A COMPARATIVE STUDY OF MACHINE LEARNING ALGORITHMS FOR SENTIMENT ANALYSIS OF GAME REVIEWS

(Date received: 13.01.2022/Date accepted: 01.04.2022)

Jie Ying Tan[1], Andy Sai Kit Chow[1], Chi Wee Tan[1]*

[1] *Faculty of Computing And Information Technology, Tunku Abdul Rahman University College, Kampus Utama,
Jalan Genting Kelang, 53300, Wilayah Persekutuan Kuala Lumpur, Malaysia.*

**Corresponding author: chiwee@tarc.edu.my*

ABSTRACT

Sentiment analysis, also known as opinion mining, is the process of analysing a body of text to determine the sentiment expressed by it. In this study, Natural Language Processing techniques and Machine Learning algorithms have been applied to create multiple sentiment analysis models customized for the gaming domain to determine the sentiment of game reviews. The dataset was collected from Steam and Metacritic through the use of web API and web scraping. This was followed by text preprocessing, data labelling, feature extraction and finally model training. In the training phase, the effects of oversampling and hyperparameter tuning on the performance of the models have been evaluated. Through comparison between Support Vector Classifier (SVC), Multi-layer Perceptron Classifier (MLP), Extreme Gradient Boosting Classifier (XGB), Logistic Regression (LR) and Multinomial Naïve Bayes (MNB), it was evident that SVC had the most superior performance.

Keywords: Sentiment Analysis, Natural Language Processing, Machine Learning, Support Vector Machine, Game Reviews

1.0 INTRODUCTION

The video game industry is a multi-billion-dollar industry that has become increasingly competitive due to the advancement of technology that has spurred the accessibility of video games and popularized it. Therefore, in order for video game companies to keep up and stay ahead in the market, it is crucial for them to understand the needs and wants of their users. Sentiment analysis helps game developers to uncover the opinions of users towards their games so that they can design and develop their games according to the users' expectations. In view of the above, the specific objectives of this study are:

1. To train multiple machine learning models to classify sentiment of game reviews and compare their performance.
2. To investigate whether oversampling and hyperparameter tuning improve the models' performance.

2.0 LITERATURE REVIEW

This section describes the supervised machine learning algorithms implemented in this study.

2.1 Machine Learning Algorithms

LR is a machine learning algorithm that is used to solve classification issues based on the concept of probability. There are a few assumptions that must be met for LR which include the dependent variable must be dichotomous and the linear relationship between the dependent and independent variable does not exist (Prabhat & Khullar, 2017). An example of real-world application of LR is in the medical field where it can predict the mortality of injured patients (Boyd *et al.*, 1987).

The Support Vector Machine (SVM) is a statistical classification method that determines a hyperplane in an N-dimensional space, where N is the number of characteristics, that categorises data points. It was thought to be the most effective text categorization approach (Xia *et al.*, 2011). It is a non-probabilistic binary linear classifier that can linearly separate classes by a considerable margin, making it one of the most powerful classifiers because of its capacity to handle infinite dimensional feature vectors (Al Amrani *et al.*, 2018). SVC is developed based on SVM and has various applications which include numerical pattern recognition, face detection, text categorization and protein fold recognition (Lau & Wu, 2003).

The Naive Bayes (NB) algorithm is a classification strategy based on the Bayes' Theorem and the assumption of predictor independence. It is mostly used to classify documents at the document level. The main concept behind the technique is that it can calculate the probability of categories given a test document using the combined probabilities of words and categories. The decision-making time for NB classifiers is computationally short and learning can be started without a large amount of data (Ashari *et al.*, 2013). There are a few variations of the NB classifier, namely Multinomial Naive Bayes (MNB), Bernoulli Naive Bayes (BNB) and Gaussian Naive Bayes (GNB).

The XGB is a variation of the Gradient Boosting Machine. The unrivalled scalability of XGB in all circumstances, which consumes considerably fewer resources than previous systems, is its main selling feature. On a single machine, the system is ten times quicker than existing popular methods, and it is scalable to billions of samples in distributed or memory-limited environments. The scalability of XGB is influenced by a number of factors, including system and algorithmic improvements.

For example, a novel tree learning technique is used to handle sparse data, and a theoretically justified weighted quantile sketch procedure is used to handle instance weights. These distinguishing characteristics have made XGB a well-known system for machine learning and data mining problems (Chen & Guestrin, 2016).

MLP is a feed-forward artificial neural network composed of perceptrons which are organised hierarchically in numerous linked layers, each of which is made up of three types of layers: input, output, and hidden. The input signal is transferred to the input layer which will then be passed to the output layer where prediction and classification are performed while the hidden layer performs computational processing in the network to generate network outputs. The goal of MLP network training is to find the optimum collection of connection weights and biases to reduce prediction error (Alboaneen *et al.*, 2017).

2.2 Related Work

Several studies on machine learning-based sentiment analysis have been carried out in the past. Chakraborty *et al.* (2018) performed sentiment analysis on game reviews obtained from Amazon and Twitter. The algorithms which include NB, SVM, LR and Stochastic Gradient Descent (SGD) were used to train sentiment analysis models and the models were evaluated in terms of their accuracies. The feature extraction method used was the Bag-of-Words method.

Zuo (2018) performed sentiment analysis on game reviews collected from Steam. The algorithms applied were NB and Decision Tree classifiers. Feature selection using information gain was carried out, followed by feature extraction through Term Frequency-Inverse Document Frequency (TF-IDF) and hyperparameter tuning of the models through grid search.

A study was conducted by Britto and Pacifico (2020) on video game acceptance by performing sentiment analysis on game reviews. The dataset used was game reviews written in Brazilian Portuguese language extracted from Steam. Feature extraction was performed using the Bag-of-Words method. The algorithms implemented were Random Forest classifier, SVM and LR.

Based on the previous studies, there exists several research gaps for sentiment analysis of game reviews using machine learning techniques. Firstly, there is a lack of implementation of XGB and MLP algorithms. Besides that, exploration on sentiment analysis for more professional and complex reviews written by game critics such as the Metacritic reviews was absent from previous studies. Furthermore, the effect of resampling techniques such as oversampling along with hyperparameter tuning of TF-IDF have not been studied before. Therefore, this study was conducted to fill the above-mentioned research gaps such that XGB and MLP models were trained to compare their performance, Metacritic reviews were included in the training data and an oversampling technique and hyperparameter tuning of TF-IDF were experimented to investigate their effects on the models' performance.

3.0 METHODOLOGY

This section presents the system framework, datasets, text preprocessing, data labelling, and feature extraction, handling of imbalanced classes, model applications and hyperparameter tuning.

3.1 System Framework

Figure 1 summarises the steps involved in preparing the data, training and optimizing the sentiment analysis models.

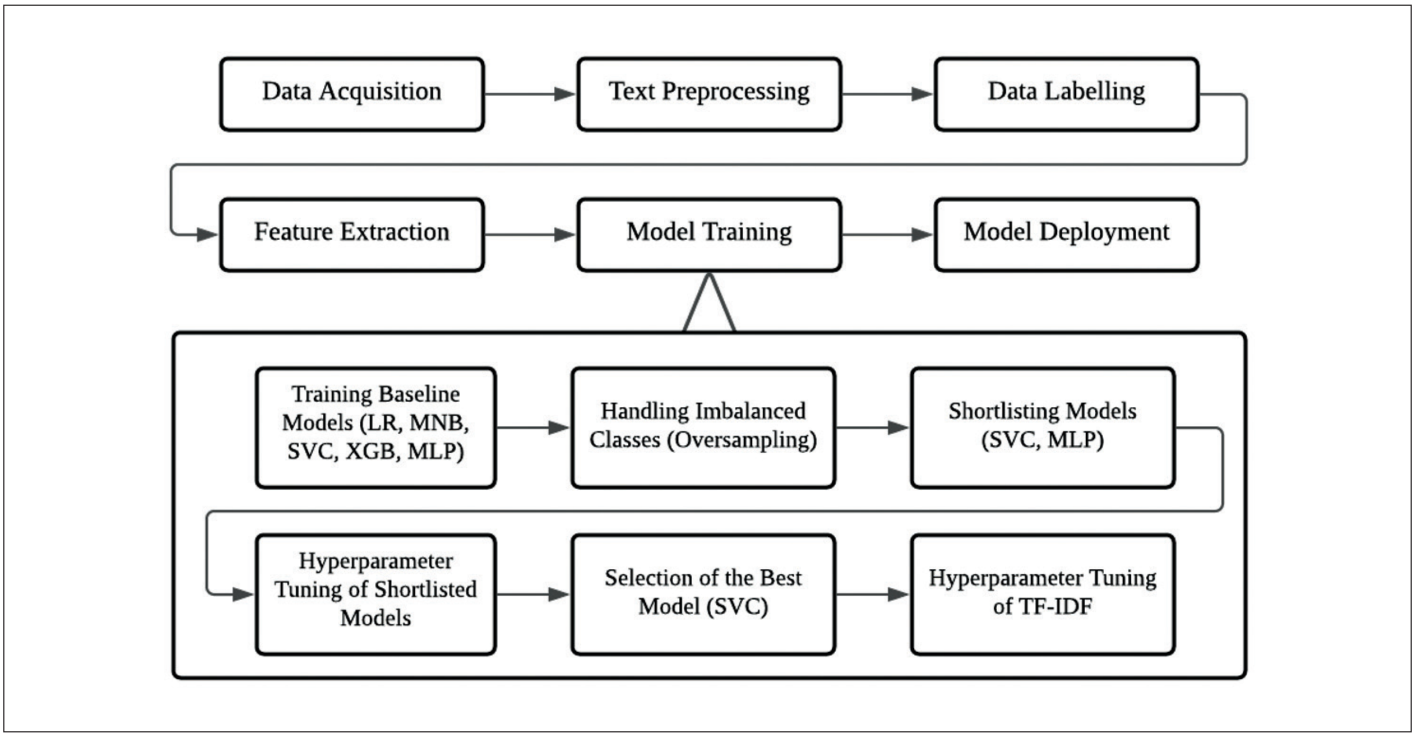

Data Acquisition
Text Preprocessing
Data Labelling
Feature Extraction
Model Training
Model Deployment
Training Baseline Models (LR, MNB, SVC, XGB, MLP)
Handling Imbalanced Classes (Oversampling)
Shortlisting Models (SVC, MLP)
Hyperparameter Tuning of Shortlisted Models
Selection of the Best Model (SVC)
Hyperparameter Tuning of TF-IDF

Figure 1: System Framework

3.2 Dataset

The dataset contains 17543 game reviews, 8363 of which were critic reviews on 300 games scraped from Metacritic's website (https://www.metacritic.com/) and 9180 were user reviews on 100 games collected from Steam (https://store.steampowered.com/) through its web API in July 2021. Table 1 shows the sample reviews obtained from Metacritic and Steam.

Table 1: Sample reviews

	Metacritic critic reviews	Steam user reviews
1	Considering how well Valve got the action down pat, I was very unpleasantly surprised that they managed to fumble the storyline so badly. HL2 starts off with such promise and ends with something akin to what you'd find in "[deleted]" which is a travesty of a mockey of a sham, if there ever was one.	This is a game that I've played on and off for several years now. Sometimes I leave because of issues I have with this game, or personal frustration, or some other reason, but I've always come back to my favorite shooty boat game in the end.
2	This jump & run is a must-buy for every fan of the genre. Even beginners should take a look because Limbo shows that you don't need DirectX 11 or a huge story - just two colors and a lot of love.	Very fun and addictive. Me and my daughter played Rise and Generations ultimate on Nintendo switch a bunch, now starting on world. 100% worth playing!!
3	After spending fifty hours with GW2, I have a lot of praise for ArenaNet's work and the way it changes up some of the typical trappings of the MMO. And yet, I find myself thinking less and less about it each day. It's not a declaration against the product, mind you, but simply a fact that this game still is very much an MMO, and your enjoyment will directly relate to how much you enjoy the genre. For many who were hoping for a clean break from MMO design philosophy, Guild Wars 2 will probably come across as a slight disappointment. It pushes the genre slightly forward, however, and could lead to even further development in the future.	Call of Duty: Black Ops III is one of the best zombies experiences I have ever had in my life. The DLC maps really pull the story line together. When Treyarch decided to add another DLC pack to the game (DLC 5/Zombie Chronicles) it truly brought nostalgic experiences to all players that have played the other games in this franchise. The use of Steam Workshop just keeps this games spirit high. If you love zombies this is the game for you. Despite the flaws (campaign) this is a great game.

3.3 Text Preprocessing

The dataset was preprocessed before it was used to train the models. Firstly, HTML tags and hyperlinks were removed. Next, the texts were converted into lowercase and contractions were expanded. Besides that, special characters were removed. This is followed by removal of numbers, single character words, extra whitespaces and stopwords, except for negations such as “no” and “not” because removal of such words would invert the sentiment of the reviews. Then, tokenization and part-of-speech (POS) tagging were performed. The POS tags were passed on to the lemmatizer so that lemmatization can be carried out based on the context of the tokens.

3.4 Data Labelling

The sentiments of the reviews were labelled as positive, negative or neutral by using pretrained sentiment analysis models of three libraries. The models used were NLTK's VADER Sentiment Intensity Analyzer, Textblob's Pattern Analyzer and Flair's TARS Classifier. A majority voting approach was used to determine the final sentiments of the game reviews. There was a total of 10426 positive reviews, 2975 neutral reviews and 2017 negative reviews. The distribution of the labelled reviews is shown in Figure 2.

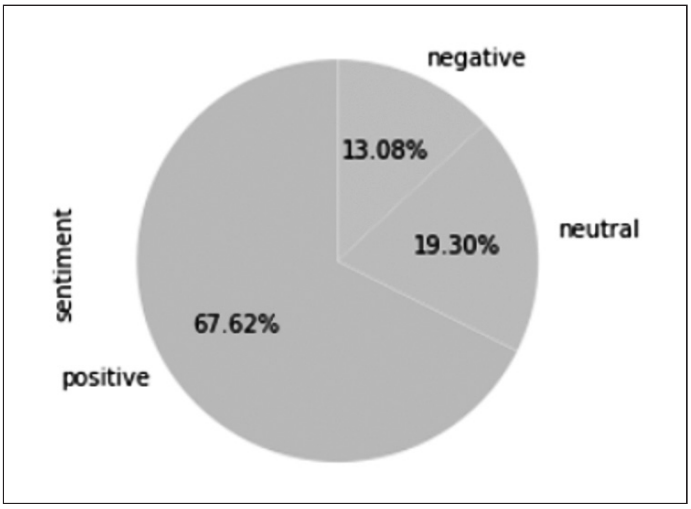


Figure 2: Labelled reviews distribution

3.5 Feature Extraction

The TF-IDF approach has been applied by using Scikit-learn's TfidfVectorizer to perform feature extraction. The “max_features” hyperparameter was set to 2500 while default values were used for other hyperparameters.

3.6 Handling Imbalanced Classes

Oversampling is an approach to deal with data with imbalanced classes by adding more samples to the minority class. Synthetic Minority Oversampling Technique (SMOTE) is one of the oversampling techniques that generates synthetic samples for the minority class. Since the data contains a significantly greater number of positive reviews than neutral and negative reviews which may affect the performance of the models, SMOTE was applied to adjust the distribution of the classes so that all classes have the same number of samples.

3.7 Model Applications

The machine learning algorithms used in this study are as follows:

a. Logistic Regression (LR)
LR is by default used for binary classification but it is extended by the Scikit-learn library to also perform multi-class classification.

b. Multinomial Naïve Bayes (MNB)
MNB is a probabilistic learning method used for classification with discrete features. Scikit-learn's MNB algorithm not only allows the use of integer feature counts, but also fractional counts obtained from TF-IDF.

c. Support Vector Classifier (SVC)
SVC is a classification algorithm that can be used to solve binary and multi-class problems. Scikit-learn's SVC algorithm uses a one-vs-one scheme to support multi-class classification.

d. Extreme Gradient Boosting Classifier (XGB)
XGB is a decision-tree-based ensemble machine learning algorithm that implements gradient boosting. The XGBoost library provides a Scikit-learn wrapper class that allows the XGB algorithm to be used the same way as other Scikit-learn algorithms.

e. Multi-layer Perceptron Classifier (MLP)
MLP is a feedforward Artificial Neural Network (ANN) algorithm that consists of multiple fully connected layers. Scikit-learn's MLP algorithm provides a regularization term that can be used to constraint the size of the weights in the neural network to prevent overfitting.

All the models were trained with their default hyperparameters to obtain their baseline performances except for MLP. Scikit-learn's MLP algorithm has a default architecture that consists of one input layer, one hidden layer with 100 neurons and one output layer, which causes the model to be computationally expensive to train. Therefore, a smaller value for the "hidden_layer_sizes" hyperparameter was set. The MLP model trained comprised 2 hidden layers, with 10 neurons in the first hidden layer and 5 neurons in the second hidden layer. The number of hidden layers and neurons were set arbitrarily as the model only acts as a baseline model before hyperparameter tuning was performed.

3.8 Hyperparameter Tuning

Hyperparameter tuning is the process of determining the combination of hyperparameters which maximizes the model's performance. In order to improve the performance of the models, a Randomized Search Cross Validation with 3 splits was carried out to find the best combination of hyperparameters. In addition, a Grid Search Cross Validation with 3 splits was also performed on TF-IDF to select the best hyperparameters for it to further improve the performance of the models.

4.0 RESULTS AND DISCUSSION

Table 2 and Table 3 show the baseline performance of the models trained on the imbalanced dataset and oversampled dataset obtained through cross validations. Weighted precision, weighted recall and weighted F1-score were used as the metrics as they take into account the number of instances in each class.

Table 2: Baseline performance of all models trained on imbalanced dataset

	Accuracy	Weighted Precision	Weighted Recall	Weighted F1-Score		
				Negative	Neutral	Positive
LR	74.8%	71.7%	74.8%	71.6%	71.1%	72.2%
MNB	68.3%	63.3%	68.3%	57.2%	56.8%	56.9%
SVC	72.9%	69.9%	72.9%	66.6%	67.4%	67.6%
XGB	75.9%	73.5%	75.9%	74.1%	72.6%	73.3%
MLP	69.4%	70.1%	69.4%	70.1%	69.1%	70.0%

Table 3: Baseline performance of all models trained on oversampled dataset

	Accuracy	Weighted Precision	Weighted Recall	Weighted F1-Score			
				Negative	Neutral	Positive	Status
LR	79.3%	79.6%	79.3%	79.1%	79.5%	79.6%	Rejected
MNB	67.4%	67.5%	67.4%	67.2%	66.7%	67.1%	Rejected
SVC	87.7%	88.7%	87.7%	87.4%	87.4%	88.0%	Accepted
XGB	79.7%	80.1%	79.7%	80.0%	79.9%	79.5%	Rejected
MLP	86.8%	87.0%	86.8%	86.5%	86.8%	86.9%	Accepted

Based on the results in Table 2 and Table 3, oversampling has significantly improved the performance of all the models except MNB which was observed to have a drop in accuracy and weighted recall. The improved performance was due to the class distribution being balanced after performing duplication of data to synthesize new data from the minority classes.

MNB performed poorer on the oversampled data and was the worst-performing model most likely due to its assumption that all features are independent which is rarely true in real-world use cases where there are a large number of features.

The most significant improvement of performance was observed in SVC and MLP. These two models worked well with the larger, balanced dataset and they were also the two best-performing models. Therefore, they have been shortlisted for hyperparameter tuning through Randomized Search Cross Validation. The best combinations of the models' hyperparameters and their performances are shown in Table 4 and Table 5 respectively.

Table 4: Best hyperparameter values of SVC and MLP

	Best hyperparameter values
SVC	kernel: rbf, gamma: scale, C: 10
MLP	"solver": "adam", "max_iter": 150, "hidden_layer_sizes": (10,), "alpha": 0.0001, "activation": "relu"

Table 5: Performance of fine-tuned SVC and MLP

	Accuracy	Weighted Precision	Weighted Recall	Weighted F1-Score			
				Negative	Neutral	Positive	Status
SVC	89.7%	90.0%	89.7%	89.2%	89.7%	90.1%	Accepted
MLP	87.0%	87.1%	87.0%	86.7%	87.0%	87.0%	Rejected

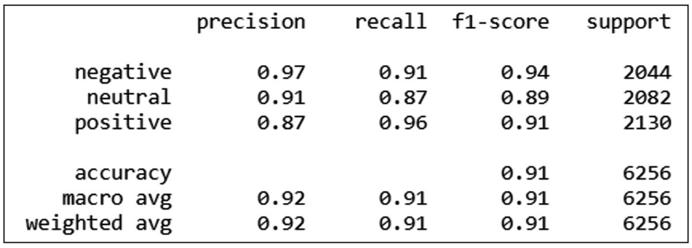
Table 5 shows that hyperparameter tuning has improved both models' performance. Hyperparameter tuning is able to improve the models' performance because it determines the best combinations of hyperparameters which produce optimal models that minimize the loss functions.

SVC outperformed MLP in terms of accuracy, weighted precision, weighted recall and weighted F1-score after the hyperparameter tuning. Hence, SVC as the best-performing model among all the models, was selected to test the effect of hyperparameter tuning of TF-IDF on its performance. The tested values of the TF-IDF hyperparameters and the best values determined by Grid Search Cross Validation are shown in Table 6.

Table 6: Tested hyperparameter values and the best values of TF-IDF

Hyperparameter	Tested values	Best value
max_features	2500, 5000, 10000	2500
max_df	0.25, 0.5, 0.75	0.25
ngram_range	(1, 1), (1, 2), (1, 3)	(1, 1)

As can be seen in Table 6, SVC had the best performance under the condition in which the top 2500 terms across the corpus ordered by term frequency were considered, terms that occurred in more than 25% of the documents were ignored and only unigrams were extracted. The performance of SVC trained with the features extracted by the fine-tuned TF-IDF is shown in Figure 3.



```
              precision    recall  f1-score   support

    negative       0.97      0.91      0.94      2044
     neutral       0.91      0.87      0.89      2082
    positive       0.87      0.96      0.91      2130

    accuracy                           0.91      6256
   macro avg       0.92      0.91      0.91      6256
weighted avg       0.92      0.91      0.91      6256
```

Figure 3: Classification report of SVC with fine-tuned TF-IDF

Based on the classification report in Figure 3, hyperparameter tuning on TF-IDF has improved SVC's performance. The model has achieved an accuracy of 91%, precision of 92%, recall of 91% and F1-score of 91%.

5.0 CONCLUSION

In conclusion, by training five machine learning models with game reviews acquired from Metacritic and Steam, this study has provided clear evidence that oversampling will lead to an improved performance for most models. In addition to that, better results are also evident after hyperparameter tuning on the models and TF-IDF have been performed.

Support Vector Classifier with an accuracy of 91 percent has emerged as the best-performing model among the five models after comparing the accuracy scores for each model. The way SVC performs classification, which is based on hyperplanes instead of probabilities, is likely to be the main contribution to its outstanding performance. It is therefore the ideal model to be used for text classification tasks with a large number of features such as sentiment analysis.

Future research into machine learning algorithms for sentiment analysis of game reviews should focus on exploring the sentiment of emojis and emoticons since it is common for users to incorporate them in their reviews. Furthermore, experimenting with ensemble methods are required to gain more insight into building a more robust sentiment analysis model.

6.0 ACKNOWLEDGEMENTS

Authors thank the Faculty of Computing and Information Technology, Tunku Abdul Rahman University College for financial support and resources to carry out this study. ■

REFERENCES

[1] Al Amrani, Y., Lazaar, M., & El Kadiri, K. E. (2018). Random forest and support vector machine based hybrid approach to sentiment analysis. Procedia Computer Science, 127, 511-520. https://doi.org/10.1016/j.procs.2018.01.150.

[2] Alboaneen, D. A., Tianfield, H., & Zhang, Y. (2017). Sentiment analysis via multi-layer perceptron trained by meta-heuristic optimisation. 2017 IEEE International Conference on Big Data (Big Data), 4630-4635. https://doi.org/10.1109/BigData.2017.8258507.

[3] Ashari, A., Paryudi, I., & Tjoa, A. M. (2013). Performance comparison between Naïve Bayes, decision tree and k-nearest neighbor in searching alternative design in an energy simulation tool. International Journal of Advanced Computer Science and Applications (IJACSA), 4(11). https://doi.org/10.14569/IJACSA.2013.041105.

[4] Boyd, C. R., Tolson, M. A., & Copes, W. S. (1987). Evaluating trauma care: the TRISS method. Trauma Score and the Injury Severity Score. The Journal of trauma, 27(4), 370-378. https://doi.org/10.1097/00005373-198704000-00005.

[5] Britto, L. F., & Pacıfico, L. D. (2020) Evaluating Video Game Acceptance in Game Reviews using Sentiment Analysis Techniques. Proceedings of SBGames 2020, 399-402.

[6] Chakraborty, S., Mobin, I., Roy, A., & Khan, M. H. (2018). Rating Generation of Video Games using Sentiment Analysis and Contextual Polarity from Microblog. 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), 157-161. https://doi.org/10.1109/CTEMS.2018.8769149.

[7] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785-794. https://doi.org/10.1145/2939672.2939785.

[8] Tan, J. Y., Chow, A. S. K., & Tan, C. W. (2021). Sentiment Analysis on Game Reviews: A Comparative Study of Machine Learning Approaches. International Conference on Digital Transformation and Applications (ICDXA2021), 209-216.

[9] Lau, K. W., & Wu, Q. H. (2003). Online training of support vector classifier. Pattern Recognition, 36(8), 1913-1920. https://doi.org/10.1016/S0031-3203(03)00038-4.

[10] Prabhat, A., & Khullar, V. (2017). Sentiment classification on big data using Naïve Bayes and logistic regression. 2017 International Conference on Computer Communication and Informatics (ICCCI), 1-5. https://doi.org/10.1109/ICCCI.2017.8117734.

[11] Xia, R., Zong, C., & Li, S. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. Information sciences, 181(6), 1138-1152. https://doi.org/ 10.1016/j.ins.2010.11.023.

[12] Zuo, Z. (2018). Sentiment Analysis of Steam Review Datasets using Naive Bayes and Decision Tree Classifier. Student Publications and Research - Information Sciences. http://hdl.handle.net/2142/100126.

PROFILES

TAN JIE YING received her Bachelor's degree of Computer Science (Honours) in Data Science from Tunku Abdul Rahman University College (TAR UC), Malaysia in 2022. She is passionate about creating value from data using Machine Learning, Natural Language Processing and Computer Vision.
Email address: tanjy-wp17@student.tarc.edu.my

ANDY CHOW SAI KIT is a Bachelor of Computer Science (Honours) in Data Science graduate from Tunku Abdul Rahman University College (TAR UC), Malaysia in 2022. His research interest is in Artificial Intelligence and Machine Learning.
Email address: andycsk-wp17@student.tarc.edu.my

DR TAN CHI WEE received BCompSc(Hons) and PhD degrees in year 2013 and 2019 respectively in Universiti Teknologi Malaysia. Currently, he is a Senior Lecturer cum Programme Leader at Tunku Abdul Rahman University College and actively involved in the Centre of Excellence for Big Data and Artificial Intelligent (CoE) and become the research group leader for Audio, Image and Video Analytics Group under Centre for Data Science and Analytics (CDSA). Dr Tan's main research areas are Computer Vision (CV), Image Processing (IP) and Natural Language Processing (NLP) and Artificial Intelligence (AI). He is an enthusiastic researcher experienced in conducting and supporting research into Image Processing. Being a meticulous and analytical researcher with Train-The-Trainer certificate of many years of educational and hands-on experience, he was invited to Université d'Artois (France) under Marie Skłodowska-Curie Research and Innovation Staff Exchange (RISE) programme for collaborative research between European countries with Southeast Asian countries on motion detection and computer vision and being involved in industry project as professional consultant.
Email address: chiwee@tarc.edu.my