INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN PROGRESSIVE EDUCATION & DEVELOPMENT

OPEN ACCESS JOURNAL

Vol 11, Issue 4, (2022) E-ISSN: 2226-6348

A Preliminary Study on Students Perception of Online Distance Learning: A Sentiment Analysis on Twitter

Ahmad Haris Marzuqi Shamsul Bahar, Norzatul Bazamah Azman Shah, Rashidah Ramle

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM), 73000 Jasin, Malacca, Malaysia

Corresponding Author Email: norzatulb@uitm.edu.my

To Link this Article: http://dx.doi.org/10.6007/IJARPED/v11-i4/15187 DOI:10.6007/IJARPED/v11-i4/15187

Published Online: 26 November 2022

Abstract

Online Distance Learning (ODL) is any learning activities in the formal, informal, and non-formal domains aided by information and communication technology, to reduce physical and psychological distance and promote interactivity and communication among learners, learning sources, and facilitators. Sentiment analysis is the process of identifying and extracting subjective information from text using natural language processing and text analysis techniques. However, students' sentiments about ODL are varied, and many factors may contribute to that. This study aims to identify students' sentiments about online distance learning experiences by gathering and analyzing Twitter. The sentiment classification model was developed using K-Nearest Neighbor (KNN) technique. The result visualized through the dashboard with different rates of accuracy results for the KNN classifier. We manage to get 94.49% of model accuracy using a confusion matrix. As a result, this preliminary study can be implied to help the universities to improve the ODL success rate and future decision-making for the ODL process.

Keywords: Online Distance Learning, Twitter, Sentiment Analysis, Perception, Data Classification

Introduction

Higher education institutions are no longer limited to traditional brick-and-mortar studying and teaching techniques because of technological advancements. ODL refers to the provision of flexible educational possibilities in terms of access and diverse forms of knowledge acquisition. A journal article stated that the new norm caused by the pandemic has changed the teaching and learning landscape. Almost all universities including Universiti Teknologi MARA (UiTM) have chosen to adapt teaching and learning in ODL for most of the programs offered (Kechil et al., 2020).

It is hard for one to find all the information about student's sentiment towards ODL in one place because social media is a vast space of various people sending messages and

interacting with each other every day for example, Twitter has a daily flow of over 500 million messages and Tweets (Fiesler, 2018). In 2019, Twitter's global viewership was estimated to be about 290.5 million monthly active users, with the number expected to rise to over 340 million by 2024. Twitter is still a successful marketing platform and one of the most prominent social networks in the world (Dixon, 2022).

Based on the research paper entitled Progress in Neural NLP: Modelling, Learning, and Reasoning, Natural Language Processing (NLP) is a branch of artificial intelligence concerned with allowing computers to comprehend and process human languages (Zhou et al., 2020). In order to discover which grouping of words and phrases belong to each other, NLP employs a hierarchy. A token, which might be a phrase or a single word, is the smallest level of text. A document is a collection of tokens, such as a paragraph or chapter. A corpus is a collection of documents, such as a book or an essay. Finally, a corpora is a collection of corpus that might include numerous books or articles that data scientists want to compare and analyze (Zhou et al., 2020). Humans have watched the rapid progress of NLP in tasks such as machine translation, question answering, and machine reading comprehension based on deep learning and a massive amount of annotated and unannotated data over the previous five years. All of these technologies from various sectors of methodology are connected in various ways, but they all play a critical role in applying their algorithms to make human jobs easier and more automated.

Sentiment analysis is an NLP task, in which the model must assess the sentiments of texts based on a machine learning training dataset. Sentiment is useful for quickly gaining insights using large volumes of text data. In today's environment where we are suffering from data overload, it is impossible to analyse it manually without any sort of error or bias. Sentiment analysis provides answers to the most important questions that arise because sentiment analysis can be automated and decisions can be made based on a significant amount of data rather than on mere intuition, which is not always correct (Alsaeedi et al., 2019). Sentiment analysis towards ODL is needed to visualize the data from social feedback, so it is easy for everyone to understand it such as university staffs, researchers or students. The KNN algorithm was used to discover the training dataset that are closest to the target object (Guo et. al, 2019). They also claimed that KNN method is a well-established theoretical tool that is very simple to apply. Furthermore, KNN has a greater predicting accuracy and makes no assumptions about the data it collects, and it is less susceptible to outliers.

Our aim was to conduct a preliminary study using sentiment analysis on twitter, to provide information on sentiment about ODL from users across Malaysia, in form of their learning experiences. Their sentiment could be either positive, neutral or negative and it could be affected by various variables. The initial data was scrapped from social media engagement, which is Twitter. From the data gathered, a sentiment analysis would be constructed through a dashboard.

Objectives

The objectives of this study as follow

- To gather and analyse the raw data from Twitter regarding individuals' and students' sentiment about ODL.

- To design a classification model based on sentiments classification using KNN technique.
- To construct a sentiment analysis that can be resourceful for universities or other researchers.

Methods

In this study, the method of Cross Industry Standard Process for Data Mining (CRISP-DM) is being adapted. It is akin to a collection of guidelines for planning, organizing, and executing a data science (or machine learning) project. It consists of business understanding, data understanding, data preparation, data modelling, evaluation, and development phase (J. Saltz et al., 2017).

Understanding workflow is a crucial process to achieve the objectives of this study. Figure 1 shows the flow of data and the sequential order of the activities. The collection of data must be done initially. Data cleaning, data transformation, stop words, tokenization, and stemming are the five sub-steps that make up data pre-processing. Pre-processing data is a crucial step in obtaining high accuracy (Prakash, 2019).

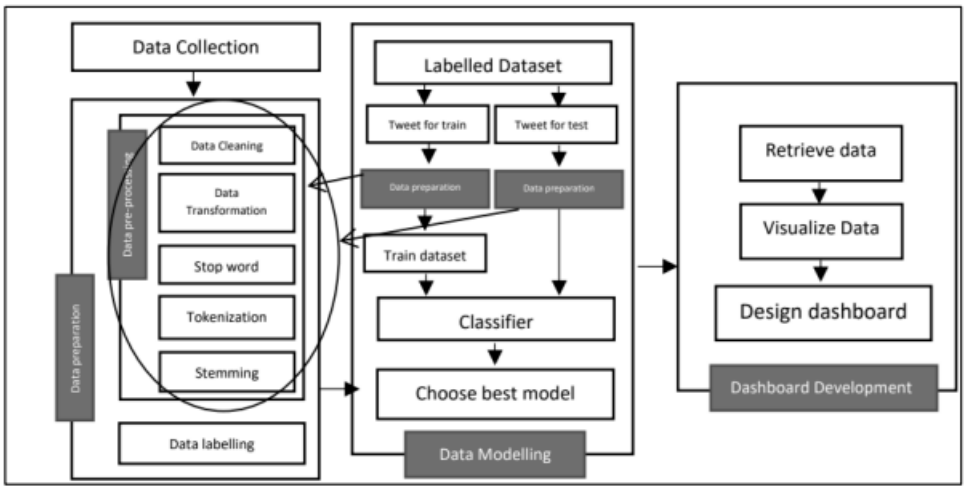

Figure 1: System Workflow

RapidMiner was used to label the data following data pre-processing. Vader is the corpora used in this study to extract sentiment. The labelled data used in data modelling was divided into two groups: training and testing data. Then, both data would be again prepared. After that, a trained dataset was used to train the model. KNN classifier was used for this purpose. Lastly, the dashboard was developed to visualize the sentiment's result.

A. Data Collection

The tweet's language considered for this study is Malay because most Malaysians tweet their opinions in Malay. The conclusions of this study will presumably be biased, erroneous, or small if it employs an English tweet as its source of data. Twint was used to scrape the twitter data. Twint is a sophisticated Python script for Twitter scraping that enables researcher to extract Tweets from Twitter accounts without utilising Twitter's API. Twint makes advantage of Twitter's search operators to enable us to scrape Tweets from certain individuals, scrape tweets associated with particular themes, hashtags, and trends, or filter out private

information from Tweets like e-mail and phone numbers. With the use of Twint's unique Twitter queries, we were able to scrape a user's followers, liked Tweets, and followers without using any authentication, API, Selenium, or browser emulation.

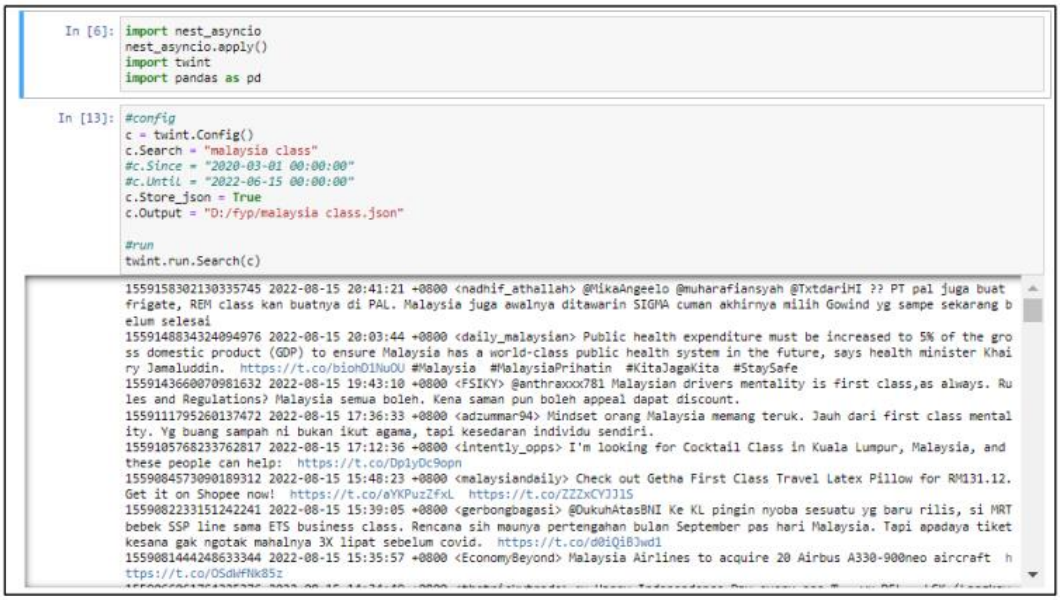

Figure 2: Sample of Twint command

As shown in Figure 2, the data was gathered between 1 March 2020 until 15 June 2022. The scraped data includes a total of 10,000 tweets using the following keywords; *online*, *kelas online, online class, google meet, zoom, UFuture, UiTM, Universiti Teknologi MARA, google classroom, Microsoft Teams, presentation online,* and *online presentation.*

Table 1
Sample of raw data

42	1.56E+18	1.56E+18	2022-08-1	########	0:45:33	800	1.05E+18	pkpmiiunu PMII Nurul Jadid	Pengurus Komisariat PMII dan Kopri Universitas Nurul Jadid ¿in	[]	[]	['https://p	0	0	1 []
43	1.56E+18	1.56E+18	2022-08-1	########	0:39:22	800	9.04E+17	oskajohnn bila anda pusing minu	Kayanya orang rumah culture shock mendengar pembicaraa¤in	[]	[]	[]	0	0	0 []
44	1.56E+18	1.56E+18	2022-08-1	########	0:35:48	800	1.55E+18	jacobikrist Kristofer.Jacobi	Webcam with Ring Light & Tripod 1080P HD Computer ¡en	[]	['https://w	[]	0	0	0 []
45	1.56E+18	1.56E+18	2022-08-1	########	0:33:16	800	9.95E+17	yubie94 silah ad-din	aku tahu lah aku ni selalu guna google meet , tapi jangan bor¡n	[]	[]	['https://p	0	0	2 []
46	1.56E+18	1.56E+18	2022-08-1	########	0:28:48	800	1.26E+18	sunghoonf â‹†Ëš¿Ëš Ÿ hani â‹†	tanak org masuk space takyah la buat space ðŸ˜ - gi la borak ¡n	[]	[]	[]	0	1	2 []
47	1.56E+18	1.56E+18	2022-08-1	########	0:25:05	800	1.37E+18	zerokinetix ezro	Everyone was talking about the email notif ng google classr¢tl	[]	[]	[]	0	0	0 []
48	1.56E+18	1.56E+18	2022-08-1	########	0:21:22	800	1.27E+08	grateroles esmelin graterol	Ãltimos Cupos. Conferencia: Los Operadores EpistemoiÃ³gi es	[]	[]	['https://p	0	2	2 ['confe

Table 1 shows only the specific parts of the data collected. The full table consists of 36 columns which include tweet id, twitter username, tweet text, geo location, language, tweet date and time, URL links and many more. The only important columns are twitter username, tweet text and date and time for this project so other columns will be deleted in the data cleaning process. Python was then used to clean the data. The odd alphabet, linkages, symbols, and other symbols were deleted. The deletion also included duplicate and empty tweets. Figure 3 shows the coding sample to remove links from the tweet dataset column.

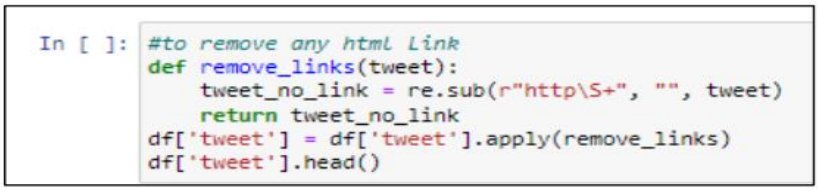

Figure 3: Removing URLs from the tweet column

B. Data Pre-processing

Pre-processing data is a data mining approach that cleans the data, combines information from several sources in a database or more, and turns raw data into standardised, tablet-friendly data. This involves data cleaning, data transformation, tokenization, stop word and data labelling. In this study, the unstructured data were used in pre-processing activities.

The data cleaning process was done initially. Hence, all data was investigated, including its size, structure, and relevance to our study requirements. For instance, dollar symbol represents money; therefore symbols, punctuation, and numbers are vital in fields like business or economics, but it did not hold much information relevant to our study. Therefore, all the symbols were eliminated. Since part of the data in this study were in Malay, the data were translated in order to extract the emotions. Therefore, the tweet's spelling must be fixed before it can be translated. The majority of the tweets were brief texts. They will give an impact on the sentiment extraction procedure that follows. The outcomes will be less precise. Therefore, MALAYA was used. Malaya is a Natural-Language Toolkit library for Malay language, driven by Deep Learning Tensorflow. Using Load Probability Speller, this coding was used to fix the tweet's spelling. The probability speller enhances and expands the capabilities of Peter Norvig's spell checker by utilising part of the normalisation of noisy texts in Malaysian online reviews systems. The word "sy," for instance, was changed to "saya." In order to avoid having a confusing translation, the sentence would be better organised.

The tweet's spelling was corrected before being translated into English. Google Translate handled the translation. With its free service, Google mechanically translates phrases, sentences, and webpages between English and more than 100 additional languages. However, some of the words were misspelt or wrongly translated, hence the translation was not perfect. Tokenization was used on the data once it had been translated and cleaned. Tokenization is a technique for breaking up a text block into smaller components known as tokens. Words, letters, or sub words can all be used as tokens. Word, character, and sub-word (n-gram character) tokenization are the three broad categories into which tokenization may be broadly subdivided.

The data was then subjected to lemmatization. Lemmatization, which attaches to suffixes, prefixes, or the roots of words known as the lemma, is a technique to reduce a word to its root verb as shown in Figure 4. Natural language processing (NLP) and natural language understanding (NLU) both benefit from stemming NLP. Processing, for instance, is not a root verb. Thus, processing of the word will follow lemmatization.

```
# Lemmetization
lemmatizer = WordNetLemmatizer()
wordnet_map = {"N":wordnet.NOUN,"V":wordnet.VERB, "J":wordnet.ADJ, "R":wordnet.ADV}

def lemmatize_words (tweet):
    pos_tagged_tweet = nltk.pos_tag(tweet.split())
    return " ".join([lemmatizer.lemmatize(word, wordnet_map.get(pos[0], wordnet.NOUN)) for word, pos in pos_tagged_tweet])
```

Figure 4: Word lemmatization

Before modelling, we should remove a stop word as the following step in Figure 5. Stop words are words that do not really add anything to a sentence's meaning. They can be safely ignored without affecting the sentence's meaning. Words like the, a, he, she, has, have are

good examples. As the irrelevant terms that the model must assess were eliminated, the model could perform better.

```
# Prepare Stop Words
stop_words = stopwords.words('english')
stop_words.extend(['from', 'https', 'twitter', 'still'])
def remove_stopwords(tweet):
    return [[word for word in simple_preprocess(str(tweet)) if word not in stop_words] for tweet in tweet]
df['tweet'] = remove_stopwords(df['tweet'])
df['tweet'].head()
```

Figure 5: Removing stop words

The next stage was labelling the data once it has been cleaned. Using RapidMiner, the data was labelled. The operator Extract Sentiment in RapidMiner was used to extract sentiments of a tweet as in Figure 6. Using a textual information attribute for either the open-source emotion dictionary or exclusive API methods, Extract Sentiment created a sentiment score. Vader was the corpora used in this study. The polarity of the text is ascertained using each corpus individually. Following the extraction of the emotion, the polarity score was then normalised to lie between - 1.0 and 1.0.

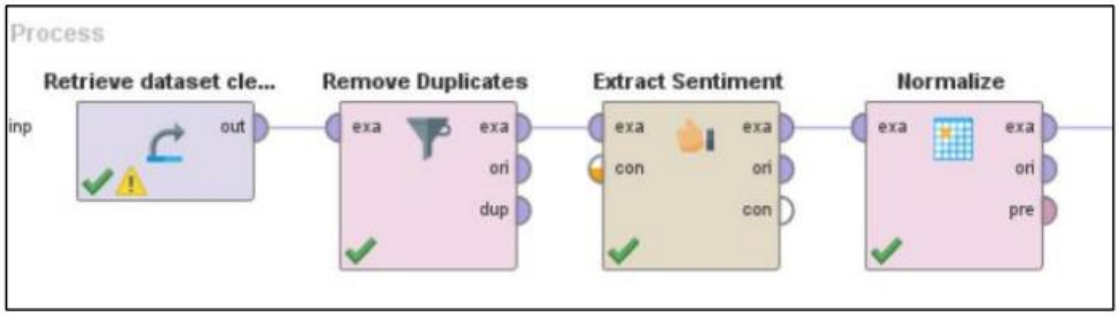

Figure 6: Extract sentiment process in RapidMiner

C. Data Modelling

A machine learning algorithm KNN was used as a classifier. We used the model on the test data. RapidMiner was used to apply the model. The dataset needs to be evenly distributed before modelling in order to remove bias. In order to obtain great accuracy, the training data must thus be evenly distributed. The under-sampling technique was used in this study to get an evenly dispersed sample of data for the model. The class with the least quantity of data, which in this study is negative value was considered while doing this. The most positive and most negative tweets were used as the primary data in this study. The number of positive tweets were chosen at random when using the Vader Random extraction approach.

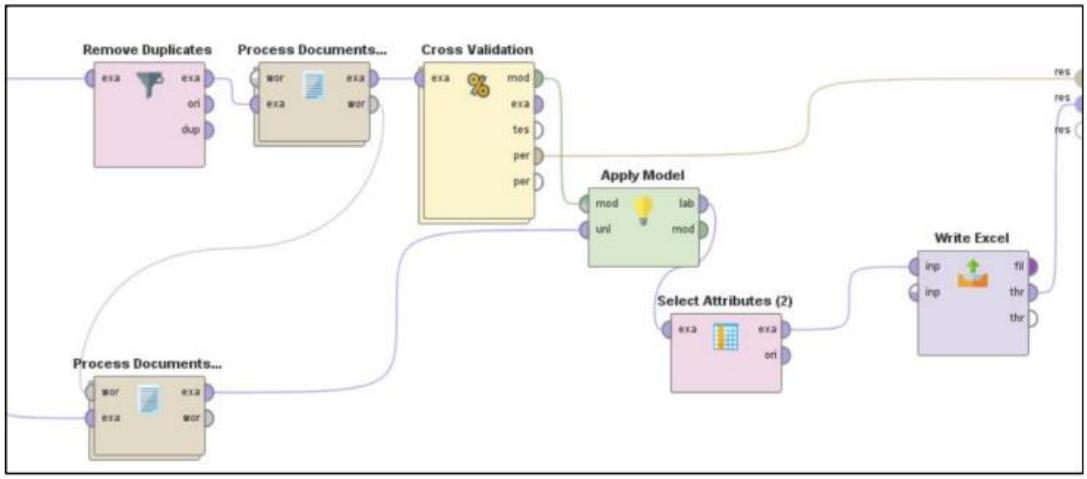

Figure 7: Process in data modelling phase.

The dataset retrieval appeared first, as seen in Figure 7. The dataset was then divided into the train data and the test data. The dataset for this study is divided into two portions: 70:30 and 80:20. The rationale behind the selection of these two ratios is that practical trials have demonstrated that employing 20–30% of the study data and the rest 70–80% yields the best outcomes. Stratified sampling was the sampling technique used for this purpose. By grouping the data according to the class labels, stratified sampling distributed the sample fairly.

The Set Role operator was used to determine the class label for the training data. The dataset's label is sentiment. The train data was then transformed from polynomial data to text using the nominal to text operator once the class label had been established. In deleting duplicates, the Remove duplicate operator was used.

After that, the process document operator carried out five tasks as illustrated in Figure 8. All tweets were converted into tiny capital letters as part of the initial procedure, called Transform Cases. The tweets would be tokenized using the Tokenize operator. The operator named Filter Stopwords then deleted the stop word from the tweets. In order to determine the terms' root verbs, stemming was also performed. Finally, an N-gram was produced using the Generate N-gram operator. Bi-gram and unigram operators were also used in this study's modelling.

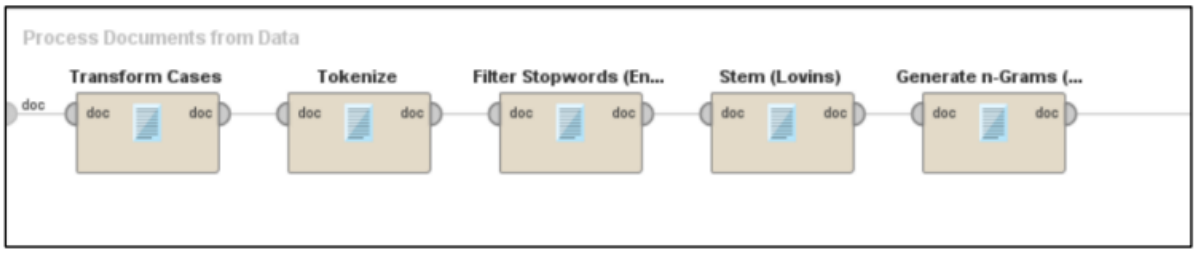

Figure 8: Processes in process documents from data

Figure 9 shows a process where KNN technique was then applied using the apply model operator. With the aim to assess the statistical effectiveness of the classification task, the performance operator was implemented. The number of folds for the cross-validation was

set to 10. In Order to test the model's performance, the train data were partitioned into 10 equal pieces, nine of which were utilised as train data and one as test data. A separate testing component was used in each of the ten iterations. K-folds = 10 was selected since it is frequently utilised in studies.

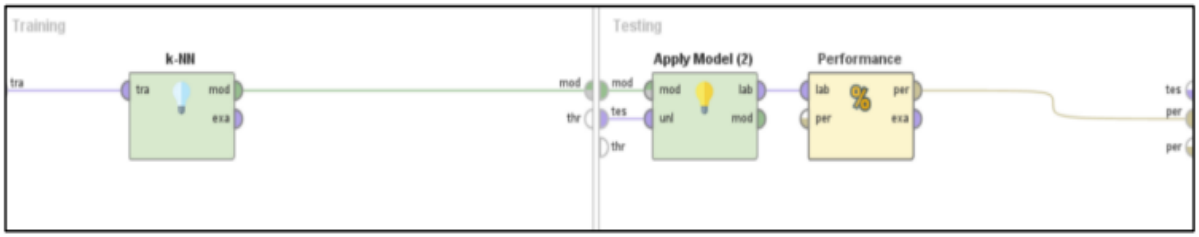

Figure 9: Process in cross-validation operator

By using the Nominal to Text Operator, the data were converted into a text datatype for testing purposes. The Remove Train Duplicates operator was used to eliminate the duplicated data. The information was subsequently handled exactly as the test information in the process document operator. The applied model function used the machine learning algorithm model from the Cross-Validation operator and the test data, to assess the sentiment of the tweet after processing the train data that was used to evaluate model performance and the test data.

Result and Discussion

A. Data Classification Result

As mentioned above, this study used KNN as the classifier. After implementing all the models with different parameters and operators, the suitable choice for the model was 70:30 percentage spilt using unigram operator, with Vader extraction method because it has the highest performance metric of them all. Table 2 depicts the data classification result from the model mentioned earlier.

Table 2
Results from KNN by using unigram with Vader extraction method

Row No.	prediction(s...	confidence(...	confidence(...	confidence(...	Score
1	negative	0.193	0	0.807	-0.385
2	neutral	0	0.589	0.411	-0.308
3	positive	0.398	0.394	0.208	-0.385
4	neutral	0	1	0	0
5	neutral	0	1	0	0
6	negative	0.201	0	0.799	-0.333
7	positive	1	0	0	0.897
8	neutral	0	1	0	0
9	neutral	0	1	0	0
10	neutral	0	1	0	0
11	neutral	0	1.000	0	0
12	negative	0.395	0	0.605	-0.359
13	neutral	0	1	0	0
14	positive	0.806	0	0.194	0.744
15	positive	0.800	0	0.200	0.333
16	positive	1	0	0	0.333

Accuracy result of 94.49% as in Table 3 can be considered as overfitted. Overfitting happens when a function fits a lot of a set of data too closely. This prevented the model from populating any other data sources, therefore it is useless for examining alternative possibilities. It describes the situation where training data is overtrained due to effective modelling. A model that is too fitted to the data is less able to make use of fresh data by adjusting to it since it is unable to discern between detail and noise from unknown data.

Table 3
Confusion matrix for 70:30 split KNN classifier extracted with Vader corpus

accuracy: 94.49% +/- 1.79% (micro average: 94.49%)

	true positive	true neutral	true negative	class precision
pred. positive	297	6	33	88.39%
pred. neutral	19	1198	20	96.85%
pred. negative	11	6	134	88.74%
class recall	90.83%	99.01%	71.66%	

There could be a few causes of overfitting such as, there was no cleaning or proper disposal of the garbage values in the training data, and there was a very limited training dataset. This issue could be overcome by going back to clean process and cleaning the data more thoroughly. Additional functions need to be implemented in data pre-processing and more suitable data should be provided for training and testing purposes.

B. Dashboard Development

The dataset acquired via data modelling was used to generate the visualisation for dashboard development. In order to create the dashboard, Power BI was used. Microsoft offers a solution for business analytics called Power BI. The collected dataset from data modelling was used to produce the visualisation for the dashboard development in this study. The dashboard was generated using Power BI Based as shown in Figure 10 below. The complete dashboard was created throughout dashboard creation to guarantee a seamless and effective inclusion.

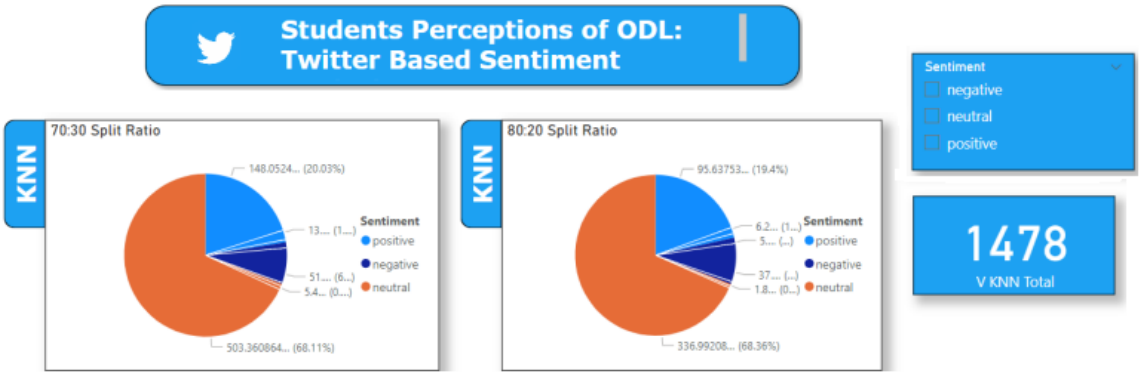

Figure 10: The dashboard

Based on the result generated, more than 60% of twitter users gave neutral comments, like ODL was okay. However, it could be a sign of concern since the user can quickly turn to the positive or negative value. There is some opportunity here for the universities and schools

to engage with their students, in sharing their perception, opinion and suggestions on ODL. They might need a little more attention, in order to form more positive results.

The number on the right side of the dashboard shows the total of sentiments with different rates of confidence in results from KNN classifier, that used the Vader extraction technique. The pie charts can be filtered with the use of a radio button at the right side of the dashboard as well. For example, if the users requested the total negative sentiment predicted by 70:30 KNN classifier, they just need to click the negative radio button on the sentiment filter. Figure 11 shows the filter result for negative sentiment, which appeared to be 107 of KNN total value.

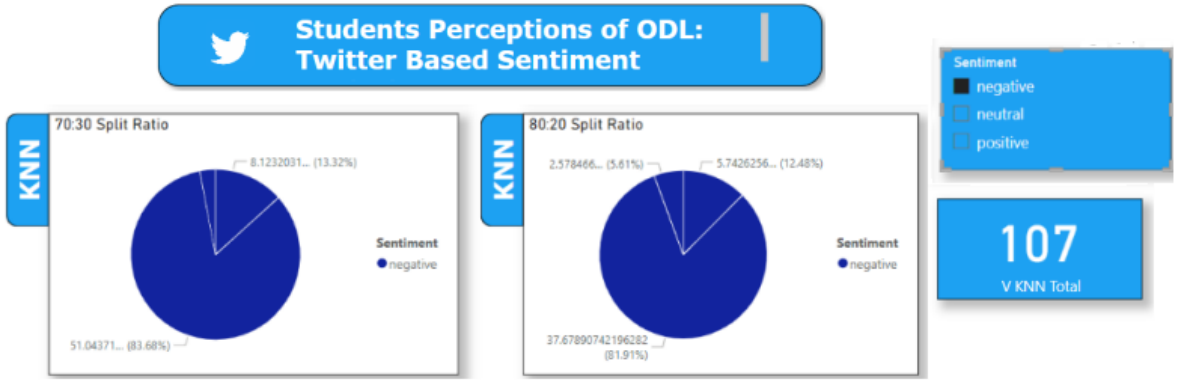

Figure 11: Example of negative filter

Conclusion

The study aims to investigate the user's sentiment regarding ODL among Malaysian twitter user. This study has examined more than 10 distinct models with KNN classifier, vader extraction method, two different percentage splits and numbers of n-grams, in order to find the optimal model. The dashboard was developed to visualise sentiment analysis for researchers who are interested to use the information. The finding shows that more than 60% of twitter users gave neutral comments indicating that ODL experience neither affected their sentiments positively or negatively. Besides, the user can easily grasp the data by utilising the dashboard, and they can get access to a comprehensive view of public sentiments towards ODL. The study suggests a real-time sentiment analysis on ODL to produce a superior result for the future work. In addition, the application of Malay language corpus for sentiment analysis would be preferable in order to reduce errors in the translation of tweets into English. Overall, the study also proposes to employ effective, advanced method or tool, and data to generate good results for future research.

References

Kechil, R., Mydin, A. M., Anisha, W., Mohammad, W., & Learning, D. (2020). Pendidikan Jarak Jauh Terbuka (Odl): Adaptasi Norma Baharu Dalam Pembelajaran. SIG: E-Learning@CS, 1(September), 31–38.

Fiesler, C., & Proferes, N. (2018). "Participant" Perceptions of Twitter Research Ethics. SAGE Journals. 1-14. https://doi.org/10.1177/2056305118763366

Dixon, S. (2022). Twitter: Number of Worldwide Users 2019-2024. Statista. https://www.statista.com/statistics/303681/twitter-users-worldwide/#:~:text=In%202019%2C%20Twitter%27s%20audience%20counted,and%20a%20popular%20marketing%20channel.

Zhou, M., Duan, N., Liu, S., & Shum, H. Y. (2020). Progress in Neural NLP: Modelling, Learning, and Reasoning. In Engineering. Elsevier Ltd. Vol. 6, Issue 3, 275–290. https://doi.org/10.1016/j.eng.2019.12.014

Alsaeedi, A., & Khan, M. Z. (2019). A study on sentiment analysis techniques of Twitter data. International Journal of Advanced Computer Science and Applications, 10(2). https://doi.org/10.14569/ijacsa.2019.0100248

Guo, G. F., Zheng, Y. H., & Hong, W. C. (2019). Application of the weighted k-nearest neighbor algorithm for short-term load forecasting. Energies, 12(5). https://doi.org/10.3390/en12050916

Saltz, J., Shamshurin, I., & Connors, C. (2017). Predicting data science sociotechnical execution challenges by categorizing data science projects. Journal of the Association for Information Science and Technology, 68(12), 2720–2728. https://doi.org/10.1002/ASI.23873

Prakash, T. N., & Aloysius, A. (2019). Data Preprocessing in Sentiment Analysis using Twitter Data. International Education Applied Research Journal (IEARJ), vol. 03, issue 07, 89-92. E-ISSN: 2456-6713