7] E. Riloff, J. Wiebe, Learning extraction patterns for subjective expressions, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2003, pp. 105–112.
[8] S.-M. Kim, E. Hovy, Identifying and analyzing judgment opinions, in: Proceedings of the Conference of North American Chapter of the Association of Computational Linguistics, 2006, pp. 200–207.
[9] R. Mihalcea, C. Banea, J. Wiebe, Learning multilingual subjective language via cross-lingual projections, in: Proceedings of Annual Meeting of Association for Computational Linguistics, 2007.
[10] X. Wan, Co-training for cross-lingual sentiment classification, in: Proceedings of the Joint Conference of the Forty-Seventh Annual Meeting of the Association for Computational Linguistics and the Fourth International Joint Conference on Natural Language Processing, 2009, pp. 235–243.
[11] K. Denecke, Using sentiwordnet for multilingual sentiment analysis, in: Proceedings of International Conference on Data Engineering Workshops, 2008, pp. 507–512.
[12] E. Cambria, J. Fu, F. Bisio, S. Poria, AffectiveSpace 2: Enabling affective intuition for concept-level sentiment analysis, in: Proceedings of AAAI Conference on Artificial Intelligence, 2015, pp. 508–514.
[13] S. Poria, A. Gelbukh, E. Cambria, A. Hussain, G.-B. Huang, EmoSenticSpace: A novel framework for affective common-sense reasoning, Knowl.-Based Syst. 69 (2014) 108–123.
[14] E. Cambria, D. Olsher, D. Rajagopal, SenticNet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis, in: Proceedings of AAAI Conference on Artificial Intelligence, 2014, pp. 1515–1521.
[15] S. Poria, E. Cambria, G. Winterstein, G.-B. Huang, Sentic patterns: Dependency-based rules for concept-level sentiment analysis, Knowl.-Based Syst. 69 (2014) 45–63.
[16] B. Pang, L. Lee, Opinion mining and sentiment analysis, Found. Trends Inf. Retr. 2 (1–2) (2008) 1–135.
[17] S. Volkova, T. Wilson, D. Yarowsky, Exploring sentiment in social media: Bootstrapping subjectivity clues from multilingual twitter streams, in: Proceedings of Annual Meeting of the Association of Computational Linguistics, 2013, pp. 505–510.
[18] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity in phrase-level sentiment analysis, in: Proceedings of Conference on Empirical Methods in Natural Language Processing, 2005, pp. 347–354.
[19] A. Cui, M. Zhang, Y. Liu, S. Ma, Emotion tokens: Bridging the gap among multilingual twitter sentiment analysis, in: Information Retrieval Technology, Springer, 2011, pp. 238–249.
[20] S. Baccianella, A. Esuli, F. Sebastiani, SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining, in: Proceedings of Language Resources and Evaluation Conference, 2010, pp. 2200–2204.
[21] L. Barbosa, J. Feng, Robust sentiment detection on twitter from biased and noisy data, in: Proceedings of the Twenty-Third International Conference on Computational Linguistics: Posters, 2010, pp. 36–44.
[22] D. Davidov, O. Tsur, A. Rappoport, Enhanced sentiment learning using twitter hashtags and smileys, in: Proceedings of the Twenty-Third International Conference on Computational Linguistics: Posters, 2010, pp. 241–249.
[23] A. Pak, P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining, in: Proceedings of Language Resources and Evaluation Conference, 2010, pp. 1320–1326.
[24] C. Banea, R. Mihalcea, J. Wiebe, A bootstrapping method for building subjectivity lexicons for languages with scarce resources, in: Proceedings of Language Resources and Evaluation Conference, 2008, pp. 2764–2767.
[25] P.D. Turney, Mining the Web for synonyms: PMI-IR versus LSA on TOEFL, in: Proceedings of the 12th European Conference on Machine Learning, 2001, pp. 491–502.
[26] P.D. Turney, Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews, in: Proceedings of Annual Meeting of the Association of Computational Linguistics, 2002, pp. 417–424.
[27] S.T. Dumais, G.W. Furnas, T.K. Landauer, S. Deerwester, R. Harshman, Using latent semantic analysis to improve access to textual information, in: Proceedings of the Special Interest Group on Computer-Human Interaction Conference, 1988, pp. 281–285.
[28] A. Bakliwal, P. Arora, V. Varma, Hindi subjective lexicon: A lexical resource for Hindi polarity classification, in: Proceedings of Language Resources and Evaluation Conference, 2012, pp. 1189–1196.
[29] G.A. Miller, WordNet: A lexical database for English, Commun. ACM 38 (11) (1995) 39–41.
[30] S. Chowdhury, W. Chowdhury, Performing sentiment analysis in Bangla microblog posts, in: Proceedings of International Conference on Informatics, Electronics & Vision, 2014, pp. 1–6.
[31] E. Cambria, A. Hussain, C. Havasi, C. Eckl, Sentic computing: Exploitation of common sense for the development of emotion-sensitive systems, LNCS 5967 (2010) 148–156.
[32] E. Cambria, P. Gastaldo, F. Bisio, R. Zunino, An ELM-based model for affective analogical reasoning, Neurocomput. 149 (2015) 443–455.
[33] Y. Xia, X. Li, E. Cambria, A. Hussain, A localization toolkit for SenticNet, in: Proceedings of IEEE International Conference on Data Mining Workshops, 2014, pp. 403–408.
[34] Y. Chen, S. Skiena, Building sentiment lexicons for all major languages, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics (Short Papers), 2014, pp. 383–389.
[35] A. Go, R. Bhayani, L. Huang, Twitter sentiment classification using distant supervision, CS224N Proj. Rep. Stanf. (2009) 1–12.
[36] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up?: Sentiment classification using machine learning techniques, in: Proceedings of Conference on Empirical Methods in Natural Language Processing, 2002, pp. 79–86.
[37] J. Wiebe, T. Wilson, C. Cardie, Annotating expressions of opinions and emotions in language, Lang. Resour. Eval. 39 (2–3) (2005) 165–210.
[38] J. Wang, P. Neskovic, L.N. Cooper, Training data selection for support vector machines, in: Advances in Natural Computation, Springer, 2005, pp. 554–564.
[39] J.L. Fleiss, Measuring nominal scale agreement among many raters, Psychol. Bull 76 (5) (1971) 378.
[40] T. Joachims, Text categorization with support vector machines: Learning with many relevant features, In: Proceedings of the 10th European Conference on Machine Learning, 1998, pp. 137–142.
[41] S.L. Lo, D. Cornforth, R. Chiong, Identifying the high-value social audience from Twitter through text-mining methods, in: Proceedings of the Asia Pacific Symposium on Intelligent and Evolutionary Systems, 2015, pp. 325–339.
[42] S.L. Lo, R. Chiong, D. Cornforth, Using support vector machine ensembles for target audience classification on Twitter, PloS One 10 (4) (2015) e0122855.
[43] C.J. Burges, A tutorial on support vector machines for pattern recognition, Data Min. Knowl. Discov. 2 (2) (1998) 121–167.
[44] P. Willett, The Porter stemming algorithm: Then and now, Program Electron. Libr. Inf. Syst. 40 (3) (2006) 219–223.
[45] X. Hu, J. Tang, H. Gao, H. Liu, Unsupervised sentiment analysis with emotional signals, in: Proceedings of the International Conference on World Wide Web, 2013, pp. 607–618.
[46] P. Fournier-Viger, A. Gomariz, T. Gueniche, A. Soltani, C.-W. Wu, V.S. Tseng, SPMF: A Java open-source pattern mining library, J. Mach. Learn. Res. 15 (1) (2014) 3389–3393.
[47] Y. Choueka, Looking for needles in a haystack or locating interesting collocational expressions in large textual databases, in: Proceedings of Recherche d'Information Assistée par Ordinateur conference, 1988, pp. 609–623.
[48] M. Sokolova, N. Japkowicz, S. Szpakowicz, Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation, in: Advances in Artificial Intelligence, 2006, pp. 1015–1021.
[49] J. Zobel, How reliable are the results of large-scale information retrieval experiments? in: Proceedings of the Annual International ACM Special Interest Group on Information Retrieval conference on Research and Development in Information Retrieval, 1998, pp. 307–314.
[50] F. Benamara, C. Cesarano, A. Picariello, D.R. Recupero, V.S. Subrahmanian, Sentiment analysis: Adjectives and adverbs are better than adjectives alone, in: Proceedings of International Conference on Web and Social Media, 2007, pp. 1–7.
[51] V. Perez-Rosas, C. Banea, R. Mihalcea, Learning sentiment lexicons in Spanish, in: Proceedings of Language Resources and Evaluation Conference, 2012, pp. 3077–3081.
[52] D. Rao, D. Ravichandran, Semi-supervised polarity lexicon induction, in: Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics, 2009, pp. 675–682.
[53] E. Boiy, M.-F. Moens, A machine learning approach to sentiment analysis in multilingual Web texts, Inf. Retr. 12 (5) (2009) 526–558.
[54] A. Balahur, M. Turchi, Improving sentiment analysis in Twitter using multilingual machine translated data, in: Proceedings of Recent Advances in Natural Language Processing, 2013, pp. 49–55.