International Journal of Electrical and Computer Engineering (IJECE)
Vol.8, No.6, December 2018, pp. 5153~5161
ISSN: 2088-8708, DOI: 10.11591/ijece.v8i6.pp5153-5161

Analysis of Mobile Service Providers Performance Using Naive Bayes Data Mining Technique

M. A. Burhanuddin[1], Ronizam Ismail[2], Nurul Izzaimah[3], Ali Abdul-Jabbar Mohammed[4], Norzaimah Zainol[5]
[1, 3, 4]Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Malaysia
[2, 5]Faculty of Science and Technology, Kolej Universiti Islam Melaka, Malaysia

Article Info

Article history:

Received Mar 11, 2018
Revised Jul 4, 2018
Accepted Jul 14, 2018

Keyword:

Big data
Data mining
Data science
Mobile services
Naive Bayes algorithm
Telecommunication services

ABSTRACT

Recently, the mobile service providers have been growing rapidly in Malaysia. In this paper, we propose analytical method to find best telecommunication provider by visualizing their performance among telecommunication service providers in Malaysia, i.e. TM Berhad, Celcom, Maxis, U-Mobile, etc. This paperuses data mining technique to evaluate the performanceof telecommunication service providers using their customers feedback from Twitter Inc. It demonstrates on how the system could process and then interpret the big data into a simple graph or visualization format. In addition, build a computerized tool and recommend data analytic model based on the collected result. From prepping the data for pre-processing until conducting analysis, this project is focusing on the process of data science itself where Cross Industry Standard Process for Data Mining (CRISP-DM) methodology will be used as a reference. The analysis was developed by using R language and R Studio packages. From the result, it shows that Telco 4 is the best as it received highest positive scores from the tweet data. In contrast, Telco 3 should improve their performance as having less positive feedback from their customers via tweet data. This project bring insights of how the telecommunication industries can analyze tweet data from their customers. Malaysia telecommunication industry will get the benefit by improving their customer satisfaction and business growth. Besides, it will give the awareness to the telecommunication user of updated review from other users.

Corresponding Author:

Ali Abdul-Jabbar Mohammed,
Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka,
Universiti Teknikal Malaysia Melaka,
Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia.
Email: p031610009@student.utem.edu.my

1. INTRODUCTION

The main regulator and governor of telecommunications and its rules in Malaysia is the Malaysian Communications and Multimedia Commission [1], [2]. Regulatory reforms and rehabilitation are very important aspectsin creating competition effectiveness among the industry of telecommunications. Correspondingly, the Malaysian telecommunications industry has been exceptional growth in recent years [3]. Therefore, this leads to produce a huge and diverse data sets i.e., big data, which is need analytics and investigation to discover hidden correlations, customer preferences, market trends, and further valuable information that may help organizations make better business decisions. Problem arises, with the growing field of big data, utilization of structured and unstructured data leads to worthy information for telecommunications industry in Malaysia to grow exponentially [4]. Consequently, issues on utilization of structured and unstructured data requires critical and analytical methods to overcome the needs of industry

growth [5], [6]. There are many challenges to be faced for finding out the best telecommunication service provider since nowadays there are too many choices of mobile communication services with a different service rates and speeds [7].

The contribution of this study is to give a solution for evaluating the performance of telecommunication service providers inthe Malaysian telecommunications industry, this is by:

- Analyzing huge and diverse data giving by the telecommunication service users using their twitter accounts daily.
- Ranking the performance of the telecommunication service providers in Malaysia based on the tweets data of their users.

2. RESEARCH METHOD

From prepping data for pre-processing until conducting analysis, the scope of this project is focusing on the process of data science itself. The method used in this study, is based on Cross Industry Standard Process for Data Mining (CRISP-DM) [8], as this model is well-known in the data mining process [9]–[11]. The complete process diagram of CRISP-DM is given in the Figure 1 and followed by the description for each process included in the model.

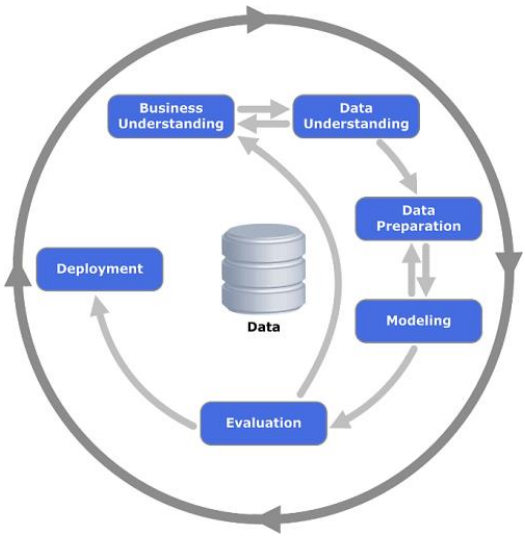


Figure 1. Cross Industry Standard Process for Data Mining (CRISP-DM) model [8]

From Figure 1, the business understanding process focuses on the purposes and requirements of the project, which comprises understanding the business objectives, success criteria, project plan, and deliveries [12]–[14]. The data understanding process starts with an initial data collection and manage to proceed with the data description and data exploration. The data preparation process includes data cleaning, sampling, normalization, and feature selection. The modeling process includes select modeling techniques, building, and training the model, in addition to make prediction. The evaluation process includes the model validation, review the results, and success criteria evaluation. Finally, the deployment process includes result visualization, and the report creation. Therefore, the method that suits our sentiment analysis for telecommunication business operation is defined in the workflow that given in Figure 2.

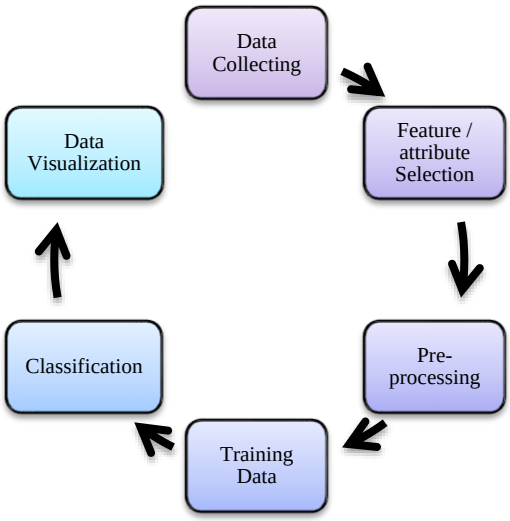


Figure 2. Sentiment Analysis Flow

The computer program in this project is written using R Studio and R language which is a programming language for statistical computing and graphics. While the data that will be used during the test, gathered from the Twitter Application Platform Interphase (API). For the user that want to access the data from Twitter API need to have the Twitter account. However, the first step before beginning the code, R studio needs an API key to synchronize it with the Twitter API. After the synchronize success, the data can be gathered freely from the Twitter API, but the R studio can access only the data within seven days before the request date.

For the big data analysis, Naïve Bayes technique is deployed in this project to obtain the result from big datato produce the most accurate result. The Naïve Bayes classifier is a supervised learning and one of the simple probabilistic classifier techniques in the Machine Learning course with strong (naive) independence assumptions between the features [15]–[17]. The Figure 3 is showing the processes flowchart of Naïve Bayes Technique.

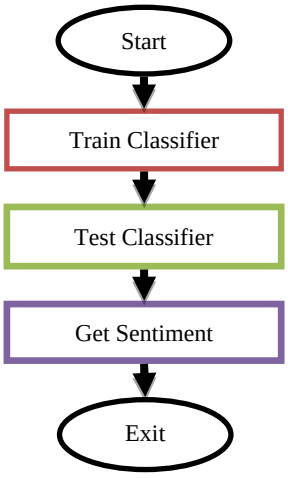


Figure 3. Naïve Bayes Technique Flowchart

The train classifier can be used for training the data to calculate Bayes-optimal estimates and make predictions of the model parameters [18]–[20]. The process flowchart of the train classifier that applied in this project is given in Figure 4.

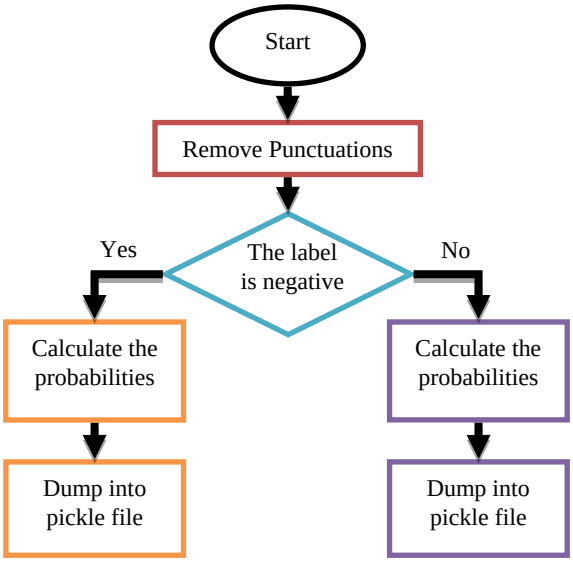


Figure 4. Train Classifier

The Figure 5 shows how Naïve Bayes works in the test set classifier for sentiment data. This is appropriately representative intended for the underlying recognition problem, that leads to worthy information for telecommunications industry in Malaysia to grow exponentially.

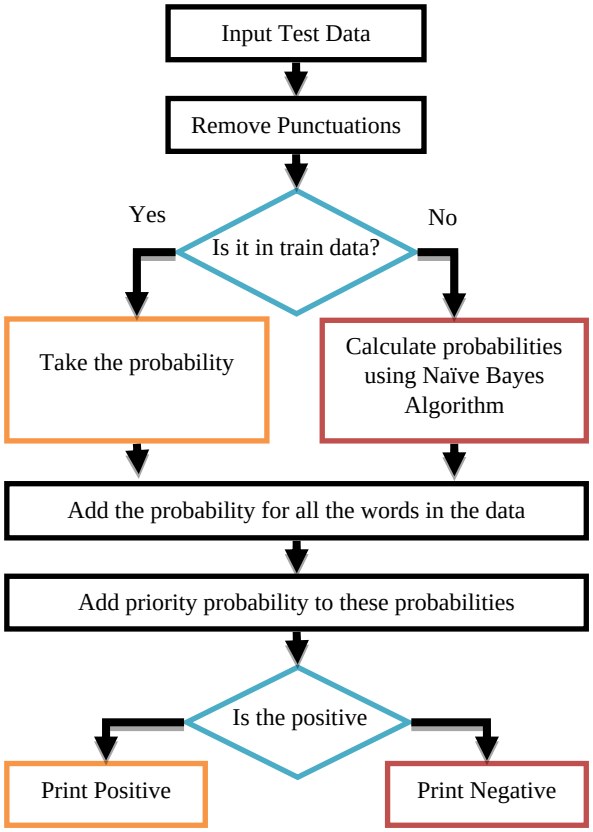


Figure 5. Test Classifier to Get Sentiment Result

The figures above show the methodology of how to get the results from our analysis. Consequently, the following is a brief explanation including step by step of how Naïve Bayes technique work. This can be detailed as:

- Step 1: determining the test set in our dataset as the followingin Table 1.

Table 1. Test Set

DOC	TEXT	CLASS
1	I loved the service	+
2	I hated the service	-
3	A great service, good service	+
4	Poor service, Poor connection	-
5	A good service, great connection	+

So, a total of 10 unique words eg. I, loved, the, service, a, great, hated, good, connection, poor.

- Step 2: converting the data into a frequency table, which is given in Table 2 as follows:

Table 2. Frequency Table

DOC	1	2	3	4	5
I	1	1			
loved	1				
the	1	1			
service	1	1	2	1	1
hated		1			
a			1		1
great			1		1
poor				1	
connection				1	1
good			1		1
Class	+	-	+	-	+

Next, look at the probabilities per outcome (+ or -)

- Step 3: Compute the priority

  P (+) = total of + class

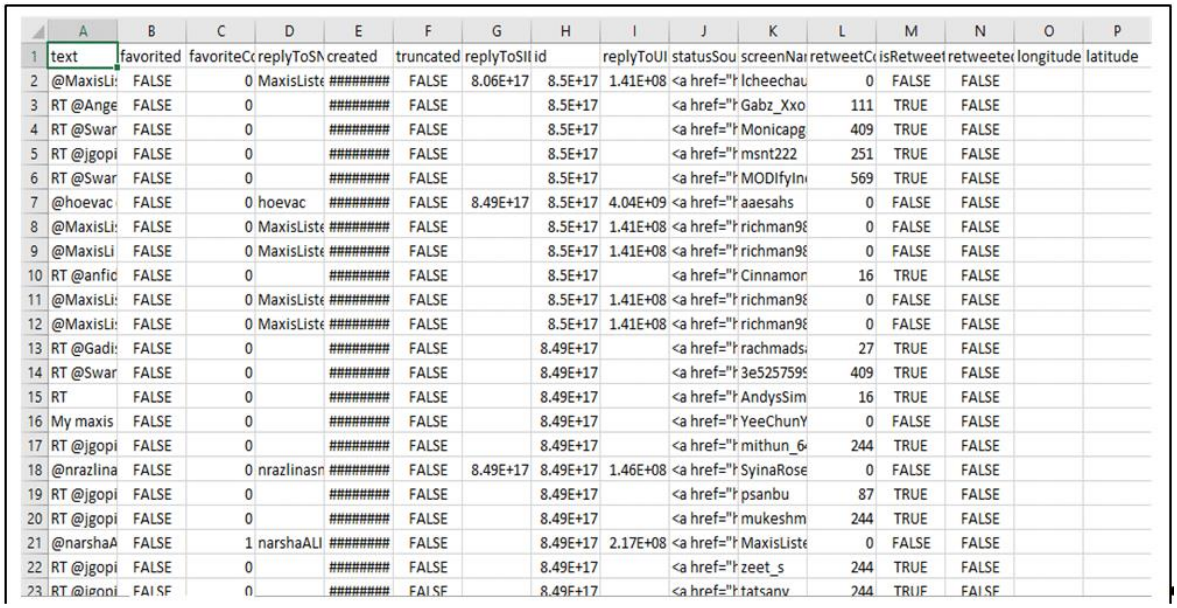
  P (-) = total of - class

- Step 4: Compute the conditional probability / possibility of each attribute

P(I|+); p(loved|+); P(the|+); P(service|+); P(a|+); P(great|+); P(good|+); P(connection |+); P(wk.|+) =
nk: number of times word k occurs in these cases (+)
n: number of words in (+) case -> 14
vocabulary: total unique words while testing, for unknown words we use nk = 0 and find its probability being both positive and negative.

3. DATA ANALYSIS

In this study, we are using a real data extracted from Twitter API, a website uses to access core Twitter data. Consequently, we save the data into .csv file format as given in Figure 6. Next, dataset is loaded in R studio for further analyses.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	text	favorited	favoriteCc	replyToSN	created	truncated	replyToSII	id	replyToUI	statusSou	screenNar	retweetCc	isRetweet	retweete	longitude	latitude
2	@MaxisLi:	FALSE	0	MaxisListe	########	FALSE	8.06E+17	8.5E+17	1.41E+08	<a href="h	lcheechau	0	FALSE	FALSE		
3	RT @Ange	FALSE	0		########	FALSE		8.5E+17		<a href="h	Gabz_Xxo	111	TRUE	FALSE		
4	RT @Swar	FALSE	0		########	FALSE		8.5E+17		<a href="h	Monicapg	409	TRUE	FALSE		
5	RT @jgopi	FALSE	0		########	FALSE		8.5E+17		<a href="h	msnt222	251	TRUE	FALSE		
6	RT @Swar	FALSE	0		########	FALSE		8.5E+17		<a href="h	MODIfyIn	569	TRUE	FALSE		
7	@hoevac	FALSE	0	hoevac	########	FALSE	8.49E+17	8.5E+17	4.04E+09	<a href="h	aaesahs	0	FALSE	FALSE		
8	@MaxisLi:	FALSE	0	MaxisListe	########	FALSE		8.5E+17	1.41E+08	<a href="h	richman98	0	FALSE	FALSE		
9	@MaxisLi	FALSE	0	MaxisListe	########	FALSE		8.5E+17	1.41E+08	<a href="h	richman98	0	FALSE	FALSE		
10	RT @anfic	FALSE	0		########	FALSE		8.5E+17		<a href="h	Cinnamon	16	TRUE	FALSE		
11	@MaxisLi:	FALSE	0	MaxisListe	########	FALSE		8.5E+17	1.41E+08	<a href="h	richman98	0	FALSE	FALSE		
12	@MaxisLi:	FALSE	0	MaxisListe	########	FALSE		8.5E+17	1.41E+08	<a href="h	richman98	0	FALSE	FALSE		
13	RT @Gadi:	FALSE	0		########	FALSE		8.49E+17		<a href="h	rachmads	27	TRUE	FALSE		
14	RT @Swar	FALSE	0		########	FALSE		8.49E+17		<a href="h	3e5257595	409	TRUE	FALSE		
15	RT	FALSE	0		########	FALSE		8.49E+17		<a href="h	AndysSim	16	TRUE	FALSE		
16	My maxis	FALSE	0		########	FALSE		8.49E+17		<a href="h	YeeChunY	0	FALSE	FALSE		
17	RT @jgopi	FALSE	0		########	FALSE		8.49E+17		<a href="h	mithun_6	244	TRUE	FALSE		
18	@nrazlina	FALSE	0	nrazlinasn	########	FALSE	8.49E+17	8.49E+17	1.46E+08	<a href="h	SyinaRose	0	FALSE	FALSE		
19	RT @jgopi	FALSE	0		########	FALSE		8.49E+17		<a href="h	psanbu	87	TRUE	FALSE		
20	RT @jgopi	FALSE	0		########	FALSE		8.49E+17		<a href="h	mukeshm	244	TRUE	FALSE		
21	@narshaA	FALSE	1	narshaALI	########	FALSE		8.49E+17	2.17E+08	<a href="h	MaxisListe	0	FALSE	FALSE		
22	RT @jgopi	FALSE	0		########	FALSE		8.49E+17		<a href="h	zeet_s	244	TRUE	FALSE		
23	RT @jgopi	FALSE	0		########	FALSE		8.49E+17		<a href="h	tatsany	244	TRUE	FALSE		

Figure 6. Data in .csv format

The dataset obtained from the Twitter API in our project is consist of 5 files of data according to 5 different mobile communication services providers, and these data files, includes:

1. Celcom Tweet Data
2. Maxis Tweet Data
3. Digi Tweet Data
4. U-Mobile Tweet Data
5. Tunetalk Tweet Data

All data files contain the same data attributes, these attributes are given in Figure 7.

```
> df <- read.csv(file="umobile_tweetsdf.csv",header=TRUE, sep=",")
> attributes(df)
$names
 [1] "text"          "favorited"     "favoriteCount" "replyToSN"     "created"       "truncated"     "replyToSID"
 [8] "id"            "replyToUID"    "statusSource"  "screenName"    "retweetCount"  "isRetweet"     "retweeted"
[15] "longitude"     "latitude"
```

Figure 7. Data attributes

Based on the obtained dataset and data attributes, not all the data have been applied in the analysis, only text attribute will be selected and will be used for modelling purposes. The purpose of the selected attributes is to see the weightage of the positive, negative and neutral word.

For the result of sentiment analysis, all the tweet texts have been scanned, and the score has been given. The score is based on their positivity and negativity words, which are based on the positive file and negative file. The Figure 8 is showing the tweets and its given scores.

	text	by	score
1	umobile no wonder why i alreadyused all my data lol	Umobile	1
2	umobile okay	Umobile	0
3	hawaxx weve replied your dm please check ya	Umobile	0
4	change to umlimited power plan guess what so fucking slow like snailalso add umi also slow damn snail faster laa s	Umobile	-3
5	herbelye hi thank you and we shall continue our discussion over there ya	Umobile	1
6	officialcelcom umobile free unlimited facebook instagram and twitter eduaubdedubue	Umobile	2
7	finally i change my number celcom to postpaid umobile dah boleh otp n facetime dgn mama every second every m	Umobile	0
8	zaimmsalmi hi we ceainly appreciate your continuous suppo do be informed that that soundcloud is not li	Umobile	1
9	mymaybank im using umobile and no have not changed recently was just working fine yesterday and other banks t	Umobile	1
10	qtbunny hahaha pakailah umobile postpaid p g call unlimited im umobile usereduaubdedubuceduaubdedubuc	Umobile	1
11	ftnallysha hi based on the screenshot on given perhaps may we suggest you to swap the u mobile sim into the fi	Umobile	0
12	hapoy me with umobile eduaubdedubueduaubdedubueduaubdedubueduaubdedubu	Umobile	0
13	suzanatahir umobile nye plan postpaid free call data gb sms kena charge rmgst	Umobile	1
14	walerjames maricrismoor digihotlink umobile celcom hahaha	Umobile	0
15	hai umobile idk whats going on but lately ur service ur line n ur everything suckme n my friends are thinking to swi	Umobile	0
16	umobile why the line so bad wei	Umobile	-1
17	lilianubung hi perhaps may we know does the interruption merely affecting the platforms mentioned may we sug	Umobile	-1
18	taufiqjuahir appreciate if you could reset your network settings as per belowtap settings gt general gt reset gt r	Umobile	1
19	edjunaidi farhanahjamil khairunnyssa the thing i hate about umobile is inconsistency across plan dia sebab tu dah	Umobile	-2
20	hello umobile can u please fix the internet line i couldnt get lte in kota tinggisebelum ni okay je okay tq	Umobile	0
21	lurveifa hiwe have reply your dmtq	Umobile	0

Figure 8. Tweets that already have score

These scores and results can be used to improve the customer experience and business growth by discovering unknown correlations, hidden patterns, customer preferences, market trends, and further valuable information that may help organizations make better business decisions. The technique that deployed in this project is the Naïve Bayes, which able to provide strong independence assumptions between the features related to the sentiment analysis. Furthermore, it gives the robust solution among telecommunication service providers [10].

4. FINDINGS AND RESULTS

After the score had given, the results graph is plot based on their negativity and positivity polarity as shown in Figure 9 below.

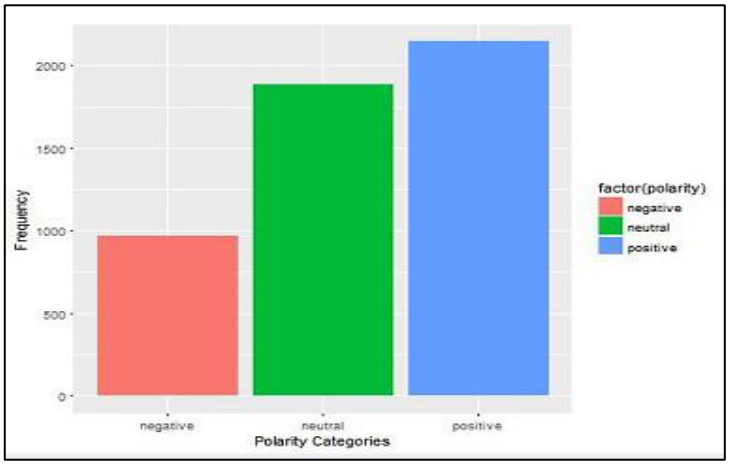


Figure 9. Polarity of the tweets

After the graph done plotting, all these results are transferred to R Shiny which is used to visualize the result in more proper and creative way. R Shiny had been chosen as its easy interphase to understand and use even for the very first-time user. Based on the Figure 10 below, we can see that there are different 5 boxes with different color and value. The value stated in the box is the amount of raw data gathered from the Twitter API that we are dealing with for this project. Based on polarity scores, telecommunication service providers ranked as Telco 1, Telco 2, Telco 3, Telco 4 and Telco 5.

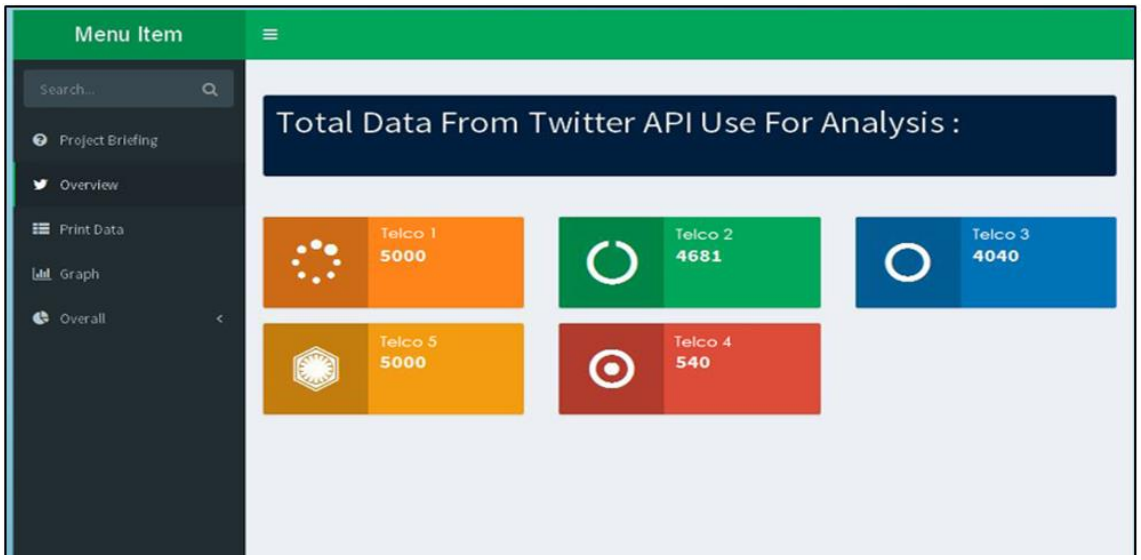


Figure 10. Overview of data

From Figure 10, highest tweet frequency come from Telco 1, which is 5000. Lowest is Telco 4, which is 540. It might be Telco 1 having highest number of customers in Malaysia. The overall module created to make a comparison between all the telecommunication service providers in Malaysia based on their positive polarity and negative polarity. The comparison is plotted in a pie chart and each of the telecommunication service providers' weightage are stated in a percentage value as shown in a Figure 11 as follows.

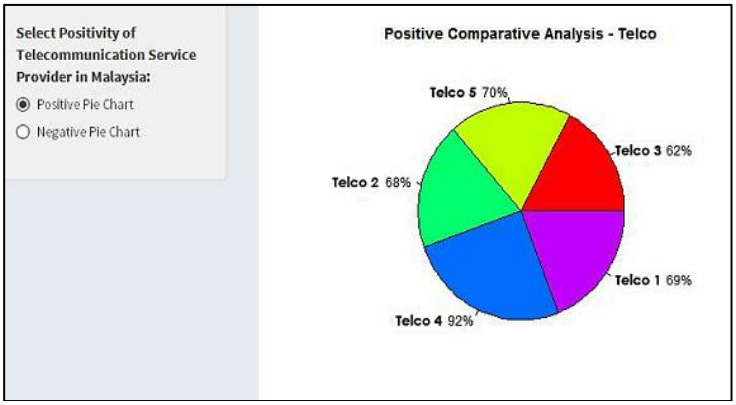


Figure 11. Summarization based on Positivity polarity

Based on the result showed in Figure 11, the telecommunication company, Telco 4 is the best, which getting 92% positive twitter comments from their customers. Lowest score is Telco 3, which is only 62% score on positive comments. By looking at this graph, telecom service providers can evaluate their performance easily from their customers' tweet data.

5. CONCLUSION

This paper shows on how to analyze and visualize tweet data, where information effectively delivered, especially towards an individual with no background in analytics or related subject. With the right visualization and graphics on time, we can improve end user understanding and at the same time creates a data interaction between the users and the information itself. Based on the project result, the service provider companies can see the graphs and their service performance from twitters. Thus, it will be able to use this project as a reference to compete with the other telecommunication service providers. However, improvement is definitely needed in every system that is developed. This is to ensure a gradual increase in user satisfaction and continues improvement of the system.

ACKNOWLEDGEMENTS

The authors would like to thank the Universiti Teknikal Malaysia Melaka, UTeM Zamalah Scheme for providing the facilities to conduct this research project and Kolej Universiti Islam Melaka for the financial support in this project.

REFERENCES

[1] S. Jayasingh and U. Eze, "An empirical analysis of consumer behavioral intention toward mobile coupons in Malaysia", *Int. J. Bus. …*, vol. 4, no. 2, pp. 221–242, 2009.
[2] "Suruhanjaya Komunikasi dan Multimedia Malaysia (MCMC)", 2017.
[3] "Axiata Annual Report", *Axiata*, 2016.
[4] N. Kshetri, "The emerging role of Big Data in key development issues: Opportunities, challenges, and concerns", *Big Data Soc.*, vol. 1, no. 2, p. 2053951714564227, 2014.
[5] N. Patil, P. Kiran, N. Kiran, N. K., "A Survey on Graph Database Management Techniques for Huge Unstructured Data", *International Journal of Electrical and Computer Engineering (IJECE)*, Vol. 8 No. 2 pp. 1140-1149, 2018.
[6] S. Borodo, S. Shamsuddin, S. H. "Big data platforms and techniques", *TELKOMNIKA Indonesian Journal of Electrical Engineering and Computer Science*, Vol. 1(1), pp. 191-200, 2016.
[7] P.S.H. Leeflang, P.C. Verhoef, P. Dahlstr??m, and T. Freundt, "Challenges and solutions for marketing in a digital era", *Eur. Manag. J.*, vol. 32, no. 1, pp. 1–12, 2014.
[8] R. Wirth and J. Hipp, "CRISP-DM : Towards a Standard Process Model for Data Mining", *Appl. Knowl. Discov. Data Min.*, no. 24959, pp. 29–39, 2000.
[9] M. Spruit, R. Vroon, and R. Batenburg, "Towards healthcare business intelligence in long-term care: An explorative case study in the Netherlands", *Comput. Human Behav.*, vol. 30, pp. 698–707, 2014.
[10] M.N.M. Ibrahim and M.Z.M. Yusoff, "Twitter sentiment classification using Naive Bayes based on trainer perception", in *2015 IEEE Conference on e-Learning, e-Management and e-Services, IC3e 2015*, 2016, pp. 187–189.
[11] P. Kalgotra and R. Sharda, "Progression analysis of signals: Extending CRISP-DM to stream analytics", in *Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016*, 2016, pp. 2880–2885.
[12] A. Tarhini, H. Ammar, T. T.-I. B. Research, and undefined 2015, "Analysis of the critical success factors for enterprise resource planning implementation from stakeholders' perspective: A systematic review", *ccsenet.org*.
[13] A. B.-I. J. of P. Management and undefined 2016, "The impact of project management (PM) and benefits management (BM) practices on project success: Towards developing a project benefits governance", *Elsevier*.
[14] C. Serra, M. K.-I. J. of P. Management, and undefined 2015, "Benefits realisation management and its influence on project success and on the execution of business strategies", *Elsevier*.
[15] P. Tsangaratos and I. Ilia, "Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size", *Catena*, vol. 145, pp. 164–179, 2016.
[16] L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S. Tiwari, "Sentiment Analysis of Review Datasets Using Naïve Bayes' and K-NN Classifier", *Int. J. Inf. Eng. Electron. Bus.*, vol. 8, no. 4, pp. 54–62, 2016.
[17] S. Shah, K. Kumar, R. S, "Sentimental Analysis of Twitter Data Using Classifier Algorithms", *International Journal of Electrical and Computer Engineering (IJECE)*, Vol. 6 No. 1, 357-366, 2016.
[18] A. McCallum and K. Nigam, "A Comparison of Event Models for Naive Bayes Text Classification", *AAAI/ICML-98 Work. Learn. Text Categ.*, pp. 41–48, 1998.
[19] S. Raschka, "Naive Bayes and Text Classification I - Introduction and Theory", *arXiv Prepr. arXiv1410.5329*, p. 20, 2014.
[20] D. Li-guo, D. Peng, L. A., "A new naive Bayes text classification algorithm", *TELKOMNIKA Indonesian Journal of Electrical Engineering and Computer Science*, Vol 12 No 2, pp. 947-952; 2014.