SENTIMENT ANALYSIS ON MALAYSIAN AIRLINES WITH BERT

(Date received: 13.01.2022/Date accepted: 03.04.2022)

Huay Wen Kang[1], Kah Kien Chye[1], Zi Yuan Ong[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 has been a popular research area in Natural Language Processing (NLP), where sentiments expressed through text data including positive, negative and neutral sentiments are analyzed and predicted. It is often performed to evaluate customer satisfaction and understand customer needs for businesses. In the airline industry, millions of people today use social networking sites such Twitter, Skytrax, TripAdvisor and more to express their emotions, opinions, reviews and share information about the aircraft service. This creates a treasure trove of information for the airline company, showcasing different points of views about the airline's brand online and providing insightful information. Hence, this paper experiments with six different sentiment analysis models in order to determine and develop the best model to be used. The model with the best performance was then used to determine the social status, company reputation, and brand image of Malaysian airline companies. In conclusion, the BERT model was found to have the best performance out of the six models tested, scoring an accuracy of 86 percent.

Keywords: Supervised Learning, Ensemble Learning, Deep Learning, Transfer Learning, Airline Sentiment

1.0 INTRODUCTION

Social media is a huge source of opinionated text data that comes in various forms such as Facebook posts, tweets, comments, replies, product reviews and others. The Internet era has altered how people express their thoughts and opinions. It is now primarily done through blog posts, online forums, product review websites, social media, and so on. It is extremely difficult to manually track multiple social media platforms filled with this data for large airlines, hence this is where natural language processing comes in useful. As a result, the goal of this project is to conduct research and analysis on Malaysian airline companies' social status, company reputation, and brand image, specifically Malaysia Airlines, AirAsia, and Malindo. In order to develop a sentiment analysis model about airlines, various methods and techniques were explored including supervised learning, ensemble learning, deep learning, and transfer learning in this research work.

2.0 LITERATURE REVIEW

Sousa *et al*. (2019) has predicted the following movements of the Dow Jones Industrial (DJI) Index in the task of stock market analysis to evaluate BERT. 582 financial news were crawled from various news websites such as CNBC, Forbes and New York Times using Selenium tool to serve as the dataset and it was manually annotated with sentiment classes such as positive, negative and neutral. WordPiece is used to tokenize and transform each document into a token sequence. Then, the parameter for attention head was set to BERT BASE as it is smaller and is able to cater to their limited computational power. A 10-fold cross validation was done before training the models using labeled data. Based on the results, BERT clearly outperformed other methods compared to Support Vector Machines, Naive Bayes and Convolutional Neural Network with the highest F1-score of 72.5 percent. This analysis was used as an indicator of falling and rising of the economy in the day.

Zhang, Wang and Liu (2018) give a detailed overview on deep learning and also a comprehensive survey of the current applications of deep learning in sentiment analysis. Through their analysis, they have concluded that many deep learning techniques have shown state-of-the-art results for sentiment analysis tasks. This paper touches upon the concept of Neural Networks, Deep Learning, Word embedding, Autoencoder and denoising autoencoder, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short-Term Memory Network (LSTM), Attention Mechanism with RNN, Memory Network (MemNN) and Recursive Neural Network (RecNN). This paper also surveys the various kinds of sentiment analysis tasks, like document-level sentiment classification, sentence-level sentiment classification, aspect-level sentiment classification, aspect extraction and categorization, sentiment composition, opinion holder extraction, temporal opinion mining, sentiment analysis with word embedding, sarcasm analysis, emotion analysis, the use of multimodal data for sentiment analysis, and resource-poor language and multilingual sentiment analysis.

Agarwal, Xie, Rambow and Passonneau (2011) proposed a method of analysis sentiment for Twitter data. They have prepared an emoticon dictionary by labeling 170 emoticons listed on Wikipedia with their emotional state, such as "Extremely-positive", "Extremely-negative", "Positive", "Negative" and "Neutral". An acronym dictionary was created for 5,184

acronyms. They consider all words found inside WordNet as English words, while the rest are stop words or non-English words. A combination of the Dictionary of Affect in Language (DAL) and WordNet to get the polarity of words. Several models were investigated, including a Unigram model, a senti-features model, a tree kernel model, a unigram + senti-features model and tree kerna + senti-features model. In the feature-based model, feature analysis was done and revealed that the most important features are those that combine prior polarity of words and their parts-of-speech tags. The unigram model, senti-features, tree kernel, unigram + senti-features and kernel + senti-feature models managed an average accuracy of 56.58 percent, 56.31 percent, 60.60 percent, 60.50 percent and 60.83 percent respectively.

2.1 Existing Method

Naive Bayes is a classifier family based on Bayes' popular probability theorem that is well suited for creating simple but powerful models, particularly in the area of textual classification (Lewis, 1998). Compared to many models, Naive Bayes do not need much data to train and tune the required parameters (Abbas *et al.*, 2019). Therefore, it is less complicated to implement, and is more reliable in comparison to more complex and slower algorithms. The Naive Bayes Classifier has multiple variations, which are Multinomial Naive Bayes (MNB), Bernoulli Naive Bayes (BNB) and Gaussian Naive Bayes (GNB). A Naive Bayes classifier is essentially a probabilistic machine learning model for classification that employs the Bayes theorem, which describes the probability of an event based on prior knowledge of the conditions that may be associated with the event. Multinomial Naive Bayes was selected instead of the other Naive Bayes types because MNB is an improved version of Multivariate Bernoulli Naive Bayes model (BNB) and considers word frequency and information and thus obtains better accuracy.

Support Vector Machine (SVM) (Cristianini & Shawe-Taylor, 2000) is a learning technique that excels at sentiment analysis because it can significantly reduce the need for labelled training instances in both the standard inductive and transductive settings (Phienthrakul, 2009). SVM creates that hyperplane by transforming data using mathematical functions known as "Kernels." Types of Kernels are linear, sigmoid, RBF, non-linear, polynomial, etc. The kernel function used in SVM determines its performance. As a result, selecting the appropriate kernel will improve classification efficiency. Previously, research on sentiment classification using SVM and multiple kernel functions yielded promising results with high accuracy. However, in order to adapt to new changes and be more flexible in today's world, SVM has evolved into various versions. Other researchers had proposed various extensions of SVM such as support vector classifier, transductive support-vector machines, multiclass SVM, structured SVM, bayesian SVM, and others during their research. Hence, it is possible that other combination methods may improve sentiment classification efficiency compared to conventional SVM. The support vector classifier is then used in this research based on this idea.

The Random Forest algorithm (Breiman, 2001) consists of a large number of individual decision trees that operate in an ensemble manner, resulting in a forest of trees. Rather than using best split among all variables to split each node, Random Forest randomly selects a subset of predictors and narrows down the best among them. The tree grows with randomly selected features and is not pruned, resulting in the Random Forest algorithm having great accuracy (Breiman and Cutler, 2004). Random Forest algorithm requires each individual tree to have low correlation with each other to perform well. Random Forest has been proven to be a suitable classifier to perform sentiment analysis in multiple research projects. Based on a sentiment analysis research on Malaysian mobile digital payment applications done by Balakrishnan *et al.* (2020), Random Forest achieved the highest accuracy of 75.62 percent and f1-score of 71.99 percent amongst other algorithms including Support Vector Machine, Naive Bayes and Decision Tree. Likewise, Hedge and Padma (2017) also managed to obtain an accuracy of 72 percent when using Random Forest to perform sentiment analysis for mobile product reviews in Kannada.

Compared to a single model, the ensemble method improves predictive performance by combining the results of several models. This eliminates the possibility of overfitting while improving overall performance (Yu *et al.*, 2010). Furthermore, ensemble methods are frequently used to solve the class imbalance problem using multi-objective optimization algorithms (Yang *et al.*, 2020) and to minimise the number of features better than existing ensemble algorithms such as AdaBoost and Gradient Boosting. In Wan and Gao's (2016) experiment with an ensemble sentiment classification system for airline service analysis, the ensemble classifier using a majority vote method with five classifiers achieved the highest accuracy of 84.2 percent when compared to other single classifiers.

The Recurrent Neural Network (RNN) (Marhon *et al.*, 2013) family of models, particularly the LSTM networks, has been demonstrated to be the most effective sequence model used in practical applications (Goodfellow, Bengio and Courville, 2016). To learn from both forward and backward time dependencies, Bi-directional LSTMs are used. Each unit in a Bi-derectional LSTM is divided into two units that share the same input and are connected to the same output. One unit is used for the forward time sequence, while the other is used for the reverse time sequence. As a result, Bi-directional LSTM is useful for learning from long-spanning time-series data, and it produces better results without increasing training time.

In 2018, the Bidirectional Encoder Representations from Transformers (BERT) (Devlin, Chang, Lee, & Toutanova, 2019) model was introduced to quickly and effectively create a high-quality model with minimal effort and training time using the PyTorch interface, regardless of the specific NLP task, and produce state-of-the-art results. Recently, a sentiment analysis using the BERT model on the impact of coronavirus on social life achieved 94 percent validation accuracy on the collected data sets (Singh *et al.*, 2021). In short, BERT is one of the most powerful NLP models available today, requiring only a small amount of data while achieving cutting-edge results with minimal task-specific adjustments for a wide range of NLP tasks such as named entity recognition, language inference, semantic similarity, question answering, and classification like sentiment analysis.

3.0 METHODOLOGY

This section talks about the datasets used and the design of the models.

3.1 Datasets

The training and testing dataset are sentiment datasets about US airlines (US Airways, United, Virgin America, Delta, Southwest, and American) from Twitter which can be obtained from Kaggle. There is a total of 9,178 negative tweets, 3,099 neutral tweets, and 2,363 positive tweets. For model training, it is divided into a train and test set with an 80:20 ratio. The models trained with this dataset are expected to be generalizable and capable of handling all airline reviews, regardless of company or country, as long as they are all written in English. The following are the 13 columns from the mentioned dataset:

1. tweet_id: Unique ID of the tweet
2. airline_sentiment: Sentiment of the tweet
3. airline_sentiment_confidence: Confidence level of the sentiment of the tweet
4. negativereason: Negative reason if sentiment is negative
5. negativereason_confidence: Confidence level of the negative reason if sentiment is negative
6. airline: Name of the airline
7. name: Username of the tweet author
8. retweet_count: Number of retweets
9. text: Text data of the tweet
10. tweet_coords: Location coordinates of the person who posted the tweet
11. tweet_created: date and time of the tweet posted
12. tweet_location: Location of the person who posted the tweet
13. user_timezone: Timezone of the person who tweeted

Table 1: Sample Data for Malaysian Airlines

	Username	Review	Airline
0	Ritesh Aggarwal	Booked flight from Delhi to Sydney for my pare...	mas
1	Tara-lets T	I was looking for affordable airfare for multi...	mas
2	w KING	I booked a flight from Delhi to KL on the 2 Ju...	mas
3	paul	Booked a flight with these thieves from Jakart...	mas
4	Mickey Pearson	With the E-Voucher Malaysia Airlines is forcin...	mas

3.2 Text Preprocessing

Data preprocessing was done on the training data before being used to train the models for traditional machine learning methods. The steps are as follows:

1. Emoticons and emojis are converted to text that expresses their textual meaning using the emoticon library (Shah & Rohilla, 2018) as emoticons and emojis carry sentiment.
2. All words in the message are converted into lowercase in order to normalize the text. To reduce noise and ease further process, symbols, punctuations, hyperlink, extra whitespace, new lines along with digits are removed.
3. Stopwords are removed and lemmatization is done to return the words back to their root word since the word would be disrupted by an irrelevant inflection like a simple plural or present tense inflection. On the other hand, BERT does not require text preprocessing such as stopword removal, lemmatization and others because the model has a specific way of dealing with out-of-vocabulary words using its own fixed vocabulary and the BERT tokenizer for text formatting.
4. Feature selection is done to transform the text into unigram word vectors using CountVectorizer.
5. Following that, resampling techniques are used to address the data imbalance between positive, neutral, and negative sentiment. This is due to the fact that class imbalance may have an impact on the accuracy and performance of the models. As a result, random undersampling is used to randomly reduce the majority class to the desired ratio against the minority class, and it is combined with SMOTE oversampling (Chawla *et al.*, 2002), which is oversampling by creating "synthetic" examples of the minority class. Both undersampling and SMOTE use the default sampling strategy, which is "not minority" for undersampling and resamples all classes except the minority, and "not majority" for SMOTE, which oversamples all classes except the majority. Both will bring the unequal classes into balance. The training dataset will have N = 21,867 samples from the initial 11,712 samples at the end of the process.

3.3 Model Applications

Naive Bayes (NB) is based on the Bayes theorem, and can be used for classification challenges. Multinomial Naive Bayes (MNB) is one probabilistic learning method variant of NB which is commonly used in Natural Language Processing (NLP). MNB performs calculation of probabilities based on probabilities of causal factors and is useful to model feature vectors where each value represents the number of occurrences of a term, where in the case of NLP a text can be considered as a particular instance of a dictionary and the relative frequency of all terms in the text provide enough information for inferring a belonging class. To avoid the possibility of getting zero probability, additive laplace smoothing parameter is used as one of the hyperparameters, whether to learn class prior probabilities or not and prior probabilities of the classes. Lower alpha values are used to avoid getting a likelihood of around 0.5, which will not be helpful for the results. The optimized set of parameters for the model after hyperparameter tuning: `MultinomialNB(alpha=0.5)`.

A simple, linear Support Vector Classifier (SVC) is proposed to classify the airline data into three different classes. SVC is a different implementation of the SVM algorithm which is implemented in terms of liblinear rather than libsvm (Asif *et al.*, 2020). It is just a thin wrapper around libsvm but has more options for penalties and loss functions and has higher scalability to larger samples. It supports both dense and sparse inputs and also handles multi classes. The classifier will consider each unique word present in the sentence, along with multiple word expressions which are suitable for text classification. A linear hyperplane is used to identify the text involved in the airline dataset and separate them into three categories by implementing this mechanism. Common features such as gamma are set at 0.5, regularisation parameter (c) of 100, and the use of probability estimates in conjunction with a balanced mode after parameter tuning. The tuned parameters for this module are as follows: `SVC (C=100, class_weight='balanced', gamma=0.5, kernel='linear', probability=True)`.

The hyperparameters of the Random Forest algorithm include the number of decision trees in the forest, number of features considered by each tree when splitting a node, number of

levels in each decision tree, min number of data points allowed in a leaf nod, min number of data points placed in a node before the node is split, and number of trees in the forest. The parameters of the random forest are the variables and thresholds used to split each node learned during training. The line below shows the optimized parameters for the model after hyperparameter tuning: `RandomForestClassifier(max_depth=100, max_features='sqrt', min_samples_leaf=2, min_samples_split=10, n_estimators=800).`

For the voting classifier, the previous three algorithms (Multinomial Naive Bayes, Random Forest and LinearSVC) with their respective optimized parameters set are passed into the classifier and used soft voting, which predicted the class with the highest summed probability from models.

Bidirectional LSTMs are an extension of traditional LSTMs in which two LSTMs are trained instead of one on the input sequence, with the second LSTM trained on the reversed copy of the input sequence, providing additional context to the network and resulting in faster and even more complete learning on the problem. The data was preprocessed before being fed into the Bidirectional LSTM. To summarise, the sentiments were converted into a binary class matrix before tokenizing the text, converting it to an encoded form, and finally padding all of the encoded text to the same length. The proposed Bidirectional LSTM model is simple, with only 7 layers, including one embedding input layer and one embedding layer, followed by one Convolutional layer, one max pooling layer, one bidirectional LSTM layer, one dropout layer, and finally one dense layer to generate 3 classification classes.

BertForSequenceClassification model with an added single linear layer on top for classification that serves for sentence classifier purposes is used in this project. After the airline data sets are fed in, the entire pre-trained BERT model and the additional untrained classification layer is trained on the specific task, that is the multi-class classification. With the exception of the batch size, learning rate, and number of training epochs, most model hyperparameters remain at their default values for fine-tuning. The author discovered a set of feasible values that work well across all NLP tasks, so the batch size is set to 32, with a 2e-5 learning rate (Adam) from the recommended range values and a maximum epoch value of 4. In addition, the model will be trained using the PyTorch framework. The model will go through the standard PyTorch training cycle and loop four times, iterating through the mini-batches, performing a feedforward pass for each batch, computing the loss, performing backpropagation for each batch, and finally updating the gradients. In short, the power of the pre-trained transformers (BERT) model and the PyTorch framework will be leveraged by this fine-tuned model.

4.0 RESULTS AND DISCUSSION

The following table shows the results of all models. Unigram features were used for Multinomial Naive Bayes, Linear Support Vector Classifier, Random Forest and Voting Classifier. The testing dataset has N = 2,928 samples, split from the US Airline dataset.

Among the traditional machine learning techniques (Multinomial Naive Bayes, Linear Support Vector Classifier and Random Forest), Multinomial Naive Bayes achieved the highest results. Hence it is further used to test using bigram and trigram features.

Table 2: Results for all models

	Accuracy	Precision	Recall	F1-Score	Status
Multinomial Naive Bayes	70.4%	77.7%	78.2%	77.9%	Rejected
Linear Support Vector Classifier	66.7%	74.0%	73.8%	73.9%	Rejected
Random Forest	66.5%	76.3%	77.3%	76.5%	Rejected
Voting Classifier	68.4%	79.2%	80.2%	78.7%	Rejected
Bidirectional LSTM	77.0%	77.0%	77.0%	77.0%	Rejected
BERT	86.0%	86.0%	86.0%	86.0%	Accepted

Table 3: Results for Multinomial Naive Bayes using different n-gram orders

	Accuracy	Precision	Recall	F1-Score
Unigram	70.4%	77.7%	78.2%	77.9%
Bigram	53.4%	64.9%	66.9%	65.6%
Trigram	41.4%	61.3%	65.8%	59.1%

According to the experimental results, the BERT model outperformed the other six models with an accuracy of 86.0 percent. It demonstrates that in airline sentiment analysis, transfer learning approaches outperform traditional machine learning algorithms. This may be due to that it is trained on available plain text corpora that are larger than reviews only, as it already encode a lot of information of generic English text as it was trained from Wikipedia and book corpus (in total 3,300 million words) (Devlin *et al.*, 2019), therefore allowing better performances. Experiments show that BERT outperforms unsupervised text classification, such as the standard preprocessing process, which includes decapitalization, punctuation removal, stopword removal, and emoji conversion to text (Kang *et al.*, 2021).

The Bidirectional LSTM was the second-best performer, with an accuracy of 77.0 percent. This result outperforms all traditional machine learning methods tested, proving the hypothesis that deep learning models outperform traditional machine learning algorithms. This could be because the Bidirectional LSTM considers both backward and forward time dependencies.

Among the various n-gram models tested for Multinomial Naive Bayes, unigram features produce the best results (70.4 percent). With each increase in n-gram, the accuracy decreases even further. This could be because as n-gram length increases, so does the frequency of any given n-gram, and it may not generalise well to a different data set, resulting in lower accuracy.

5.0 CONCLUSION

In conclusion, this paper investigated various sentiment analysis models and discovered that Multinomial Naive Bayes is the best machine learning model out of the four chosen, with an accuracy of 70.4 percent. It was also demonstrated that the Unigram model of the Multinomial Naive Bayes outperforms the Bigram

and Trigram models. On the other hand, deep learning methods outperform traditional machine learning methods, as evidenced by the results, with both the Bidirectional LSTM (77 percent accuracy) and BERT model (86 percent accuracy) outperforming all traditional machine learning methods. Finally, we discovered that the BERT model outperforms all six models tested.

A high-accuracy sentiment analysis model has been successfully developed and applied to a new airline reviews dataset. According to the crawled data, the majority of the 1371 reviews were classified as negative (N=1118, 81.55 percent), 226 positive (16.48 percent), and 27 neutral (1.97 percent). This suggests that the majority of Malaysian Airlines' customers are dissatisfied with the airline's services. Companies can get better directions and make better business decisions based on sentiment analysis insights, such as providing better training for their flight crews if negative sentiments toward their services are discovered.

However, the crawled data used in this project is relatively small; such a small and sparse dataset may not be truly representative of the larger population, and it also limits several types of data analysis, such as trend analysis, due to data scarcity. As a result, this project can be improved by crawling more data for the dataset and reducing the sparsity of the data, making it more representative of the actual population and allowing for more complex analysis.

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 project. ■

REFERENCES

[1] Abbas, M., Memon, K., Jamali, A., Memon, S. and Ahmed, A. (2019). 'Multinomial Naive Bayes Classification Model for Sentiment Analysis'. International Journal of Computer Science and Network Security, 19(3), pp.62-67.

[2] Agarwal, A., Xie, B., Rambow, O. and Passonneau, R. (2011). 'Sentiment Analysis of Twitter Data'. Proceedings of the Workshop on Languages in Social Media.

[3] Asif, M., Ishtiaq, A., Ahmad, H., Aljuaid, H. and Shah, J. (2020). 'Sentiment analysis of extremism in social media from textual information'. Telematics and Informatics, 48, p.101345.

[4] Balakrishnan, V., Selvanayagam, P.K., Yin, L.P. (2020). 'Sentiment and Emotion Analyses for Malaysian Mobile Digital Payment Applications'. ACM International Conference Proceeding Series. Association for Computing Machinery, pp. 67–71.

[5] Breiman, L. (2001). Machine Learning, 45(1), 5-32. https://doi.org/10.1023/a:1010933404324

[6] Breiman, L., Cutler, A. (2004). 'RFtools for Predicting and Understanding Data'. Interface Workshop 1–62.

[7] Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal Of Artificial Intelligence Research, 16, 321-357. https://doi.org/10.1613/jair.953

[8] Cristianini, N., & Shawe-Taylor, J. (2000). 'An Introduction to Support Vector Machines and Other Kernel-based Learning Methods'. https://doi.org/10.1017/cbo9780511801389

[9] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv, abs/1810.04805.

[10] Goodfellow, I., Bengio, Y. and Courville, A. (2016). 'Deep Learning'. MIT Press.

[11] Hegde, Yashaswini & Padma, S.K. (2017). Sentiment Analysis Using Random Forest Ensemble for Mobile Product Reviews in Kannada. 777-782. IEEE 7th International Advance Computing Conference (IACC) 2017, Hyderabad, India. IEEE, 10.1109/IACC.2017.0160.

[12] Kang, HW., Chye, KK., Ong, ZY., Tan, CW. (2021). 'The science of emotion: Malaysian airline sentiment analysis using BERT Approach', International Conference On Digital Transformation And Applications 2021, ICDXA 2021, Kuala Lumpur, Malaysia.

[13] Lewis, D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. Machine Learning: ECML-98, 4-15. https://doi.org/10.1007/bfb0026666

[14] Marhon, S., Cameron, C., & Kremer, S. (2013). Recurrent Neural Networks. Intelligent Systems Reference Library, 29-65. https://doi.org/10.1007/978-3-642-36657-4_2

[15] Neel Shah & Shubham Rohilla (2018). Description of the emot library, v2.2. Github [Online]. Available at: https://github.com/NeelShah18/emot (Accessed: 19 April 2021)

[16] Phienthrakul, T., Kijsirikul, B., Takamura, H. and Okumura, M. (2009). 'Sentiment Classification with Support Vector Machines and Multiple Kernel Functions'. Neural Information Processing, pp.583–592.

[17] Singh, M., Jakhar, A.K. and Pandey, S. (2021). 'Sentiment analysis on the impact of coronavirus in social life using the BERT model'. Social Network Analysis and Mining, 11(1).

[18] Sousa, M.G., Sakiyama, K., Rodrigues, L.S., Moraes, P.H., Fernandes, E., Matsubara, E.T. (2019). 'BERT for stock market sentiment analysis', International Conference on Tools with Artificial Intelligence, ICTAI. IEEE Computer Society, pp.1597–1601.

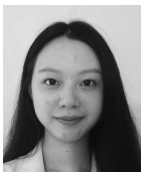
[19] Wan, Y., Gao, Q. (2016). 'An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis'. 15th IEEE International Conference on Data Mining Workshop, ICDMW 2015. Institute of Electrical and Electronics Engineers Inc., pp. 1318–1325.

[20] Yang, K., Yu, Z., Wen, X., Cao, W., Chen, C, Wong, H., You, J. (2020). 'Hybrid Classifier Ensemble for Imbalanced Data'. IEEE Transactions on Neural Networks and Learning Systems 31, 1387–1400.

[21] Yu, H., Lo, H., Hsieh, H., Lou, J., McKenzie, T., Chou, J., Chung, P., *et al.* (2010). 'Feature engineering and classifier ensemble for KDD cup 2010'. JMLR: Workshop and Conference Proceedings 1–12.

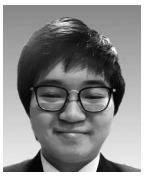
[22] Zhang, L., Wang, S. and Liu, B. (2018). 'Deep learning for sentiment analysis: A survey'. WIREs Data Mining and Knowledge Discovery, 8(4).

PROFILES



HUAY WEN KANG received BCompSc (Hons) Computer Science in Data Science from Tunku Abdul Rahman University College (TAR UC) in 2022. She is passionate about Artificial Intelligence (AI) and Data Science and aspires to be a professional data scientist. Her research interests involved Data Analytics, Code Mixed Sentiment Analysis and Computer Vision.
Email address: kanghw-wp17@student.tarc.edu.my

KAH KIEN CHYE graduated from Tunku Abdul Rahman University College (TAR UC) in 2022 with a BCompSc (Hons) (Computer Science in Data Science). He is a recent graduate who is passionate about applying data science and machine learning skillsets to SAP systems, an enterprise software to manage business operations and customer relations in which 77% of the world's transaction revenue touches an SAP system. His research interests include natural language processing, image processing, data mining, process mining, as well as advanced analytics.
Email address: chyekk-wp17@student.tarc.edu.my



ZI YUAN ONG received his Bachelor's degree of Computer Science (Honours) in Data Science from Tunku Abdul Rahman University College (TAR UC) in 2022. He is a fresh graduate who is passionate about Artificial Intelligence (MI), Machine Learning (ML) and Data Science. His research interests include artificial intelligence, natural language processing, image processing, data analytics as well as computer vision.
Email address: ongzy-wp17@student.tarc.edu.my

DR CHI WEE TAN 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