Sentiment Evaluation of Public Transport in Social Media using Naïve Bayes Method

NurKhaleeda Othman, MasnidaHussin, Raja Azlina Raja Mahmood

Abstract:*Nowadays, there is a trend in business organization to use social media as a medium to get feedback from customers. This gives advantage in improving the business values such as increasing customers' satisfactions and building better company reputation. However, the response and feedback from the customers are varies and hold different perspectives. It might be led to ambiguous answer.In this work, we utilized Naïve Bayes machine learning approach for analyzing sentiment at social media on transportation services. We collected all feedback from Facebook and Twitter about transportation services. From the unstructured comments and feedback, we classified accordingly to determine the related scope of the sentiment. By using the Naïve Bayes method those massivecomments and feedback are presented in appropriate way and easier to understand.*

Keywords: *Naïve Bayes machine learning, sentiment analysis, social networking, and transportation service.*

I.INTRODUCTION

Social networking that inherit from the Internet technology turn out to be the most favorite communication tool that every user up to. There are several favorite social media platforms such as Facebook, Twitter, Instagram, etc. The trend of using social media by business organizations as their customer service medium had started in early 2000 when there is perceptionon the importance of customer engagement helps in increasing company profit(Saragih & Girsang, 2017). By concerning on customers' requirements and expectation, business or organization can be planned for better company's prospect also strategize marketing and service delivery. In arise of social media era, such platform is been shifted from manual to online response mechanism. The customers seemly are happy with the changes due to it is easier for them to put comments and feedbacks towards company's' services (DhahiAlgaithi et.al, 2017; Cenni, D. et.al, 2017). The sentiment analysis is a process of identification the emotional content of opinions that voiced by customers (Goel, A. et.al,2016). The clear and transparent information can be a basis for effective decision-making process.

This further gives advantage in improving the business values such as increasing customer satisfaction and building better company reputation.Due to the online platform is generally open and free communication, it makes the comments and feedbacks become numerous and massive. Further, respondents are might not be the actual customers of the company's services. It leads to ambiguous answer that makes the online information inaccurate.

In our work, we focus on the public transportation service where it is one of important infrastructure in community. In Malaysia, Prasarana Malaysia Berhad is the main public transport-service provider thatprovided bus and rail services in Peninsular Malaysia. However, in this work we emphasis on Rapid KL services where theservices are operated only atKlang Valley area. Due to the public transportation bus, railway or taxi is significant to the country development; the customers' comments and feedback are necessity to effectively stress out. Therefore, in our work, the collection of comments and feedbacks from the Facebook and Twitter about Rapid KL by using Naïve Bayes method is realized.The remainder of this paper is organized as follows. Section 2 describes related work on existing sentiment analysis of public feedbacks from social media. Section 3 details the Naïve Bayes method used in the paper. Experimental settings and results are presented in Section 4. Finally, Section 5 concludes the paper.

II.LITERATURE REVIEW

The significance of customer engagement in business operation becomes a new perspective in organization prospect. In this section we investigate from two different perceptions.

Sentiment Analysis over Social Media Platform

There are many researchers workssuch as in (Baj-Rogowskaet.al, 2017; Cenni, D. et.al, 2017; Windasari et al., 2017)that are conducted the sentiment analysis towardsthe company's transportation services. The authors in (Windasari et al., 2017)had conducted the sentiment analysis on GoJekby using Twitter dataset and analyzed it using Support Vector Machine (SVM) method. In their research, it reveals thaton average 86% in year 2000 most of comments is in Indonesian language. The authors in (Baj-Rogowskaet.al, 2017) had conducted the sentiment analysis of customer engagement on Uber service through Facebook. The analysis results show that social media is an efficient platform (fast with high number respondents) to get information on customers feedback. Meanwhile, the authors

Revised Manuscript Received on October 30, 2019.

* Correspondence Author

NurKhaleeda Othman, Department of Communication Technology and Network, Faculty of Computer Science and Information Technology, University Putra Malaysia, Serdang

Masnida Hussin, Department of Communication Technology and Network, Faculty of Computer Science and Information Technology, University Putra Malaysia, Serdang

Raja Azlina Raja Mahmood, Department of Communication Technology and Network, Faculty of Computer Science and Information Technology, University Putra Malaysia, Serdang

in(Saragih & Girsang, 2017)performed the survey on Facebook and Twitter fan page. It is related to transportation services from several companies such as Uber, GoJek, and Grab by using data mining method. They considered number of followers and analyzed the comments of such transportation services. There is a result that revealed the fan page in Facebook have few positive comments but more on unbiased (neutral) and negative comments.

Sentiment Analysis using Machine Learning

There are researchers (e.g., Thakkar, H.,2013; Yang, P., & Chen, Y., 2017; Joyce, B., & Deng, J., 2017) that used Naïve Bayes method for classifying sentiments and calculating a score for developing a sentiment polarity. Naïve Bayes (NB) method has high scalability where it can perform well in classifying a limited or large-scale data. The results in (Yang, P., & Chen, Y., 2017) determined that NB method has good performance in term of accuracy and learning speed. In (Thakkar, H.,2013) createdthe hybrid NB method for better data classifying. The results demonstrated that NB reachedthe highest scalability in dealing with diverse set of data. The authors in (Joyce, B., & Deng, J., 2017) conducted sentiment analysis on the US Presidential Election year 2016 by using lexicon and Naïve Bayes machine learning method to calculate the sentiment political tweets that collected 100 days before the election. They have investigated that the accuracy of Naïve Bayes method is much higher than lexicon. It reveals that the sentiment analysis becomes an important instrument to analyze public opinions. In our work, we analyzed sentiment of the Rapid KL customers about the transportationservices that collected from social media platform.

III.SENTIMENT ANALYSISPROCESS

In this section, we describe on how the Naive Bayes algorithm is been used to study raw data until it turns to meaningful information. We also explained for each step of sentiment analysis process.

Naïve Bayes for Sentiment Data

Naive Bayes methodis one of machine learning approach that handling for simple and effective text classification strategy. It is purposely chosen to minimize the computational complexity where the social media data is already enormous and hectic (Thakkar, H.,2013; Yang, P., & Chen, Y., 2017; Goel, A. et.al,2016). Naïve Bayes classifiers also can perform very well with independent assumption in limited and small size of data. Consequently, by using Naïve Bayes method, it is not necessary to have large dataset for training purposes. This method gives benefits in our study, due to the real datasets from Rapid KL Facebook and Twitter is slightly in medium-scale size. In the Eq. 1 it is given the fundamental notion of Bayes' theorem. In response to Bayes' theorem, we embedded the element of sentiment analysis in context text similarity. Hence the probability score of customers' sentiments for identifying the right expression either positive or negative sentiment is given in Eq. 2.

$$\mathrm{P(A|B)} = \frac{\mathrm{P(B|A)\ x\ P(A)}}{P(B)}[1]$$

$$\mathrm{P(label\ sentiment)} = \frac{P(sentiment|label)xP(label)}{P(sentiment)}[2]$$

We used Multinomial Naïve Bayes (MNB) modelas our classifier because it is used for discrete counts; like our dataset features. The MNB model will regularly count on how often the text/word had occurred in dataset while reading the data. Specifically, by using Eq. 2 we analysis each raw data and classify accordingly. Each output from the text classification process is then used to identify either it falls into positive or negative sentiment. Each class of sentiment is then processed again using the earlier process (that with MNB) to ensure the sentiment resides in the right class/output.

Methodology for Sentiment Analysis

Generally, sentiment analysis methodology consists of five stages which are text preprocessing, feature generation, initial classification, hyper-parameter tuning, and final classification (Ghiassi, M.et.al., 2016; Li, T. et.al.2018).We mapped the element of Naive Bayes method for predicting and investigating the raw data to form the sentiment labels.

Observation/Discussion

We separated the two main phases are training and testing (Fig. 1) for thoroughly analyzing the dataset to become meaningful information. Prasaranais required to know better on their customers feel towards their services. We had conducted several discussion and meeting with them to clearly understand their expectation and requirementsthrough the social media on Rapid KL services.

Data Collection

For train and test the learning machine, the data are collected and extracted from Rapid KL's Facebook and Twitter. At the Rapid KL Facebook page, it is simply pushing the text and form into appropriate document. Meanwhile for Twitter, data are collected from various account such as @AskRapidKL, @GrabMY, @MyCar_Malaysia and @ktm_berhad by executing through Python program. Because of the data is lively and comes from various source of customers, there is data that irrelevant for this analysis purposes. Thus, we excluded those data and it comes useful 500 data from both social media source

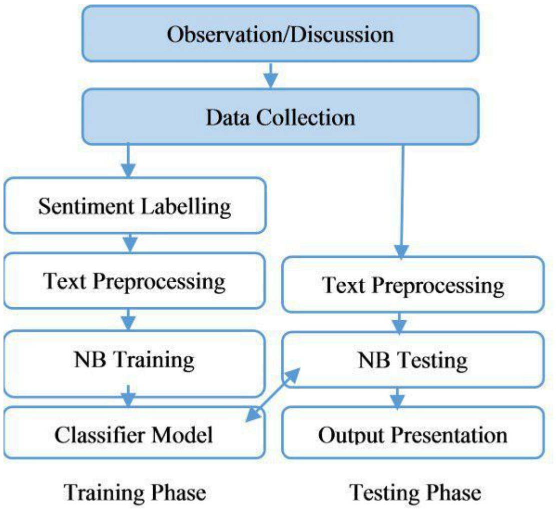

Fig. 1 Sentiment Analysis Process

Sentiment Labelling

The label training dataset refers to labelling process where we formed as positive and negative labels. In order to perform labeling process, thosesentiments from the collected data is measured. In particular, the sentiments are categorized based on emotion detection group. There are emotion groups that identified in prior are happiness, love, satisfy action, frustration, anger, and annoyed.

Text preprocessing

The text preprocessing aims to eliminate any information that may affect the analysis. In this work, the case folding, and punctuation removal are been performed in the preprocessing step. Note that, all words are converted to lowercase and substitute punctuation with space. It helps improving the accuracy of sentiment prediction. NB model is it treats each word in sentence individually due to NB assumed that all words are independent.

NB Training

As we implement Naïve Bayes method, training is a must step because we need to train our learning machine before it is able to predict polarity of actual data (final dataset). For this stage, we initially used 200 data consists of 100 positive and 100 negative data.

NB Classifier Model

Multinomial NB method used as classifier model where it will be counting regularly on how often the word occurs in the document. Note that this model will be classified the sentiments according to the training dataset. Later, it will be used as reference for predicting the final/testing dataset.

NB Testing

In this step, we tested our learning machine using testing dataset. Testing dataset consists of 400 data sourced from Rapid KL Facebook and Twitter account with 200 data for each sourced. These data do not have positive and negative label yet. We practically calculate the sentiment prediction score using Naïve Bayes method by referring to Multinomial NB. If the prediction score is below 0.5, it refers negative sentiment and vice versa.

Output Presentation

We indicate service issues are;i) driver attitude; ii) time management; iii) technical; iv) facilities; and v) customer service to represent the final data for Prasarana management team. We also perform comparison for each category based on the positive and negative sentiments to determine either the issue is the major concern or not.

IV. RESULTS ANALYSIS

Analysis Sentiment in Social Media

Based on Fig. 2, it shows that most of customers convey negative comments and feedbacks in all services (i.e., bus, train, others). The result also shows that on average 64.5% the rail services have more negative comments compared than the bus services (61%). We also can see that people tends to share their feedback and opinion in their Facebook posts. The results do not mean that the Prasarana provides bad services. Implicitly, it proves that the public can posts and comments anything about the available services. In the bright side, the company can investigate details the sentiment analysis data as part of their justification for future improvement.

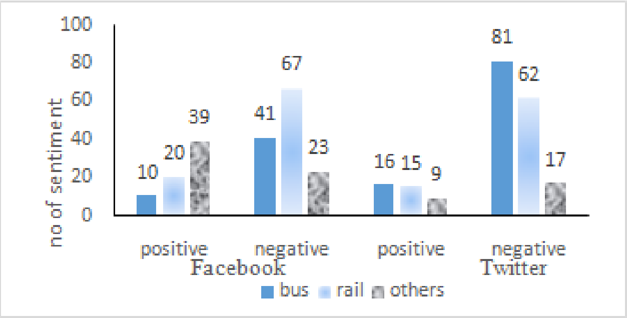

Fig. 2 Sentiment analysis on social media

Categorization of Sentiment

Based on the comments that collected from the social Medias, we categorized into several service issues are driver attitude, time management, technical issues, facilities and customer services. According to different social media platforms (i.e. Facebook and Twitter) the comments are relatively show in diverse pattern of responses towards the service issues. As shown in Fig.3, the technical issue becomes the highest response with negative sentiment while in other issues showing relatively 22number of sentiments; on average negative sentiments. However, from the comments in Twitter that can be seen in Fig. 4, on average 32 negative sentiments responded. Also, in both social media platforms, the technical issues remain as the maximum negative sentiments revealed.It shows that the customers prefer to complaints on Twitter. It might be because the Twitter is straight forward way of posting where the customers can easily mention about @AskRapidKL on their posts. In Facebook platform, customers will complaint by commenting on posts that update in Rapid KL page which is not as fast as in Twitter.

We assumed that the customers choose Twitter platform because they want to get quick respond and further attention from Rapid KL team.

IV.CONCLUSION

There are millions of social media users shared their opinion and feeling on their posts. Due to there are massive feedback and comments that available in the social media, make the data analysis process is a tough task. Furthermore, if all opinions and complaints that posted need to be analyzed manually, it takes a lot of time. The automation sentiment analysis can review raw and massive data into meaningful information automatically. In this work, we focused on analyzing the feedback of public transport service by collecting the data from Facebook and Twitter. We used Naïve Bayes methodfor developing the prediction and classification algorithm. The sentiment analysis assists the company in making decision for improving their strengths while overcoming the weaknesses.

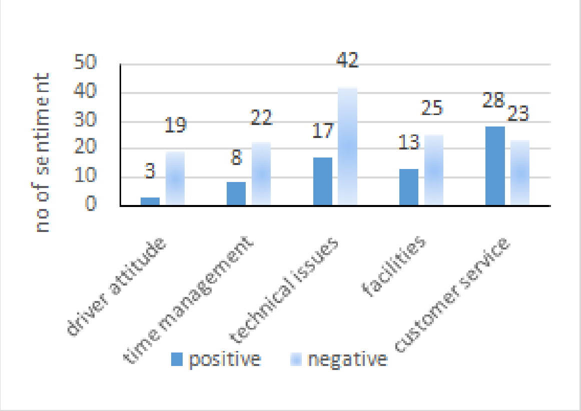

Fig. 3 Sentiment issues in Facebook

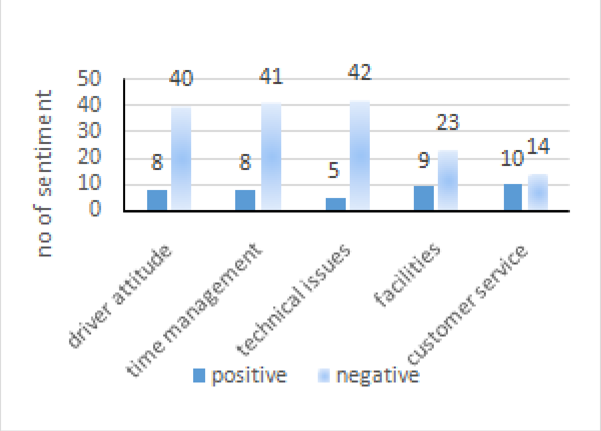

Fig. 4 Sentiment issues in Twitter

ACKNOWLEDGMENT

This work is supported by a Research Management Center (RMC) University Putra Malaysia under GeranPutra 9621200 at Feb/2018.

REFERENCES

1. Baj-Rogowska, A. (2017). Sentiment analysis of Facebook posts: The Uber case. In 2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS) (pp. 391–395).
2. Cenni, D., Nesi, P., Pantaleo, G., &Zaza, I. (2017). Twitter vigilance: A multi-user platform for cross-domain Twitter data analytics, NLP and sentiment analysis.IEEE SmartWorld (pp. 1–8).
3. DhahiAlgaithi, Amna& Ali Al-Shihi, Amal& Yahya Al Baloushi, Hajar& Khan, Firdouse. (2017). Impact of Social Media on Customers Satisfaction: Bank Muscat – A Case Study. International Journal of Recent Research Review. 1. 05-2017.
4. Goel, A., Gautam, J., & Kumar, S. (2016). Real time sentiment analysis of tweets using Naive Bayes. 2nd International Conference on Next Generation Computing Technologies (NGCT) (pp. 257–261).
5. Li, T., Li, J., Liu, Z., Li, P., &Jia, C. (2018). Differentially private Naive Bayes learning over multiple data sources. Information Sciences, 444, 89-104.
6. Joyce, B., & Deng, J. (2017). Sentiment analysis of tweets for the 2016 US presidential election. IEEE MIT Undergraduate Research Technology Conference (URTC) (pp. 1–4).
7. MyRapid Your Public Transport Portal. (n.d.). Retrieved April 10, 2018, from https://www.myrapid.com.my/
8. Saragih, M. H., &Girsang, A. S. (2017). Sentiment analysis of customer engagement on social media in transport online. International Conference on Sustainable Information Engineering and Technology (SIET) (pp. 24–29).
9. Thakkar, H. (2013). Twitter Sentiment Analysis using Hybrid Naïve Bayes.
10. Windasari, I. P., Uzzi, F. N., &Satoto, K. I. (2017). Sentiment analysis on Twitter posts: An analysis of positive or negative opinion on GoJek. 4th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE) (pp. 266–269).
11. Yang, P., & Chen, Y. (2017). A survey on sentiment analysis by using machine learning methods. IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC) (pp. 117–121).