Malaysia Journal of Science 37 (2): 137 - 153 (2018)

TEXT MINING AND SENTIMENT ANALYSIS ON REVIEWS OF PROTON CARS IN MALAYSIA

Yap Bee Wah[1], Nazira Abdullah[2], Shuzlina Abdul-Rahman[3], Michael Loong Peng Tan[4]

[1,2]Advanced Analytics Engineering Centre (AAEC), Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, MALAYSIA
[3] RIG Intelligent Systems, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, MALAYSIA
[4] School of Electrical Engineering, Faculty of Engineering, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia.
*Corresponding Author: beewah@tmsk.uitm.edu.my
Received: 21st May 2018 Revised: 28th October 2018 Accepted: 1st November 2018
DOI: https://doi.org/10.22452/mjs.vol37no2.5

ABSTRACT Social networks play an important role in commercial products, and thus knowing the emotions and opinions of users is very useful in improving services, sales, business and marketing strategies. This paper illustrates the text mining and sentiment analysis approach to gain valuable insights into consumer perceptions towards Proton cars in Malaysia. In the first case study, a total of 5000 comments posted on Paultan's Automotive News Facebook page were retrieved and analyzed using R, a statistical open source programming software. The final dataset consists of 277 posts with 2964 comments. Data cleaning was carried out to remove non-English word. A word cloud was then generated and the frequently mentioned words are "China", "like", "price" "good" and 'SUV'. In the second case study, we perform sentiment analysis on a total of 5330 comments using SAS Enterprise Miner 14.1. Out of the 60 documents, 39 (65%) posts were positive comments from the public while 21 (35%) posts were negative comments. The frequently mentioned words in the positive documents are "look good", "buy", "better" and "nice". Chi-Square and Information Gain were compared in selecting the meaningful terms of features. The selected features were used to evaluate the performance of Support Vector Machine (SVM) in classifying the posts as positive or negative. Five-fold cross validation results showed that SVM using linear kernel function has the highest accuracy (73.3%), sensitivity (76.7%) and F-measure (0.805).

Keywords: Facebook; Proton; Social Networks; R; Text Mining; Word Cloud

ABSTRAK Rangkaian sosial memainkan peranan penting dalam sesuatu produk komersil. Justeru itu, memahami emosi dan pendapat pengguna adalah penting untuk memajukan perkhidmatan, jualan, perniagaan dan strategi pemasaran. Kertas kerja ini menggambarkan pendekatan perlombongan teks dan analisa sentimen dan melihat sejauh mana persepsi pengguna terhadap kereta Proton di Malaysia. Kajian kes pertama melibatkan 5000 komen yang telah dipos pada Berita Mukabuku Paultan Automotif. Pos ini telah diambil dan dianalisa dengan menggunakan R, sejenis perisian pengaturcaraan sumber terbuka. Set data yang akhir mengandungi 277 pos yang melibatkan 2964 komen. Pembersihan data telah dijalankan untuk mengeluarkan perkataan bukan Bahasa Inggeris. Sekelompok awan perkataan telah dihasilkan dan perkataan yang sering disebut ialah "China", "like", "price", "good" dan "SUV". Manakala untuk kajian kes kedua, analisa sentimen telah dilakukan pada 5330 komen menggunakan perisian SAS Enterprise Miner 14.1. Sebanyak 39 (65%) pos dari sejumlah 60 dokumen adalah komen positif dan selebihnya iaitu 21 (35%) pos adalah komen negatif. Analisa mendapati perkataan yang seringkali disebut ialah "look good", "buy", "better" dan "nice". Dua teknik pengurangan atribut iaiatu *Chi-Square* dan *Information Gain* telah digunakan untuk pemilihan atribut. Atribut yang dipilih digunakan untuk mengukur prestasi pengkelasan *Support Vector Machine* (SVM) dalam mengklasifikasikan pos positif atau negatif. Lima-lipatan validasi silang menunjukkan SVM dengan fungsi kernel linear memberikan ketepatan pengkelasan (73.3%), sensitiviti (76.7%) dan *F-measure* (0.805) yang tertinggi.

Kata kunci: Mukabuku; Proton; Rangkaian Sosial; R; Perlombongan Teks; Awan Perkataan

INTRODUCTION

With advances in computer technology, various software has been developed for analyzing and visualizing important information from both structured and unstructured data. Unstructured data from Facebook, Twitter or blogs can now be analyzed to reveal information that may be important for decision-makers. IBM Content Analytics, SAS Text Miner, SAS Sentiment Analysis, and WEKA are software that can perform text mining or text analytics. R a statistical programming software also has the capability of text mining and visualization of unstructured data using Word Cloud.

In this modern technological era, consumers and users post their feelings and comments on various social media platforms such as Facebook, Twitter, or blogs. Thus, the internet has become an important source of data that can be harnessed by organizations to improve their services or business performance. Text mining, sentiment analysis and social analytics are some new technology developed for analyzing unstructured data.

Text mining (TM) is the process of deriving high-quality information from texts in documents, Facebook or Twitter. TM includes text categorization, text clustering, sentiment analysis, document summarization, and entity relation modeling (*i.e.*, learning relations between named entities). With the advances in big data technology, recent focus is on gaining insights from the unstructured data from social media.

Sentiment analysis (SA) involves analyzing an individual's attitude (judgment or evaluation) and affective state (emotions) expressed in a document or sentence. Social media networks such as Facebook and Twitter provide new ways of sharing updates, news and opinions. These social medium networks enable users to share and give their opinions in the form of comments or discussions (Choi & Lee, 2015).

Consequently, many companies and manufacturers use social networking to sell their products or to obtain feedback from consumers. The number of social media users has increased tremendously. In fact, Facebook has announced that 1.25 billion out of its 1.44 billion users are active (Oeldorf-Hirsch and Sundar, 2015). Therefore, social media plays an important role in promoting commercials products. Determining the emotions and opinions of users is useful not only in education but also in marketing and economics. This new technology trend enables individuals to express their ideas on any topics through social media platforms.

The automotive industry is one of the most important industries in Malaysia's manufacturing sector, and it significantly contributes to the country's drive towards becoming an industrialized nation. The main car production companies in Malaysia are Perusahaan Otomobil Nasional Sdn. Bhd. (PROTON) and Perusahaan Otomobil Kedua Sendirian Berhad (PERODUA). The former was established in 1983 as the sole national car company for almost 10 years until the advent of the latter in 1993. The Malaysian automotive website, paultan.org, has a Facebook page called Paultan's Automotive News. Users post their comments about Proton cars on this Facebook page.

With 500,000 unique visitors per day, paultan.org is currently the most popular automotive news website among Malaysian internet users. This website provides a platform for readers to interact with each other through its comment section in each post. This platform has led to the unlimited access of Internet users to express their opinions and feelings toward automotive-related issues. People can now look for and understand the opinion of others through information technology (Pang and Lee, 2008).

The aim of this paper is to illustrate text mining using R to gain valuable insights into consumer perceptions of Proton cars in Malaysia. The comments posted on Paultan's Automotive News Facebook page were retrieved and analyzed to gain insights into consumers' perception on Proton cars. We provide the syntax necessary to do text mining using R, a statistical programming software which is as popular as Python programming in big data analytics. We also illustrate Sentiment Analysis using SAS Enterprise Miner software.

This paper is organized as follows: Section 2 presents some related studies on sentiment analysis. In Section 3 we present the methodology followed by the results and discussions for two case studies. Finally, Section 4 concludes the study.

LITERATURE REVIEW

Current technology trends enable individuals to express their opinions or thoughts about any topic through social media platforms. Online forums, blogs, Facebook, and Twitter are among the well-known platforms where individuals can express themselves. The ability to accurately predict public sentiment about a particular issue, brand, or product can be valuable in marketing research and even in detecting cyber risks and security threats (Bai, 2011; Mostafa, 2013).

SA has been widely used by researchers from various areas including affective computing, linguistics, and psychological studies (Yasavur et al., 2014). Recently, SA through

the text mining approach has attracted the interest of many researchers because of its widely available subjective texts from blog posts, online forums, and social network site postings. SA is expressed as opinion mining, a sub-discipline of data mining and computational linguistics that refers to the computational techniques for extracting, classifying, understanding, and assessing the opinions expressed in various online news sources, social media comments, and other user-generated content (Chen and Zimbra, 2010). SA is generally conducted at document, sentence, or entity levels. Document-level SA focuses on the task of classifying a textual review on a single topic as a positive or a negative sentiment. It considers the whole document as a basic information unit that addresses one topic (Liu, 2012). There are two approaches available in the classification of sentiments in textual comments: machine learning-based approach and lexicon-based approach. The machine learning-based text classification can be divided into supervised and unsupervised learning while the lexicon-based classification can be grouped into dictionary-based approach and corpus-based approach (Nor Nadiah et al., 2015; Wang et al., 2014).

SA has been performed on open-ended questions in Quality of Life surveys (Cerrito, 2011), surveys on consumer confidence and political opinion in the 2008–2009 period (O'Çonnor et al., 2010) and identifying the political orientation by commenters' sentiment patterns towards political news articles and consequently their political inclinations from the sentiments expressed in the comments (Park et al., 2011). Netzer et al. (2012) used customer reviews from Web 2.0, the gathering places for internet users in blogs, forums and chat rooms to gain consumers' insight towards sedan cars and diabetes drugs without having to interview them. SA on Twitter postings evaluated consumers' sentiment towards well-known brands such as Nokia, T-Mobile, International Business Machines Corporation (IBM), Koninklijke Luchtvaart Maatschappij (KLM) and Dalsey, Hillblom & Lynn (DHL) (Mostafa,2013) and twitter postings of movie reviews and classify them into two classes, good and bad movies (Abd Samad et al, 2013). Asyraf et al. (2017) used SA of online financial reviews to predict stock price.

METHODOLOGY

A. CASE STUDY 1-Text Mining

This section presents the Facebook word cloud analysis using R via the Facebook application programming interface (API). The methodology comprises three phases: Text retrieval, text pre-processing, and word cloud corpus generation.

Text Retrieval

The first phase involves obtaining the textual data. The data used for this study was retrieved from a Facebook page of Malaysia's leading automotive website, paultan.org (https://www.facebook.com/paultanautonews). A sample post is shown in Figure 1.

Figure 1. Sample post on Paultan's Auto News Facebook page

This website contains various reviews pertaining to the automotive scene in ASEAN. Paultan.org is an interactive website, whereby every user can post their comments, and they are free to express their opinion whether it is positive or negative. Each post on the website will be linked to this Facebook page, and every Facebook users are free to express their comments and opinions through this platform. Thus, this page has become a suitable medium for car lovers to interact with each other regarding what they like or dislike about anything posted in the website. We selected every comment regarding PROTON cars posted from April 2016 until June 2017. This time frame was selected as such to capture any postings regarding car models which were currently in production line in Shah Alam and Tajung Malim. The car models are Saga, Persona, Exora, Preve, Suprima S, and Iriz.

A total of 5000 comments were extracted from the Facebook using R. First, App IP and App Secret were generated from the Facebook API at *https://developers.facebook.com*. These App ID and App Secret are shown in Figure 2.

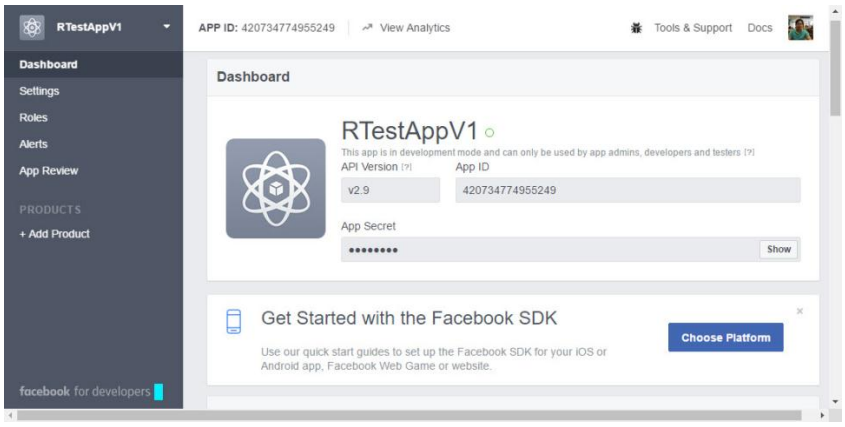

Figure 2. Facebook app dashboard

The access to the Facebook Graph API can be done using R package Rfacebook which is available on CRAN and GitHub. The personalized App ID and App Secret were used in the R script for the extraction process. Out of 5000 posts only 277 were identified to be related to Proton and the results are exported into an excel file with 12 variables as illustrated in Figure 3. Only comments posted in English were filtered for the study to reduce the complexity when handling multilingual text data. Figure 3 shows a sample of the extracted posts.

	A	B	C	D	E	F	G	H	I	J	K	L
1		from_id	from_nan	message	created_time	type	link	id	story	likes_cou	comment	shares_count
2	5	14779190194	Paul Tan's	Yes, that's a Proton Saga being driven around Kuala Lumpur... in reve	2017-07-03T05:30:	link	https://pa	147791901	#N/A	320	85	94
3	9	14779190194	Paul Tan's	The Mellors Elliot Motorsport-prepared Proton Iriz has taken on the f	2017-07-03T03:30:	link	https://pa	147791901	#N/A	243	12	25
4	22	14779190194	Paul Tan's	Brought to you by the outfit which prepared the title-winning Satria '	2017-06-30T06:52:	link	https://pa	147791901	#N/A	694	40	67
5	58	14779190194	Paul Tan's	A Proton SUV about the size of a Honda CR-V, and priced at under RM	2017-06-28T07:40:	link	https://pa	147791901	#N/A	5202	641	671
6	71	14779190194	Paul Tan's	Geely already has some rough ideas of what it has in store for Proton	2017-06-23T11:15:	link	https://pa	147791901	#N/A	836	67	48
7	72	14779190194	Paul Tan's	Geely has big plans for the Proton brand, and has "reserved" the ASE	2017-06-23T10:45:	link	https://pa	147791901	#N/A	441	20	10
8	73	14779190194	Paul Tan's	With the Proton-Geely definitive agreement signed, DRB-Hicom has	2017-06-23T10:15:	link	https://pa	147791901	#N/A	424	20	68
9	74	14779190194	Paul Tan's	Take a good look at the Geely Boyue, which will form the basis of Pro	2017-06-23T09:42:	video	https://w	147791901	#N/A	4245	458	1531
10	76	14779190194	Paul Tan's	The Geely Boyue will become Proton's first SUV.https://paul.my/Gee	2017-06-23T08:34:	video	https://w	147791901	#N/A	1180	202	182
11	78	14779190194	Paul Tan's	Despite both having similar brand positioning, Geely will not encroa	2017-06-23T05:32:	link	https://pa	147791901	#N/A	975	54	58
12	79	14779190194	Paul Tan's	The Proton-Geely strategic partnership has officially begun with tod	2017-06-23T03:53:	link	https://pa	147791901	#N/A	316	40	21
13	80	14779190194	Paul Tan's	Proton's seven-seater has received revisions for 2017, with detail ch	2017-06-23T03:22:	link	https://pa	147791901	#N/A	1501	152	124
14	81	14779190194	Paul Tan's	The Geely Boyue - the base for Proton's first ever SUV - has made its	2017-06-23T02:37:	link	https://pa	147791901	#N/A	4804	500	723
15	82	14779190194	Paul Tan's	LIVE: The signing ceremony between DRB-Hicom (Proton) and Geely,	2017-06-23T01:41:	video	https://w	147791901	Paul Tan's	1168	1631	438
16	121	14779190194	Paul Tan's	Proton is offering free safety inspections at its service centres and al	2017-06-20T08:35:	link	https://pa	147791901	#N/A	46	1	7
17	141	14779190194	Paul Tan's	The government says Proton's strategic partnership with Chinese aut	2017-06-19T07:00:	link	https://pa	147791901	#N/A	137	5	3
18	251	14779190194	Paul Tan's	In conjunction with Hari Raya Aidilfitri celebrations, Proton is offerin	2017-06-09T07:00:	link	https://pa	147791901	#N/A	32	0	0
19	305	14779190194	Paul Tan's	A more streamlined variant line-up, improved NVH and revised equi	2017-06-05T10:01:	link	https://pa	147791901	#N/A	638	46	52
20	313	14779190194	Paul Tan's	This is the Proton Bayu, a design offering a take of what the national	2017-06-05T04:30:	link	https://pa	147791901	#N/A	3673	310	703

Figure 3. Sample of extracted post in Excel

Text pre-processing

Textual datasets are unstructured data that must be converted into a structured form. Text cleaning is a task of eliminating the irrelevant words that may not be meaningful to the documents, namely punctuation, numbers, website links, black spaces, special and characters. In addition to text cleaning, the removal of English stop words and Bahasa Malaysia words were also performed in R. The text mining process is summarized in Table 1.

Table 1. Text mining process

Step	Action	Description	Tools
1	Text Retrieval	Fetching data from Facebook and transform it into a structured dataset to be used in pre-processing stage.	• R • Facebook API • Ms Office Excel (2007)
2	Text Processing	Decompose textual data, and generate a quantitative representation that is suitable for wordcloud analysis. The process includes: • text cleaning • removal of stop words • removal of Bahasa Melayu wods	• R • Ms Office Excel (2007)
3	Wordcloud Corpus	Corpus is generated	• R

Generating word cloud using R

Text mining can be easily carried out using R (Silge, 2017; Graham, 2016). For this case study, the necessary steps for generating word cloud are as follows:

Step 1: Install text mining and other related packages

```
install.packages("twitteR")
install.packages("wordcloud")
install.packages("ROAuth")
install.packages("tm")
install.packages("NLP")
install.packages("RColorBrewer")
install.packages("SnowballC")
```

Step 2: Access paultanautonews Facebook comments related to Proton

```
fb_page <-
getPage(page="paultanautonews",
token=fb_oauth, n=5000,
          since='2016/04/01',
until='2017/06/31')

# number of post
nrow(fb_page)


# number of comments
sum(fb_page$comments_count)

# row with Proton
grep("Proton", fb_page$message)

# post with Proton
ProtonSubset <-
fb_page[grep("Proton",
fb_page$message), ]

# number of post with Proton
nrow(ProtonSubset)

# number of comments with Proton
sum(ProtonSubset$comments_count)

# Proton post
write.xlsx(ProtonSubset,
paste0(getwd(),"/proton_post.xlsx")
)
```

Step 3: Clean text function by removing punctuation, @, numbers, links, English stopwords and Bahasa Melayu words

```
# remove punctuation
x = gsub("[[:punct:]]", "", x)

# remove @ using gsub
x = gsub("@\\w+", "", x)

# remove numbers using gsub
x = gsub("[[:digit:]]", "", x)

# remove links http using gsub
x = gsub("http\\S+", "", x)
#match any non white space after
http\\
```

```
# remove tabs using gsub
x = gsub("[ |\t]{2,}", "", x)

# remove blank spaces at the
beginning using gsub
x = gsub("^ +", "", x)

# remove blank spaces at the end
using gsub
x = gsub(" +$", "", x)

# remove special characters using
gsub
x =  gsub("[^[:alnum:]' ]", "", x)

# tolower case
x = tolower(x)

myStopwords <-
c(stopwords('english'), addstop)
mystring <- readLines("bm.txt")
```

Step 4: Build Corpus

```
r_stats_text_corpus <-
Corpus(VectorSource(r_stats_text))
```

Step 5: Generate Wordcloud

```
wordcloud(r_stats_text_corpus,
scale=c(8,0.7), max.words=80,
random.order=FALSE, rot.per=0.2,
use.r.layout=FALSE, colors=pal2)
```

CASE STUDY 1: Results & Discussions

Figure 4 shows the top eight terms and their frequency in the comments based on the Paultan's post. Many of the commentators are intrigued by the upcoming Proton SUV which will be based on the Chinese Geely Boyue. At the moment, Geely Boyue SUVs are in production and on sale in China. It has a wide range of good specifications and design that made the netizen curious about the price. As such, words namely, 'China', 'like', 'price' and 'SUV are trending in the comments.

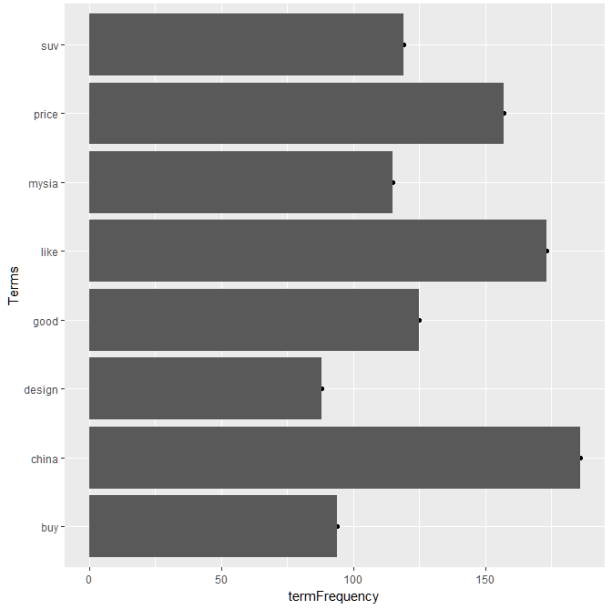

Figure 4. Frequency of top eight words

The top five words with its frequency value are shown in Table 2. A word cloud is a graphical representation of word frequency. The word cloud created

using R for posts on Proton is shown in Figure 5. Thus analysis of unstructured data reveals that price and design are being emphasized in discussions about Proton cars.

Table 2. Top five words in FB posts on Proton cars

Word	Frequency
China	186
like	173
price	157
good	125
SUV	119

Fig. 5. Word Cloud associated with Proton Comment Post

B. CASE STUDY 2: Sentiment Analysis

Text Retrieval

In Case Study 2, comments regarding PROTON cars posted from July 2011 until July 2015 were selected for sentiment analysis. This time frame was selected as such to capture any postings regarding car models which were currently in production line in Shah Alam and Tajung Malim. The car models are Saga, Persona, Exora, Preve, Suprima S, and Iriz. A total of 5330 comments were extracted from the Paultan website. Only comments posted in English were used for the study to reduce the complexity when handling multilingual text data.

Data Pre-processing

Textual datasets are unstructured data that must be converted into a structured form before classification can be performed. Data pre-processing is a task of eliminating irrelevant words that may not be meaningful. It consists of three main steps which are tokenization, removal of stop words and stemming. All three processes were performed using the Text Parsing node in SAS Enterprise Miner 14.1. The text mining algorithm is available in SAS Enterprise Miner, under the text mining tab. The text mining process is summarized in Table 3.

Table 3: Text mining process

Step	Action	Description	Tools
1	Text Retrieval	Fetching data from Facebook and transform it into a structured dataset to be used in pre-processing stage.	• Facepagers • SAS 9.3 • Ms Office Excel (2016)
2	Data Pre-processing	Decompose textual data, and generate a quantitative representation that is suitable for data mining purposes. The process includes: • stemming • tokenization • removal of stop words • part-of-speech tagging • synonyms identification	• SAS E-Miner 14.1 • Ms Office Excel (2016)
3	Feature Selection	Selecting important features by using Chi-Square and IG	• SAS E-Miner 14.1
4	Document Classification	• Assigning sentiments polarity to each document (positive/negative) • Applying SVM classifier	• SAS E-Miner 14.1

Feature Selection

Text filtering involves the process of transforming the pre-processed text data into a more compact form by reducing its dimensions. The filtering step is the feature selection stage. Feature selection (FS) is the process of extracting the main terms of the text. For a classification algorithm to perform well, it is important to identify the features that are descriptive of the text. The data pre-processing stage is the process of cleansing the textual data from the words that is considered as "noise" while the FS stage is to further filter the pre-processed text particularly to increase the classification performance.

Terms reduction were performed by mapping the terms extracted with a library of sentiment words which was compiled by (Hu and Liu, 2004). This library of sentiment words, or also known as opinion lexicon consists of 2006 positive words and 4783 negative words that are commonly found in product reviews. By mapping this list with the extracted terms from the Paultan's Auto news comments, a set of 128 terms that convey sentiment orientation specifically in

Proton car reviews from Paultan's Auto News Facebook page was obtained. The feature extraction process is shown in Figure 6.

For further term reduction, the FS were performed to identify the words that express sentiments. We used two feature selection methods which are Information Gain (IG) (Rogers et al., 2015) and Chi-Square statistics (CHI) (Firmino, et al., 2014). IG is one of the popular feature selection methods in sentiment classification. It was used to measure the obtained information for a category prediction by knowing the presence or absence of a term in a document. IG is calculated for each term depending on how much more information is gained with respect to the class (positive or negative). Chi-square statistics is also widely used in feature selection and text classification.

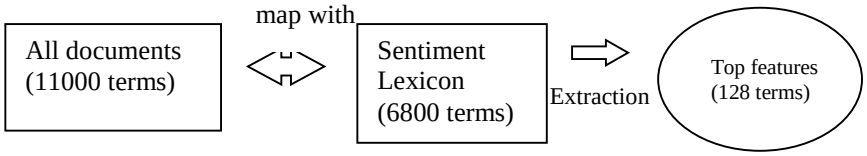

Figure 6. Feature extraction process

Document Classification

To assign the sentiment polarity to each document, the same pre-processing procedure was performed on each document. The terms extracted for each document were then mapped with the 128 top features. In assigning sentiment polarity, if the frequency of positive terms is higher than negative terms, then will be considered as a positive document and vice versa. Out of the 60 documents, 39 (65%) posts were positive comments from the public while 21 (35%) posts were negative comments.

Support Vector Machine (SVM) was used as the classifier to assign the pre-processed documents into the sentiment categories (positive/negative). SVM is a supervised machine learning method which performed classification analysis by constructing a set of hyperplanes that maximize the margin between two sentiment categories using kernel functions. There are four kernel functions, namely linear function, polynomial function, radial basis function as well as sigmoid function. This study compared the SVM (Bui et al; 2016) classification performance for the linear and radial basis function (RBF) kernel. The CHI and IG feature selection methods are available in the Decision Tree node. RBF is the most popular kernel function because it can work well in case of nonlinear separation between datasets (Flyer, 2014). The classification flow using SAS Enterprise Miner 14.1 is shown in Figure 7. The data source node was connected to the Decision Tree node (here CHI and IG can be set) for feature selection. The CHI and IG were then connected to SVM node for classification of the sentiments based on the selected features. We also evaluated the performance of SVM when there was no feature selection.

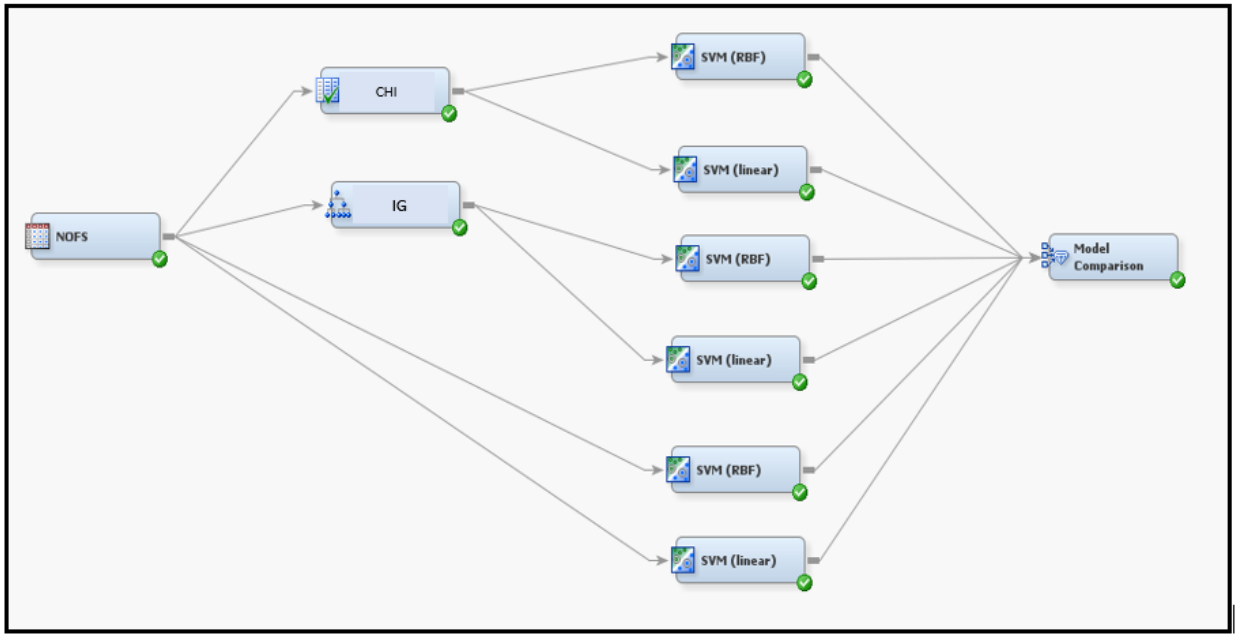

Figure 7. Classification using SAS Enterprise Miner 14.1

Model Evaluation Phase

We used five-fold cross-validation to evaluate the performance of SVM. In five-fold cross-validation method, the sample was partitioned into five sections, where one set was used for testing and the remaining four sets for training the model. The process was repeated five times, and the average of the performance measures were obtained. The performance metrics used in this study were accuracy, recall, precision and F-measure value. Accuracy rate represents the percentage of correctly classified cases, whereby the higher accuracy rate indicates better classification model performance. Other than accuracy, the information regarding true positive rate (sensitivity or recall) and specificity (true negative rate) can also be obtained from the confusion matrix. Another standard evaluation metric used in text categorization studies is F-measure (Chen and Chen, 2014; Hamouda, 2013). The F-Measure is the harmonic mean of precision and recall (Janez et al., 2011). Precision refers to the ratio of true positive to all instances predicted as positive. Recall on the other hand is the ratio of true positive to all instances that are actually positive (also known as sensitivity).

CASE STUDY 2: Results And Discussions

The retrieved comments from Facebook were transformed into a structured form to enable the performance of the classification task. The dimension of the structured dataset is 60 × 128 (60 documents × 128 features). Feature selection was then applied to further reduce the data dimensionality. Two types of feature selection methods were explored in this study: CHI and IG selection methods. The significant features for CHI were selected using 5% significance level, while the significant features for IG were selected by using the entropy splitting criteria of the decision tree. The CHI method selected eight features while IG selected only six features. The result of the feature selection stage is summarized in Table 4.

Table 4. Selected Features using CHI and IG methods

FS method	CHI	IG
Selection criteria	Chi-Square statistics, α=0.05	Entropy = $-[\sum_{i=1}^{2} p_i \log_2(p_i)]$ Information Gain=1 - Entropy
Selected features	Beg, jam, kill, nice, bull, hell, spacious, top	Ugly, proud, jam, wrong, scrap, easy

The results in Table 5 show that the highest classification accuracy and F-measure were IG selection method and by applying linear kernel as a mapping function for the SVM. The SVM classifier with RBF kernel using the IG selection method also results in generally high accuracy and average F-measure value compared with other combinations This experimental result implies that a dataset with IG FS is linearly separated and that a linear kernel function provides higher classification accuracy. The findings also reveal the SVM classification performance works well with Feature Selection methods. Tan and Zhang (2008) also reported that IG performs the best for sentimental terms selection and SVM exhibits the best performance for sentiment classification.

Table 5. Classification Performance using the Paultan Dataset

FS method	SVM kernel	Accuracy (%)	Precision	Recall	F-measure
CHI	Linear	66.67	0.711	0.821	0.762
	RBF	68.33	0.672	1.000	0.804
IG	Linear	**73.33**	**0.767**	**0.846**	**0.805**
	RBF	70.00	0.723	0.872	0.791
Without feature selection	Linear	68.33	0.700	0.897	0.787
	RBF	65.00	0.650	1.000	0.788

All the comments were passed into an online word-cloud generator Tagul.com, and the words are displayed Figure 8 and Figure 9. Words that are frequently used in the positive documents are "look good", "buy", "better" and "nice". In the negative documents, the frequent words are "value", "problem" and "price".

We need to take note that bias is nearly always present in any published study due to sampling, data collection or in the process of data analysis. In analyzing unstructured Facebook data, bias may result from comments by a specific group of people with special satisfaction status or with specific good (promoting) or bad (rivalry) intention of posting their comments. This bias can be overcome by checking on the frequency and consistency of the person posting the comments. Comment that are prejudice and with clear bias intentions should be removed before performing text or sentiment analysis as the aim is to get information from the sentiments of the general public about a certain product.

Figure 8. Positive Documents

Figure 9. Negative Documents

CONCLUSION

The automotive scene in Malaysia has become very competitive, and Proton is no longer Malaysians' first choice when purchasing cars. Thus, the need to understand consumers' brand preference and perception has become increasingly important. Text mining and sentiment analysis are useful techniques for analyzing product reviews or comments posted in social media. Classifiers such as SVM are useful in classification of sentiments. Overall, there were more positive than negative comments toward Proton cars which indicate that consumers were quite happy with Proton's new car models. Car manufacturers should capitalize on sentiment analysis to gain insights on the consumers' feelings and opinions as such information are valuable and can assist in decision making on designs, production and sales of their car models. The results of the text and sentiment analysis show that the comments focus on price, design, value and problem with regards to Proton cars. These are important factors for car manufacturers to focus on in order for them to increase the sales and good perception of Proton cars. Car manufacturers should capitalize on text and sentiment analysis on product reviews or customer service survey to optimize their sales, profits and customer satisfaction.

ACKNOWLEDGMENT

The authors would like to thank IRMI (Institute of Research Management and Innovation), Universiti Teknologi MARA for the financial support under the REI research grant (600-RMI/DANA 5/3/REI (16/2015)).

REFERENCES

Abd Samad, H. B., Hussin, B., Ananta, I. G. P. & Zeniarja, J. (2013). Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization Procedia Engineering, 53, 453–462.

Asyraf, A. S., Abdul-Rahman, S., & Mutalib, S. Mining textual terms for stock market prediction analysis using financial news. In *Soft Computing in Data Science - 3rd International Conference, SCDS 2017, Proceedings* (Vol. 788, pp. 293-305). (Communications in Computer and Information Science; Vol. 788). Springer Verlag. DOI: 10.1007/978-981-10-7242-0_25

Bai, X. (2011). Predicting consumer sentiments from online text. Decision Support Systems, 50(4), 732-742.

Bui, D. T., Tuan, T. A., Klempe, H., Pradhan, B. & Revhaug, I. (2016). Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. *Landslides*, 13(2), 361-378.

Cerrito, P. (2011). Sentiment mining using SAS® text miner®. In SAS Global Forum, 2011 .

Chen, G. & Chen, L. (2014). Recommendation based on contextual opinions. In User Modeling, Adaptation, and Personalization (pp. 61-73). 2014. Springer International Publishing.

Chen, H. & Zimbra, D. (2010). AI and Opinion Mining, IEEE Intelligent Systems, 25(3), 74-80.

Choi, J. & Lee, J. K. (2015). Investigating the effects of news sharing and political interest on social media network heterogeneity. Computers in Human Behavior, 44, 258-266.

Firmino Alves, A. L., Baptista, C. D. S., Firmino, A. A., Oliveira, M. G. D., & Paiva, A. C. D.A (2014). Comparison of SVM versus naive-bayes techniques for sentiment analysis in tweets: a case study with the 2013 FIFA confederations cup. In Proceedings of the 20th Brazilian Symposium on Multimedia and the Web, 2014, pp. 123-130.

Flyer, N., Wright, G. B. & Fornberg, B. (2014). Radial basis function-generated finite differences: A mesh-free method for computational geosciences. Handbook of Geomathematics, Springer, Berlin.

Graham, W. (2016). Hands-on Data with R Text Mining. Available at: http://onepager.togaware.com/TextMiningO.pdf

Hamouda, S. B. & Akaichi, J. (2013). Social networks' text mining for sentiment classification: The case of Facebook' statuses updates in the 'Arabic Spring' era. International Journal Application or Innovation in Engineering and Management, 2(5), 470-478.

Hu, M. & Liu, B. (2004). Mining and Summarizing Customer Reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 168-177.

Janez,B., Mladenić,D. & Grobelnik,M. (2011). Feature Construction in Text Mining. Encyclopedia of Machine Learning. Springer US, 397-401.

Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

Mostafa, M. M. (2013). More than words : Social networks ' text mining for consumer brand sentiments. *Expert Systems With Applications*, *40*(10), 4241–4251.

Netzer, O., Feldman, R., Goldenberg, J. & Fresko, M.(2012). Mine your own business: Market structure surveillance through text mining. Marketing Science, 31(3), 521-543.

Nor Nadiah Yusof, Azlinah Mohamed & Shuzlina Abdul Rahman. (2015). Reviewing Classification Approaches in Sentiment Analysis. Soft Computing in Data Science, SCDS2015, 545,43-53. Springer CCIS.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. ICWSM, 11(122-129), 1-2.

Oeldorf-Hirsch, A. & Sundar, S. S. (2015). Posting, commenting, and tagging: Effects of sharing news stories on Facebook. Computers in Human Behavior, 44, 240-249.

Pang, B. & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1-2), 1-135.

Park, S., Ko, M., Kim, J., Liu, Y., & Song, J. (2011). The politics of comments: predicting political orientation of news stories with commenters' sentiment patterns. In Proceedings of the ACM 2011 conference on Computer supported cooperative work, pp. 113-122.

Rogers, B., Qiao, Y., Gung, J., Mathur & T., Burge, J. E. (2015). Using text mining techniques to extract rationale from existing documentation. In

Design Computing and Cognition'14, 457-474. Springer International Publishing.

Silge, J. Robinson, D. (2017) Text Mining with R: A Tidy Approach. Available at: http://tidytextmining.com/index.html

Tan S. & Zhang, J. (2008). An empirical study of sentiment analysis for Chinese documents. Expert Systems With Applications, 34(4):2622–2629.

Wang, H., Liu, L., Song, W., & Lu, J. (2014). Feature-based sentiment analysis approach for product reviews. Journal of Software, 9(2), 274-279

Yasavur, U., Travieso, J., Lisetti, C. L., & Rishe, N. D. (2014). Sentiment Analysis Using Dependency Trees and Named-Entities. In Proceedings of the 27th International InFLAIRS Conference, 134-139.