Applied Mathematics and Computational Intelligence
Volume 11, No.2, Dec 2022 [360-372]

Sentiment Analysis on TikTok Using RapidMiner

Nurul Shahazira Rosli[1], Muhammad Firdaus Mustapha[2*], Maira Madihah Mohamed Azmee[3], Nur 'Aisyah Mohd Samsudin[4]

[1,2,3,4]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Cawangan Kelantan, Bukit Ilmu, 18500 Machang, Kelantan, Malaysia

[*] Corresponding author: mdfirdaus@uitm.edu.my

Received: 28 September 2022; Accepted: 7 November 2022; Available online (in press): 22 November 2022

ABSTRACT

Users commonly provide feedback on certain applications. Users can provide either positive, negative or neutral reviews. To determine whether the reviews are positive, negative or neutral, this study use sentiment analysis through various methods of text mining and materials. In this study, a sentiment analysis application for TikTok analysis was conducted using RapidMiner. This project is conducted based on three issues from TikTok which are account review, sound review and video review. These issues are analyzed using Decision Tree, Naive Bayes and k-NN. RapidMiner is used throughout the process to ensure that the data is accurately performed. Then, the result is gathered by checking the accuracy of data based on the three methods. To analyze the data and obtain an exact performance of the outcome, the process of visualization and modelling is required. The analysis of the reviews from the users shows that majority reviews were positive compared to the negative and neutral reviews especially on video issue.

Keywords: RapidMiner, Sentiment Analysis, TikTok

1 INTRODUCTION

TikTok is a short-form video hosting website owned by the Chinese company ByteDance. TikTok is known as Douyin in China. It hosts various user videos from many genres such as stunts, pranks, tricks, dance, entertainment, and jokes [1], [2] with a video length of 15 seconds to 10 minutes [3], [4]. TikTok was initially made available on iOS and Android in 2017 outside of the Chinese mainland. It became available worldwide when it merged with Musical.ly on 2 August 2018. Malaysia's population is currently estimated to be around 33 million people [5] and as expected demand on social media is expanding especially when the pandemic struck in our country. When the movement control order is executed, everyone is just at their home without going anywhere so they are enjoying their life with social media including TikTok. TikTok is one of the social media platforms that has been growing rapidly over the past few years, and its popularity has only recently peaked as a result of the global pandemic. Through variety of video contents in TikTok, including dances, trends and challenges, most people have used this platform to express themselves creatively and connect with people around the world. It takes a short time to build up a large TikTok community, especially in difficult times like these to deal with grief or hardship that many people are going through. When there is a fast-growing community, there are also those with influence within those communities, especially when it comes to content creators on social media [6].

Therefore, this study is mainly to identify the user's opinions or reviews of TikTok applications. Sentiment analysis is used to identify the reviews about account, video and sound. In order to make this study happened, the data about account, video and sound reviews are needed. These data were obtained from Kaggle website. The methodology that has been used is Decision Tree, Naive Bayes and K-Nearest Neighbor(k-NN). The result of this study will be shown by using RapidMiner.

2 LITERATURE REVIEW

Opinion mining is a type of computer research that aims to recognize and reflect sentiment, evaluation, behavior, emotion, subjectivity, assessment, or viewpoint in a current text [7]. The practice of investigating the options of text on a certain object is known as sentiment analysis. This research is carried out to discover whether a person's predisposition toward TikTok is positive, neutral, or negative. Sentiment analysis is also one of the Natural Language Processing (NLP) fields, dedicated to the exploration of subjective feelings or opinions collected from various sources about a specific subject. When mentioning NLP or being involved in Machine Language (ML), it has supervised and unsupervised ML. In supervised ML, a set of text documents are annotated or tagged with examples of what the machine could look for and how it should interpret the aspect [8]. These documents are used to train and build a statistical model that capable to analyze the given un-tagged text. Meanwhile, unsupervised ML includes training a model without annotating or pre-tagging. Some of these methods are simple to comprehend.

Researchers have noticed the scientific studies and potential applications of sentiment analysis and they have predicted an upcoming trend of applying it to TikTok. As TikTok sentiment analysis has gained popularity in recent years, Decision Tree, K-NN and Naive Bayes models have been applied and used in sentiment analysis. K-NN is the simplest machine learning algorithm. It is based on the principle that the samples that are similar, generally lie in close vicinity [9]. It is an instance-based learning method and also called lazy-learning algorithms that require less computation time during the training phase than eager learning algorithms [10]. Naive Bayes is a simple Statistical Bayesian Classifier and it is based on Bayesian Theorem [11]. It is known as naive because it makes the assumption that all factors influence classification and are associated with one another. It is used for high input dimensions [12]. Meanwhile, Decision Tree algorithm is a data mining induction technique that recursively partitions a dataset of records using breadth-first approach or depth-first greedy approach until all the data items belong to a specific class. A decision tree structure consists of root, internal and leaf nodes. It is a flow chart like tree structure, where every internal node represents a test condition on an attribute, each leaf node (or terminal node) is assigned with a class label, and each branch denotes result of the test condition [10].

The three algorithms which are Decision Tree, Naïve Bayes and K-NN are chosen as the methodology to compare which algorithms provide the best result because they are the most popular classification methods and they are not in the same group so the comparison can be done. Naïve Bayes and K-NN are both examples of supervised learning. Because Naive Bayes is a linear classifier, it frequently performs more quickly when used with big data.

Many researchers have conducted studies on the TikTok application. For an example, Zeng et al. [13] proposed the perspectives on TikTok and its legacy applications. They briefly reviewed why TikTok is successful and managed to catch many eyes of people in the entire world. But, in order to be a successful application, TikTok also has a bumpy road ahead. That is the reason why not all of the reviews are positive for the TikTok application. They still have negative and neutral reviews. The worldwide lockdowns during the COVID-19 outbreak in 2020 further sparked the enormous growth and variety of TikTok's user groups. Based on the trend, many users downloaded TikTok in the early stage of the global crisis. TikTok was able to attract older audiences and adults from all professions to the platform during the epidemic by positioning itself as a key source for both entertainment and learning resources.

Therefore, this study proposes to create a classification model to identify the best accuracy of the account review, sound review and video review using the three methods which are Decision Tree, Naive Bayes and k-NN.

3 METHODOLOGY

This study uses RapidMiner Studio to analyze the data such as structured and unstructured data. RapidMiner provides data mining and machine learning procedures including data preprocessing and visualization, data loading and transformation (ETL), predictive analytics and statistical modelling, evaluation, and deployment. It is written in the Java programming Language. The purpose of this study is to identify the review of the TikTok application based on three issues which are review on the user account, review on the video and lastly is review on the sound which has positive, neutral or negative feedback. Data from Kaggle was collected where 6736 data for video review, 3083 data for sound review, and 6996 data for account review. This study applies Decision Tree model, Naive Bayes model and k-NN model to analyze the sentiment of the review from the users.

3.1 Data Preparation

After getting the review of the TikTok dataset, this study then removes unwanted data for the three issues. Figure 1 shows the “Select Attributes” operator that is used to choose which attributes will appear on the output. This will obviously help the data to be well organized. This study chooses ‘review_text’ attributes to analyze the sentiment.

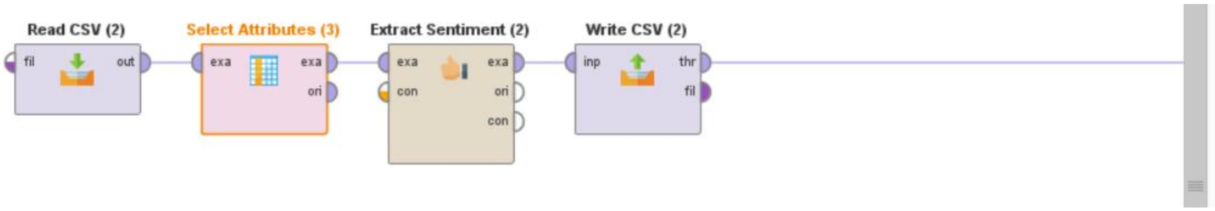


Figure 1: Processing the Data

After that, this study uses “Extract Sentiment” to create a sentiment ‘Score’ attribute. Then “Write CSV” is used to save the clean data into CSV.

3.2 Data Modelling

Modelling is the stage of final sub-process data from preparation results to obtain required information as shown in Figure 2. This study applies the same operator before starting the modelling process using a different model. Firstly, retrieve data and then use the “Set Role” operator to set the

‘Score’ attribute as the ‘special attribute’. Next is the “Nominal to Text” operator which convert all the nominal attributes to string attributes. After that “Process Document from Data” is used and in the operator contains “Tokenize”. Tokenization is a pre-processing method that splits text streams into tokens, which can be words, symbols, phrases, or other significant elements. Tokenization is mostly used to recognize significant keywords. “Transform Cases” operator changes all characters in a document to lowercase or uppercase, depending on the case. Transform case is to eliminate confusion between comparable terms that are different in uppercase or lowercase and “Filter Stopwords (English)” operator filters English stopwords from a document by eliminating every token which equals a stopword from the built-in stopword list. Figure 2 shows all the elements used in the “Process Document from Data”.

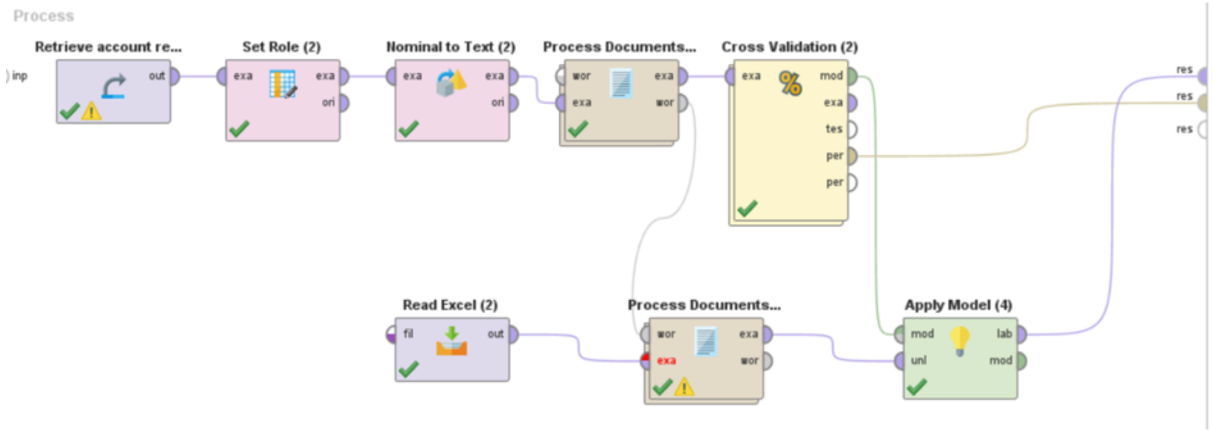


Figure 2: Process

After that, this study uses the “Cross Validation” operator to use the three models. Next is “Read Excel” and connect it to the second “Process Document” before “Apply Model”.

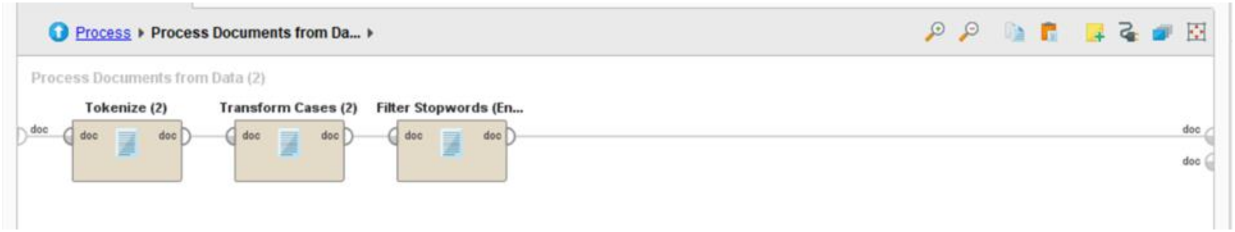


Figure 3: Process Document from Data

3.2.1 *Decision Tree Model*

A decision tree is a tree-like collection of nodes that is uses to make a decision on values affiliation to a class or an estimate of a numerical target value [14]. Figure 4 displays the classification diagram for decision tree while Figure 5 displays tree structure.

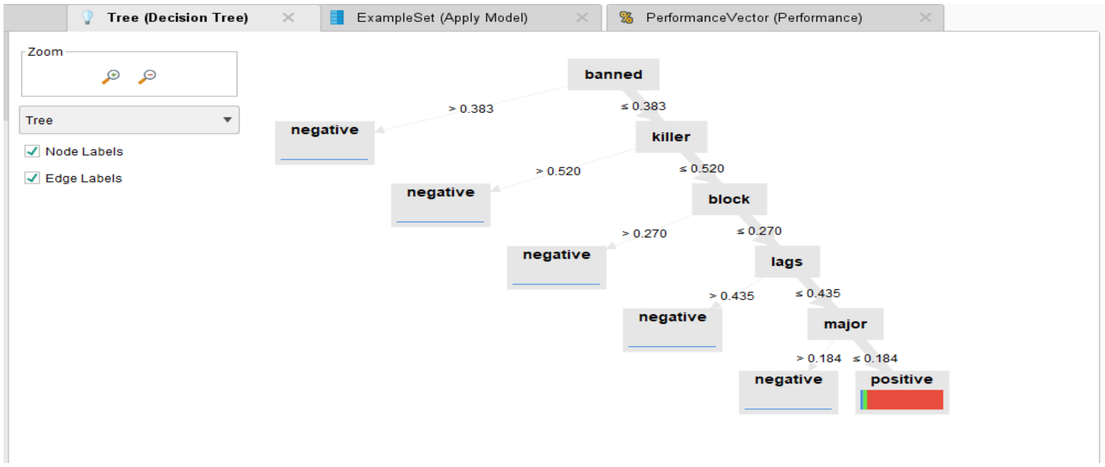
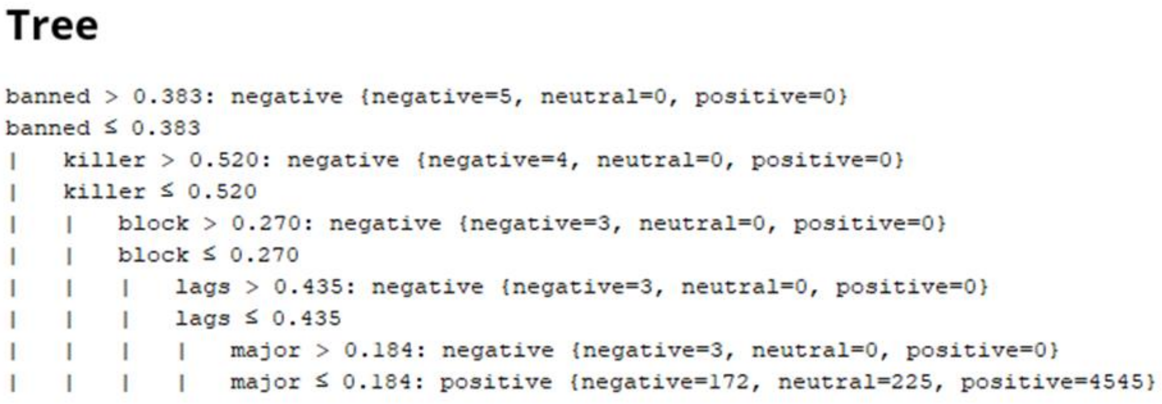


Figure 4: Classification Diagram



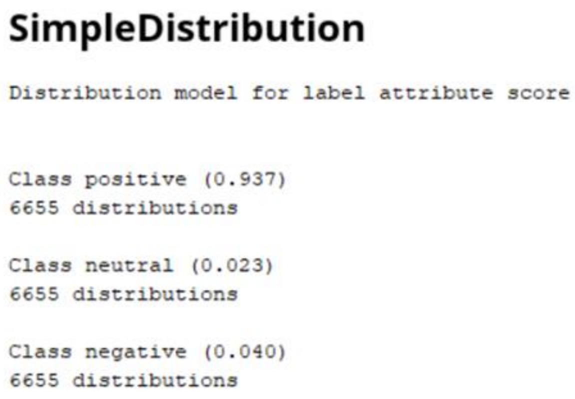
Tree

```
banned > 0.383: negative {negative=5, neutral=0, positive=0}
banned ≤ 0.383
|   killer > 0.520: negative {negative=4, neutral=0, positive=0}
|   killer ≤ 0.520
|   |   block > 0.270: negative {negative=3, neutral=0, positive=0}
|   |   block ≤ 0.270
|   |   |   lags > 0.435: negative {negative=3, neutral=0, positive=0}
|   |   |   lags ≤ 0.435
|   |   |   |   major > 0.184: negative {negative=3, neutral=0, positive=0}
|   |   |   |   major ≤ 0.184: positive {negative=172, neutral=225, positive=4545}
```

Figure 5: Tree Structure

3.2.2 Naïve Bayes

Naive Bayes is a low-variance and high-bias classifier. It can create a good model even with a small dataset [15]. It is computationally inexpensive and simple to use. Common use cases include text categorization such as sentiment analysis, recommender systems, and spam detection. Simple distribution of Naive Bayes is illustrated in Figure 6.



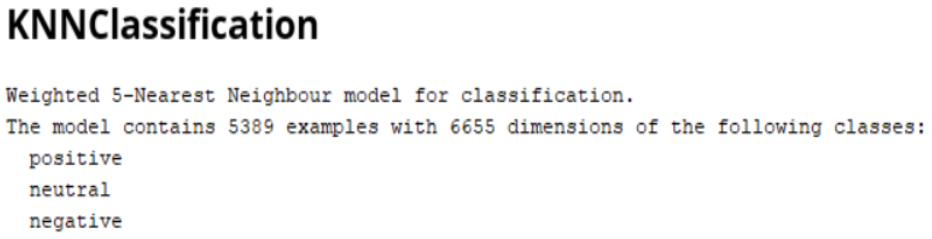
SimpleDistribution

```
Distribution model for label attribute score


Class positive (0.937)
6655 distributions

Class neutral (0.023)
6655 distributions

Class negative (0.040)
6655 distributions
```

Figure 6: Simple Distribution of Naive Bayes

3.2.3 K-NN

K-NN is a supervised ML algorithm that is simple and easy-to-implement. It is commonly used to solve problems related to regression and classification [16]. Figure 7 depicts the k-NN classification.



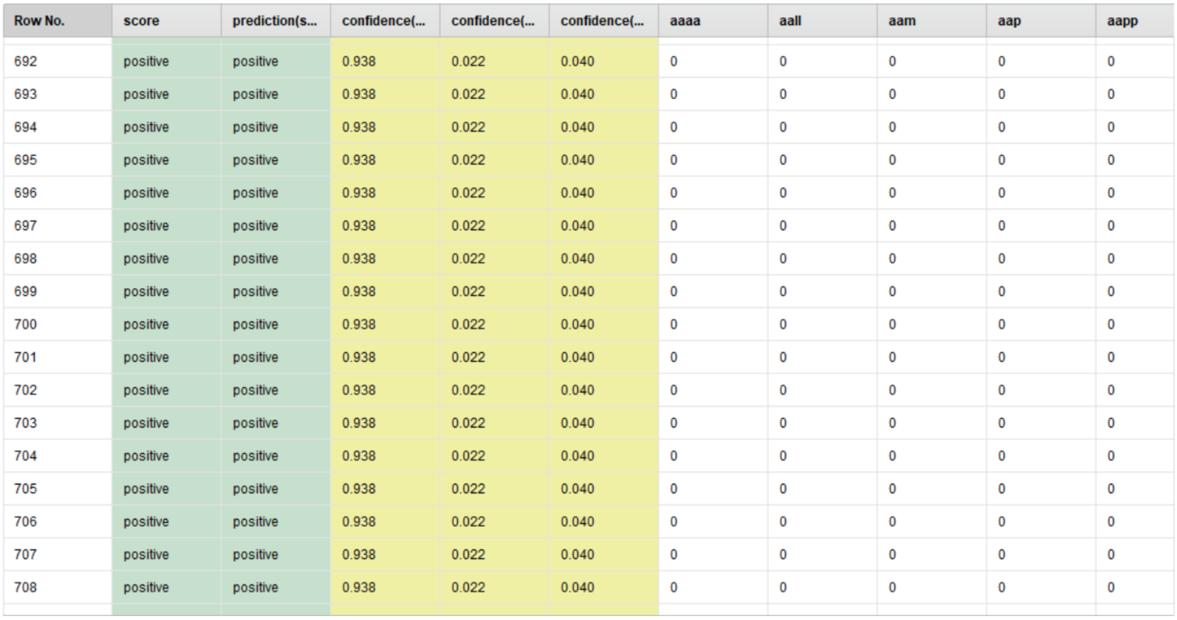
KNNClassification

```
Weighted 5-Nearest Neighbour model for classification.
The model contains 5389 examples with 6655 dimensions of the following classes:
  positive
  neutral
  negative
```

Figure 7: KNN Classification

3.3 Applying the Model

After that, this study needs to predict the accuracy of the dataset. Here, operators such as "Decision Tree", "k-NN", "Naive Bayes" and "Apply Model" are used. After run the process as shown in Figure 8, 'score' is the original and 'prediction' is the prediction of the score. The confidence column shows the precision value of each of the predictions.



Row No.	score	prediction(s...	confidence(...	confidence(...	confidence(...	aaaa	aall	aam	aap	aapp
692	positive	positive	0.938	0.022	0.040	0	0	0	0	0
693	positive	positive	0.938	0.022	0.040	0	0	0	0	0
694	positive	positive	0.938	0.022	0.040	0	0	0	0	0
695	positive	positive	0.938	0.022	0.040	0	0	0	0	0
696	positive	positive	0.938	0.022	0.040	0	0	0	0	0
697	positive	positive	0.938	0.022	0.040	0	0	0	0	0
698	positive	positive	0.938	0.022	0.040	0	0	0	0	0
699	positive	positive	0.938	0.022	0.040	0	0	0	0	0
700	positive	positive	0.938	0.022	0.040	0	0	0	0	0
701	positive	positive	0.938	0.022	0.040	0	0	0	0	0
702	positive	positive	0.938	0.022	0.040	0	0	0	0	0
703	positive	positive	0.938	0.022	0.040	0	0	0	0	0
704	positive	positive	0.938	0.022	0.040	0	0	0	0	0
705	positive	positive	0.938	0.022	0.040	0	0	0	0	0
706	positive	positive	0.938	0.022	0.040	0	0	0	0	0
707	positive	positive	0.938	0.022	0.040	0	0	0	0	0
708	positive	positive	0.938	0.022	0.040	0	0	0	0	0

Figure 8: Prediction Analysis

3.3.1 Training and Validation

For training and validation, this study used cross-validation [17] operators. When the cross-validation operators were double clicked, there were 2 processes involves which were training and testing as shown in Figure 9, 10 and 11. For training, the "Decision Tree" operators were used while for testing, this study used the "Apply Model" and "Performance" operators. The number of folds used in this process was 10 which means that the datasets were divided into 10 parts. Only one part was used in the testing process and the rest of it was used for the training process.

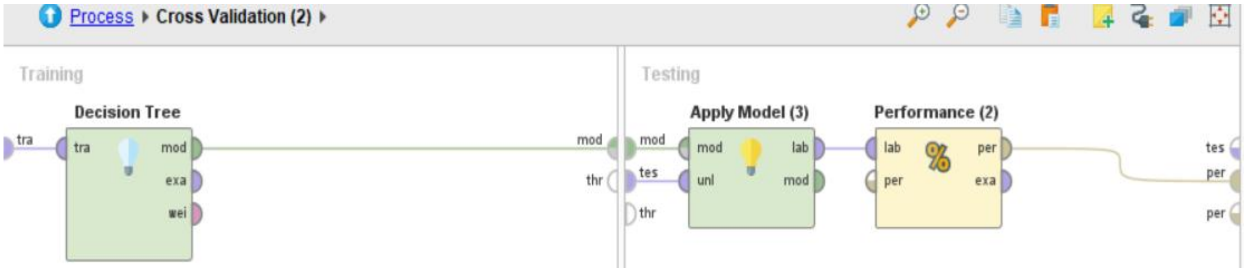


Figure 9: Cross Validation for Decision Tree

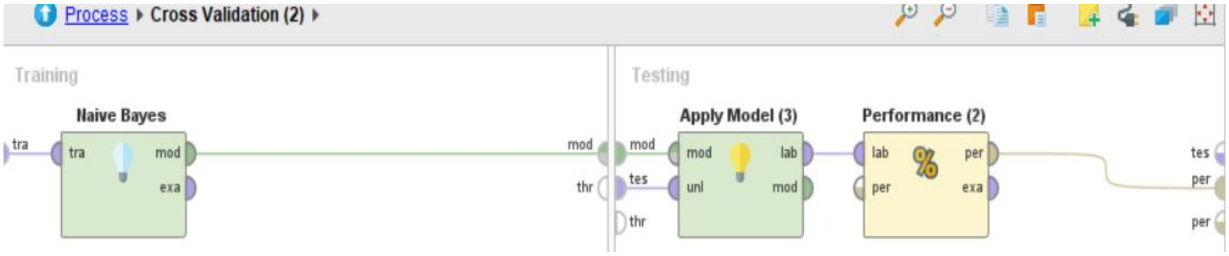


Figure 10: Cross Validation for Naive Bayes

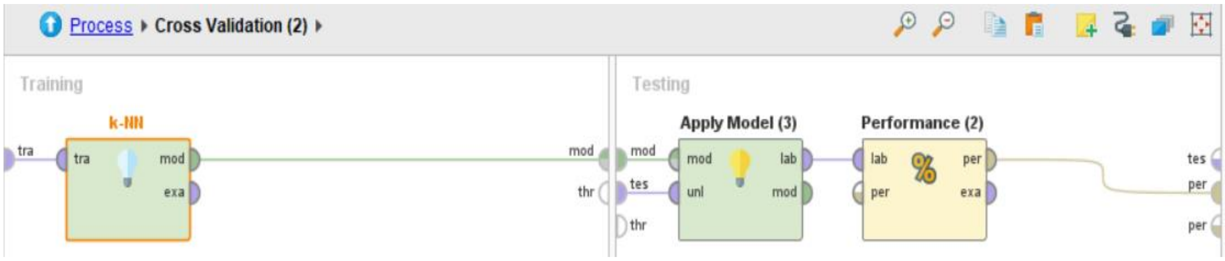


Figure 11: Cross Validation for k-NN

3.4 Evaluation

Evaluation is done to identify the accuracy of the three models. By using the same operators from the previous step and adding the “Performance” operator, it will show the accuracy. Nevertheless, the obtained result is not the final result as it only gives the output of the given dataset for testing. So, a validation process is required for an unknown dataset and to identify the actual accuracy for the models.

4 RESULTS AND DISCUSSION

From the result, we identify the user’s opinions or reviews of TikTok applications. Sentiment analysis is used to identify the reviews about account, video and sound. From the different classification that has been applied, the accuracy result differs on RapidMiner and it generates the output as shown in Table 1. Thus, this study concludes that for all three issues, the highest accuracy is using Decision Tree model which is for video review, the accuracy is 93.71% and ±0.16% standard deviation, for account review, the accuracy is 88.16% and ±0.13% standard deviation and the last issue which is sound review has the accuracy 89.94 and ±0.41% standard deviation. The most important when evaluating a model is, the smaller standard deviation value, the more stable the model. When the higher accuracy and precision, the more stable the model and it is good as it will produce a smaller range of best and worst cases. In addition, this study needs to consider the quality of the prediction. Last but not least, this study needs to make a decision on how to proceed with the obtained results. Based on the calculation of data mining, the Decision Tree model is the most stable since it has higher accuracy and lower standard deviation.

Table 1: Visualization Accuracy for Each Review

ISSUE	MODEL	STANDARD DEVIATION	ACCURACY
ACCOUNT	**Decision Tree**	**±0.13%**	**88.16%**
	k-NN	±0.69%	87.78%
	Naïve Bayes	±1.62%	61.52%
SOUND	**Decision Tree**	**±0.41%**	**89.94%**
	k-NN	±1.25%	89.23%
	Naïve Bayes	±2.31%	76.42%
VIDEO	**Decision Tree**	**±0.10%**	**93.71%**
	k-NN	±0.16%	93.60%
	Naïve Bayes	±1.47%	67.04%

Table 2 displays the accuracy of Video Review using Decision Tree. The accuracy which includes reviews that are positive, neutral, or negative, is shown in Table 2. There were 6315 total positive reviews, 152 total neutral reviews, and 269 total negative reviews. It can be seen that 6312 predicted positive reviews actually were positive, 0 expected neutral reviews actually were neutral, and 0 predicted negative reviews actually were negative. This study can learn about the percentage of accurate predictions from the class precision column. The model was able to predict 99.95% for positive reviews and 0.00% for neutral reviews from the class recall row. The predicted value for a negative review was 0.00%.

Table 2: Accuracy for Decision Tree model of Video Review

accuracy:93.71%+/-0.10% (micro average:93.71%)

	True positive	True neutral	True negative	Class precision
pred.positive	6312	152	269	93.75%
pred.neutral	3	0	0	0.00%
pred.negative	0	0	0	0.00%
Class recall	99.95%	0.00%	0.00%	

Table 3 displays the accuracy of Sound Review using Decision Tree. The accuracy which includes reviews that are positive, neutral, or negative, is shown in Table 3. There were 2783 total positive reviews, 170 total neutral reviews, and 130 total negative reviews. It can be seen that 2771 predicted positive reviews actually were positive, 0 expected neutral reviews actually were neutral, and 2 predicted negative reviews actually were negative. This study can learn about the percentage of accurate predictions from the class precision column. The model was able to predict 99.57% for positive reviews and 0.00% for neutral reviews from the class recall row. The predicted value for a negative review was 1.54%.

Table 3: Accuracy model Decision Tree of Sound Review

accuracy:89.94%+/-0.41%(micro average:89.94%)

	True positive	**True neutral**	**True negative**	**Class precision**
pred.positive	2771	169	128	90.32%
pred.neutral	0	0	0	0.00%
pred.negative	12	1	2	13.33%
Class recall	99.57%	0.00%	1.54%	

Table 4 displays the accuracy of Account Review using Decision Tree. The accuracy which includes reviews that are positive, neutral, or negative, is shown in Table 4. There were 6171 total positive reviews, 202 total neutral reviews, and 623 total negative reviews. It can be seen that 6168 predicted positive reviews actually were positive, 0 expected neutral reviews actually were neutral, and 0 predicted negative reviews actually were negative. This study can learn about the percentage of accurate predictions from the class precision column. The model was able to predict 99.95% for positive reviews and 0.00% for neutral reviews from the class recall row. The predicted value for a negative review was 0.00%.

Table 4: Accuracy model Decision Tree of Account Review

accuracy:88.16%+/-0.13%(micro average:88.16%)

	True positive	**True neutral**	**True negative**	**Class precision**
pred.positive	6168	202	623	88.20%
pred.neutral	3	0	0	0.00%
pred.negative	0	0	0	0.00%
Class recall	99.95%	0.00%	0.00%	

Based on the Figure 12, it shows that a lot of people have a TikTok account and use it for daily fun. For video review from users, the positive responses are 6315, 269 negative responses followed by 152 neutral. For account review, the positive responses are 6171, 623 negative responses and 202 neutral. While for sound review from users, the positive responses are 2783, 130 negative responses followed by 170 neutral.

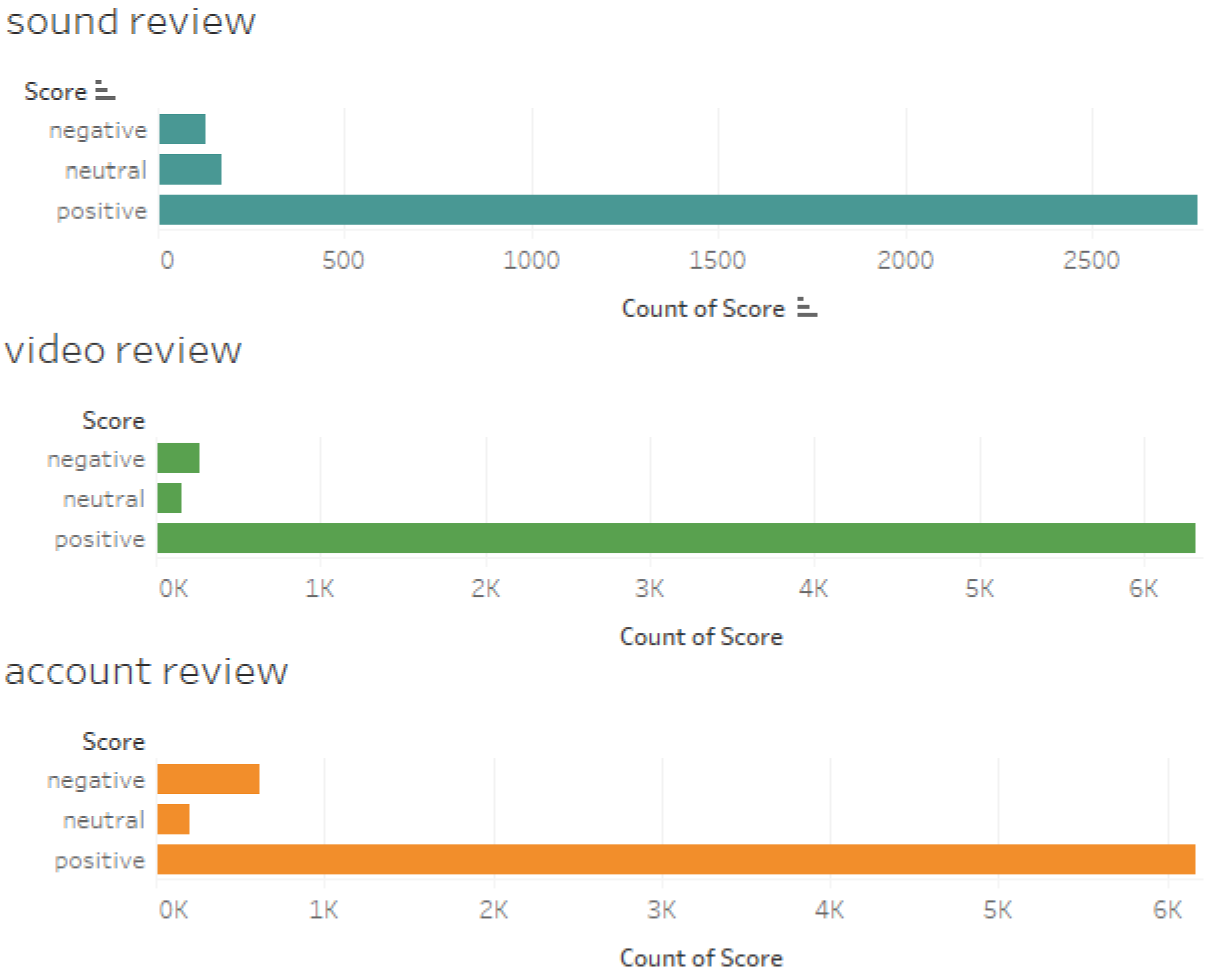


Figure 12: Visualization Sentiment Analysis Sound, Video and Account Review (Horizontal Bars)

Since this study wants to find which users have the least problems with TikTok, so the precision of the best model for each data needs to be considered. The class precision is the ratio of true positives to all positives predicted. From the decision tree model, video review precision is 93.75%, account review is 88.20% and sound review is 90.32%. It can conclude that most TikTok users do not have a problem with the video of TikTok rather than accounts and sounds. In the recall class is the ratio of true positives to all actual positives. By decision tree model, it shows that video and account review give the same value of class recall which is 99.95% while class recall for sound review is 99.57%. By naive bayes model, the class recall for video review is 64.35%, for account review is 69.88% and for sound review is 81.03%. By k-NN model, the class recall for video review is 99.76%, for account review is 98.23% and for sound review is 97.59%. In the nutshell, from all of the data collected, the model decides most TikTok users do not have a lot of problem with the sound based on their positive feedback for video on TikTok.

5 CONCLUSION

The sentiment analysis research on TikTok reviews has been performed using RapidMiner. This work involves text mining and predicting sentiment categories which are positive, negative and neutral. The performance of three types of models namely k-NN, Decision Tree and Naive Bayes were tested and compared for predicting sentiment task. Of all the issues analysed, video reviews got the highest score for all the models used which is score the highest positive prediction. Based on the results, this study concludes that the video-based TikTok application has a more positive effect on giving pleasure to users. TikTok is a medium that people can use to express themselves, gain followers and also build a community based on their interests. Meanwhile, the decision tree model shows the best performance compared to naive bayes and k-NN for all the issues reviewed. In detail, the decision tree model scores 88.16% for the account issue, 89.94% for the sound issue, and 93.71% for the video issue. In addition, the decision tree model is also the most stable in each decision with the lowest score for standard deviation. In the future, further comprehensive analysis on this topic needs to be done by analysing more data to improve prediction accuracy.

ACKNOWLEDGEMENT

Highly appreciation to Universiti Teknologi MARA, Kelantan Branch, Malaysia for financial support and to all reviewers for their constructive comments.

REFERENCES

[1] "Top categories on TikTok by hashtag views 2020," *Statistica*, 2021. [Online]. Available: https://www.statista.com/statistics/1130988/most-popular-categories-tiktok-worldwide-hashtag-views/. [Accessed Nov. 26, 2021].

[2] J. Bailey, "The five key genres found in the world of TikTok," *The Sydney Morning Herald*, 2020. [Online]. Available: https://www.smh.com.au/culture/art-and-design/the-five-key-genres-found-in-the-world-of-tiktok-20200303-p546ji.html. [Accessed Nov. 26, 2021].

[3] J. KASTRENAKES, "TikTok is rolling out longer videos to everyone," *Theverge*, 2021. [Online]. Available: https://www.theverge.com/2021/7/1/22558856/tiktok-videos-three-minutes-length. [Accessed Jul. 01, 2021].

[4] A. Hutchinson, "TikTok Confirms that 10 Minute Video Uploads are Coming to All Users," *Social Media Today*, 2022. [Online]. Available: https://www.socialmediatoday.com/news/tiktok-confirms-that-10-minute-video-uploads-are-coming-to-all-users/619535/. [Accessed Mar. 24, 2022].

[5] "Malaysia Population," *Worldometer*, 2022. [Online]. Available: https://www.worldometers.info/world-population/malaysia-population/. [Accessed Jan. 31, 2022].

[6] Chris-Seto, “The Trending Growth & Impact of TikTok in Malaysia,” *LinkedIn*, 2021. [Online]. Available: https://www.linkedin.com/pulse/trending-growth-impact-tiktok-malaysia-chris-seto. [Accessed Jul. 14, 2022].

[7] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014, doi: https://doi.org/10.1016/j.asej.2014.04.011.

[8] “Machine Learning (ML) for Natural Language Processing (NLP),” *Lexalytics*, 2022. [Online]. Available: https://www.lexalytics.com/blog/machine-learning-natural-language-processing/. [Accessed Mar. 24, 2022].

[9] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967, doi: 10.1109/TIT.1967.1053964.

[10] S. D. Jadhav and H. P. Channe, “Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques,” *Int. J. Sci. Res.*, vol. 5, no. 1, pp. 1842–1845, 2016, doi: 10.21275/v5i1.nov153131.

[11] R. O. Duda and P. E. Hart, *Pattern Classification And Scene Analysis*. New York: Wiley, 1973.

[12] S. S. Nikam, “A Comparative Study of Classification Techniques in Data Mining Algorithms,” *Orient. J. Comput. Sci. Technol.*, vol. 8, no. 1, pp. 13–19, 2015.

[13] J. Zeng, C. Abidin, and M. S. Schäfer, “Research Perspectives on TikTok & Its Legacy Apps,” *Int. J. Commun.*, vol. 15, pp. 3161–3172, 2021.

[14] “Decision Tree,” *RapidMiner*, 2022. [Online]. Available: https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/trees/parallel_decision_tree.html.

[15] “Naive Bayes,” *RapidMiner*, 2022. [Online]. Available: https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/bayesian/naive_bayes.html.

[16] S. U. Hassan, J. Ahamed, and K. Ahmad, “Analytics of machine learning-based algorithms for text classification,” *Sustain. Oper. Comput.*, vol. 3, no. July 2021, pp. 238–248, 2022, doi: 10.1016/j.susoc.2022.03.001.

[17] R. Kohavi and S. Edu, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” *Proc. 14th Int. Jt. Conf. Artif. Intell.*, vol. 2, pp. 1137–1143, 1993.