9-2017

An unsupervised multilingual approach for online social media topic identification

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

Raymond CHIONG

David CORNFORTH

Expert Systems With Applications 81 (2017) 282–298

Contents lists available at ScienceDirect

Expert Systems With Applications

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

An unsupervised multilingual approach for online social media topic identification

Siaw Ling Lo*, Raymond Chiong, David Cornforth

School of Electrical Engineering and Computing, The University of Newcastle, Callaghan, NSW 2308, Australia

ARTICLE INFO

Article history:
Received 12 July 2016
Revised 13 March 2017
Accepted 14 March 2017
Available online 21 March 2017

Keywords:
Topic identification
Multilingual analysis
Unsupervised learning
Social media

ABSTRACT

Social media data can be valuable in many ways. However, the vast amount of content shared and the linguistic variants of languages used on social media are making it very challenging for high-value topics to be identified. In this paper, we present an unsupervised multilingual approach for identifying highly relevant terms and topics from the mass of social media data. This approach combines term ranking, localised language analysis, unsupervised topic clustering and multilingual sentiment analysis to extract prominent topics through analysis of Twitter's tweets from a period of time. It is observed that each of the ranking methods tested has their strengths and weaknesses, and that our proposed 'Joint' ranking method is able to take advantage of the strengths of the ranking methods. This 'Joint' ranking method coupled with an unsupervised topic clustering model is shown to have the potential to discover topics of interest or concern to a local community. Practically, being able to do so may help decision makers to gauge the true opinions or concerns on the ground. Theoretically, the research is significant as it shows how an unsupervised online topic identification approach can be designed without much manual annotation effort, which may have great implications for future development of expert and intelligent systems.

1. Introduction

While it is a known fact that social media sharing is of huge volume and high velocity, it is not often realised that the content shared is of varied structures, which include images, videos, and text. With the openness and free form of expression, it is also not surprising to find textual content written in a mixture of languages, including informal terms and phrases that hardly follow any proper grammatical rules. Effective mining of social media data therefore can no longer be focused on a single language, but it is essential to embark on a multilingual approach to fully comprehend the sentiment and content shared online.

In this study, our aim is to identify topics that are of value to a community, in the form of opinions, concerns, news and so on, within the mass of social media data. Following previous studies (e.g., Aiello et al., 2013; Zhao et al., 2011) that have used content from main stream media as coverage comparison with social media, we are interested to assess if topics shared on social media are the same as main stream media in our study. We have chosen Twitter as the base of our investigation because of its ability to propagate hot topics in a very short duration and to a wide audience. In addition, the degree of variations in languages found on Twitter can be immense. Most of the past studies (e.g., see Vicient and Moreno, 2015; Zhao et al., 2011) have focused only on analysing English tweets, even though English is used by just 28.6% of the Internet users.[1]

Besides the linguistic variations found in tweets, it is observed that mixed languages and use of a localised lingual range are commonly seen for expressing emotion online, especially in a multicultural environment (Zielinski et al., 2012). One such example is Singlish, the colloquial Singaporean English that has incorporated elements of some Chinese dialects and the Malay language (Leimgruber, 2011). One of the main reasons of the prevalent use of such a unique 'language' is because a native or localised vernacular can resonate with the local community better than a formal language.[2] This leads us to the decision of carrying out multilingual analysis on Singlish tweets, to see if we could detect the concerns of or interesting news from the local Singaporean community. However, this can be very challenging because of limited resources available for an informal language like Singlish.

Topic detection and tracking have been extensively studied to identify new topics in a temporally-ordered news stream (Allan, 2012). Topic modelling (Lu, 2015; Zhao et al., 2011), temporal

* Corresponding author.
E-mail addresses: siawling.lo@uon.edu.au (S.L. Lo), raymond.chiong@newcastle.edu.au (R. Chiong), david.cornforth@newcastle.edu.au (D. Cornforth).

[1] http://www.internetworldstats.com/stats7.htm.
[2] http://mypaper.sg/top-stories/officials-use-singlish-dialects-reach-out-20150211.

http://dx.doi.org/10.1016/j.eswa.2017.03.029
0957-4174/

segmentation (Benhardus & Kalita, 2013; Lu, 2015), classification of event and non-event (Becker, Naaman, & Gravano, 2011), unknown event identification (Psallidas, Becker, Naaman, & Gravano, 2013), hashtags (Vicient & Moreno, 2015) and exemplar-based topic detection (Elbagoury, Ibrahim, Farahat, Kamel, & Karray, 2015) are some of the approaches used in this research area. Temporal segmentation is important for detecting topics, and various studies in the literature have adopted different kinds of metrics, such as minutes (Benhardus & Kalita, 2013), hours (Becker et al., 2011; Benhardus & Kalita, 2013), days (Lu, 2015), and weeks (Lu, 2015). While most of the topic detection analyses have used specific events and annotated datasets for evaluation (e.g., see Aiello et al., 2013; Becker et al., 2011; Elbagoury et al., 2015), we propose a multilingual analysis through high-value term comparison using a Singlish dataset (see Section 3.2) to discover interesting topics via a candidate day selection process with minimal annotation effort. In addition, instead of manually categorising the vast amount of tweets for evaluation purposes, an online web service is adopted to automatically classify the ground truth dataset (see Section 3.3) for more comprehensive evaluation in this study.

In terms of methods, we propose a Peak Identification algorithm and compare it to Term Frequency-Inverse Document Frequency (TFIDF) and Term Frequency (TF). We also consider three topic clustering methods – Twitter Latent Dirichlet Allocation (LDA) (Zhao et al., 2011), K-Means (MacQueen, 1967) and the Dirichlet Process Mixture Model (DPMM) (Antoniak, 1974). Our intention is to identify terms and topics that are relevant and of high-value to a local community in an unsupervised manner. To ascertain if the proposed approach can identify topics that are of high-value, tweets have been clustered using days for assessment across different datasets. Besides that, the top topics identified are subjected to multilingual sentiment analysis to uncover sentiments on the ground.

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 attempt to identify topics from tweets with consideration of its multilingual nature through unsupervised learning without the use of any external knowledge base to decipher the context of tweets.
- Tweets in a localised language (i.e., Singlish as used in this study) can be leveraged for identifying relevant and important topics that are of interest or concern to the local community. The comparison of top terms discovered with the Singlish dataset succeeded in choosing appropriate candidate days with minimal annotation effort.
- Our proposed approach of using the DPMM clustering method and a 'Joint' term ranking method has consistently performed well in the topic recall and precision@10 evaluation metrics.
- From the observation of our results, it is essential to find optimal parameters for the DPMM even though there is no predefined topic number required for DPMM clustering (as opposed to Twitter LDA and K-Means).
- Our multilingual sentiment analysis has uncovered mixed-language tweets that were not detected when using an English polarity detection algorithm. This finding is important, since it highlights the necessity of considering the multilingual nature of online sharing to ensure a more comprehensive analysis.
- It is observed that both the social and main stream media platforms would share the same main topics if there are prominent events on the day. However, this observation does not hold for 'ordinary' days. Our approach of ranking the high-value topics therefore plays a crucial role in understanding/gauging the interests or concerns of the local community.

The rest of this paper is structured as follows. We review related work in Section 2, and describe the details of datasets and resources constructed in Section 3. Section 4 presents the methods and experimental setups, which include candidate day selection, term ranking methods, topic clustering methods, evaluation metrics and multilingual analysis. In Section 5, we list our findings and explain the results. Section 6 discusses our approach and observations before the conclusions are drawn in Section 7.

2. Related work

2.1. Topic detection

Broadly speaking, two main types of data sources have been used in evaluating topic detection approaches: labelled/curated and unlabelled data sources. The former mainly relies on annotated datasets of specific topics for identification or classification, while the latter attempts to cluster relevant topics based on features and information found in tweets without labelled data.

Aiello et al. (2013) adopted standard natural language processing techniques, n-grams, co-occurrence and a variant of TFIDF to detect topics on three manually annotated Twitter datasets. Becker et al. (2011) focused on differentiating the event and non-event items on Twitter. Their classification was based on a list of features, which include temporal, social, topical and Twitter-centric features. The output of a Support Vector Machine (SVM) was assigned to logistic regression models to obtain probability estimates of the class assignment. However, prior clustering was required before the classification and substantial manual annotations were needed for the classification job. Elbagoury et al. (2015) used the concept of exemplar tweets to detect similar topics. The data source was based on specific topics using keywords and hashtags. A similarity matrix of tweets was constructed and the tweet with the highest variance was selected in an iterative manner to form the exemplar tweet for each topic. Psallidas et al. (2013) used user-tagged Flickr data to implement an online clustering framework that leverages multiple features (e.g., temporal, topical, platform behaviours, hashtags) to identify an unknown event. Ensemble learning methods were used to learn and assign the weight and threshold to each feature before applying the knowledge to extract relevant events. Magdy and Elsayed (2016) made use of a set of predefined queries to train classifiers for extracting relevant microblogs belonging to specific topics. These seeding queries were selected to train the classifiers frequently in order to adaptively filter content related to the topic. Vicient and Moreno (2015) proposed a semantic approach through linking WordNet and Wikipedia to cluster tweets based on curated hashtags. While a semantic linkage can enhance the context of tweets, it often relies upon an external knowledge base such as WordNet or Wikipedia.

A study by Vavliakis, Symeonidis, and Mitkas (2013) integrated named entity recognition and dynamic topic map discovery with the help from LDA and topic clustering for event identification in unlabelled blogposts. They made use of open linked data corpora (DBpedia.org and Geonames.org) to discover and summarise interesting events. Benhardus and Kalita (2013) used different temporal segmentations, which include minutes, days and weeks, on unlabelled tweets to identify trending topics through TFIDF and relative normalised TF analysis. The results were compared to human-annotated topics for verification. Zhao et al. (2011) developed a Twitter-LDA model to discover topics from a representative sample of tweets and adopted text-mining techniques to compare the topics with traditional news media. It was found that Twitter could be a faster news feed, and that it covers the same number of topics as traditional media. A recent variant of LDA is the Probit-Dirichlet Hybrid Allocation topic model (Lu, 2015) that incorporates each

document's temporal features to detect the dynamic of short-term cyclical topics.

To avoid the need of extensive manual annotation, we have opted to use unlabelled data in our work. The motivation is to develop an approach that can be adopted in a real-world environment (where source data is mostly unlabelled and untagged) and see if we can leverage emotion-rich localised content for topic detection. Besides that, we also do not rely on external resources like Vavliakis et al. (2013) and Vicient and Moreno (2015) did, given the dynamism of the tweets shared and the informal structure detected, which require another layer of ontology mapping or deciphering before any meaningful information can be derived. Although an idea similar to the concept of exemplar tweets has been used in our approach (see Section 4.4.2), we rely on three methods to identify top terms, namely the Peak Identification algorithm, TFIDF and TF ranking. This is to ensure that the exemplar tweets that we have chosen are representative through the best ranking algorithm. Different from Elbagoury et al. (2015), the tweets identified by our term ranking methods are not used directly for identifying topics but for evaluating results with ground truth topics extracted to ascertain the best approach in topic identification.

The approach proposed by Benhardus and Kalita (2013) is interesting, as it is able to identify events from an unlabelled dataset using a relatively simple approach of TFIDF term weighting and normalised TF analysis together with temporal analysis. We have included TFIDF and TF ranking in our analysis, but our results show that the two methods do not work well on their own. Instead, we propose using a 'Joint' term ranking method that can consistently identify relevant terms with high precision. Another work based on unlabelled Twitter data is that by Zhao et al. (2011), in which they compared the topics found in Twitter data to those from main stream media. However, they used Twitter data from the United States (U.S.), which is quite different from that of Singapore. They analysed only the English language and did not take into consideration the effect of localised languages in the tweets. Table 1 summarises the pros and cons of different topic detection approaches proposed in the field.

As shown in the table, we have used the DPMM instead of LDA or its variants in our study. The DPMM has recently gathered much attention due to the fact that it is a clustering method that does not require a pre-defined number of clusters, unlike other approaches based on the LDA or K-Means model. The DPMM has been adopted in topic identification for short text streams (Wang, Yuan, Wang, & Xue, 2011) and multi-object tracking (Luo, Stenger, Zhao, & Kim, 2015). Wang et al. (2011) extended the DPMM to identify topics for streaming texts by managing topic segmentation, topic detection and tracking simultaneously, while Luo et al. (2015) applied the DPMM to topic discovery from video sequences by treating them as documents. We use the DPMM for clustering tweets, in order to uncover topics that can be used for verification purposes.

2.2. Multilingual analysis

Research on multilingual analysis mainly concentrates on translation engines (Gao, Zhou, Diao, Sorensen, & Picheny, 2002; Mitamura, 1999) and sentiment analysis (Balahur & Perea-Ortega, 2015; Cui, Zhang, Liu, & Ma, 2011; Volkova, Wilson, & Yarowsky, 2013). In the area of automated translation research, Mitamura (1999) used a controlled vocabulary through a controlled language checker for multilingual machine translation, while Gao et al. (2002) developed a statistical semantic parser to do automatic translation between spoken English and Chinese languages.

For research efforts on social media multilingual sentiment analysis, Volkova et al. (2013) proposed an approach to bootstrap subjectivity clues from Twitter data and evaluated the approach on English, Spanish and Russian Twitter streams. The proposed approach uses the multi-perspective question answering lexicon (Wilson, Wiebe, & Hoffmann, 2005) to bootstrap sentiment lexicons from a large pool of unlabelled data using a small amount of labelled data to guide the process. Balahur and Perea-Ortega (2015) and Cui et al. (2011) worked on Twitter sentiment analysis instead of a subjectivity study. Balahur and Perea-Ortega (2015) used multilingual machine-translated data to improve sentiment classification. It was found that joint classifiers from languages with similar structures help to achieve improvement over monolingual classifiers through eliminating noisy features and reinforcing valuable ones. Cui et al. (2011) did not use a translation machine but focused on building emotion tokens (SentiLexicon) using emoticons, repeating punctuations and repeating letters. A comparative evaluation using SentiWordNet (Baccianella, Esuli, & Sebastiani, 2010) indicated that emotion tokens are helpful for both English and non-English Twitter sentiment analysis.

Most of the multilingual sentiment analysis studies have focused on single languages (Volkova et al., 2013) or relied on a rich-resource language to infer other languages (Balahur & Perea-Ortega, 2015). Recently, the emphasis has shifted to include multiple languages when analysing online content, and this includes localised languages with limited resources available. It has been shown that by including these languages, more comprehensive analyses can be done (Lo, Cambria, Chiong, & Cornforth, 2016a). Furthermore, a survey on social media streams by Bontcheva and Rout (2014) has identified the multilingualism of social media as one of the outstanding challenges, since most methods surveyed were developed and tested on English content only. The survey paper acknowledges the scarceness of resources for such languages and suggests to do crowdsourcing to overcome the limitation.

It is worth highlighting that none of the related work discussed above had dealt with multilingual types of tweets for relevant term identification and topic detection. Also, none of them had used a localised language like Singlish as a seed to extract highly relevant candidate days for topic clustering analysis. Recently, a study was carried out by Kim, Weber, Wei, and Oh (2014) on sociolinguistic analysis of Twitter in multilingual societies and they found that users who speak localised languages have a stronger influence over others who mainly use a single language on the platform. This group of users has the tendency to express informative/political/debatable topics in a localised language rather than using English. Even though Kim et al. (2014) did not discuss about topic identification on Twitter, they have observed an interesting outcome that multilingual users are more likely to use localised languages to put across topics of concerns. This finding is important since it concurs with the basis of our research to leverage localised languages for identifying topics of interest or concern to the local community.

3. Details of datasets

3.1. Twitter dataset collection

In order to enable unsupervised topic detection, continuous tweet extraction through a period of time was carried out. Since there was no Twitter dataset with Singlish content available for analysis, we collected the Twitter dataset used in this study by following a list of Twitter users who were tweeting topics relevant to Singapore and its regions. Given that the location information of Twitter users is typically not verified, we used Twitter's location information only as a reference and relied predominantly on the content shared through verification with news sources and topics from forums. As a result, 1498 users were consolidated for data collection purposes. Twitter's Search API was used to extract

Table 1
A comparison of topic detection approaches.

	Data source	Pro	Con
Aiello et al. (2013)	Twitter (curated)	Leveraged n-gram co-occurrences and time-dependent ranking to identify bursty events on three different datasets.	Data was crawled based on specific topics using keywords and hashtags that may not be reflective of the heterogeneous nature of streaming tweets.
Becker et al. (2011)	Twitter (streaming, labelled)	Detailed feature analysis was done. An SVM was used for classification before assigning the output to logistic regression for probability estimation.	Labelling of event and non-event can be labour intensive.
Elbagoury et al. (2015)	Twitter (curated)	Usage of exemplar tweets to detect similar topics using a similarity matrix and iterative selection via a variance value.	Data was crawled based on specific topics using keywords and hashtags that may not be reflective of the heterogeneous nature of streaming tweets.
Psallidas et al. (2013)	Flickr (labelled)	Multiple features were used in ensemble learning and the online clustering framework to identify unknown events.	The source data was manually tagged by users with event IDs. The approach may not be feasible on other platforms without tagged data.
Magdy and Elsayed (2016)	Twitter (curated)	Predefined queries were used for content filtering on broad and dynamic topics using classifiers.	It can be challenging to detect unknown topics since an initial set of tweets was used to retrieve relevant content for training a binary classifier.
Vicient and Moreno (2015)	Twitter (curated)	A semantic linkage based on external resources enables clustering of tweets using hashtags.	Tweets were hash-tagged and extracted from specific sites. The approach may not be feasible without tagged data. External resources were used – WordNet and Wikipedia.
Vavliakis et al. (2013)	Blogposts (unlabelled)	The approach integrates named entity recognition and dynamic topic map discovery for topic detection so that semantically rich representations of events can be extracted.	External resources required (DBpedia.org and Geonames.org).
Benhardus and Kalita (2013)	Twitter (streaming, unlabelled)	The proposed method is relatively simple using TFIDF term weighting, normalised TF analysis together with temporal analysis.	Human validation on relevancy was in place to compare Twitter trending topics as results from the precision score were not satisfactory. This may imply that it is not straightforward to identify correct terms for topic identification.
Zhao et al. (2011)	Twitter (streaming, unlabelled)	An unsupervised Twitter-LDA model was developed to discover topics and it has been shown that the content of Twitter is comparable to traditional news media.	A semi-automatic topic categorisation was done on tweets to compare results with the ground truth topics that may not be scalable.
Lu (2015)	Yahoo Finance WalMart Message Board, New York Times, Reuters-21,578 (unlabelled)	The ability to include a document's temporal features to capture a topic's short term cyclical dynamics in an unsupervised topic model.	The computational cost of incorporating short-term cyclical dynamics has multiplied the processing time when the topic number is large.
Our approach	Twitter (streaming, unlabelled)	We leverage the uniqueness of localised languages for topic identification without relying on external resources. Our topic clustering was done using DPMM clustering without the need of a prior topic number assignment.	Limited resources on localised languages to address the word sense disambiguation problem.

tweets and relevant meta-data from September 2013 to March 2014, and a total of 3125,600 records were collected.

Since the source of our dataset is of mixed language nature, it is of interest to understand the distributions of different languages in the extracted Twitter dataset. A detailed language analysis was done using the Language Detection Library for Java.[3] The analysis showed that 24% of the data collected contains mixed languages, and 46% of it is detected as having English language content, while 19% falls under the category of Indonesian/Malay languages. This is evidence that the content of the dataset is heterogeneous, and that conducting analysis using a single language such as English is unable to fully comprehend the diversity of its content.

3.2. Singlish dataset construction

As there was no available de-facto Singlish dictionary, manual construction of a Singlish dictionary has been done by consolidating several Internet resources. Resources used to construct this Singlish dictionary include the Dictionary of Singlish and Singapore English,[4] Coxford Singlish Dictionary,[5] and Wikipedia Singlish vocabulary.[6] Since we are interested to assess if a localised language, i.e., Singlish, can be used to identify relevant topics, each Singlish term has been given a simple English description instead of elaborated explanation. The finalised list of our Singlish-English dictionary contains 978 unique Singlish expressions. These unique terms have been used to extract Singlish tweets from the Twitter dataset collected (see Section 3.1), and a total of 517,350 tweets were identified to form what we called the Singlish dataset in this paper.

3.3. Ground truth dataset construction

For the purpose of evaluating results of unsupervised topic clustering, there is a need to construct a ground truth topics dataset of the same period. A previous study by Zhao et al. (2011) on news from the U.S. found that Twitter can be regarded as a faster news feed than traditional media, with both covering the same

[3] http://code.google.com/p/language-detection/.

[4] http://www.singlishdictionary.com/.

[5] http://www.talkingcock.com/html/lexec.php.

[6] https://en.wikipedia.org/wiki/Singlish_vocabulary.

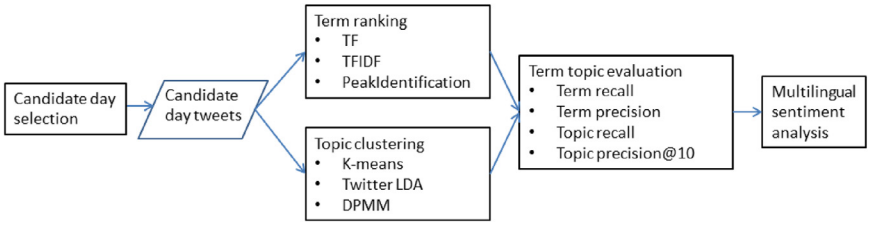

Fig. 1. The overall architecture of our unsupervised multilingual topic identification approach.

topics. Since there was no known annotated Twitter dataset with Singlish content, news headlines reported from Singapore's main stream media such as The Straits Times and Channel News Asia were extracted to produce the ground truth dataset for this study. In view of the real-time nature of Twitter, contents shared by Twitter accounts of main stream media (including @STcom, @ChannelNewsAsia, @SGnews, @sgbroadcast, @TODAYonline) were also consolidated in the ground truth dataset.

4. Methods and experimental setups

In this section, we present the main components of our unsupervised multilingual topic identification approach. The overall architecture can be found in Fig. 1. As we can see from the figure, the process starts with candidate day selection - we analyse tweets over a period of seven months (as mentioned in Section 3.1) and extract candidate days that may have content that can resonate well with the local community. After that, tweets from the candidate days are extracted for further analysis. Terms from the Twitter dataset (see Section 3.1) as well as the Singlish dataset (see Section 3.2) of a candidate day are compared and ranked by three methods, namely, TF, TFIDF and our proposed Peak Identification algorithm. A top term list is then created for each ranking method for subsequent procedures such as identifying relevant topics for clustering methods and creating word vectors for K-Means clustering. Three clustering methods have been used in this study: K-means, Twitter LDA of 10 to 40 (with an interval of 10) topics, and the DPMM. In order to select topics for evaluation, top terms discovered by each of the ranking methods are used to extract the top topics. The results are evaluated using the ground truth dataset (see Section 3.3) based on both recall and precision for terms and topics. Lastly, multilingual sentiment analysis is carried out on tweets belonging to the top topic discovered by the best performing topic clustering method.

4.1. Candidate day selection

Since the assumption is that tweets with Singlish terms are more expressive and hence may contain information that is of interest or concern to the local community, candidate terms clustered by day from both the Twitter and Singlish datasets are compared and matched terms are extracted. The details are depicted in Fig. 2. As shown in the figure, tweets from both the Twitter and Singlish datasets are first pre-processed before terms are extracted for ranking. Potential candidate terms are then used to select candidate days that contain the relevant topics. In order to aid in the selection of a candidate day, term frequencies of the matched terms from both datasets are summed together and used as a metric for ranking the candidate day.

4.2. Term ranking

As mentioned before, it is common for tweets to have informal languages mixed with linguistic variations, purposely misspelled words or repetitions as signs of emphasis (e.g., "perrfeeect"). It is therefore necessary to conduct a noise removal procedure through term frequency analysis before any term ranking is done. Common terms (terms that have appeared in 95% of the days) and infrequent terms (terms found in less than 5% of the days) will be removed through this procedure. In addition, the following pre-processing steps are also needed for each tweet:

- Removal of URLs and Twitter usernames.
- Preserving hashtags but the # symbol is removed.
- Removal of common stop words, such as "a", "the" and Twitter's special character "RT" (stands for retweet).

It is also not unusual to find informal or expressively 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 corresponding word is reduced to two occurrences. For example, "gooood" is converted to "good" and "hahahaha" is amended to "haha" . This process also applies to punctuations so "?????" is transformed into "??" for the sake of consistency.

As mentioned at the beginning of Section 4, the term ranking methods are used to generate a top term list for each ranking method on each candidate day so that relevant topics can be identified from the clustering results. Since the Peak Identification ranking method (see Section 4.2.1 for details) extracts a finite number of terms, this same cut-off threshold will be applied to TF and TFIDF so that none of the ranking methods has more terms assigned and each ranking method has the same number of terms.

4.2.1. Peak identification

The Peak Identification algorithm works on the basis that if an unusually high frequency or a peak is found within a range of values, that peak is often an incident that is worth investigating. However, to adapt the idea for analysing tweets, where noise is a big issue, pre-processing of tweets, as mentioned earlier, is essential to minimise the identification of irrelevant peaks.

After the noise removal and pre-processing steps are done, each of the existing terms will be analysed through its frequency distribution. A previous study by Shamma, Kennedy, and Churchill (2011) used the top peak for modelling microblog conversations. In this study, we keep top three peaks for term ranking purposes. We first cluster the tweets based on days. Each term and its peaks are later consolidated and mapped back to their respective dates. Any dates containing peaks are dates with candidate terms that are of higher frequency.

4.2.2. TF and TFIDF

TF measures the frequency of a term in a document. In this study, we consider the normalised form and hence term frequency *tf* is defined as

$$tf_{ij} = n_{ij}/N_j \tag{1}$$

where n_{ij} is the number of times term i occurs in tweet j and N_j is the total number of terms in tweet j.

TFIDF (Salton & Buckley, 1997) is commonly used in information retrieval for assessing the importance of a term in a document or

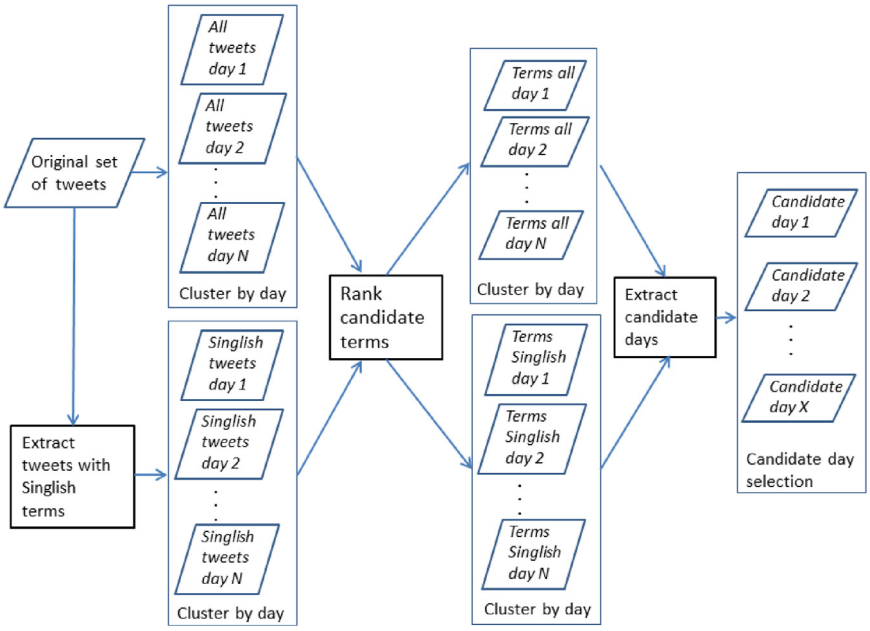

Fig. 2. Candidate day selection.

a corpus. As it has been used to identify trending topics on Twitter (e.g., see Benhardus & Kalita, 2013; Elbagoury et al., 2015) and differentiate event and non-event items (Becker et al., 2011), we would like to see if it is suitable to be used for identifying relevant topics for a local community. The TFIDF value of a term used in this study is the product of two components, term frequency *tf* and inversed document frequency *idf*. The calculation of *idf* is as follows:

$$idf_i = \log(D/d_i) \tag{2}$$

where d_i is the number of tweets that contain term i and D is the total number of tweets.

4.3. Topic clustering

As the purpose of this study is to discover high-value topics through unsupervised learning, three clustering methods are used. K-Means and Twitter LDA are commonly used in various clustering tasks (e.g., see Elbagoury et al., 2015; Lo, Chiong, & Cornforth, 2016), but there is a need to fix the number of clusters a priori. On the other hand, the DPMM does not require us to pre-define a total number of clusters as a parameter before performing the analysis.

4.3.1. K-Means

K-Means (MacQueen, 1967) is a simple unsupervised learning algorithm that can be used to solve clustering problems. It starts with defining the number of clusters required - assuming it to be k clusters - before randomly placing k points or centroids into the space represented by feature vectors. Due to the linguistic variations or noises in tweets, a word list is used to minimise the sparseness of term occurrence feature vectors. This word list is a joint list of top terms extracted by the three ranking methods mentioned in Section 4.2. After which, each tweet represented as an object is assigned to the group that has the closest centroid through Euclidean distance calculation. When all the objects have been assigned, we recalculate the positions of the k centroids. Repeat the assignment process until there is no change in the position of centroids.

Due to the need to pre-define a k or number of clusters, a range of values were tested in this study to find a suitable k for evaluating the performance of K-Means topic clustering. Four k values of 10 to 40 (with an interval of 10) have been used, as it is reasonable to consider a value around 20 topics (derived from analysing the ground truth dataset) when analysing the number of topics in a day.

4.3.2. Twitter LDA

LDA (Blei, Ng, & Jordan, 2003), a renowned generative probabilistic model for topic discovery, has been used in various social media studies (e.g., see Yang and Rim (2014); Zhao et al. (2011)). LDA uses an iterative process to build and refine a probabilistic model of documents, each containing a mixture of topics. However, standard LDA may not work well with Twitter, as tweets are typically very short. If one aggregates all the tweets of a follower to increase the size of the documents, this may diminish the fact that each tweet is usually about a single topic. Moreover, our previous study has shown that it is essential to represent each individual tweet as a single topic, as combining all the tweets to extract representative topics do not perform well in the context of SVM classification (Lo, Cornforth, & Chiong, 2015). We therefore have adopted the implementation of Twitter LDA (Zhao et al., 2011) for unsupervised topic discovery.

Twitter LDA requires a pre-defined topic number n for generating the n-topic model. Similar to K-Means, four different topic models ranging from 10 to 40 (with an interval of 10) topics have been used in this study. We generated a list of topics after running 100 iterations of Gibbs sampling while keeping the other model parameters (Dirichlet priors) constant: $\alpha = 0.5$, $\beta_{word} = 0.01$, $\beta_{background} = 0.01$ and $\gamma = 20$.

4.3.3. DPMM

The DPMM (Antoniak, 1974) is a Bayesian nonparametric model that can be constructed as a single mixture model, where the number of mixture components is infinite. As a result, the DPMM does not need to use any pre-defined number of clusters from the beginning. Fig. 3. is the graphical model of DPMM and its generative process. A topic $\theta_i = \{\theta_{ij}\}_{j=1}^{j=|V|}$ (a multinomial distribution over words belonging to the vocabulary V) is first sampled for t_i according to a Dirichlet Process (DP) $G \sim DP(\alpha, G_0)$. $G_0 = Dir(\vec{\beta})$ is the base distribution of DP while G is a random distribution over Θ parameter space sampled from the DP that assigns probabilities

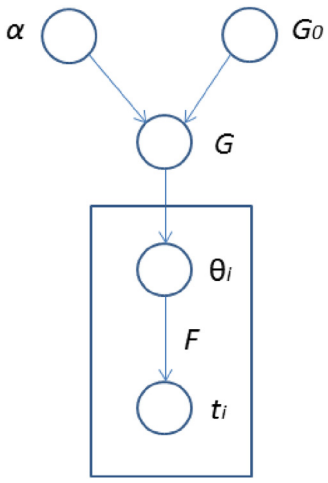

Fig. 3. A graphical model of DPMM.

to parameters. The generative distribution F is parameterised by θ_i and t_i is considered as a bag of words, given the topic θ_i. θ_i can also be seen as latent variables on t_i with information on clusters assigned to t_i. Gibbs sampling is used to estimate the cluster assignments of the model.

The implementation of DPMM specified by Hu, Li, Li, Shao, and Wang (2015) has been used in this study. Even though we initially set the default values of two main parameters, α(alpha) and β(beta), to 1 in the implementation, a study by Yin and Wang (2014) showed that alpha performs best at 0.1. On the other hand, beta is commonly set at 0.01 in the literature (Griffiths & Steyvers, 2004) but the study by Yin and Wang (2014) suggested that a beta value of 0.1 is more suitable for short texts. In order to find the optimal alpha and beta values for our study, grid search on a range of values with respect to perplexity (Blei, Ng, & Jordan, 2003) and topic numbers was conducted. 100 iterations of Gibbs sampling were run for each DPMM experiment (see Section 5.3 for details).

4.4. Evaluation

Even though it is common to use perplexity (Blei et al., 2003) as an evaluation metric for a clustering task, we focus on recall and precision for terms and topics when assessing the quality of clustering, as shown in Fig. 1. This is mainly because our aim is to identify high-value topics and thus it is essential to assess how well each clustering method performs on the selected candidate days. Perplexity does not play a direct role in evaluating high-value topic identification but it is used for DPMM parameter selection in our work.

4.4.1. Perplexity

Most of the time there is no label or annotation available for a clustering dataset. Perplexity, as defined by Blei et al. (2003), is typically used to measure the ability of a clustering model in generating unseen data. In general, the lower the perplexity, the better the model performs for the dataset. In this study, however, perplexity is not used directly to evaluate the performance of the three clustering methods. Instead, it is used as a guide to select suitable values for DPMM parameters, namely, alpha and beta, for clustering tweets.

The perplexity equation used in this study is listed below:

$$Perplexity(T_{unseen}) = \exp\left\{-\sum_{t\in T_{unseen}} \log p(w_t) \Big/ \sum_{t\in T_{unseen}} N_t\right\} \quad (3)$$

where T_{unseen} represents the unseen tweets, $p(w_t)$ is the probability of generating all the words in an unseen tweet $t \in T_{unseen}$, and N_t denotes the number of words in tweet t.

4.4.2. Term recall and precision

We have considered two aspects of topic evaluation in this study. The first is to consider the relevancy of terms that are extracted by the topics, and the second is to assess the topics identified. Term-based evaluation is covered in this section while topic-based assessment is explained in the next section. A term-based topic detection evaluation metric proposed by Aiello et al. (2013) uses term recall (the number of correct terms over the total number of terms in the ground truth topics) and term precision (the number of correct terms in the detected topics over the total number of terms in these detected topics).

We have constructed a ground truth dataset for the selected days from news headlines reported by main stream media (see Section 3.3). In order to standardise the term extraction process and eliminate any manual annotation effort, the ground truth dataset was first categorised by the OpenCalais web service[7] to extract topics. Relevant terms of each topic that fall within the top 30% of their TFIDF scores were used as terms representing the ground truth topic. A list of ground truth terms was then generated from the combination of terms from all the topics.

In our study, the calculation of term recall uses content of matching tweets. This is to ensure none of the potential relevant terms is being filtered by any of the term ranking methods. To construct this set of exemplar tweets for matching, the top terms generated by each ranking method are used to extract tweets from the matching topics generated by each clustering method. As a result, nine sets of exemplar tweets were generated. They are TLDA_Peak, TLDA_TFIDF, TLDA_TF, KMeans_Peak, KMeans_TFIDF, KMeans_TF, DPMM_Peak, DPMM_TFIDF and DPMM_TF. The equation for term recall is as follows:

$$term\_recall = matched\_count/size\ of(ground\_truth\_terms) \quad (4)$$

where *matched_count* is a match of any ground truth terms in the exemplar tweets and *ground_truth_terms* is the list of combined terms from all the ground truth topics.

Term precision uses representative terms extracted from the exemplar tweets. This set of representative terms was generated using TF by taking top 70% of the terms and considering only terms with frequencies more than five. These terms are named as exemplar tweet terms. The equation for term precision is as follows:

$$term\_precision = matched\_count/size\ of(exemplar\_tweets\_terms) \quad (5)$$

where *matched_count* is a match of any ground truth terms with exemplar tweet terms and *exemplar_tweets_terms* are the representative terms of an exemplar tweet set.

4.4.3. Topic recall and precision@10

Aiello et al. (2013) also proposed a topic recall evaluation metric and defined it as the percentage of topics successfully retrieved. Topic precision was not used, as it was found that not all the topics covered by Twitter would appear in main stream media. The datasets collected by Aiello et al. (2013) correspond to two distinct events in the U.S., and the topics curated were manually annotated. This is in contrast to our datasets as we focus on analysing tweets of selected candidate days, which most likely would contain many diverse topics. Besides that, since we are exploring an unsupervised approach with minimal annotation effort, we have adopted the OpenCalais web service to automatically categorise the content

[7] https://www.opencalais.com/open-calais-api/.

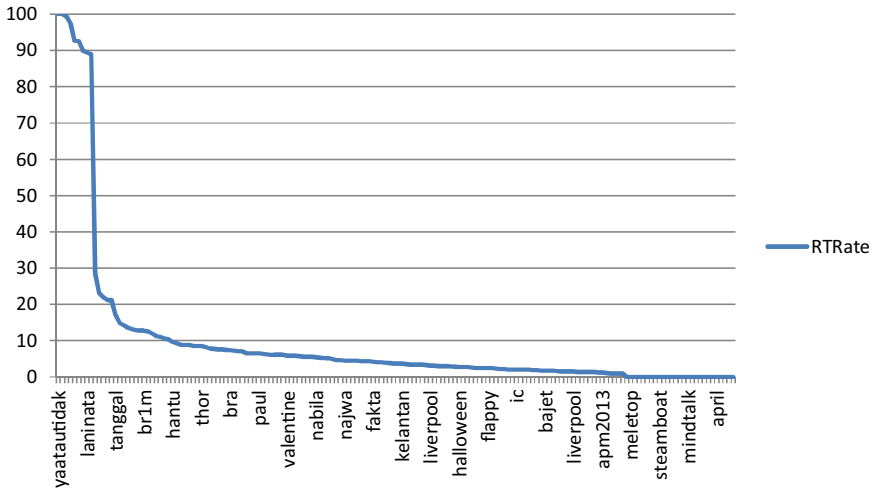

Fig. 4. Distribution of RTRate with respect to terms.

of tweets based on its internal classifications. As a result, the topic recall we used in this study is an adapted version that considers the percentage of cluster topics that fall within the categories defined by OpenCalais.

Given that it is not feasible to do topic precision, we consider precision@10, which is the correct number of topics found within the top-10 topics. This metric is applicable in our study since we are looking at high-value topics. The list of topics is ranked by each of the term ranking methods before calculating the topic precision@10 value.

4.5. Multilingual sentiment analysis

Even though this study focuses on identifying high-value topics from Twitter, it is of interest to further analyse the sentiments of topics discovered so that useful insights or feedback can be gathered. Due to the mixed language nature of our datasets, we carried out multilingual sentiment analysis through the use of a polarity detection algorithm (Lo et al., 2016a). Briefly, this algorithm incorporates a few knowledge-based lexicons, which include the Singlish dictionary constructed in Section 3.2, and n-gram Singlish sentic patterns together with Malay and English polarity lexicons. Besides that, sentic patterns such as negation, adversative (e.g., "but") and Twitter's retweet structure are integrated in the sentiment analysis process for multilingual polarity detection.

5. Results

In this section, we first describe the results of candidate days identified and their top matched terms before going into the details of ground truth data analysis. After which, we discuss the results of DPMM parameter selection and evaluate the three term ranking and three topic clustering methods using recall and precision of terms and topics. Lastly, we present the findings of multilingual sentiment analysis on the top topic of each candidate day.

5.1. The list of matched terms and candidate days

As per Section 4.2, ranked candidate terms generated through the Twitter and Singlish datasets were compared and matched. Further analysis was done on the matched terms and it was observed that some of the terms were retweeted many times without carrying much meaning. A likely cause of this is the use of a username without using the convention of @username. This would not be captured in the pre-processing step mentioned earlier.

Consequently, we have to consider Twitter's retweet and discussion rates before identifying candidate days. The retweet rate (RTRate) is calculated as the percentage of tweets with RT found in them, including the term of interest. The discussion rate measures the number of users sharing tweets containing the term. In this study, an inverse discussion rate (InversedDiscussionRate) was used, calculated from the number of tweets containing the term divided by the number of users. This means the bigger the InversedDiscussionRate's value becomes, the smaller the number of users found sharing tweets containing the term is. The distribution of RTRate can be found in Fig. 4, and the threshold is set at 20. Similarly, a distribution for the InversedDiscussionRate is presented in Fig. 5. From Fig. 5, terms with InversedDiscussionRate values less than 20 are taken into consideration since 20 is a reasonable cut-off point when overlaid with the RTRate. The terms filtered off are omitted from the selection of top-10 candidate days.

Table 2 shows the top-10 candidate days with their matched terms and consolidated term frequency. From the table, we can see that the matched terms can be used to identify days of interest. Some of the terms are easily understandable, as they represent international or important local incidents, such as the death of Paul Walker due to an accident on 1-Dec-2013,[8] the death of Nelson Mandela on 6-Dec-2013,[9] and the 2014 Singapore budget announcement by Minister of Finance Mr Tharman on 21-Feb-2014.[10] There are also days that are of celebration or festive nature, such as 14-Feb-2014 the Valentine's day (with terms - valentine, valentines); 31-Jan-2014 (with terms - cny, gong, xi, cai), which was the first day of Chinese New Year in 2014; and 15-Oct-2013 (with terms – adha, lembu, korban, kambing, rendang), which was the first day of Hari Raya or New Year Celebration for the Malays in 2013.

However, mixed language terms extracted that may not make sense include "emazing" and "thor" on 9-Nov-2013, "bangla" on 9-Dec-2013, "nabila", "hantu" and "seram" on 18-Nov-2013, "pmr" on 1-Oct-2013, and "pesawat" on 8-Mar-2014. The term "emazing" is the name of an award given to a world-wide performer through voting, of which the winner is crowned as the most emazing star.

[8] http://www.nytimes.com/2013/12/02/movies/paul-walker-screen-actor-is-dead-at-40.html.

[9] http://edition.cnn.com/2013/12/05/world/africa/nelson-mandela/.

[10] http://www.straitstimes.com/singapore/singapore-budget-2014-relief-for-the-elderly-cpf-boost-for-all-workers.

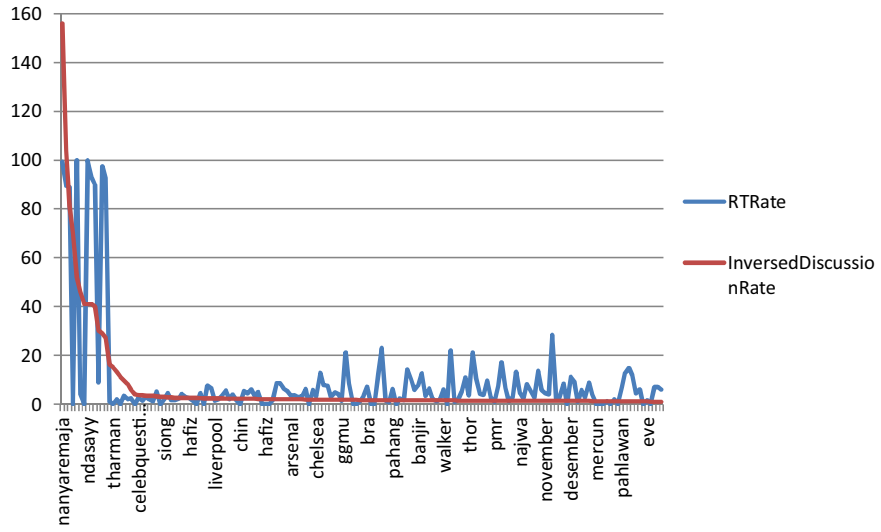

Fig. 5. Distribution of InversedDiscussionRate with respect to terms and overlaying with the RTRate.

Table 2
Top-10 candidate days with their matched terms and consolidated term frequency.

Candidate day	Matched terms *	Frequency
1-Dec-2013	paul, walker, december, rip, desember (December)	2176
21-Feb-2014	sgbudget, budget, tharman	927
14-Feb-2014	valentine, valentines	920
9-Nov-2013	emazing, thor	850
31-Jan-2014	cny (Chinese New Year), gong, xi, cai	799
9-Dec-2013	riot, bangla	763
18-Nov-2013	villa, nabila, hantu (ghost), seram (scream)	751
6-Dec-2013	mandela, nelson	740
15-Oct-2013	adha, lembu (cow), korban (sacrifice), kambing (goat), rendang (Malay dry curry)	717
1-Oct-2013	pmr, october, oktober	615
8-Mar-2014	prayformh370, plane, pesawat (plane)	593

*translated English words can be found in the brackets

Netizens were buzzing on 9-Nov-2013 to garner votes for EXO, a favourite k-pop boy band, for the award.[11] "thor" is the name of a popular action movie, and it captured much attention at that time.[12] "bangla" is a Singlish expression of construction workers, and on the night of 8-Dec-2014[13] a riot broke out in the Little India district of Singapore involving these workers. It was initially a surprise to see terms like "nabila", "hantu" (ghost) and "seram" on 18-Nov-2013. After verifying with the news source,[14] it was found that the terms refer to the tale of a creepy, abandoned place named Villa Nabila in Johor, a state of the neighbouring country Malaysia. Many netizens shared their stories about the urban legend, but most of the stories were believed to be rumours. The term "pmr" found on 1-Oct-2013 refers to the public examination of lower secondary schools in Malaysia that was scheduled to start on 2-Oct-2013.[15] Finally, the term "pesawat" means planes in Malay, and it was used in relation to the disappearance of the ill-fated Malaysian Airline flight MH370[16] on 8-Mar-2014.

It is worth noting that although the riot broke out on 8-Dec-2013, the term "riot" was found only on 9-Dec-2013. Besides that, the disappearance of MH370 on 8-Mar-2014 had created a myriad of controversies. Consequently, these three days have been selected for further analysis to better understand the topics found based on their top ranking terms. The ranking results can be found in Table 3.

Taking a close look at the table, it has become clear that the term "riot" was not identified by the Peak Identification algorithm as a top term, although it was found as the top term in the Singlish dataset on 8-Dec-2013 (not shown in the table). A detailed analysis of the top three peaks on the term "riot" discovered by Peak Identification can be found in Fig. 6. As shown in the figure, the term only has one significant peak on 9-Dec-2013, followed by not-so-significant ones on 17-Dec-2013 and 20-Dec-2013, in the Twitter dataset. However, this term is consistently found across the three days from 8 to 10-Dec-2013 in the Singlish dataset. This explained why "riot" is not found among the matched terms on 8-Dec-2013 and consequently not identified as a potential candidate day in the Twitter dataset on that day. Another reason could be that, as the riot happened at night, the buzz on social media was not captured as strongly on 8-Dec-2013 compared to 9-Dec-2013.

It is clear from Table 3 that the top topics on 9-Dec-2013 and 8-Mar-2014 are the Little India riot and the disappearance of MH370, respectively. While all three ranking methods managed to identify the relevant terms as top terms, Peak Identification found more relevant terms compared to TFIDF and TF. TFIDF found more hashtags than others while TF identified more generic terms such as "love" and ":)" . Even though the quality of TF could be improved by including these generic terms in the stop word list (and thus

[11] http://omonatheydidnt.livejournal.com/12083463.html.
[12] http://www.boxofficemojo.com/intl/singapore/yearly/?yr=2013&p=.htm.
[13] http://www.theguardian.com/world/2013/dec/09/singapore-riots-decades-migrant-workers.
[14] http://www.thestar.com.my/news/nation/2013/11/18/villa-nabila-tales-from-the-net/.
[15] http://kerajaanrakyat.blogspot.sg/2013/06/jadual-waktu-peperiksaan-menengah.html.
[16] http://www.theguardian.com/world/malaysia-airlines-flight-mh370.

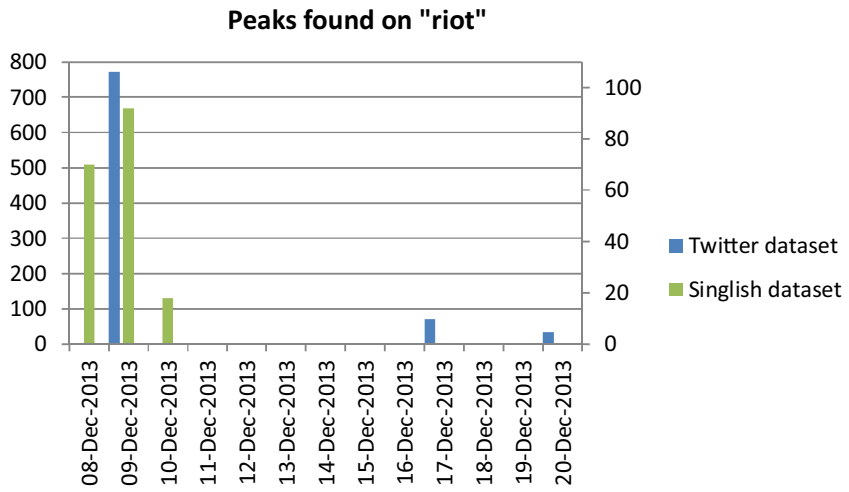

Fig. 6. Top three peaks identified on "riot" by Peak Identification in both the Twitter and Singlish datasets. Due to the different volumes of the two datasets, a dual-axes chart is used. The left axis is for the Twitter dataset and the right axis is for the Singlish dataset.

Table 3
Results of top 10 terms of each ranking method based on the Twitter dataset.

Date/Ranking method	Top terms
8-Dec-2013	
Peak Identification	aiman, md, muhammad, ujian, max5imum, remajaindonesia, laos, prata, kavenyou, sek
TFIDF	filmviqhrilshippuden, wewantchertobringherswagtoitaly, max5imum, boysrepublicinmy, skuadrasoontobe, riot, mandela, kavenyou, assaidi, laos
TF	singapore, love, riot,:), happy, india, :(, orang, night, omg
9-Dec-2013	
Peak Identification	riot, littleindiariot, riots, everton, bangla, officers, arrested, sgriot, december, injured
TFIDF	littleindiariot, sgriot, viqhrilforharuka, riot, riots, rioting, littleindiariots, sokangin, nodahitamhariantikorupsi, paranoidah
TF	riot, singapore, india, love, littleindiariot, police,:), orang, work, omg
8-Mar-2014	
Peak Identification	prayfromh370, plane, airlines, malaysiaairlines, vietnam, passenegers, crashed, aircraft, pesawat, berita
TFIDF	mh370, prayformh370, malaysiaairlines, fthxsg, swc3seoul, 26thtaeyeonday, ilightmarinabay, zed, minswagday, missingmh370
TF	mh370, prayformh370, singapore, malaysia, love, missing, plane, flight, happy,:)

omitting them from being analysed), some of these terms play a critical role in sentiment analysis and hence they are kept for understanding the polarity of the content in Section 5.6.

From the results in this section, it is clear that multilingual analysis taking localised languages into consideration plays a pivotal role in identifying topics of interest within the local community. Omitting words like "bangla" or "pesawat" may mislead our findings.

5.2. Ground truth dataset analysis

To minimise annotation effort, ground truth data for the three candidate days was categorised by the OpenCalais web service. The list of categories found is presented in Table 4. From Table 4, we can see that both 9-Dec-2013 and 8-Mar-2014 have a dominant category in "war_conflict" and "disaster_accident", respectively. However, this is not found on 8-Dec-2013 – the categories found on the day are heterogeneous with no distinct category. This may explain the diverse terms found on 8-Dec-2013 in Table 3.

5.3. Results of DPMM parameter optimization

We tested a range of values in order to find the optimal settings for both alpha and beta of the DPMM. The default value of 1 was used for beta while tuning the alpha parameter. Fig. 7 shows the perplexity generated from different alpha values. From the figure, it is obvious that different alpha values have little impact on the perplexity. Additional tests with the same range of values were also done on data from different days and similar results were observed.

Likewise, we fixed alpha to 1 in our attempt to find the optimal value for beta. A range of beta values were used, and the results can be found in Fig. 8. In contrast to Fig. 7, different beta values have noticeable influences on both the perplexity and topic numbers. Fig. 8 clearly shows that the lower the beta value, the lower the perplexity but the higher the topic numbers. Since the dataset used consists of tweets gathered from a single day, it is unreasonable to consider topic numbers with too high a number. Based on the ground truth dataset, it is reasonable to consider a topic number of around 20 topics. As a result, a beta value of 0.1 that generated 25 topic numbers has been adopted in this study. As shown in Fig. 7, the perplexity generated by a wide range of alpha values is consistent with respect to different beta values. Hence, it can be concluded that the value of alpha does not directly affect the perplexity. All subsequent experiments in this study set alpha at 1 and beta at 0.1.

5.4. Evaluation via term recall and precision

Since there are four topic numbers selected for both K-Means and Twitter LDA, clustering tasks were run for each topic number on the three selected candidate days – 8-Dec-2013, 9-Dec-2013 and 8-Mar 2014 – to choose a best performing model to represent the two clustering methods. The results of term recall and precision for the four topic numbers can be found in Figs. 9 and 10, respectively. From the figures, we observe that, in general, K-Means clustering has a higher term recall value but scores lower on term precision. While the results of term recall generated by K-Means clustering are dependent on the topic numbers and nature of data from different candidate days, K-Means clustering typically performs better with smaller topic numbers. In contrast, the values of term precision remain low for all the topic numbers and three candidate

Table 4
OpenCalais results for the three candidate days.

8-Dec-2013 (249 records)	**9-Dec-2013** (566 records)	**8-Mar-2014** (363 records)
politics 70	war_conflict 207	disaster_accident 133
sports 57	social issues 146	politics 69
social issues 50	politics 125	environment 69
war_conflict 40	disaster_accident 94	business_finance 57
human interest 38	business_finance 80	human interest 50
entertainment_culture 30	law_crime 78	law_crime 47
disaster_accident 29	environment 65	hospitality_recreation 43
hospitality_recreation 27	human interest 64	sports 35
business_finance 25	religion_belief 58	social issues 34
environment 25	sports 53	war_conflict 32
religion_belief 23	hospitality_recreation 53	health_medical_pharma 24
law_crime 20	health_medical_pharma 41	entertainment_culture 23
technology_internet 18	entertainment_culture 41	technology_internet 17
education 11	labor 38	religion_belief 15
weather 9	technology_internet 31	labor 14
health_medical_pharma 8	weather 13	education 10
labor 4	education 11	weather 7
other 1	other 5	other 3

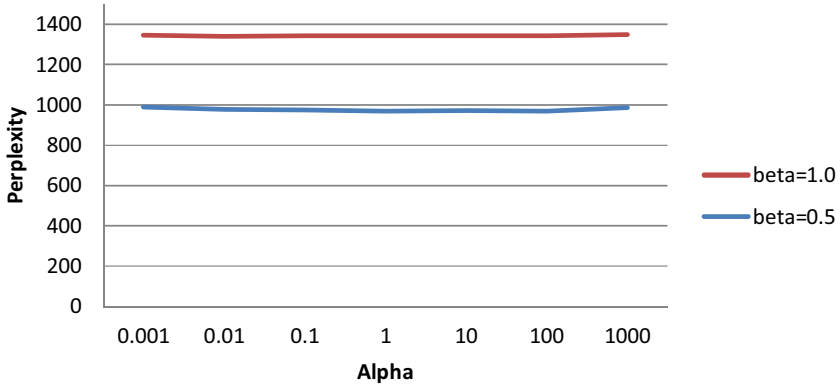

Fig. 7. The perplexity value generated by a range of alpha values.

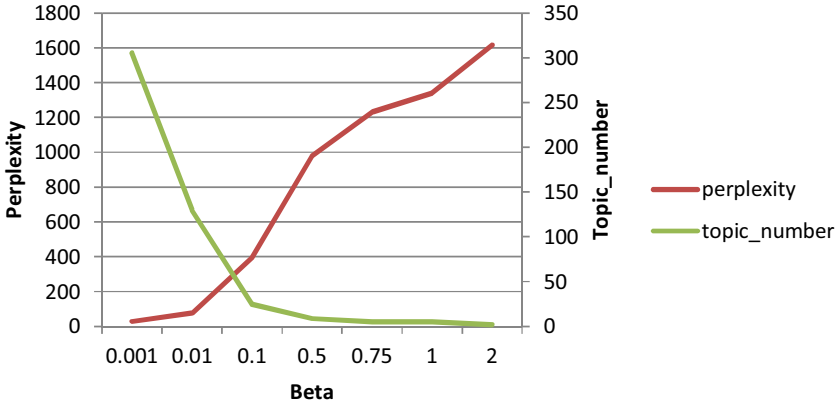

Fig. 8. The perplexity value and topic numbers generated by a range of beta values.

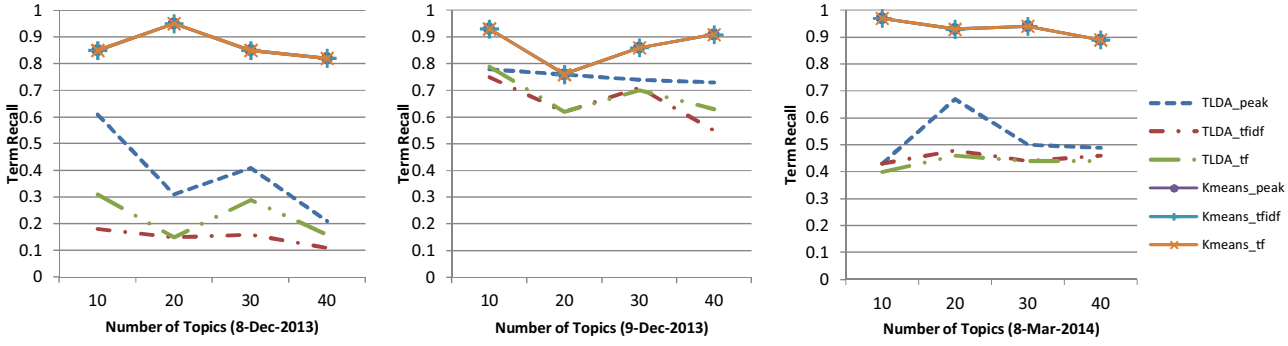

Fig. 9. Results of term recall for Twitter LDA and K-Means clustering methods.

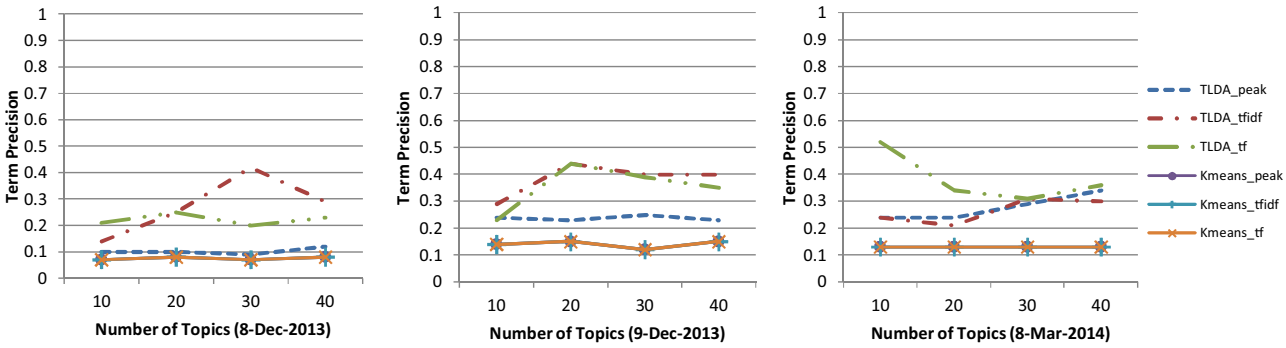

Fig. 10. Results of term precision for Twitter LDA and K-Means clustering methods.

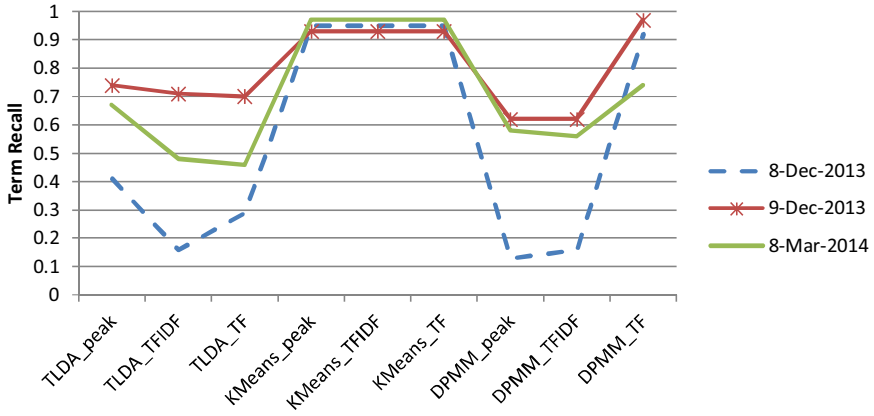

Fig. 11. Results of term recall for the three clustering methods.

days. As a result, the K-Means model of topic number 20 has been chosen for 8-Dec-2013 while topic number 10 has been selected for both 9-Dec-2013 and 8-Mar-2014.

The results are quite different for Twitter LDA. Both term recall and precision values are heavily dependent on topic numbers and data from the candidate days. There is no common trend observed, although Twitter LDA has a better term recall value with smaller topic numbers and a higher term precision value with higher topic numbers for some of the data. Since there is a trade-off between term recall and precision values, the following Twitter LDA models have been chosen considering both term recall and precision results: Twitter LDA with 30 topic numbers for 8 and 9-Dec-2013 whereas Twitter LDA with 20 topic numbers for 8-Mar-2014. These models together with the selected K-Means models are used to compare with the results of DPMM in Figs. 11 and 12.

It can be seen from Fig. 11 that K-Means clustering has the highest term recall value, but the figure also shows that there is no difference between the term ranking methods when used with K-Means. However, the term ranking methods do impact directly on the performances of Twitter LDA and DPMM clustering. Peak Identification achieves the best result with Twitter LDA, while TF performs better with the DPMM.

Fig. 12 shows a different view with K-Means scoring low term precision values with all the three term ranking methods. The DPMM achieves best term precision values for all three candidate days, with Peak Identification performing the best among the ranking methods. TFIDF appears to have better term precision values with Twitter LDA compared to other ranking methods.

The diversity of topics found on 8-Dec-2013 (as analysed in Section 5.2) is likely to be the reason causing the lower recall and precision values in Figs. 11 and 12 for all the clustering methods. In comparison, DPMM clustering is the best performing method for the other two candidate dates. This is encouraging since there is no pre-defined topic number required for the DPMM and hence, this approach can be considered truly unsupervised without too much human intervention in selecting a suitable topic number.

5.5. Evaluation via topic recall and precision@10

Since topic recall calculation is done using categories defined by the OpenCalais web service, the recall value is calculated based on the number of topics found in the 18 categories listed in Table 4. The precision@10 value, on the other hand, is the correct number of topics found within the top-10 topics. Table 5 shows the results of topic recall and precision@10 for the three clustering methods together with the three ranking methods.

As shown in Table 5, the results of topic recall and precision@10 are highly dependent on the term ranking methods and candidate days. In view that there is a need to select a representative clustering method for the subsequent multilingual sentiment analysis, we propose a 'Joint' ranking method that combines the top terms found by the three ranking methods for a more direct comparison on the three clustering methods. Figs. 13 and 14 present the results of the topic recall and precision@10 for the three clustering methods together with the three ranking methods and the 'Joint' method.

Fig. 13 clearly shows that the content of candidate days plays a direct role in the results of topic recall. It is understandable that candidate days 9-Dec-2013 and 8-Mar-2014 have better topic recall, with both days having a dominant topic, as reported in main stream media (see Table 4). However, it is not the same for 8-Dec-2013, since there is no focused topic found. This resulted in a lower topic recall value. Also, Peak Identification performs the worst on

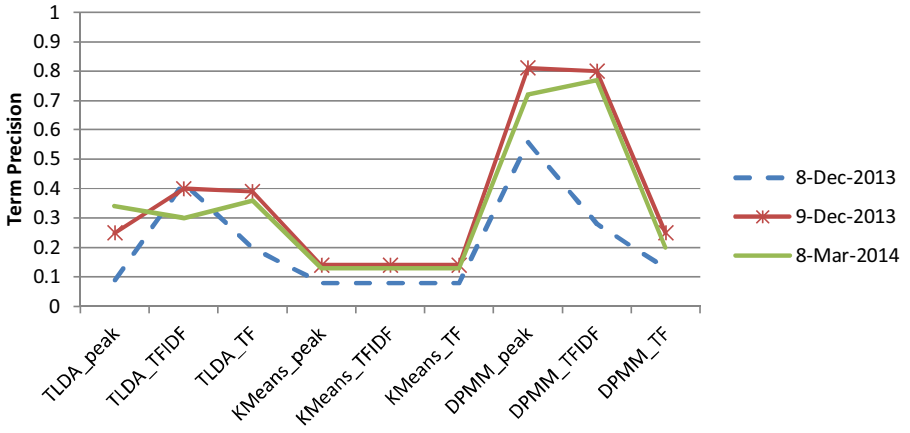

Fig. 12. Results of term precision for the three clustering methods.

Table 5
Results of topic recall and precision@10.

	8-Dec-2013		9-Dec-2013		8-Mar-2014	
	Precision@10	Topic Recall	Precision@10	Topic Recall	Precision@10	Topic Recall
TLDA_Peak	0/10	0/18	7/10	17/18	4/10	12/18
TLDA_TFIDF	3/10	5/18	6/10	16/18	4/10	15/18
TLDA_TF	4/10	8/18	5/10	17/18	6/10	16/18
KMeans _Peak	0/10	0/18	4/10	15/18	3/10	11/18
KMeans _TFIDF	4/10	5/18	5/10	16/18	2/10	11/18
KMeans_TF	2/10	7/18	4/10	17/18	2/10	13/18
DMM_Peak	1/10	1/18	7/10	17/18	2/10	12/18
DMM_TFIDF	3/10	3/18	6/10	16/18	4/10	18/18
DMM_TF	5/10	11/18	5/10	17/18	7/10	16/18

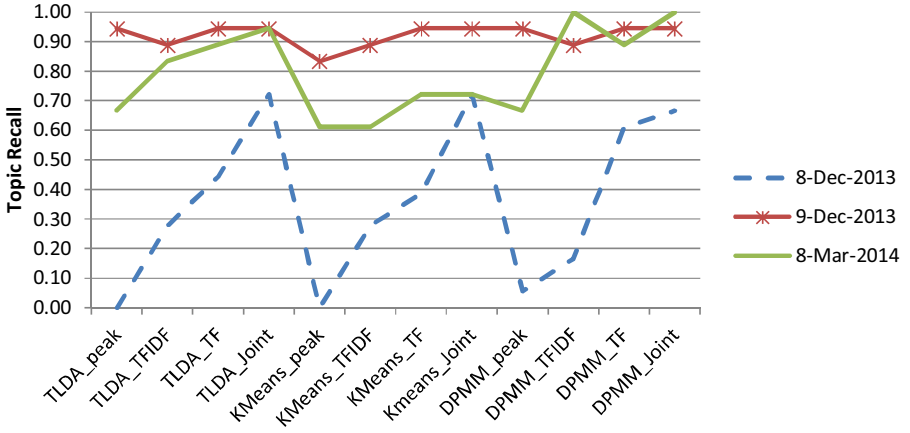

Fig. 13. Results of topic recall for the three clustering methods.

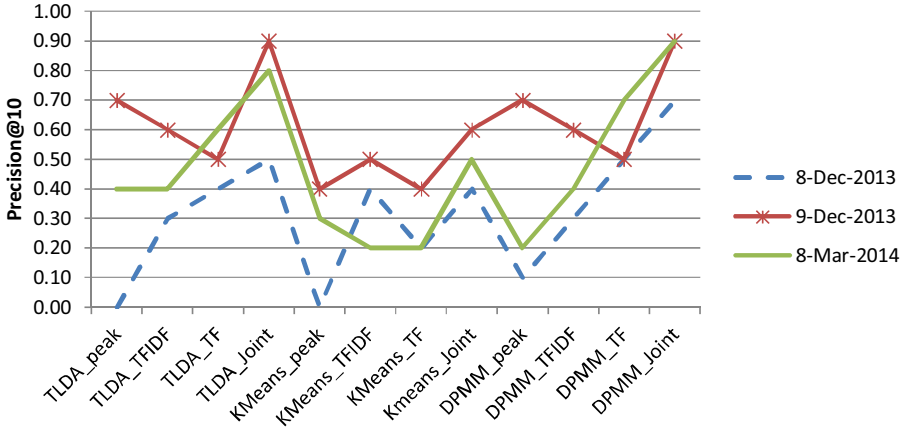

Fig. 14. Results of precision@10 for the three clustering methods.

Table 6
Results of multilingual and English sentiment analysis.

Date	Cluster no: Top five terms	Multilingual	English
8-Dec-2013	1: tired, kavenyou, boysrepublicinmy, liverpool, rebecca,	Positive: 34% Negative: 27%	Positive: 29% Negative: 27%
9-Dec-2013	0: riot, littleindiariot, riots, injured, arrested	Positive: 20% Negative: 27%	Positive: 18% Negative: 27%
8-Mar-2014	2: mh370, missing, plane, prayformh370, malaysiaairlines	Positive: 27% Negative: 33%	Positive: 9% Negative: 20%

8-Dec-2013, although it is one of the better ranking methods for 9-Dec-2013. This observation suggests that Peak Identification is only useful in picking up relevant topics in a dataset with a dominant topic. TFIDF is the best scoring method with the DPMM on 8-Mar-2014. Even though TF is one of the simplest ranking methods, it has consistently scored well in terms of topic recall. As expected, the 'Joint' method achieves the best results for all the three clustering methods. This is not surprising, since it combines the top terms of the three ranking methods and hence contains the most comprehensive term list for identifying relevant topics. This finding is important since the results suggest that future analysis should use the 'Joint' method instead of trying individual ranking methods.

The precision@10 results presented in Fig. 14 indicate that there is no obvious trend or regular pattern for the ranking methods. While the content of different candidate days does influence the precision@10 results to some extent, it is not as clear-cut as in Fig. 13. Selecting a distinctively best performing ranking method is thus not possible, as each of the three methods performs differently on different candidate days. The only consistent observation is the best results achieved by the 'Joint' method.

From Figs. 13 and 14, we see that both Twitter LDA and the DPMM perform better than K-Means. Although the DPMM has a better showing in topic recall, it is not clear which is the preferred clustering method in terms of precision@10. Considering that the DPMM's topic number is determined by the content of a dataset and not from a pre-defined number, plus the better results observed in Section 5.4, we have decided to use the DPMM for the subsequent multilingual sentiment analysis.

5.6. Results of multilingual sentiment analysis

For comparison purposes, tweets from the top topics identified by the DPMM on the three selected candidate days were analysed using a polarity detection algorithm, in both multilingual and English versions. The percentage of positive and negative results together with the top five terms can be found in Table 6.

From the table, it is obvious that the sentiment is negative for candidate days 9-Dec-2013 and 8-Mar-2014. However, there are also positive records found in the results. A further analysis showed that there were tweets praising the police (e.g., "*Singapore Police, SCDF draw praise for handling of Little India riot*") and asking for reconciliation (e.g., "*Following #sgriot, netizen rallying pple to go #LittleIndia tonight to hand out flowers for peace & reconciliation*") on 9-Dec-2013. While most of the tweets on 9-Dec-2013 were about the Little India riot, the same cannot be said of 8-Mar-2014. Measuring the lexical diversity of words in the processed tweets of 9-Dec-2013 produces a score of 0.178, while a higher diversity score of 0.205 is found for 8-Mar-2014. This implies that the content on 8-Mar-2014 is more heterogeneous and diverse, which may explain the lower topic recall value in Fig. 13. These lexical diversity scores were calculated by taking unique tokens (i.e., words) of the text divided by the total number of tokens (Russell, 2013). Even though there were positive tweets from 8-Mar-2014, those related to MH370 were mostly about praying for its safety (e.g., "*Hope everything is doing well, selamatkan lah mereka ya Allah. #PrayForMH370*" (English: save them please God)).

The top five terms on 8-Dec-2013 in Table 6 indicate that the Little India riot was not identified. Instead, a mixture of topics including entertainment (e.g., "*Rebecca Black's 'Saturday' is an epic fail. Just like the previous song, 'Friday', They said Malaysian fans are very passionate! #BoysRepublicinMY #KAvenyou*") and soccer (e.g., "*A good treat for Liverpool fans:)*") were extracted. A detailed analysis showed that the Little India riot was listed at the top-3 topic. This is in line with the finding observed in Section 5.1 – the Little India riot did not buzz on 8 Dec 2013 since it happened at night.

Comparing the percentages of tweets' sentiments, those detected by the English polarity detection algorithm are clearly lower compared to those detected using the multilingual polarity detection algorithm. It is thus of interest to analyse these tweets to better understand the detection process. Table 7 lists some random samples found by the multilingual polarity detection algorithm but not found by the corresponding English version. From the table, it is understandable why these tweets were not detected by the English polarity detection algorithm, as most of the tweets contain either Singlish or Malay polarity terms. To sum up, multilingual sentiment analysis is essential in this case due to the mixture of languages used in tweets we focused on.

6. Discussion

Previous studies along this line of research (e.g., Aiello et al., 2013; Becker et al., 2011; Shamma et al., 2011) used annotated and curated datasets for evaluation. It is undeniable that if strict annotation rules are enforced in the datasets employed, accurate and domain-related topics can be discovered and identified. In this study, we avoided manual annotation effort in two aspects: domain selection and topic annotation. Most studies pre-defined specific events (Aiello et al., 2013; Elbagoury et al., 2015) in domain selection but in this study, we showed that the mixed-language content of tweets can be a treasure trove to extract relevant topics that are meaningful to a local community. Specifically, tweets with localised languages such as Singlish can be leveraged for selecting important or meaningful events from a vast amount of tweets through candidate day selection. As for topic annotation, instead of manually assigning a category to a huge amount of tweets (which is not feasible in practice), we used the OpenCalais web service to categorise news headlines from main stream media. This serves as a guide in understanding the content shared on Twitter.

Even though the previous study by Zhao et al. (2011) based on news from the U.S. found that Twitter and traditional media both cover the same topics, our observation in this study does not appear to validate such a view. While major news could easily be found on both platforms, content in tweets typically consists of more opinions and views on events or incidents that are of interest or concern to the local community. It can be argued that the Twitter dataset we collected is not representative of the whole

Table 7
Samples of random tweets detected by the multilingual polarity detection algorithm but not by the corresponding English-based polarity detection algorithm.

8-Dec-2013	Sample tweets (remarks)
positive	• "*Homed..great lepak session..*" (English: great relaxing session) • "*Suarez seriously bagus seyh*" (English: seriously good)
negative	• "*back.from.camp.tired.shit.*" • "*Raining non-stop in the East today. Feeling meh.*" (English: feeling terrible)
9-Dec-2013	
positive	• "*Aiyah this little india riot is sooner or later one.*" • "*SIAL. WAHHHH. Haha RT @nyingqi: Police cannot chase them, cause bangala-dash*" (A joke on the workers)
negative	• "*Wah rabakliao the riot in little india*" (English: The riot in little india has gone crazy) • "*Ambulance terbakar?*" (English: Is the ambulance burnt?)
8-Mar-2014	
positive	• "'*Wangi tak?' hahahaha haha ha ha:3*" (English: Is it a nice smell?) • "*Tweet nak doa for the people dalam flight #MH370 tu acah je. Action speaks louder than words. Pray for them in your doa, not in twitter.*" (English: Tweet to pray for people in Flight #MH370 is just for show. Pray for them in your prayer, not in twitter)
negative	• "*Berita hari ni,membuatkan aku takut nak naik kapal terbang.*" (English: After today's news, I am afraid to take aeroplane) • "*Ya Allah harap diorang selamat:(#PrayForMH370*" (English: Dear God, hope the people are safe)

platform. However, our results clearly show that interesting observations and relevant topics can be identified via our unsupervised approach. It is always beneficial to be able to uncover topics or incidents shared online, in order to better gauge the sentiment on the ground.

In this study, we have extracted tweets from a period of time for topic identification analysis. Although it can be considered as a corpus-based topic detection and tracking problem, the term ranking methods tested have the potential to be extended to real-time tracking by using thresholds learned from collected TF, TFIDF or peak values. Besides that, as shown in Figs. 13 and 14, the 'Joint' ranking method that combines top terms identified by the three ranking methods consistently performs best in topic recall and precision@10. We suggest adopting the 'Joint' ranking method to leverage the strengths of different ranking mechanisms to achieve a more accurate topic identification outcome.

Even though the Peak Identification algorithm does show promises with its simplicity (especially in selecting the candidate days), it is important to highlight its limitations too. In particular, it did not detect the term "littleindiariot" as one of the top three peaks on 8-Dec-2013. The main reason being that the period selected for peak generation is critical to the accuracy of this ranking method. Shamma et al. (2011) used a top peak term of a time period and compared it to the rest of the corpus to find interesting topics. If an interesting topic is discussed at several different times, their approach could miss the topic as the term may be found in abundant in the whole corpus. Even though we used the top three peaks in our study, we may face a similar challenge. A possible remedy is to consider a smaller time slot (e.g., hours instead of days) and compare each slot to the immediate previous slot for better accuracy.

From Figs. 11 and 12, term ranking methods have no impact on K-Means' term recall and precision results. Further analysis revealed that it was because K-Means grouped most of the relevant tweets in its largest cluster and as a result, none of the ranking methods has any impact on the performance. This may be partly due to the sparseness of feature vectors created by the huge amount of unique terms found in tweets. In view of the results, it may not be feasible to use K-Means as a clustering method when analysing tweets.

One of the limitations or critiques of the DPMM is that it does not consider the order and context of words in a sentence for clustering. While it is undeniable that the order and context of words are extremely important for deciphering the content of a sentence, the DPMM seems to work well for tweets, especially in a multilingual environment where understanding the context and concept is limited by the available resources. Nonetheless, bigram or n-gram terms can be considered for future work, in order to preserve the word order. In addition, even though the assumption of the DPMM in assigning each document to one topic cluster can be a disadvantage for a document with many latent topics, it fits in well with tweets, where most of them usually contain one topic. It is important to highlight that while a pre-defined topic number is not required for the DPMM, parameter selection is critical since using different values in a parameter, especially the beta parameter (see Fig. 8) can lead to very different results. It is suggested to test run the DPMM in order to find suitable parameters. In our study, we have found that setting the alpha value at 1 and the beta value at 0.1 works well in identifying high-value topics.

Although the top topics discovered by the DPMM and 'Joint' ranking method on 9-Dec-2013 and 8-Mar-2014 are consistent with the ground truth dataset (see Table 6), the top topic of 8-Dec-2013 cannot be found in the main stream media. This is mainly because Twitter is used by many to share not just breaking news but also concerns and interests of a community (e.g., entertainment news and soccer as mentioned in Section 5.6). In view of the findings, it is suggested to analyse not only the top topics but also other topics in the top-10 ranking in order to gather a more complete picture on topics that are of concern to the local community. In fact, top-3 and top-4 topics of 8-Dec-2013 matched the ground truth dataset (e.g., top-3 on the Little India riot and top-4 on Nelson Mendela). Considering the more heterogeneous content found in such days, it is worthwhile to adopt techniques like n-gram analysis (Aiello et al., 2013), co-occurrence (Aiello et al., 2013) and the time-decay function (Psallidas et al., 2013) for more effective clustering on these top ranking topics. It is undeniable

that news buzz found on main stream media is often emphasised on social media sharing, especially when it is relevant to the local community, as can be seen with the cases of the Little India riot and MH370. However, it is observed that more homogeneous and relevant content was found on 9-Dec-2013 covering the Little India riot compared to that of the MH370 disaster (see Table 2 and Section 5.6). The key difference of the two topics is the Little India riot incident happened in Singapore and it resonated better with the local community. Hence it has a higher frequency value in Table 2 with more coherent sharing on the topic.

Multilingual sentiment analysis is gaining popularity (Lo, Cambria, Chiong, & Cornforth, 2016b). However, none of the previous multilingual sentiment analysis studies had taken into consideration the multiple languages found in a tweet. Instead, the focus is typically on studying the effects of different languages on a Twitter platform (Cui et al., 2011; Volkova et al., 2013) or leveraging available resources of one language for sentiment analysis of another language (Balahur & Perea-Ortega, 2015; Balahur & Turchi, 2013). Tables 6 and 7 show that multilingual analysis is essential when analysing social media content of a multi-cultural community. It is clear that there are tweets containing mixed languages that have not been detected by the English polarity detection algorithm. Omitting these tweets therefore may run the risk of not able to analyse sentiments more comprehensively.

7. Conclusion and future work

In this paper, we have demonstrated that it is feasible to identify high-value terms and topics from a vast amount of Twitter data through an unsupervised multilingual approach. We have also shown that by leveraging multilingual analysis and the Peak Identification algorithm, highly relevant topics that are of concern to the local community can be extracted through candidate day selection. From the observation of our results, the DPMM with our proposed 'Joint' ranking method has consistently performed well in selected candidate days. While the top topics of a candidate day with prominent events matched those of the main stream media, it may not be the case for other 'ordinary' candidate days. Thus, it is essential to identify and rank the topics so that the top few high-value topics can be further analysed in order to fully decipher the sentiments and opinions shared. Our approach is robust and has the potential to be adopted in a real-world application, as it does not rely on any external knowledge base for inferences to identify high-value topics. This is important considering the dynamism of social media content shared every day. The use of localised languages and unsupervised clustering can help to detect and identify topics that otherwise may go unnoticed within the vast amount of tweets.

Future directions of our research include building a comprehensive domain-specific, concept-level knowledge base that can be used for more accurate public policy sensing and multilingual sentiment analysis to understand concerns on the ground. This is essential to address the word sense disambiguation problem, so that the ambiguity of polarity of a word in different domains can be identified and assigned with the correct sentiment. For example, "crowded" should be given a 'negative' sentiment if it is used on bus capacity in the transport domain, but the same word could carry the opposite sentiment, i.e., positive, if it is found in describing a rally in the politics domain. Another area of interest is topic evolution based on localised languages to assess the potential of forecasting trending topics so that sufficient efforts can be engaged to mitigate negative sentiments. Finally, an issue that requires further investigation is on fake news or topic detection that is gaining popularity considering the impact of recent major events in the world (e.g., the 2016 U.S. election). The ability to identify the source of such news or topics plays a paramount role in determining the quality of future topic identification on social media.

References

Aiello, L. M., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., et al. (2013). Sensing trending topics in Twitter. *IEEE Transactions on Multimedia, 15*(6), 1268–1282.

Allan, J. (2012). *Topic detection and tracking: Event-based information organization*. Springer Science & Business Media.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics, 2*(6), 1152–1174.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of language resources and evaluation conference: Vol. 10* (pp. 2200–2204).

Balahur, A., & Perea-Ortega, J. M. (2015). Sentiment analysis system adaptation for multilingual processing: The case of tweets. *Information Processing & Management, 51*(4), 547–556.

Balahur, A., & Turchi, M. (2013). Improving sentiment analysis in Twitter using multilingual machine translated data. In *Proceedings of recent advances in natural language processing* (pp. 49–55).

Becker, H., Naaman, M., & Gravano, L. (2011). Beyond trending topics: Real-world event identification on Twitter. In *Proceedings of international conference on weblogs and social media: Vol. 11* (pp. 438–441).

Benhardus, J., & Kalita, J. (2013). Streaming trend detection in twitter. *International Journal of Web Based Communities, 9*(1), 122–139.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3*, 993–1022.

Bontcheva, K., & Rout, D. (2014). Making sense of social media streams through semantics: A survey. *Semantic Web, 5*(5), 373–403.

Cui, A., Zhang, M., Liu, Y., & Ma, S. (2011). Emotion tokens: Bridging the gap among multilingual twitter sentiment analysis. In *Proceedings of Asia information retrieval societies conference* (pp. 238–249).

Elbagoury, A., Ibrahim, R., Farahat, A. K., Kamel, M. S., & Karray, F. (2015). Exemplar-based topic detection in Twitter streams. In *Proceedings of AAAI conference on web and social media* (pp. 610–613).

Gao, Y., Zhou, B., Diao, Z., Sorensen, J., & Picheny, M. (2002). MARS: A statistical semantic parsing and generation-based multilingual automatic translation system. *Machine Translation, 17*(3), 185–212.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*, 5228–5235.

Hu, L., Li, J., Li, X., Shao, C., & Wang, X. (2015). TSDPMM: Incorporating prior topic knowledge into Dirichlet process mixture models for text clustering. In *Proceeding of the conference on empirical methods in natural language processing* (pp. 787–792).

Kim, S., Weber, I., Wei, L., & Oh, A. (2014). Sociolinguistic analysis of twitter in multilingual societies. In *Proceedings of the 25th ACM conference on hypertext and social media* (pp. 243–248). ACM.

Leimgruber, J. R. (2011). Singapore english. *Language and Linguistics Compass, 5*(1), 47–62.

Lo, S. L., Cambria, E., Chiong, R., & Cornforth, D. (2016a). A multilingual semi-supervised approach in deriving Singlish sentic patterns for polarity detection. *Knowledge-Based Systems, 105*, 236–247.

Lo, S. L., Cambria, E., Chiong, R., & Cornforth, D. (2016b). Multilingual sentiment analysis: From formal to informal and scarce resource languages. *Artificial Intelligence Review*. doi:10.1007/s10462-016-9508-4.

Lo, S. L., Chiong, R., & Cornforth, D. (2016). Ranking of high-value social audiences on Twitter. *Decision Support Systems, 85*, 34–48.

Lo, S. L., Cornforth, D., & Chiong, R. (2015). Identifying the high-value social audience from Twitter through text-mining methods. In *Proceedings of the Asia Pacific symposium on intelligent and evolutionary systems: Vol. 1* (pp. 325–339).

Lu, H.-M. (2015). Detecting short-term cyclical topic dynamics in the user-generated content and news. *Decision Support Systems, 70*, 1–14.

Luo, W., Stenger, B., Zhao, X., & Kim, T.-K. (2015). Automatic topic discovery for multi-object tracking. In *Proceedings of AAAI conference on artificial intelligence* (pp. 3820–3826).

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Berkeley symposium on mathematical statistics and probability: Vol. 1* (pp. 281–297).

Magdy, W., & Elsayed, T. (2016). Unsupervised adaptive microblog filtering for broad dynamic topics. *Information Processing & Management, 52*(4), 513–528.

Mitamura, T. (1999). Controlled language for multilingual machine translation. In *Proceedings of machine translation summit VII* (pp. 46–52).

Psallidas, F., Becker, H., Naaman, M., & Gravano, L. (2013). Effective event identification in social media. *IEEE Data Eng. Bull., 36*(3), 42–50.

Russell, M. A. (2013). *Mining the social web: Data mining Facebook, Twitter, LinkedIn, Google+, GitHub, and more*. O'Reilly Media, Inc.

Salton, G., & Buckley, C. (1997). Improving retrieval performance by relevance feedback. *Readings in Information Retrieval, 24*(5), 355–363.

Shamma, D. A., Kennedy, L., & Churchill, E. F. (2011). Peaks and persistence: Modeling the shape of microblog conversations. In *Proceedings of the ACM conference on computer supported cooperative work* (pp. 355–358).

Vavliakis, K. N., Symeonidis, A. L., & Mitkas, P. A. (2013). Event identification in web social media through named entity recognition and topic modeling. *Data & Knowledge Engineering, 88*, 1–24.

Vicient, C., & Moreno, A. (2015). Unsupervised topic discovery in micro-blogging networks. *Expert Systems with Applications, 42*(17), 6472–6485.

Volkova, S., Wilson, T., & Yarowsky, D. (2013). Exploring sentiment in social media: Bootstrapping subjectivity clues from multilingual twitter streams. In *Proceedings of annual meeting of the association of computational linguistics* (pp. 505–510).

Wang, C., Yuan, C., Wang, X., & Xue, W. (2011). Dirichlet process mixture models based topic identification for short text streams. In *Proceedings of international conference on natural language processing and knowledge engineering* (pp. 80–87).

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of conference on empirical methods in natural language processing* (pp. 347–354).

Yang, M.-C., & Rim, H.-C. (2014). Identifying interesting Twitter contents using topical analysis. *Expert Systems with Applications, 41*(9), 4330–4336.

Yin, J., & Wang, J. (2014). A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 233–242).

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., et al. (2011). Comparing twitter and traditional media using topic models. In *Advances in information retrieval* (pp. 338–349).

Zielinski, A., Bügel, U., Middleton, L., Middleton, S., Tokarchuk, L., Watson, K., et al. (2012). Multilingual analysis of twitter news in support of mass emergency events. In *Proceedings of European geosciences union general assembly conference: Vol. 14* (pp. 8085–8089).