Manuscript Submitted	28.10.2022
Accepted	17.1.2023
Published	30.6.2023

A Facebook Sentiment Analysis based on Malay Text for UTeM Services

Abdul Syukor Mohamad Jaya & Aida Hazirah Abdul Hamid
Fakulti Teknologi Maklumat dan Komunikasi
Universiti Teknikal Malaysia Melaka
syukor@utem.edu.my, aida.hazirah@utem.edu.my

Abstract

In recent years, many large companies around the world have used information from social media to identify customer needs, in addition to obtain immediate feedback from marketed products or services. Feedbacks in the form of text, voice and pictures are classified into sentiment to identify the percentage of customer satisfaction before any decision is made. However, responses written in short text and native language make the classification process become difficult. The aim of this study is to classified social media text data from Universiti Teknikal Malaysia Melaka (UTeM) Facebook which is mostly written in Malay, into positive, neutral, or negative sentiments. Facepager software that facilitates Facebook Page data extraction through Facebook's Application Programming Interface (API) is used to extract page posts, user comments to posts, replies to comments, and engagement. The model is developed using an exactly established technique that includes pre-processing ticket data, stemming words, feature vectorization, creating training tickets, and tuning machine learning algorithms. We use variety of techniques such as Naïve Bayes, Linear SVC, and Logistic Regression to demonstrated the highest accuracy among the chosen model by using of two datasets retrieved from UTeM Facebook data, (i) dataset 1 (cleaned dataset) consists of 7,005 Facebook data, and (ii) dataset 2 (resampled dataset) consist of 6,519 Facebook data. The experimental findings revealed that the model used Linear Support Vector Classifier achieved the highest performance with 93% accuracy. To visualize the sentiment, we also built an interactive dashboard to monitor the positive, neutral and negative sentiment in each post.

Keywords: *Facebook, machine learning, classification, Linear Support Vector Machine.*

1. Introduction

Nowadays, social media has become a platform to share almost all types of information. Social media plays a vital role in human communication and improves business applications (Flack & D'Sauza, 2014; Chen & Huang, 2021; Mashiah, 2021). The data generating among these social media applications are mammoth, from tweets to swipes, likes to shares, the online world is exploding as Terabytes to Petabytes (Wanniarachchi et al., 2020). This trend leads into big data whereby data is growing fast along the day because data is created by everyone and for everything from mobile devices, call centers, web servers, and social networking sites, etc. The world of big data is increasingly being defined by the 4 Vs. i.e. these 'Vs' become a reasonable test as to whether a big data approach is the right one to adopt for a new area of analysis. The data with a huge volume, variety and veracity structure called the big data. The social media data also belongs to the big data because of its variety, volume and veracity in nature (Vatrapu et al., 2016; Balaji et al., 2021).

Text is the biggest way of communication and more than 18.2 million text messages are transmitting in a minute. The hundreds of billions of tweets give it "volume", its hundreds of millions of tweets a day give it "velocity" and its mix of text, imagery and video offer "variety". Veracity refers to the complexity or dependability of the data. In addition, because of the massive volume of big data, traditional methods to extract and analyze the huge data are not very useful as these will not provide an accurate result for decision making etc.

The challenge of having big data is that it is extremely large, extremely fast and difficult to handle for traditional database and current technologies. Many organizations gather the huge numbers of data generated from high-volume transactions like call centers, sensors, web logs, and digital images. Therefore, using the right tools to manage the data is very important. There are three categorized of big data tools which is computing tools, storage tools and support technologies. Computing tools for big data are fundamental tools that can be used to process big data at different levels. Applying machine learning techniques to big data give much-sophisticated information (Witten et al., 2016; Tran et al., 2023).

Next, storage tools in big data have a double purpose and they offer an infrastructure on which is possible to run analytics tools and, simultaneously, a place to store and query big data. The most relevant variables in choosing a big data storage tool include the existing environment, current storage platform, growth expectations, size and type of files, database and application mix (Behrens, 2017). Languages and support technologies is needed to process the big data. JSON is a universal format that is very convenient for exchanging information between applications through various protocols and RESTful is an API that allows the communication between a web-based client and server that employs representational state transfer (REST). In addition, SQL and NoSQL offer mechanisms for storage and retrieval data. NoSQL databases have emerged as an answer to the limitations of the traditional relational databases, as a provision of performance, scalability, and flexibility, required in modern applications, besides can store different types of data; such as structured, semi-structured, unstructured, and polymorphic data (Rafique et al., 2020; Chen at al., 2022). They are increasingly used in big data and real-time web applications, due to their simplicity of design and scalability. Therefore, applying the big data analytics technique on the data is a desired way to improve the analysis.

Universiti Teknikal Malaysia Melaka uses social media to post any news and information related to university such as event and important announcement. UTeM’s management need to analyze the sentiment from netizen especially UTeM students for future improvement. It is a hard work to analyze the sentiment manually, and most of the texting words are in Malay language. This study aims to identify and synthesize the user’s sentiment in UTeM Facebook. For that, the analytic dashboard based on user’s sentiment is created. Finally, the overall opinion on a particular trending topic posting by UTeM is determined.

2. Method

In the first stage, understanding the activity is key to ensure its success and the first phase for every analytic project. We need to understand the business domain, determine how deep is the scope of the project and make sure it can be solved using the business understanding data preparation, or exploration modelling evaluation deployment available dataset. UTeM is the business user in giving their understanding and opinion of what is necessary based on business operational of the organization. It is aim is to come out with business problem. Then we identify assess the resources; identify the people who joins, technology to use, tools, systems and data. The purpose is to make sure there is sufficient support to successfully complete the project. We study the structure of the raw data to gain a quick overview and to get the high-level understanding. This sentiment project mainly studies the opinions which express or imply positive or negative or neutral sentiments. Thus, with the analytic dashboard of user sentiment analysis helps admin to identify the positive and negative comment received via their official social media account.

Data collection process also known as Exploratory Data Analysis (EDA). EDA is a well-established statistical tradition that provides conceptual and computational tools to discover patterns in a data science context (Mehmood et al., 2017). EDA is typically characterized by an emphasis on: (i) a substantive understanding of data; (ii) graphic representations of data; (iii) tentative model building in an iterative/interactive process; (iv) flexibility regarding which is the best method to apply. The final aim is to discover patterns within data. The data consists of 7,005 records which are generated from UTeM Facebook post. The quantity of data we used in our experiment covers the time span of 1st January 2020 until 31st December 2021. We can observe that there is sufficient text data in order to run the algorithms for sentiment analysis. The dataset contains 7 columns as shown in Table 1.

Table 1: Dataset definition

Column	Description	Record
`level`	The category level of the Facebook post.	7005
`object_id`	The id of the Facebook post.	
`created_time`	The time created for Facebook post.	
`message`	The text input for user post in Facebook.	
`reaction`	The total number of people react to Facebook post.	
`YEAR`	The year of Facebook post.	
`MONTH`	The month of Facebook post.	

Initially, our datasets contained valueless attributes, missing instances, inadequate attributes' data types and other problems that raise the necessity of preparing it first before feeding it to the analysis phase. Therefore, the datasets are passed through the following preparation stages.

2.1 Setting Social Media API

The social media API (application programming interfaces) has been gained by creating the social media application. An authorized API client is developed to monitor the live streaming data regarding UTeM issues from social media application such as Facebook by requesting the Facebook API to fetch tweets for a particular query. The first step before data can be retrieved from a social media platform such as Facebook is to register an application on the platform to access the APIs key of the application. By default, application only access public information on Facebook. Certain endpoints, such as those responsible for sending or receiving Direct Messages require extra consents from the user before other users can access to the information.

2.2 Extract Social Media Data

Facepager, a third-party software that facilitates Facebook Page data extraction through Facebook's Application Programming Interface (API) is used to extract HKPF Page posts, user comments to posts, replies to comments, and engagement (i.e. likes and shares). Data collection is staggered with a gap of 14 days between date of post and date of extraction. For example, data from the November 1 post is extracted on November 15. Data from 5 October 2020 to 31 March 2022 is collected.

2.3 Data Preparation

Cleaning and converting raw data before processing and analysis is known as data preparation. Prior to processing, it is a crucial stage that entails reformatting data, altering data, and fusing data sets to enrich data. Standardizing data formats, enhancing source data, or reducing outliers, for instance, are all common data preparation techniques. Before training and testing a machine learning model, raw data must be processed. A data mining approach called data pre-processing comprises transforming raw data into a format that may be used. The machine learning process involves the following data pre-processing steps:

Step 1: Obtain the dataset
Step 2: Import the libraries
Step 3: Import the dataset
Step 4: Check out the missing values
Step 5: See the categorical values
Step 6: Splitting the data-set into training and test set

2.4 Pre-Processing

In this phase, as illustrated in Figure 1, we pre-processed the gathered data in order to get it ready for text mining techniques. A technique for analyzing text is called text mining, also referred to as text analytics. Natural Language Toolkit (NLTK), a robust Python package featuring a selection of natural language processing methods, was implemented by research team. It is thoroughly documented, open source, free, simple to use, and has a sizable user base. Algorithms for tokenizing, part-of-speech tagging, stemming, sentiment analysis, subject segmentation, and named entity identification are all part of the Natural Language Toolkit (NLTK).

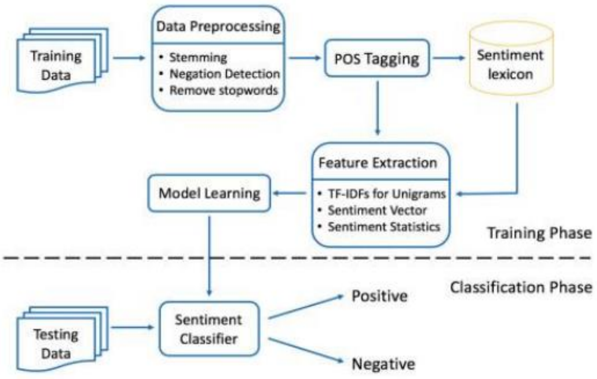
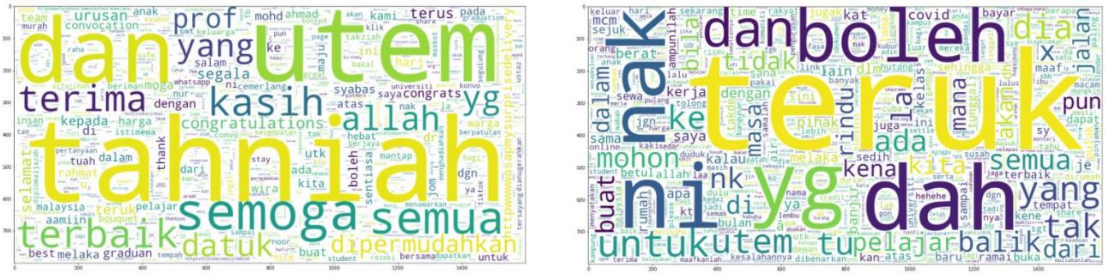


Figure 1: Pre-processing phase.

2.5 Data Exploration

Diverse data set formats may be used as input (.XLS, .TXT, .CSV, JSON). It is simple to load data from any source because of Python's basic syntax and the availability of prepackaged tools like Pandas. Figure 2 illustrates the findings about the positive and negative comments left on Facebook posts. This neutral sentiment mainly involves frequently used terms, according to an analysis of neutral words that is the most common words to post. Words with conjunctions are examples. Thus according Figure 3, the target sentiment appears to be unbalanced because the majority of the sentences are from neutral sentiment.



(a)

(b)

Figure 2: (a) Positive, and (b) negative words form Facebook

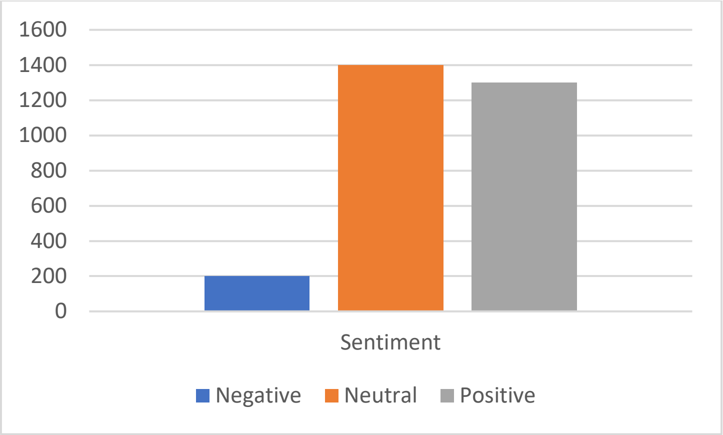


Figure 3: Distribution of sentiment

We have utilized stratified random sampling to further resolve the imbalance in the target class. We used stratified random sampling because it enables researchers to collect a sample population that most accurately reflects the whole population under study while guaranteeing that each relevant subgroup is represented. In a stratified random sample, the entire population is split into identical groups known as strata. The next step is to select random samples from each stratum. The highest sentiment value, neutral, is compared to "other" first, followed by all subsequent sentiment values, positive and negative combined. To evaluate the value just between positive and negative sentiment, we eliminate neutral sentiment. We resampled the class to a size of 1269, which represents the maximum values of positive sentiment, for this sampling. Positive and negative sentiments would be oversampled, while neutral sentiments would be under sampled.

Before generating a machine learning or deep learning model, numerous preprocessing operations are necessary when working with text data. It is a supervised approach. Depending on the requirements, we must preprocess the data using different techniques. Data analysts use data visualization and statistical tools to determine dataset characterizations including size, number, and correctness in order to better comprehend the data during the first phase of data analysis, known as data exploration. To visually explore and identify relationships between various data variables, the structure of the dataset, the presence of outliers, and the distribution of data values in order to reveal patterns and points of interest, both manual analysis and automated data exploration software solutions are used. This allows data analysts better understand the raw data.

3. Classification Algorithm

This project aims to identify the machine learning technique that offers the most accurate model for classifying tickets. In this project, we used classification techniques and, to get things going, we used the fundamental machine learning algorithms Bernoulli Naive Bayes (BernoulliNB), Linear Support Vector Classification (Linear SVC), and Logistic Regression (LR). We utilised the Bernoulli Naive Bayes classifier, which is a decent place to start for discrete data with only binary features. The Naive Bayes method uses a set of predetermined parameters or assumptions to facilitate the machine's learning process. The amount of training data has no bearing on the number of parameters used in parametric algorithms. The properties of a dataset are presumably completely unconnected to one another. The Bayes theorem is used in this conditional probability classification model to predict the class of unknown datasets.

We also utilized logistic regression because it is an easy-to-use classification method that works well with a wide range of classes. A categorical dependent variable can be predicted using this technique using a number of independent variables. To forecast the likelihood of a target variable, logistic regression is a supervised learning classification technique. There are only two categories because the aim or dependent variable is dualistic. Observations can be categorized using a variety of types of data using logistic regression, which can also quickly pinpoint the most beneficial parameters for categorization. In this style of classification, the dependent variable may have three or more possible unordered types or types without any numeric significance. Examples of what these variables might stand for include "negative," "neutral," and "positive."

In order to explain how a group of independent variables affects the outcome of a dependent variable, a type of analysis called logistic regression is used. The main limitations of the logistic regression approach are that it can only be applied when the predicted variable is binary, there are no missing values in the data, and the predictors are independent of one another. Additionally, the best-suited hyperplane for dividing or classifying our data is produced using Linear SVC, which is created to fit to the data we offer. Once we have the hyperplane, we can next input some attributes to our classifier to see what the expected class is. The Linear SVC method, which is effective with large datasets, classifies data using a linear kernel function. The Linear SVC model has more parameters than the SVC model, such as the loss function and penalty normalization ('L1' or 'L2'). The kernel method cannot be changed because linear SVC is based on the kernel linear approach. This makes this method particularly suitable for our needs, however it may also be applied in other circumstances.

4. Result and Discussion

The Facebook dataset, which was extracted from the Facepager software, is used to construct the training data. The performance of the models is compared using the following datasets because the sentiment categorization is unbalanced. We used two datasets: dataset 1, which is clean data with no sampling and sentiment classification, and dataset 2, which is resampled data with sentiment classification of positive, neutral, and negative.

Before implementing machine learning methods, we also turn the training data samples into word vectors using Keras tokenizer text to sequence. This method allows us to victories a text corpus by converting each text into either a series of numbers (each integer being the index of a word in a dictionary) or a vector with a binary coefficient for each token. The datasets are split 80:20 into training and test sets, as indicated in Table 2. In this instance, X and Y have been given as inputs to the train-test split algorithm, which appropriately divides X and Y into 20% testing data and 80% training data among X train, X test, Y train, and Y test. Dataset 1 is divided into a train dataset with 5,604 records and a test dataset with 1,401 records. Dataset 2 is divided into a train dataset with 5,215 entries and a test dataset with 1,304 records.

Table 2: Distribution of dataset

Dataset	Number of Samples	Number of Labels	Train Dataset	Test Dataset
1	7,005	7,005	5,606	1,401
2	6,519	6,519	5,215	1,304

For the first training Dataset 1 is comprised of 7,005 Facebook data that labeled with their sentiment. The Facebook data is divided into the following three classification of sentiment which are positive, neutral and negative. As a result, Table 3 depicts the accuracy of the corresponding classification models utilizing the earlier algorithms and test data sets consisting of 1,401 observations. Accordingly, the three models' prediction accuracy wasn't very impressive. The imbalance in the dataset makes training ineffective because the majority of sample instances are basic examples that don't provide any useful signal. If the simple cases predominate throughout training, degenerate models will develop. A straightforward performance metric is accuracy. The ratio of all correctly anticipated cases, whether positive or negative, to all cases in the data, is what determines the value. Precision is a positive prediction statement that has been shown to be accurate. Recall is the correct identification of the real positive rate proposition.

The results of the four models—Bernoulli Naive Bayes, Logistic Regression, and Linear SVC—are tabulated in Table 3. The table demonstrates that the Linear SVC model performs better across the board. With accuracy of 50.86%, precision of 49.65%, recall of 50.86%, and an F1 value of 49.93%, the Linear SVC model performs best. The Bernoulli Naive Bayes indicators, however, have the lowest value among the three models, with accuracy, precision, recall, and F1 values of 39.69%, 26.73%, and 28.38% respectively. The uneven data may be a factor in the indicators' poor value. As a result, accuracy is a subpar indicator of how well a categorization model is performing.

We resampled the dataset to 7,005 to allow for more accurate testing of the model. We retrained all of the algorithms' classifiers using the training Dataset 2's 6,519 members. In compared to training with Dataset 1, the median accuracy of the Linear Support Vector Classifier performed better. Table 4 summarizes accuracy statistics for the three linear regression and two Bernoulli Naive Bayes models. The findings show that the two models, Logistic Regression and Linear SVC, perform equally well in terms of prediction, while Bernoulli Naive Bayes has the lowest accuracy prediction score. It demonstrates that the value has increased from the Dataset 1 prediction accuracy. The best result is achieved by the Logistic Regression model and Linear SVC, which has a 93.00% accuracy, 93.00% precision, 93.00% recall, and 93.00% F1 value.

Table 3: Model comparison on dataset 1.

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Bernoulli Naïve Bayes	0.3969	0.2673	0.3969	0.2838
Logistic Regression	0.4588	0.4283	0.4588	0.3540
Linear SVC	0.5086	0.4965	0.5086	0.4993

Table 4: Model comparison on dataset 2.

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Bernoulli Naïve Bayes	0.7900	0.8200	0.9700	0.7900
Logistic Regression	0.9300	0.9300	0.9300	0.9300
Linear SVC	0.9300	0.9300	0.9300	0.9300

The classification report, which is displayed in Table 5, has an evenly balanced value, indicating that the majority of the sentiment is appropriately classified within that sentiment. The precision scores for each category are as follows, as shown in Figure 4.4: positive is 91%, neutral is 91%, and negative is 98%. Neutral recollection scores are 91%, neutral 89%, and negative 90%. Negative scores 98%, neutral scores 100%, and negative scores 99% on the F1 scale.

Table 5: Classification report for each sentiment

	Precision	Recall	F1-score	Support
0	0.91	0.91	0.91	423
1	0.91	0.89	0.90	442
2	0.98	1.00	0.99	439
Accuracy			0.93	1304
Macro avg	0.93	0.93	0.93	1304
Weight avg	0.93	0.93	0.93	1304

5. Conclusion

This study's objective is to categories social media text data from UTM's Facebook, which is primarily written in Malay, into positive, neutral, and negative sentiments. Manually matching each message to the appropriate sentiment took a lot of time, and mistakes could occur due to human error. A model based on supervised machine learning methods is suggested in this work to automatically classify the sentiment of analysis. To extract features vectors, we applied the Keras tokenizer technique. To prepare text files for deep learning, use Keras' tokenizer class. On either plain text or text documents that have been integer-encoded, the tokenizer is created and installed. In order to transform each text in a collection of texts into an integer sequence, we utilized the text to sequences function. Every word in the text was changed to its corresponding integer value from the word index dictionary. To assess the model performances, three different supervised classification algorithms—Linear SVC, Logistic Regression, and Bernaulli Naive Bayes—are used. As a result, when compared to other models, the Linear SVC had the highest accuracy.

Acknowledgement
We would like to thank Universiti Teknikal Malaysia Melaka (UTeM) for facilities and financial support.

References

Balaji, T.K., Annavarapu, C. S. R. & Bablani, A. (2021). Machine learning algorithms for social media analysis: A survey, *Computer Science Review*, 40: 100395.

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychology Methods*, 2(2), 131-160.

Chen, Y.-F., & Huang, S.-H. (2021). Sentiment-influenced trading system based on multimodal deep reinforcement learning, *Applied Soft Computing*, 112: 107788.

Chen, L., Davoudian, A. & Liu, M. (2022). A workload-driven method for designing aggregate-oriented NoSQL databases, *Data & Knowledge Engineering*, 142: 102089.

Flack, J. C. & D'Souza, R. M. (2014). The digital age and the future of social network science and engineering. *Proceeding of IEEE*, 102(12), 1873-1877.

Mashiah, I. (2021). "Come and join us": How tech brands use source, message, and target audience strategies to attract employees, *The Journal of High Technology Management Research*, 32(2): 100418.

Mehmood, N., Culmone, R. & Mostarda, L. (2017). Modeling temporal aspects of sensor data for MongoDB NoSql database. *Journal of Big Data*, 4:8.

Rafique, A., Van Landuyt, D., Beni, E. H., Lagaisse, B. & Joosen, W. (2020). CryptDICE: Distributed data protection system for secure cloud data storage and computation. *Information Systems*. 96: 101671.

Tran, M.-Q., Doan, H.-P., Q.Vu, V. & T.Vu, L. (2023). Machine learning and IoT-based approach for tool condition monitoring: A review and future prospects, Measurement, 207: 112351.

Vatrapu, R., Mukkamala, R. R., Hussain, A. & Flesch, B. (2016). Social set analysis: A set theoretical approach to big data analytics. *IEEE Access*, 4, 2542-2571.

Wanniarachchi, V. U., Mathrani, A., Susnjak, T. & Scogings, C. (2020). A systematic literature review: What is the current stance towards weight stigmatization in social media platforms? *International Journal Human-Computer Studies*, 135:102371

Witten, I. H., Frank, E., Hall, M.A. & Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques. *Morgan Kaufmann.*