The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0263-5577.htm

Gaining customer knowledge in low cost airlines through text mining

Bee Yee Liau and Pei Pei Tan
Department of Applied Statistics, University of Malaya, Kuala Lumpur, Malaysia

Abstract

Purpose – The purpose of this paper is to study the consumer opinion towards the low-cost airlines or low-cost carriers (LCCs) (these two terms are used interchangeably) industry in Malaysia to better understand consumers' needs and to provide better services. Sentiment analysis is undertaken in revealing current customers' satisfaction level towards low-cost airlines.

Design/methodology/approach – About 10,895 tweets (data collected for two and a half months) are analysed. Text mining techniques are used during data pre-processing and a mixture of statistical techniques are used to segment the customers' opinion.

Findings – The results with two different sentiment algorithms show that there is more positive than negative polarity across the different algorithms. Clustering results show that both K-Means and spherical K-Means algorithms delivered similar results and the four main topics that are discussed by the consumers on Twitter are customer service, LCCs tickets promotions, flight cancellations and delays and post-booking management.

Practical implications – Gaining knowledge of customer sentiments as well as improvements on the four main topics discussed in this study, i.e. customer service, LCCs tickets promotions, flight cancellations or delays and post-booking management will help LCCs to attract more customers and generate more profits.

Originality/value – This paper provides useful insights on customers' sentiments and opinions towards LCCs by utilizing social media information.

Keywords Customer relationship management, Malaysia, Clustering, Airlines, Sentiment analysis, Text mining

Paper type Research paper

1. Introduction

Low-cost carriers (LCCs) improves consumer welfare following the airline deregulation (refer to Borenstein, 1992 and the references therein). A LCC or commonly known as budget airline provides its services to the public at a relatively lower price with fewer comforts compared to the traditional airlines. In 1978, the USA implemented The Airline Deregulation Act, by removing government control over the entry of new airlines into market including their pricing and flight routes, and allowing airlines to reconfigure their flight routes to maximize the utilization of their capacity. This has resulted in the dramatic drop of airfares. With the increasing of middle class population in China, Southeast Asia and India, the demand for air travel has increased tremendously. According to the Boeing Company, the total number of airlines in Asia is estimated to increase by 65 per cent (14,750 in number) by year 2032, and nearly half of the world's air traffic growth will be driven by travels to, from or within the Asian region for the next 20 years (The Asia Foundation, 2014). Therefore, it is important for a low-cost airline to strengthen its customer relationship management (CRM) in sustaining its industrial competitiveness. Social media can aid in this aspect by providing useful information such as customer reviews quickly. In this 21st

Industrial Management & Data Systems
Vol. 114 No. 9, 2014
pp. 1344-1359

0263-5577
DOI 10.1108/IMDS-07-2014-0225

Information Age, social media is by far the most powerful form of marketing. For that reason, information communication technology service providers are changing strategies to attract customer and to reduce customer churn rate (see Lee *et al.*, 2013, 2014). There will be no exceptions for airlines industry too. The utilization of social media information of social media in text analytics has several business advantages. First, it can serves as a marketing tool as reported in Sita (2013), where 70 per cent of the airlines surveyed will make use of social media for marketing purposes, reservations, on-line check-in and CRM by year 2016. Second, it improves the brand awareness, loyalty and recognition through text mining. Gilbert *et al.* (2001) reported the need of branding in LCCs in this competitive environment. Third, it helps to gain competitive advantage. A survey carried out by the Kelsey Group concluded that review sites are greatly influencing consumers' shopping behaviour and that online reviews are not only impacting the online sales but also the offline purchase decisions (ComScore, 2007)[1].

With the increasing competition in the LCCs industry, this study fill the gap in the literature by focusing on the customer reviews of LCCs in Malaysia, in which AirAsia, the best low-cost airline since 2009 is included in the study. A survey has been conducted by O'Connell and Williams (2005) on passengers' perceptions of LCC and full service carriers. Two major airlines in Malaysia, Malaysia Airlines (MAS) and AirAsia were chosen to be studied. The stud's results showed that young people are more likely to be attracted by the LCCs. Compared with MAS, AirAsia crew productivity level is three times higher and the airplane utilization rate is five hours more a day. They also found that MAS and AirAsia customer segments are different with the former used for business trip purposes and the latter is more likely to be chosen for recreation purposes. Sentiment analysis on Croatia Airlines by Jakopović and Preradovic (2014) showed that the airline perceived more positively than negatively despite its poor financial performance and employees strike in 2010. Adeborna and Siau (2014) and Sreenivasan *et al.* (2012) both studied the airline using microblogging data. The former discussed about the airline quality using sentiment analysis and the latter discussed on the types of communication exchange between airlines and their consumers. Saha and Theingi (2009) conducted a survey on low-cost airlines in Thailand and concluded that passenger satisfaction was an important driver of behavioural intentions. Dobruszkes (2006) concluded that European LCCs have acquired a significant place in Western Europe but the markets have yet to reach the healthy level. Liberalization and point-to-point routes have boosted the creation of new routes in the industry (Dobruszkes, 2006). Dresner *et al.* (1996) studied the impact of LCCs on airport and route competition and concluded that the presence of LCCs on both new and competitive routes have led to a decrease in airfare and increase in air traffic. The spillover on competitive routes caused by the entrance of LCCs into the markets has indirectly forced LCC to pay more attention to customers' welfare.

The immense data available in textual form in databases and the World Wide Web, manual analysis and extraction of useful information are not possible. Text mining, also known as intelligent text analysis is a computer-driven automated technique used to discover significant and non-trivial patterns of information from the unstructured texts. This technique has created a strong industrial impact in decision making especially in customer-focused companies such as those in the retail, financial, communications and marketing industries. Businesses use text mining applications to analyse customer demographics, to predict future trends, to gain knowledge of competitors' developments and to make proactive and knowledge-driven decisions.

With the explosive growth of social media, there is an increase in number of consumers who actively engage in blogs, Facebook, Twitter, discussion groups and forums to share their opinions, experiences, desires and expectations. This has made the text analysis becomes more desirable and more powerful in analysing real time customers feedback compared to the traditional data analysis methods which is more time consuming. Text mining also can be used for sentiment, which it uses natural language processing application to determine the customers' preferences and detecting customers' dissatisfactions to a product. Data generated via social networks is a valuable asset as it can be used to generate information for decision making. Product reviews generated by consumers are having a significant impact on consumer buying decision (Chevalier and Mayzlin, 2006). Consumers in general are more interested on the product reviews by other similar consumers than the information provided by the vendors, as it is more credible. The sentiments from online reviews were shown to have significance influence on others in decision making (Liu, 2010). Analysis on consumer reviews on the internet has now become an essential part for any businesses to survive.

In this study, the tweets available in the Twitter accounts of LCCs in Malaysia is used to unravel the consumers' opinion and concern towards the LCCs to improve the current operational efficiency. In Malaysia, there are a total of five LCCs, namely AirAsia, Berjaya Air, FireFly, MASwings and Malindo Air. AirAsia (one of the largest LCC in Malaysia) has played a significant role in the country's economic growth by fostering the tourism industry through cheaper flight than those of prestigious airlines. AirAsia has received several awards such as world's best low-cost airline continuously for five years from 2009 to 2013. Thus far, no research has been done on the microblogging in LCCs. The main objective of this study is to comprehend consumer opinion towards the LCC industry in Malaysia by using text mining. Specifically, this study uses text mining techniques in conjunction with statistical techniques in:

(1) examining the customers' sentiments towards LCCs in Malaysia;

(2) identifying consumers' segment towards the LCCs for business decision purposes; and

(3) identifying customer sentiment in each consumers' segment in providing better CRM.

This study sheds light on the current literature to solve the three issues mentioned above by studying LCC customers' needs using social networks such as Twitter in the case of Malaysia. The results obtained from this study can be used by other airlines providers as a model to improve their customer services. Since Malay is the national language in Malaysia, many domestic consumers express their opinions in Malay. Therefore, a Malay lexicon is built. A list of stop-words for Malay is compiled in this study. Being data independent, the newly created Malay lexicon can be easily ported to other languages to be used in future.

Section 2 discusses the research methodology and conceptual framework, as well as data collection and statistical methods used in the analysis. Section 3 presents the results of customers' sentiments of the LCCs in Malaysia. In this section, the clustering results of consumer opinion on the LCCs using two unsupervised learning approaches are also discussed. Section 4 concludes the paper with some discussions and recommendations for the LCCs in Malaysia. The limitations of this study are also discussed in Section 4.

2. Methodology

Previous studies view sentiment analysis and clustering as two separate issues. Some of the studies only utilized sentiment analysis to discuss consumer sentiment but not understanding the user review on the airline. The techniques used in the previous studies on airline industry are on SentiStrength analysis, simple manual category grouping of microblogging data or conducting a survey. This study utilizes machine learning techniques such as unsupervised learning in solving the research issues.

2.1 Data preprocessing

Microblogging is a broadcast medium in the form of blogging. Twitter which was formally launched on 13 July 2006, is a microblogging service that allows users to post short updates. Each blog, or "tweet" is limited to 140 characters, equivalent to the size of the newspaper headline. It is shown that more than 80 per cent of Twitter users update their status daily (Thelwall *et al.*, 2011). Thus, Twitter data are chosen to be used in this research. Twitter is also a good source of collecting consumer opinion due to the heterogeneity of users. Twitter users are from different social backgrounds, from ordinary people to professionals, organization representatives, celebrities and politicians. Thus, the tweets collected are the words of users with different interest groups and this makes it a very valuable online source of opinion. The analysis covers a period of two and a half months (from 1 February 2014 to 15 April 2014), which comprised of 31,535 tweets. Tweets with both hashtag (#) and mention (@) of the five LCCs (AirAsia, FlyFirefly, Wingmates, berjaya_air and MalindoAir) in Malaysia are chosen as the sample in this analysis. Hashtag (#) is useful in tweets categorization and it also help in simplifying the process of tweets search. Mention (@) on the other hands is used to address another follower in the tweet. Retweets (tweets start with "RT") are treated as duplicates and are removed from both sets of data. Tweets with other languages than English and Malay are also deleted prior to the analysis. Tweets are then cleaned by removing punctuation, special characters, digits and uniform resource locators links. All the emoticons (such as emoji) are removed so that the data set contains only words. Following this, tokenization and words stemming are carried out. Tokenization is the process of breaking up a sequence of strings into pieces called tokens. The aim of tokenization is to explore the words in a sentence and identify meaningful keywords. Punctuation was removed in the process of tokenization. Tokens can be made up of characters, numeric or alphanumeric. Following this, stop-words are removed from the tweets. Stop-words are words from non-linguistic view that do not carry information. Prepositions (such as "from", "to", "after", etc.), articles (such as "a", "an" and "the") and pronouns (such as "I", "you", "she", "he", etc.) can be treated as stop-words. Eliminating stop-words helps to improve text processing performance. Next, word stemming is executed. Word stemming is a process of transforming words into their roots. Many words in English have different forms of the same words, for example "stemming", "stemmed" and "stems" have the same root word of "stem". Lastly, capital letters are converted into lower case. About 10,895 tweets are left after the data cleansing process. Tweets are then being converted to a corpus. Corpus is a large and structured set of texts. Also, all the keywords associated with a particular LCC such as "Tony Fernandes", "Zest" and "BigCard" are removed as well. The two subsections below discuss the two main techniques that are used to summarize the customers' opinions of LCCs in Malaysia.

2.2 Sentiment analysis

Sentiment analysis classifies customers' opinions into positive, neutral or negative. The sentiment polarity of a tweet is determined by comparing all the opinion words in the tweet against the subjective words in the dictionary and aggregating these words to give a final opinion to each feature. We obtain our initial opinion lexicon from Hu and Liu (2004), which contains of 2,006 positive words and 4,783 negative words. Opinion words in a lexicon can be categorized as positive and negative words. Positive opinion words are used to describe some desired state whereas negative opinion words are used to describe some undesired state. Examples of positive opinion words are good, awesome and comfortable; examples of negative opinion words are bad, frustrate and awful. Since Malay is included in the analysis and no opinion lexicon in Malay has been created, a Malay version opinion lexicon is computed manually. A Malay lexicon with a total of 54 positive and 109 negative sentiment words and a list of 181 stop words was successfully created. Table I shows some of the examples of the Malay lexicons with its polarity and the stop words. We form a lexicon of two classes, i.e. positive opinion words and negative opinion words, by extracting those words from the data set. We include also misspelled words into the lexicon as those words are common enough and appear frequently in social media contents. Also, we include as well word with its affix. For example "baik" and "terbaik" have the same root word of "baik" which mean good in English.

Positive and negative opinion words have a sentiment score of "$+1$" and "-1", respectively. In naïve algorithm sentiment score calculation, each tweet is scored by subtracting the number of occurrences of negative words from the number of occurrences of positive words. The sum of positive scores for a particular tweet indicates a positive expression of sentiment and the sum of negative score indicates a negative expression of sentiment. A tweet with score zero implies a neutral expression of sentiment. For example, the tweet "thanks for your help with this I managed to amend booking via online chat consultant was lovely and super helpful" will have a sentiment score of "4" and "bad counter service flights delayed more than an hour miscellaneous charges this got to be the worst airlines ever" is given a score of "-4". The occurrences of tweets with very positive (score $\geqslant 2$) and very negative (score $\leqslant -2$) sentiment scores are calculated. The percentage of these "extreme" tweets which are positive are calculated by dividing the total number of very positive tweets from the sum of both very positive and very negative tweets.

In addition, SentiStrength is used as an additional tool to compare the sentiment analysis results obtained from the conventional approach discussed above. SentiStrength (Thelwall *et al.*, 2011). is a lexicon-based sentiment analysis tool used to classify the sentiment strength. Sentiment strength assesses both the strength of negative and positive sentiments in a tweet, with the assumption that both the sentiment polarity can coexist within a tweet. Sentiment strength applies a scaling system on a -5 to $+5$ scale on the opinion words; from most negative to most positive, which is different from the ordinary sentiment analysis that gives equal weight to each of the opinion words.

Table I. Malay lexicon and stop words

Words polarity	Examples
Positive words	baik, terbaik, cekap, ramah, efektif
Negative words	biadab, boikot, melampau, bengong
Stop words	agar, akan, untuk, utk, daripada, jika

A score of $+1$ and -1 indicates a neutral term. The objective here is to detect the sentiment expressed by the consumers rather than the tweet's overall sentiment polarity (Thelwall *et al.*, 2011). This application considers the existence of emoticons, negation words, booster words, question words, exclamation marks, repeated letters in an opinion word and repeated punctuation in its calculation of sentiment polarity. Booster words, exclamation marks, repeated letters and punctuation alter the strength of the opinion word, whereas negation words flip the emotion from positive to negative or vice versa.

2.3 Clustering

Clustering is the fundamental task in modern data analysis. In the context of text mining, it has great practical interest as it can organize, discover and summarize latent information from unstructured text documents automatically. Clustering unstructured data begins with the creation of vector space known as bag-of-words. Using words as features, the data set is represented as a high dimensional vector. The corpus is then being converted to a term-document matrix, a mathematical matrix where its rows correspond to the tweets in the corpus and columns correspond to the appearance frequency of the terms in the corpus. With 10,895 tweets in the corpus, the term-document matrix tends to get very large. To reduce the size of the matrix without losing important information, sparse terms were removed from the matrix. Sparse terms refer to those words that occur only in a few tweets. In this research, we remove words with at least 99.5 per cent of sparsity. The selection of different parameters such as the number of desired clusters, (k) will lead to different clusters of data. The partitioning schemes will not be optimal if the selection of the input parameters is improper. Rather than choosing the k number of clusters randomly, we applied the Calinski and Harabasz (CH) index to determine the optimal value of k. A well-defined cluster should have small variance within the cluster and significantly vary between clusters. The larger the CH index, the better the data partition is.

K-Means clustering aims to group each data point into k-clusters that are fixed a priori by minimizing the distance from the data points to the cluster. K-Means clustering updates centroids in a batch mode, i.e. the existence of all data points in each cluster will be reviewed before the centroids are being updated (Zhong, 2005). It is an iterative process, whereby the membership of each term in a cluster is re-evaluated based on the centroid of each existing cluster. With $x^1, x^2, \ldots, x^m$ data points available in our data set, the steps are:

- Step 1: determine the k-number of clusters.
- Step 2: choose μ_k centroids randomly as the initial means.
- Step 3: determine the distances of each data point to the centroids using Euclidean distance, $d(x^i,\mu_j) = ||x^i - \mu_j||^2$.
- Step 4: assign each data point x^m to the closest cluster centroid, μ_j based on the minimum distance. Cluster centre, $c^i = \arg\min_j = ||x^i - \mu_j||^2$.
- Step 5: move each cluster centroid, μ_j to the mean of the points assigned to it.
- Step 6: compute new cluster centroid.
- Step 7: process stops when there is no change in cluster centroid or no object changes its clusters.

To enhance the speed of the algorithm, Spherical K-Means (SK-Means) is also being considered to cluster the customers' opinion. SK-Means clustering is a modification of K-Means clustering with cosine similarity that works on the vectors that lie on the unit sphere. Cosine similarity uses the cosine of the angle between the vectors, effectively measure the similarity between the tweets to retrieve information. Every centroid is represented as high-dimensional unit-length vector. The direction of document that is represented as a TF-IDF vector is said to be more important than the magnitude, hence each document vector is normalized to be of unit length, i.e. $||x^i|| = 1$ (Zhong, 2005). The SK-Means algorithm aims to maximize the mean cosine similarity as cost function, by calculating the inner product between the data points and their nearest centroid, $L = \Sigma_x x^T \mu_{k(x)}$ where $k(x) = \arg\ max_k\ x^T\ \mu_k$. With N unit length data vectors in our data set, $X = \{x_1, x_2, \ldots, x_N\}$ in $\mathbb{R}^d$ and K number of clusters, we wish to get a partition of the data vectors given by the cluster identity vector $Y = \{y_1,\ y_2, \ldots, y_N\}$, $y_n \in \{1,2, \ldots, k\}$ by following the below steps:

- Step 1: select the unit-length centroid vectors.
- Step 2: set $y_n = \arg\max x_n^T \mu_k$ for each data vector x_n.
- Step 3: estimate the cluster centroid for cluster k.
- Step 4: process stops when there is no change in vector.

The main difference between K-Means clustering and SK-Means clustering is the normalization of the unit-length on the re-estimated mean vectors. Strehl *et al.* (2000) show that cosine similarity measurement outperforms the Euclidean distance calculation. Cosine similarity is widely applied in text mining because it is easy to interpret and the computation for sparse vectors is simple (Salton and McGill, 1983). SK-Means clustering is one of the most efficient algorithms in terms of speed (Zhong, 2005). Also, SK-Means works well with high-dimensional data sets. With the application of cosine similarity, the algorithm makes use of the sparsity of vectors and has high efficiency especially in a large data set (Dhillon and Modha, 2001). Besides, the SK-Means algorithm can be efficiently parallelized and converges to local maxima quickly (Dhillon and Modha, 2000). Another advantage of the SK-Mean algorithm is that concept vector generation can be a model which allows it to be re-used in future classifications (Dhillon and Modha, 2001).

3. Results

3.1 Sentiment analysis

Social media allow airlines to establish stronger bonds with their customers by enabling two-way communication with their customers. Social CRM engages with the customers proactively with the objective to improve customers' overall experience. It can also be used for targeting management, customer information management and service customization purposes (Öztaysi *et al.*, 2011). The airline industry is the second most socially devoted industry where 55 per cent of reviews posted on social media were responded by the airlines (Socialbakers, 2012). A lexicon-based sentiment analysis depends greatly on the sentiment (or opinion) words stored in the dictionary.

We start the sentiment analysis with the naïve method by counting the occurrences of positive and negative opinion words. A score is assigned to each tweet by subtracting the number of occurrences of negative opinion words from the number of positive opinion words. Figure 1 shows the sentiment score distribution of Malaysia's LCCs. In total, more positive sentiments are identified compared to the negative

however overall 51.73 per cent of the tweets have a neutral sentiment. A sentiment analysis has been conducted using Croatia Airline, similar findings have been found which positive sentiments are more negative sentiment (Jakopović and Preradovic, 2014).

Next, the extreme positive tweets which are calculated using the method as discussed in the methodology section is compared to the American Customer Satisfaction Index (ACSI). ACSI evaluates the quality of products and services of more than 230 companies in 43 industries by interviewing approximately 7,000 households annually (CNBC, 2014). The criteria used by ACSI for airline industry evaluation includes, passenger's flight experiences, preferences of flight schedules, satisfaction with reservation processes, check-on luggage options, baggage handling, flight punctuality, in-flight services, seat comfort, flight crew courtesy, loyalty programmes and web site (ACSI, 2014). Customer satisfaction indexes are reported on a scale from 0 to 100. The ACSI benchmark gives LCCs a valuable insight on customer satisfaction and allows LCCs to compare results with other peers in the industry. Malaysia's LCCs' sentiment score of 67 is slightly lower than the benchmark of ACSI for airline industry, 69. The ACSI report released in April 2014 showed that poor in-flight amenities and bad seat comfort are the two major drawbacks for LCCs. However, LCCs showed an overall better performance than the traditional airlines.

The sentiment polarity of consumers towards Malaysian LCCs using SentiStrength tool is further analysed. The positive and the negative sentiment rating from the SentiStrength analysis are presented in Figures 2 and 3, respectively. On the positive sentiment scale, we rephrase the five-point scales to be neutral (0), positive (1), moderately positive (2), very positive (3) and extremely positive (4). Overall, 50.17 per cent of the total tweets are neutral; 49.83 per cent of the tweets have at least a rating

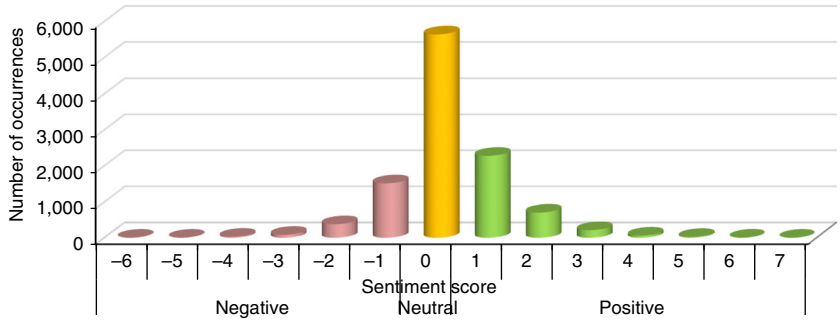

Figure 1.
Malaysia LCCs' sentiment scores using naïve algorithm

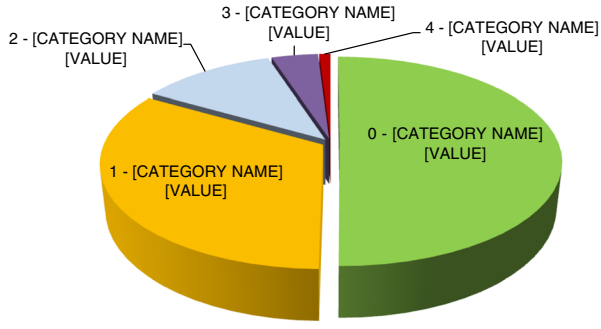

Figure 2.
Positive sentiment rating

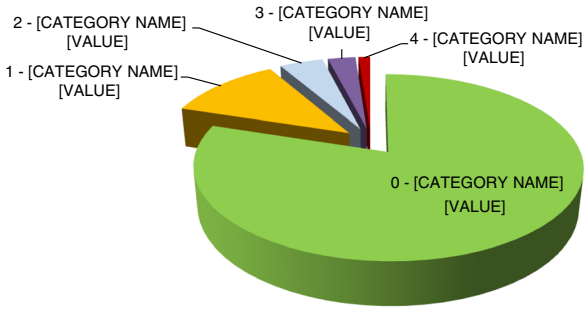

Figure 3.
Negative sentiment rating

of 2; 33.10 per cent of the total tweets are positive; 11.68 per cent are moderate positive and 4.06 per cent are very positive. The percentage of extremely positive sentiment is the lowest at only 0.99 per cent. On the negative sentiment scale, we rephrase the five-point scales to be neutral, negative, moderately negative, very negative and extremely negative. In general, Malaysia's LCCs are not perceived as extremely negative, as only <10 per cent of the total tweets are rated -3 and above. The most frequent rating is the neutral sentiment which accounted for 79.90 per cent of the total tweets. The percentage of extremely negative sentiment is the lowest at 1.18 per cent. The cross-tabulation of positive and negative sentiments indicates that Malaysia LCCs succeeded in maintaining a rather positive image among consumers. In total, 40.11 per cent of the total tweets show a positive sentiment (positive rating of 2 and above and negative rating of 1) and 10.38 per cent expressed a negative sentiment (negative rating of 2 and above and positive rating of 1).

3.2 Cluster analysis

K-Means clustering and SK-Means clustering were used in our analysis. First, we determine the ten words for each cluster. CH index suggest four optimal clusters. Tables II and III report the clustering results using K-Means clustering and SK-Means clustering, respectively. Both algorithms deliver very similar results. Cosine similarity works better than Euclidean similarity for document feature vectors as the former adjusts for differences in document length. In this case however, all tweets are short and limited to 140 characters[2].

Both results yield similar clusters with different orders only. For the first cluster, comments from consumers are mostly on the customer service of each LCC, whether or

Table II.
Clusters results using K-Means clustering

Cluster 1	Cluster 2	Cluster 3	Cluster 4
Service	Booking	Flight	Give
Customer	Thank	Booking	Flight
Flight	Airport	Change	Booking
Thanks	Ticket	Cancel	Free
Good	Travel	Time	Back
Bad	Check	Thanks	Ticket
Airport	Promo	Airport	Number
Call	Time	Email	Call
Ticket	Email	Delay	Refund
Airline	Day	Check	Thanks

Cluster 1	Cluster 2	Cluster 3	Cluster 4
Thanks	Airport	Flight	Booking
Email	Travel	Booking	Ticket
Reply	Promo	Cancel	Check
Flight	Airline	Change	Time
Love	Ticket	Time	Change
Booking	Day	Delay	Number
Response	Good	Airport	Payment
Assistance	Plane	E-mail	Like
Direct	Pilot	Ticket	Refund
Appreciate	Indonesia	Travel	Call

Table III.
Clusters results using SK-Means clustering

not they are satisfied with the service. The saying "customer is always right" suggests that good customer service is essential for all businesses including LCCs to maintain its business and to attract new customers as well as retaining customer loyalty. Bad customer service will have an adverse effect on a business as the LCC industry is a very competitive industry. With more and more people turning to social media, information posted online is shared instantly and thus negative feedback posted on the internet is detrimental to the business. Words spread in split seconds, especially when it comes to expressing negative experiences to the world. Therefore it is essential for companies nowadays to understand and incorporate social media into their strategic management. For example, companies should dedicate time to respond to posts and comments on social media platforms and be transparent when addressing issues with customers. In order to improve customer service, the first step the management should take is to train the employee to listen to customers' complaints. By taking the time to hear through the entire complaint, the customer feels that he is being cared for. Good customer service will not only gain trust with current customers, who will also become the referrals for future customers, especially with the development of electronic word of mouth (eWoM) that allows the positive statements to spread at a faster pace. This study is consistent with the findings of Sreenivasan *et al.* (2012).

LCCs use a low-pricing strategy to win customers over conventional airlines and it is very effective as consumers are naturally attracted to lower prices. Results in cluster 2 show that consumers like to discuss about tickets promotions by the LCCs. From this cluster it is obvious that consumers are more likely to choose LCCs for vacation purposes which are in line with the survey done by O'Connell and Williams (2005) on passengers' perceptions on low-cost airlines and full service carriers. There are also some negative feedbacks on airfare promotions found which are mainly on customers that had issues during online flight bookings, particularly during promotion periods when they encountered high traffic on the web site.

The third cluster mainly discusses flight cancellations and flight delays. From the sentiment polarity discussed in the previous section, 10.27 per cent (417 out of 4,062) of the total negative sentiment polarity are due to flight delays or flight cancellations. Flight delays and cancellations can result in financial loss for an airline. For example AirAsia pays compensations up to RM200 for flights which are delayed for more than two hours (AirAsiaInsure, 2014). LCCs also bear the cost of accommodation and transportation of passengers if a flight is delayed to the next day. Flight cancellations or delays also have an impact on flight turnaround time which is what LCCs' low-cost pricing relies on. It was estimated that airline companies in the USA lost $8.3 billion in

2007 due to the additional expenses incurred following flight delays or cancellations (Ball *et al.*, 2010). Passengers on the other hand also face financial loss of up to $16.7 billion for the time lost and missed flight connections. Schedule buffering was introduced in the airline industry where additional times are allocated over the unimpeded schedule flight times to account for flight delays. Federal Aviation Administration, an agency of the United States Department of Transportation estimated that the implementation of schedule buffers cost airlines $3.7 billion in 2007 (USA Today, 2013). The additional costs charged due to flight delays or cancellations are against the main principle of a low-cost airline of minimizing the unit costs and optimizing flying time for each plane. The decision on purchasing a flight ticket can be influenced by the track records on the tendency of flight delays or cancellations for a particular LCC. A study has indicated that flight delays have an upward impact on air fares and also decrease the willingness of consumers to travel by air (Ball *et al.*, 2010). From the examples of feedbacks given below, it is evident that LCCs were not able to efficiently handle flight delays or cancellations, resulting in customers' dissatisfaction.

With the rise in the usage of social networking in recent years, all LCCs are using social media platforms in particular Facebook and Twitter, as part of their marketing and customer service. In the fourth cluster, we found that passengers are making use of Twitter in managing their bookings such as to request for amendment of personal details, flight changes and refunds requests in hope that the airline can respond immediately through these platforms. AirAsia handles these requests for assistance by using direct messages, rather than an open reply. This helps in maintaining customer's privacy while at the same time limiting negative feedbacks from becoming widely known. It was found that only 77 out of 168 airlines are active Twitter users. AirAsia replies to 40 per cent of the total tweets on average, and this is almost double the industry average of 24 per cent (Shashank, 2011). This explains why most of the data collected in this research are tweets concerning AirAsia. AirAsia makes use of social media as one of their customer service platforms to enhance the brand experience, making the experience with AirAsia more personal and at the same time reduces marketing expenditures. Customers hate slow responses. Keeping the customers who call in to call centres waiting to talk to a representative is very annoying. From the tweets, most passengers complain about the slow responses from call centres. They also questioned the efficiency in responding e-mails, especially in processing payment refunds. AirAsia did not commit more resources to its call centre as the company believes that this would inevitably add to operation costs and to higher fares. However, the call centre is still the preferred point of contact for some Malaysian consumers. Until more consumers are able to adapt to the internet or the use of social media, AirAsia should improve the efficiency of its call centre in this transition period. Table IV shows some examples of tweets according to the above four clusters.

3.3 Sentiment analysis and clustering

By combining both results from sentiment analysis and clustering in Sections 3.1 and 3.2, some interesting results are obtained. Customer service, booking management and ticket promotion gain more positive opinions than negative opinions, while flight cancellation receives more negative sentiments. Table V shows the details of each cluster in conjunction with sentiment analysis. Customers perceived positively ticket promotion and only a minority of this cluster have negative sentiments. The ratio of positive to negative feedback is about 28:9, with the number of customers giving positive feedback in this cluster three times those who gave negative feedback.

Number of Clusters	Type of Feedbacks	Examples
Cluster 1 – customer service	Positive feedbacks	Morning, I was very satisfied with your flight attendants services they so kindness good job!
		Both customer services really good not all budget airlines follow the all nasty all the time model
		Thanks! amazing prompt accurate service, really appreciate that thanks now everyone can fly
	Negative feedbacks	So true! one month since my flight and still no reply!
		Terrible customer service!! More like no customer service :(
		Countless eforms, e-mails, calls to change my last name on my ticket and no response so frustrating!!! Bad customer service
		Your customer service is terrible!! I had a bag badly damaged from a flight on 3 February and still no one has attended to my complaint
Cluster 2 – ticket promotions	Positive feedbacks	Thank you very much I will go to some place with special promo price and it could be amaze
		54 days left to vacation travel I got promo ticket!!
		This time we flew and they were great and cheap stopover in KL, check out for promotions you can get it really cheap
	Negative feedbacks	Why your Surabaya prices are so expensive during June and July promotions doesn't seem to be true
		Payment server has caught a bug!!! WTF!!! I wonder if big points redemption promo is ever real!!
		Susah nak cari promo time cuti skolah sentiasa murah sepanjang sesi persekolahan (it is difficult to get cheap promo during the school holiday period, cheap tickets only available during schooling period)
		Trying to buy ticket but error message keep mentioning how come we want to buy ur free ticket
Cluster 3 – flight cancellations or delays	Negative feedbacks	Bad counter service, flights delayed more than an hour, miscellaneous charges, and this got to be the worst airlines ever!
		Really really sucks!!!! Cancelled flights without informing the customers!!! Gosh!!! Worse service please helps to pass
		Flight got cancelled, rescheduled the only notification I received less than 24hrs from is just a SMS!
		Very irresponsible, they cancelled flights less than one month notice and we are stranded now, very upsetting
Cluster 4 – post-booking management	Neutral feedbacks	I accidentally typed the wrong spelling on for my friends booking, kindly assist me changing it on the ticket tq
		Charges for call centre calls! And the chat queue is huge!
		Do you ever get back to your customers? Bad service
		I want to modify my flight due to Kelud eruption why is really hard to contact call centre even the live chat
		Received five separate e-mails from your CRMS telling me my refund is being processed since October 2013
		Made numerous phone calls to your call centre inquiring about the status of my refund but no one seems to know the status

Table IV.
Examples of tweets according to the four clusters of customer service, LCCs tickets promotions, flight cancellations and delays, and post-booking management

The same applies to booking management with a 5:2 ratio. Although there is no major differences between positive and negative feedback in the customer service cluster, positive feedback still outnumbered negative feedback. LCCs' customers are more uncomfortable with the flight cancellation as they perceived this issue negatively, despite the fact that the LCCs have been trying to please the customers by giving away monetary compensation. However, LCCs is perceived more positively than negatively in most of the dimensions of services such as customer services, booking management and ticket promotion.

The results show that each cluster provides both good and bad feedback. The negative feedback in customer service is mainly due to the failure of customer service or the call centre to respond on time. The negative segment from booking management show there are still rooms for improvement. The failure of airlines in handling flight cancellation promptly causes customer anger. Lastly, the heavy web traffic during promotional period upsets the customers. This information basically provides an idea for a management on how efficient prompt reply will solve most of the problem discussed in each cluster above.

4. Conclusion and implications

Companies are continually refining insights into customer needs, preferences, experiences and opinions on their products or services. First, by studying customer sentiments towards LCCs in Malaysia, we found more positive than negative polarity across the different algorithms. Hence, consumers in general are satisfied with the services provided by Malaysian LCCs. Second, the four main topics that are discussed by Twitter users are customer service, LCCs tickets promotions, flight cancellations and delays and post-booking management. Although these four topics are commonly discussed, this research is able to enable a deeper understanding of the sentiment under those four topics.

Consumers use Twitter as a platform to express their emotion and the problems they faced in using LCCs. Despite the minimal difference between positive and negative polarity, businesses should put more effort in improving the efficiency of their customer service by providing better training for employees especially on how to deal with the flight cancellation issue and how to respond promptly to customer requests. With eWoM, good customer service is important as it serves as an advertisement medium. As the internet is a vital part of marketing, LCCs should improve their web servers in order to cope with heavy traffic during promotion periods as the findings show LCCs customer are in general happy with the promotion and the only upset that they feel is the heavy traffic on the web. Because flight delays or cancellations create loss for both parties. Customer support play a crucial role. With enhancement of the database management system, the customer support staff can offer alternative flight to fulfil the customer's needs. This can be done by cooperating with other airlines on sharing their databases. As the results show, customers are turning to social media platforms such as Twitter for their post-booking management and also for

Table V. Sentiment analysis based on clusters

Clusters	Positive (%)	Negative (%)	Ratio
Customer service	35.22	27.13	9:7
Booking management	33.89	13.78	5:2
Flight cancellation	22.22	42.96	1:2
Ticket promotion	10.40	3.36	28:9

complaining. In particular, management should be aware that although promotion might make customers happy it may turn off the customers if the web continues to have the traffic issues. In conclusion, with any flight delay, prompt reply and a responsive web site will please the customers. LCCs would have a better opportunity in attracting more customers and generating more profits if all these problems can be addressed.

This study allows LCCs to have a better understanding on customers' sentiment towards the services provided by the companies. The result is believed to be fairer and unbiased as compared to surveys conducted where no interview is involved. Also, this study gives a clearer picture to the LCCs on topics which are broadly discussed on the social network. LCCs can identify valuable customers by making proactive decisions based on the predictions of customers' future behaviours. If the airlines are able to analyse customer tweet data in real time by classifying customers' feedback, it would help management to facilitate strategic operational activities.

During the research, a few common problems well documented in text mining field were encountered. Lexicon-based sentiment analysis can at times fail to recognize sarcasm or provocative words. Also, there were certain words that reflected a positive and neutral sentiment at the same time. The word "thank" shows a positive sentiment when a customer show appreciation to the service provided by LCCs but it is not regarded as a positive sentiment when a question is asked or help is requested. Future research should focus on rectifying this drawback. We also had difficulty in removing morphological affixes from Malay words, leaving only the word stem, therefore this should be included in future research. Consumer sentiments from MAS were excluded in our analysis due to the missing MH370 flight and with its current financial crisis. We believe that these events would have resulted in bias in the data collected. Future studies should consider including MAS to compare with LCCs in Malaysia. We also excluded hashtag in our analysis as more than 70 per cent of the total hashtags in the data set were non-sentiment related. Future studies may consider investigating hashtag that contribute to consumer sentiments. Furthermore we have also excluded emoji in the sentiment analysis. Emoji is a set of small, cartoon-like characters that are often used when typing on a mobile device. The use of emoji on Twitter was recently introduced in April 2014. Future research can consider including emoji in the sentiment analysis. Despite these limitations, we believe that our study has contributed to this subject especially on the consumer behaviour towards Malaysia's LCC industry.

Notes

1. Studies on product review can be found in Xu *et al.* (2011) and Liu *et al.* (2005).
2. Short documents limit the accuracy in computing similarities (Tagarelli and Karypis, 2013).

References

ACSI (2014), "Unique benchmarking capability", available at: www.theacsi.org/the-american-customer-satisfaction-index (accessed 7 May 2014).

Adeborna, E. and Siau, K. (2014), "An approach to sentiment analysis – the case of airline quality rating", *PACIS 2014 Proceedings, Paper 363, Chengdu, 24-28 June*.

AirAsiaInsure (2014), "AirAsia travel protection", available at: www.airasiainsure.com/info_centre/faq_coverage.php#content (accessed 12 May 2014).

Ball, M., Barnhart, C., Dresner, M., Hansen, M., Neels, K., Odoni, A. and Zou, B. (2010), "Total delay impact study: institute of transportation studies", working paper, University of California, Berkeley, CA, November.

Borenstein, S. (1992), "The evolution of US airline competition", *The Journal of Economic Perspectives*, Vol. 6 No. 2, pp. 45-73.

Chevalier, J.A. and Mayzlin, D. (2006), "The effect of word of mouth on sales: online book reviews", *Journal of Marketing Research*, Vol. 43 No. 3, pp. 345-354.

CNBC (2014), "ACSI report: customer satisfaction with airlines remains low", available at: www.cnbc.com/id/101601372 (accessed 7 May 2014).

Comscore (2007), "Online consumer-generated reviews have significant impact on offline purchase behavior", available at: www.comscore.com/Insights/Press_Releases/2007/11/Online_Consumer_Reviews_Impact_Offline_Purchasing_Behavior (accessed 13 April 2014).

Dhillon, I.S. and Modha, D.S. (2000), "A data-clustering algorithm on distributed memory multiprocessors", in Mohammed, J.Z. and Ching-Tien, H. (Eds), *Large-Scale Parallel Data Mining*, Springer, Berlin, pp. 245-260.

Dhillon, I.S. and Modha, D.S. (2001), "Concept decompositions for large sparse text data using clustering", *Journal of Machine Learning*, Vol. 42 Nos 1-2, pp. 143-175.

Dobruszkes, F. (2006), "An analysis of European low-cost airlines and their networks", *Journal of Transport Geography*, Vol. 14 No. 4, pp. 249-264.

Dresner, M., Lin, J.-S.C. and Windle, R. (1996), "The impact of low-cost carriers on airport and route competition", *Journal of Transport Economics and Policy*, Vol. 30 No. 3, pp. 309-328.

Gilbert, D., Child, D. and Bennett, M. (2001), "A qualitative study of the current practices of 'no-frills' airlines operating in the UK", *Journal of Vacation Marketing*, Vol. 7 No. 4, pp. 302-315.

Hu, M. and Liu, B. (2004), "Mining opinion features in customer reviews", paper presented at the AAAI, Washington, 22-25 August, available at: www.aaai.org/Papers/AAAI/2004/AAAI04-119.pdf (accessed 5 May 2014).

Jakopović, H. and Preradovic, N.M. (2014), "Mining web and social networks for consumer attitudes towards government-owned Croatian national airline", *International Journal of Computers*, Vol. 8, pp. 128-135.

Lee, S.G., Trimi, S. and Kim, C. (2013), "Innovation and imitation effects' dynamics in technology adoption", *Industrial Management & Data Systems*, Vol. 113 No. 6, pp. 772-799.

Lee, S.G., Trimi, S. and Yang, C.G. (2014), "ICT service providers strategies and customer migration", *Industrial Management & Data Systems*, Vol. 114 No. 7, pp. 1036-1049.

Liu, B. (2010), "Sentiment analysis and subjectivity", in Nitin, N. and Fred, J. (Eds), *Handbook of Natural Language Processing*, Chapman & Hall, London, pp. 627-666.

Liu, B., Hu, M. and Cheng, J. (2005), "Opinion observer: analyzing and comparing opinions on the web", *Proceedings of the 14th International Conference on World Wide Web, ACM, New York, NY*, pp. 342-351.

O'Connell, J.F. and Williams, G. (2005), "Passengers' perceptions of low cost airlines and full service carriers: a case study involving Ryanair, Aer Lingus, Air Asia and Malaysia Airlines", *Journal of Air Transport Management*, Vol. 11 No. 4, pp. 259-272.

Öztaysi, B., Sezgin, S. and Özok, A.F. (2011), "A measurement tool for customer relationship management processes", *Industrial Management & Data Systems*, Vol. 111 No. 6, pp. 943-960.

Salton, G. and McGill, M.J. (1983), "Introduction to modern information retrieval", in Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M. (Eds), *Lexicon-Based Methods for Sentiment Analysis, Computational Linguistics*, McGraw-Hill, New York, NY, pp. 267-307.

Saha, G.C. and Theingi (2009), "Service quality, satisfaction, and behavioural intentions: a study of low-cost airline carriers in Thailand", *Managing Service Quality: An International Journal*, Vol. 19 No. 3, pp. 350-372.

Shashank, N. (2011), "Eezeer and SimpliFlying launch airline monthly twitter report – Delta leads Twitter use, and customer service is in", available at: http://simpliflying.com/2011/eezeer-and-simpliflying-launch-airline-monthly-twitter-report-delta-leads-twitter-use-and-customer-service-is-in/ (accessed 13 May 2014).

Sita (2013), "Airline passengers in India ready for 'revolution'", available at: www.sita.aero/content/airline-passengers-india-ready-revolution (accessed 21 May 2014).

Socialbakers (2012), "Top 6 most socially devoted industries and brands", available at: www.socialbakers.com/blog/656-top-6-most-socially-devoted-industries-and-brands (accessed 13 May 2014).

Sreenivasan, N.D., Lee, C.S. and Goh, D.H.-L. (2012), "Tweeting the friendly skies: investigating information exchange among Twitter users about airlines", *Program: Electronic Library and Information Systems*, Vol. 46 No. 1, pp. 21-42.

Strehl, A., Ghosh, J. and Mooney, R. (2000), "Impact of similarity measures on web-page clustering", paper presented at the Workshop on Artificial Intelligence for Web Search (AAAI 2000), Austin, Texas, 30 July-1 August, available at: www.aaai.org/Papers/Workshops/2000/WS-00-01/WS00-01-011.pdf (accessed 5 May 2014).

Tagarelli, A. and Karypis, G. (2013), "Document clustering: the next frontier", in Aggarwal, C.C. and Reddy, C.K. (Eds), *Data Clustering: Algorithms and Applications*, Taylor & Francis Group, Boca Raton, FL, pp. 305-338.

The Asia Foundation (2014), "Economic growth in ASEAN drives demand for low-cost air carriers", available at: http://asiafoundation.org/in-asia/2014/02/26/economic-growth-in-asean-drives-demand-for-low-cost-air-carriers/ (accessed 30 March 2014).

Thelwall, M., Buckley, K. and Paltoglou, G. (2011), "Sentiment in Twitter events", *Journal of the American Society for Information Science and Technology*, Vol. 62 No. 2, pp. 406-418.

USA Today (2013), *Airlines Pad Flight Schedules to Boost On-time Records*, available at: www.usatoday.com/story/travel/flights/2013/02/14/airlines-flights-early-arrivals/1921057/

Xu, K., Liao, S.S., Li, J. and Song, Y. (2011), "Mining comparative opinions from customer reviews for competitive intelligence", *Journal of Decision Support Systems*, Vol. 50 No. 4, pp. 743-754.

Zhong, S. (2005), "Efficient online spherical k-means clustering", *the Neural Networks, 2005, IJCNN'05, Proceedings 2005 IEEE International Joint Conference, 31 July-4 August*, pp. 3180-3185.

Further reading

Forman, G. and Kirshenbaum, E. (2008), "Extremely fast text feature extraction for classification and indexing", *Proceedings of the 17th ACM conference on Information and Knowledge Management, ACM, New York, NY, 26-30 October*, pp. 1221-1230.

Liu, F. and Xiong, L. (2011), "Survey on text clustering algorithm", *IEEE 2nd International Conference, Software Engineering and Service Science (ICSESS), Beijing, 15-17 July*, pp. 901-904.

Moreno-Ortiz, A. and Hernández, C.P. (2012), "Lexicon-based sentiment analysis of Twitter messages in Spanish", *Journal of Procesamiento del Lenguaje Natural*, Vol. 50, pp. 93-100.

Corresponding author

Dr Pei Pei Tan can be contacted at: peipei@um.edu.my