JOURNAL OF STATISTICAL METHODS AND DATA SCIENCE

Journal of Statistical Methods and Data Science
Volume 01 Nomor 1 June 2023
https://jurnalmipa.unri.ac.id/jsmds

TWITTER SENTIMENT ANALYSIS OF ELECTRIC VEHICLE SUBSIDY POLICY USING NAÏVE BAYES ALGORITHM

Agung Satrio Wicaksono[1*]
Universitas Sultan Ageng Tirtayasa, Banten, 42163, Indonesia

***e-mail**: *agungsatriow@untirta.ac.id*

Article Info:
Received: 31-05-2023
Accepted: 15-06-2023
Available Online: 30-06-2023

Keywords:
electric vehicle
naïve bayes
sentiment analysis,
subsidy policy
twitter

Abstract: *This research aims to apply the Naïve Bayes classifier in Indonesian-language sentiment analysis, regarding electric vehicle subsidy policies using Twitter data with the query 'subsidi kendaraan listrik'. The stages of analysis include data pre-processing, tokenization, stemming, forming the Naïve Bayes model, and evaluating model performance using accuracy, precision, and recall. The SMOTE technique is used to deal with class imbalances, in which the majority of negative sentiments towards the policy is 66%. The results obtained from the 10-fold Cross Validation with the binary classification (positive and negative sentiment) show that the accuracy value of the model is 69.49%, with precision and recall values of 53.27% and 74.26%.*

1. INTRODUCTION

The electric vehicle subsidy policy has been implemented in Indonesia starting March 20, 2023. One of its goals is to encourage the use of renewable energy-based vehicles that are more environmentally friendly. Environmentally friendly and low emission vehicle innovation is one of the solutions considered to be able to reduce Greenhouse Gas (GHG) emissions and correlate with controlling climate impacts (Nur & Kurniawan, 2021). However, the success of this policy is also determined by the level of public acceptance and perception of it. Therefore, a sentiment analysis is needed that can measure and understand people's opinions and attitudes towards the electric vehicle subsidy policy.

In the era of social media development, Twitter has become an important platform for gaining insight into public sentiment regarding electric vehicle subsidy policies (Zhu et al., 2013). Twitter's data gives access to real-time public opinion and comments. In the context of sentiment analysis, Twitter can be a valuable source of data to understand public reactions and views on the electric vehicle subsidy policy. Therefore, the application of the Naïve Bayes algorithm to sentiment analysis using Twitter data can provide a deeper understanding of public sentiment regarding this policy.

The Naïve Bayes algorithm is a classification algorithm that is popularly used in sentiment analysis (Wati, 2016). By utilizing the characteristics and words contained in the tweet text, the Naïve Bayes algorithm can classify sentiments as positive, negative or neutral. In the context of electric vehicle subsidy policies, the application of the Naïve Bayes algorithm to Twitter sentiment analysis can help identify and understand public opinion on the policy.

Sentiment analysis using Twitter data is able to identify general trends and patterns of sentiment regarding electric vehicle subsidy policies. This information can help governments and other stakeholders make better decisions in the formulation and implementation of future energy and environmental policies. In addition, sentiment analysis can also reveal concerns or issues that need to be addressed to increase public acceptance and support for the policy (Afif & Pratama, 2021).

The application of the Naïve Bayes algorithm to sentiment analysis allows researchers to carry out quantitative measurements of people's views and attitudes towards electric vehicle subsidy policies. This study aims to measure the performance of the Naïve Bayes algorithm in sentiment analysis of the electric vehicle subsidy policy by using data in the form of tweets, responses and opinions from the public regarding this policy. By analyzing public sentiment, this research can provide insight into general views and attitudes towards electric vehicle subsidy policies, which can be used as a basis for evaluating existing policies and improving future policies.

2. LITERATURE

2.1. Sentiment Analysis

Sentiment Analysis is the process of collecting, extracting and modeling the opinions, sentiments or emotions contained in text or other data. Sentiment analysis is also known as opinion mining, which is the process of extracting an opinion or opinion from a document for a particular topic (Kurniawan & Susanto, 2019). The aim is to understand and analyze the sentiments expressed by individuals or groups in certain contexts. These sentiments can then be classified as positive, negative, or neutral, or they can be described in terms of certain emotional dimensions such as happy, sad, angry, or disappointed (Ramadhon, 2020). The use of sentiment analysis can provide valuable insights about perceptions, preferences, and public satisfaction regarding policies issued by the government.

2.2. Naïve Bayes Algorithm

The Naïve Bayes algorithm is a probabilistic classification based on the use of the Bayes Theorem with a strong assumption that each feature is mutually independent or independent. The Naïve Bayes model is an easy-to-use model, with no parameter estimates or complex iterations, making it usable in a wide range of classification cases, including sentiment analysis (Aggarwal, 2022). The basic concept of Bayes' Theorem is to calculate the posterior probability $P(c|x)$ of $P(c)$, $P(x)$, and $P(x|c)$. Naïve Bayes is based on the simplifying assumption that attribute values are conditionally independent when given output values. In other words, given the output values, the probabilities of observing together are the product of the individual probabilities (Kawani, 2019).

The posterior probability is obtained through the following equation (Vembandasamyp et al., 2015).

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \tag{1}$$

with:

$P(c|x)$: probability of target class posterior given predictor (attribute);
$P(c)$: prior probability of the class;

$P(x|c)$: likelihood which is the probability of a given predictor class;
$P(x)$: prior probability of the predictor.
where c and x are two occurrences or events. The Naïve Bayes algorithm uses probability theory to find the most likely classification of unclassified instances.

2.3. Synthetic Minority Oversampling Technique (SMOTE)

In classification modeling, class imbalance is a common problem that can occur (Thabtah et al., 2020). The method commonly used in handling data imbalance cases is the Synthetic Minority Oversampling Technique (SMOTE). SMOTE makes synthetic data for less or minority classes, so that the number of observations in the minority class becomes more balanced with other more dominant or majority classes (Elreedy & Atiya, 2019). The formation of synthetic data for minority observations is illustrated in Figure 1.

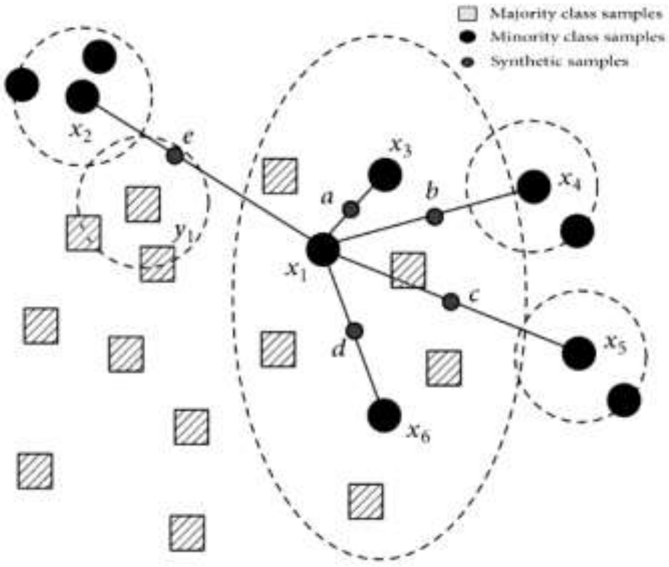

Figure 1. Illustration of synthetic data generation in SMOTE

The procedure for creating synthetic data in SMOTE is as follows.

i. Find the k nearest neighbors, which neighbors are also minority class observations;
ii. Choose randomly j observations from k observations in point (i), where the value of j depends on the amount of oversampling that has been determined;
iii. Generate new synthetic observations on the minority observation straight line with the selected neighbours.

2.4. k-Fold Cross Validation

k-Fold Cross Validation is a method commonly used in predictive model building and model performance evaluation (Berrar, 2019). The data set is divided into k subsets or folds using this technique, and each fold is alternately utilized as a training data set and a test data set. Measurement of the developed model's performance on data is the goal of k-Fold Cross Validation. We may acquire a more precise estimate of the model's general performance by using k-Fold Cross Validation, which combines several training data sets to build the model and test data sets to determine the model's efficacy.

The procedure for the k-Fold Cross Validation method is as follows.

i. Split the data set into k subsets or folds;

ii. For each fold, use that fold as the test data set and incorporate the other folds into the training data set;
iv. Train the model using the training data set and evaluate the performance of the model on the test data set;
v. Repeat steps (ii) and (iii) until each fold becomes validation data alternately;
vi. Compute and aggregate the measures of model goodness (accuracy, precision, recall) from each iteration of the fold.

3. METHODOLOGY

3.1. Data and Sources

The data used in this study is data obtained from Twitter with the keyword 'subsidi kendaraan listrik. The tweet data was taken using the RapidMiner software. The data taken is tweet data in Indonesian. The variables used in this study are the content (*Text*) and sentiment category (*Sentiment*) of each tweet.

3.2. Method

The data analysis method used in this study is the data classification method using the Naïve Bayes algorithm. Each tweet is cleaned first and then identification of the sentiment in the form of positive or negative is carried out. Furthermore, by using k-Fold Cross Validation, the data is divided into 10 subsets or sections with the same proportions for each part, each of which takes turns becoming a test data set while at the same time the others join to become a training data set. From the training data set, data imbalances were handled using SMOTE, and then a classification model was created using the Naïve Bayes algorithm. Each model is tested with a test data set, and so on until each subset or part of the data takes turns becoming a test data set. The final result obtained is the aggregation of each measure of the accuracy of the classification obtained from the 10 repetitions. Modeling is done with the help of Software Python.

3.3. Data Analysis

The steps of data analysis are as follows.

i. Collecting data from Twitter: In this study, the data collected was tweets with the keyword ‘subsidi kendaraan listrik’ obtained in May 2023;
ii. Data cleaning: The tweet data obtained is then processed to clean the text from special characters, punctuation, and URLs. This step is also taken to eliminate retweets, mentions, hashtags, and data duplication;
iii. Pre-processing and formation of TF-IDF: This step includes repairing abbreviations, tokenizing, removing stopwords, and stemming to prepare the text for processing. The tweet text is then converted into a vector representation using the Term Frequency-Inverse Document Frequency (TF-IDF) method. This is done to calculate the weight of the words in the text and highlight the most important words in the sentiment analysis;
iv. Data exploration: All tweets are combined into one string of text and presented in the wordcloud.
v. Data labeling: Tweet data is sentiment labeled by positive or negative category;
vi. k-Fold Cross Validation: Tweet data is divided into 10 subsets or sections with the same proportions in each section;
vii. Training data clusters and test data clusters: Each subset or part alternately becomes a test data cluster, while the rest combine to become a training data cluster. The training data

set is used to train the Naïve Bayes model, while the test data set is used to test the model's performance;

viii. Handling of data imbalance: Handling of data imbalance cases for the training data set is done by oversampling technique using SMOTE for observations with minority classes;

ix. Establishment of the Naïve Bayes model: The Naïve Bayes model is trained using pre-prepared training data sets. The model will learn the relationship between features and sentiments in the training data set;

x. Prediction: The model that has been created is then tested using a test data set. The prediction results are then used to measure the goodness of the model;

xi. Measuring model goodness: The performance of the Naïve Bayes model is measured using a test data set. Measurement of the goodness of the model is measured from the value of accuracy, precision, and recall. All three are used to measure how well the model can correctly classify sentiments;

xii. Repeat steps (vii) to (xi). This repetition is carried out until each subset or part of the data turns into a test data set;

xiii. Calculates the aggregation of the goodness of the model: the average accuracy, precision, and recall of each iteration.

4. RESULT AND DISCUSSION

4.1. Data Collection and Cleaning

The results of a search for data from Twitter using the RapidMiner software with the keyword 'subsidi kendaraan listrik' found 5,000 tweets in May 2023. After cleaning irrelevant data and duplicate data, we obtained 822 tweets that were ready to be processed. The results of the sample data before and after the data is cleaned are shown in Table 1.

Table 1. Sample Data Before and After Cleaning

Before	After
Sri Mulyani Tanggapi Kritik DPR soal Subsidi Kendaraan Listrik https://t.co/xh4CwLGb92	Sri Mulyani Tanggapi Kritik DPR soal Subsidi Kendaraan Listrik
@aki_jupi2 Menurutku sih gak akan terjadi kesenjangan sosial kalo subsidi cuman sekali aja, bukan seperti BLT. Lagian sekarang ini pendapatan per kapita Indonesia lagi tinggi2nya, sangat perlu didorong untuk mulai pindah ke mobil listrik yg komponennya masih mahal	Menurutku sih gak akan terjadi kesenjangan sosial kalo subsidi cuman sekali aja, bukan seperti BLT. Lagian sekarang ini pendapatan per kapita Indonesia lagi tinggi2nya, sangat perlu didorong untuk mulai pindah ke mobil listrik yg komponennya masih mahal
Yg bikin kebijakan soal subsidi mobil listrik ini emang konyol, ga liat apa tu moda transport macam TJ & KRL kembang kempis butuh modal	Yg bikin kebijakan soal subsidi mobil listrik ini emang konyol, ga liat apa tu moda transport macam TJ KRL kembang kempis butuh modal
Dokter Orang Pinter Di PRANK2 Apalagi Petani Kecil Pupuk Subsidi Dikurangi BBM di Naikin Meski Harga Dunia Turun Harga Kebutuhan Hidup Naik, APBN Duit Pajak Rakyat Buat Subsidi Konglomerat Mobil Listrik Semua Diancam UU Omnibus Law Kritik Di Polisikan Dgn UU ET Oleh Buszer Rp https://t.co/qK8IOsaPCK	Dokter Orang Pinter Di PRANK2 Apalagi Petani Kecil Pupuk Subsidi Dikurangi BBM di Naikin Meski Harga Dunia Turun Harga Kebutuhan Hidup Naik, APBN Duit Pajak Rakyat Buat Subsidi Konglomerat Mobil Listrik Semua Diancam UU Omnibus Law Kritik Di Polisikan Dgn UU ET Oleh Buszer Rp

4.2. Pre-processing

Improvements to abbreviations, changes to lowercase, tokenize processes, removal of stopwords, and stemming are carried out at this stage with the help of Python software to prepare text for processing. The tweet text is then converted into a vector representation using the Term Frequency-Inverse Document Frequency (TF-IDF) method. After that, the word order is re-combined into a sentence. The results of the sample data before and after pre-processing are shown in Table 2.

Table 2. Sample Data Before and After *Pre-processing*

Before	After
Sri Mulyani Tanggapi Kritik DPR soal Subsidi Kendaraan Listrik	sri mulyani kritik fraksi dpr bijak subsidi kendara listrik
Menurutku sih gak akan terjadi kesenjangan sosial kalo subsidi cuman sekali aja, bukan seperti BLT. Lagian sekarang ini pendapatan per kapita Indonesia lagi tinggi2nya, sangat perlu didorong untuk mulai pindah ke mobil listrik yg komponennya masih mahal	turut sih gak senjang sosial kalo subsidi cuman aja blt dapat kapita indonesia tinggi dorong pindah mobil listrik komponen mahal
Yg bikin kebijakan soal subsidi mobil listrik ini emang konyol, ga liat apa tu moda transport macam TJ KRL kembang kempis butuh modal	bikin bijak subsidi mobil listrik emang konyol ga liat tu moda transport tj krl kembang kempis butuh modal
Dokter Orang Pinter Di PRANK2 Apalagi Petani Kecil Pupuk Subsidi Dikurangi BBM di Naikin Meski Harga Dunia Turun Harga Kebutuhan Hidup Naik, APBN Duit Pajak Rakyat Buat Subsidi Konglomerat Mobil Listrik Semua Diancam UU Omnibus Law Kritik Di Polisikan Dgn UU ET Oleh Buszer Rp	dokter orang pinter prank tani pupuk subsidi rang bbm naikin harga dunia turun harga butuh hidup apbn duit pajak rakyat subsidi konglomerat mobil listrik ancam uu omnibus law kritik polisi dgn uu et buszer rp

4.3. Data Exploration

Data exploration using wordcloud is one of the popular methods in text analysis to provide a visual picture of the words that appear most often in the dataset. By visualizing words proportionally sized based on their frequency, wordcloud helps identify key words or topics that are dominant in the data. This allows analysts to quickly spot important patterns, identify trends, or highlight relevant issues that arise in text. Wordcloud is a useful tool for presenting visual summaries of text data and providing initial insights that can be used as a basis for further exploration and analysis. Figure 2 shows the wordcloud from the data.

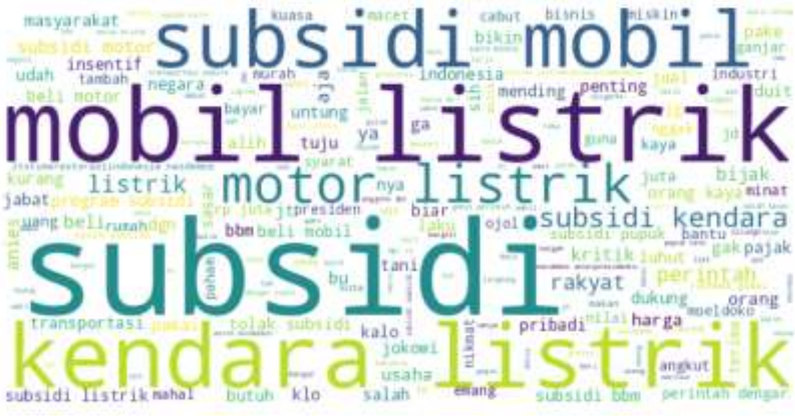

Figure 2. Wordcloud of Electric Vehicle Subsidy Policy

Figure 2 shows that there are several words that are larger than other words, such as "mobil listrik", "motor listrik", "subsidi", “listrik”, and "kendara". The interpretation of this wordcloud is that in discussions about electric vehicle subsidy policies on Twitter, these words appear with high frequency and become the focus of attention of Twitter users. This shows that issues related to the accuracy of policy targets, related industries, and impact on traffic are becoming important topics and attracting the attention of Twitter users. This is also reinforced by words such as "orang kaya", "pajak", "masyarakat", "industri", and "macet" which appear in a smaller size.

4.4. Labeling

The data that is ready to be processed is then labeled. In this study, sentiment labels were determined automatically using Python with nltk vader lexicon sentiwords_id. The results of data labeling are shown in Table 3.

Table 3. Sample Results of Data Labeling

Tweet	Sentiment
sri mulyani kritik fraksi dpr bijak subsidi kendara listrik	Positive
turut sih gak senjang sosial kalo subsidi cuman aja blt dapat kapita indonesia tinggi dorong pindah mobil listrik komponen mahal	Positive
bikin bijak subsidi mobil listrik emang konyol ga liat tu moda transport tj krl kembang kempis butuh modal	Negative
dokter orang pinter prank tani pupuk subsidi rang bbm naikin harga dunia turun harga butuh hidup apbn duit pajak rakyat subsidi konglomerat mobil listrik ancam uu omnibus law kritik polisi dgn uu et buszer rp	Negative

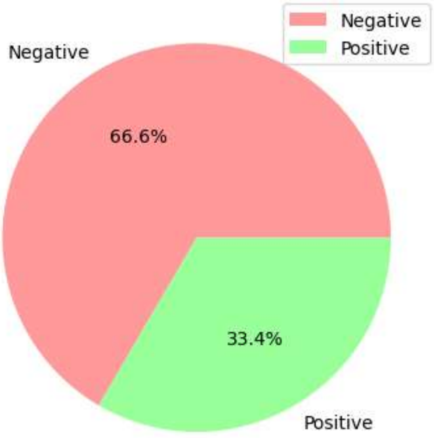

Figure 3. Pie Chart Sentiment of Electric Vehicle Subsidy Policy

Figure 3 shows that the response from the majority of the public has a negative sentiment. There are cases of imbalance in the data that need to be addressed. Cases of data imbalance in this study will be handled by the SMOTE method, which is applied to each training data, which aims to increase the measurement of the accuracy of the classification in the model.

4.5. Modeling

In this study, the Naïve Bayes classification algorithm used in modeling with a k-fold Cross Validation of 10 folds. The k-fold Cross Validation method is used to see the stability of the model results, which is measured by how far the available data is around the average. At each repetition, data attacks are handled against the training group data using the SMOTE method. The measurement of classification accuracy from modeling results is measured through the results of accuracy, precision, and recall. Accuracy is the ratio of correct predictions (positive and negative) to the entire data. Precision is the ratio of positive correct predictions compared to the overall positive predicted results. Meanwhile, recall is the ratio of positive correct predictions compared to all positive correct data. The results of the measurement of classification accuracy at each repetition are shown in Table 4.

Table 4. Classification Accuracy Measurement

Fold	Accuration	Precision	Recall
1	0.686747	0.521739	0.857143
2	0.771084	0.621622	0.821429
3	0.650602	0.486486	0.642857
4	0.734940	0.575000	0.821429
6	0.578313	0.414634	0.607143
7	0.722892	0.567568	0.750000
8	0.731707	0.564103	0.814815
9	0.621951	0.447368	0.629630
10	0.731707	0.575758	0.703704
Average	**0.694946**	**0.532691**	**0.742593**
Standard Deviation	**0.060045**	**0.064658**	**0.091060**
Maximum	**0.771084**	**0.621622**	**0.857143**

The results of the classification accuracy measure in Table 4 show that the average accuracy of the resulting model is 69.49%, with precision and recall values of 53.27% and 74.26%, respectively. The highest accuracy and precision was obtained on the 2nd iteration, namely 77.11% and 62.16%. While the highest recall value was obtained in the 1st repetition, which was 85.71%. The result of the standard deviation is used to measure how far the data is spread around the average.

5. CONCLUSION

The results show the application of the Naïve Bayes algorithm in sentiment analysis related to electric vehicle subsidy policies. Modeling using the Naïve Bayes algorithm is able to predict the results of the accuracy or ratio that is suspected to be correct with an overall data of 69.49% with a standard deviation of 6%. The model is also able to show the precision or ratio of true positive predictions compared to the overall positive predicted results of 53.26% with a standard deviation of 6.46%. Judging from its sensitivity (recall), the model is able to measure the ratio of correct positive predictions compared to all correct positive data of 74.26% with a standard deviation of 9.1%. The sentiment results obtained show that the majority of people have negative sentiments towards the policy (66.6%). This negative sentiment can be seen from the perspective of the interests of policy makers. The community is of the view that these subsidies will only be enjoyed by those who can afford them, even though subsidy funds should be allocated to other, more important sectors.

Further understanding and analysis from the perspective of public policy towards this negative sentiment is important to do, so as to produce better insights on the various existing perspectives. Future research can use other social media as data sources. In addition, the application of multiclass classification can also be applied to measure sentiment in more detail and be classified as a whole.

REFERENCES

Afif, A. S., & Pratama, A. R. (2021). Analisis Sentimen Kebijakan Pendidikan di Masa Pandemi COVID-19 dengan CrowdTangle di Instagram. *Automata*. https://journal.uii.ac.id/AUTOMATA/article/view/19429

Aggarwal, C. C. (2022). An Introduction to Text Analytics. In *Machine Learning for Text* (pp. 1–17). Springer.

Berrar, D. (2019). *Cross-Validation.*

Elreedy, D., & Atiya, A. F. (2019). A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. *Information Sciences, 505*, 32–64.

Kawani, G. P. (2019). Implementasi Naive Bayes. *Journal of Informatics, Information System, Software Engineering and Applications (INISTA)*, *1*(2), 73–81. https://doi.org/10.20895/inista.v1i2.73

Kurniawan, I., & Susanto, A. (2019). Implementasi Metode K-Means dan Naïve Bayes Classifier untuk Analisis Sentimen Pemilihan Presiden (Pilpres) 2019. *Eksplora Informatika*, *9*(1), 1–10. https://doi.org/10.30864/eksplora.v9i1.237

Nur, A. I., & Kurniawan, A. D. (2021). Proyeksi Masa Depan Kendaraan Listrik di Indonesia: Analisis Perspektif Regulasi dan Pengendalian Dampak Perubahan Iklim yang Berkelanjutan. *Jurnal Hukum Lingkungan Indonesia*, *7*(2), 197–220. https://doi.org/10.38011/jhli.v7i2.260

Ramadhon, M. I. (2020). *Analisis Sentimen Terhadap Pemindahan Ibu Kota Indonesia Pada Media Sosial Twitter Menggunakan Metode Algoritma K-Nearest Neighbor (K-Nn).*

Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. *Information Sciences*, *513*, 429–441.

Vembandasamyp, K., Sasipriyap, R. R., & Deepap, E. (2015). Heart Diseases Detection Using Naive Bayes Algorithm. *IJISET-International Journal of Innovative Science, Engineering & Technology*, *2*(9), 1–4. www.ijiset.com

Wati, R. (2016). Penerapan Algoritma Genetika Untuk Seleksi Fitur Pada Analisis Sentimen Review Jasa Maskapai Penerbangan Menggunakan Naive Bayes. *Jurnal Evolusi*, *4*(1), 25–31.

Zhu, Z., Blanke, U., Calatroni, A., & Tröster, G. (2013). Human activity recognition using social media data. *12th International Conference on Mobile and Ubiquitous Multimedia, MUM 2013*. https://doi.org/10.1145/2541831.2541852