ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 12, Issue 03, Desember 2023

STANCE ANALYSIS OF POLICIES RELATED TO EMISSION TEST OBLIGATIONS USING TWITTER SOCIAL MEDIA DATA

Dwi Retnoningrum[1], Dea Annisayanti Putri[2], Indra Budi[3], Aris Budi Santoso[4], Prabu Kresna Putra[5]

[1,2,3,4,5]Faculty of Computer Science, University of Indonesia

email: dwi.retnoningrum@ui.ac.id[1], dea.annisayanti@ui.ac.id[2], indra@ui.ac.id[3], aris.budi@ui.ac.id[4], prabu.kresna@ui.ac.id[5]

Abstract

Social media is currently widely used to disseminate various kinds of information, whether expressing feelings, or opinions. Public opinion is no exception regarding government policies and the implementation of emission tests, which describe the conditions that exist in society. Information on public opinion data obtained through social media in real time can assist the government in evaluating policies and improving the quality of currently implemented policies, particularly evaluating the implementation of emission tests on motorized vehicles. In this research, the application of stance analysis is used to evaluate emission test policies based on public opinion.In addition, this research aims to combine several machine learning methods and feature extraction methods to find the best combination based on accuracy, training time, and prediction time based on emission test policies. The best model based on the level of accuracy is a combination of Decision Tree and BERT, which reaches a value of 66%. Meanwhile, based on training time, the model that has the advantage is the Ridge Classifier with fasttext text representation. Based on prediction time, there are 3 combination models, namely Decision Tree with word2vec, SVM with Word2Vec, and Logistic Regression with fasttext text representation.

Keywords : Emission Test Policy, Social Media, Stance Analysis, Machine Learning, Feature Extraction.

Received: 03-11-2023 | **Revised:** 04-01-2024 | **Accepted:** 08-01-2024
DOI: https://doi.org/10.23887/janapati.v12i3.69004

INTRODUCTION

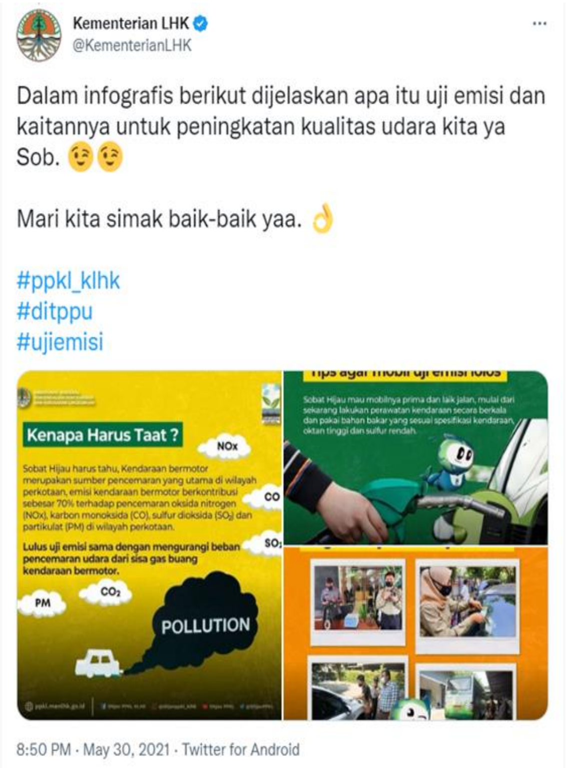
The development of technology and information has increased rapidly and has entered various aspects of human life, such as social, cultural and economic activities [1]. Through technology, social activities that previously required physical contact can now be carried out remotely via social media [2]. This shift is happening all over the world, 60% of the world's population is connected to the Internet and 53% of the world's population has access to social media[3].

Social media is currently widely used to disseminate various kinds of information, whether expressing feelings, opinions or opinions. For consumers, social media can be used to convey positive feedback about the products used [4]. For the government, social media can be used to find out the public's response to an issue.

Twitter is the social media with which a lot of data is collected as research material. Apart from using a lot of data, Twitter is also a social media that has the most active users in Indonesia. According to [3], approximately 108 million Indonesians are active Twitter users. Corporate communication facilities, customer service, and product campaigns are conducted via Twitter. Research topics are also very broad, such as politics [5], education [6], and film [7].

With the availability of important information for companies on social media such as Twitter, the analysis is not only done manually but also utilizes data mining techniques in the textual form which will be grouped into several classes [8]. The stance detection technique was widely used before the creation of social media, where this technique was used to analyze an issue whose data was taken from websites, blogs, or surveys. Stance detection can be used to determine the condition of certain aspects of society. Various approaches or methods are used in stance analysis, a popular example is research by [9] which discusses the use of Word2Vec and LSTM to carry out stance classification, [10] who utilizes Word2vec and SVM, and [11] who uses Glove, Word2Vec, and CNN.

Not only private companies have taken advantage of social media, but government agencies have also opened official accounts on the Facebook, Twitter, and Instagram platforms. The government through the related Ministries distributes information and appeals to the public, one of which is through social media channels. No exception is related to one of the government programs in handling global warming, namely emission test regulations test Figure 1



Kementerian LHK
@KementerianLHK

Dalam infografis berikut dijelaskan apa itu uji emisi dan kaitannya untuk peningkatan kualitas udara kita ya Sob. 😉😉

Mari kita simak baik-baik yaa. 👌

#ppkl_klhk
#ditppu
#ujiemisi

Kenapa Harus Taat ?

Sobat Hijau harus tahu, Kendaraan bermotor merupakan sumber pencemaran yang utama di wilayah perkotaan, emisi kendaraan bermotor berkontribusi sebesar 70% terhadap pencemaran oksida nitrogen (NOx), karbon monoksida (CO), sulfur dioksida (SO2) dan partikulat (PM) di wilayah perkotaan.

Lulus uji emisi sama dengan mengurangi beban pencemaran udara dari sisa gas buang kendaraan bermotor.

Tips agar mobil uji emisi lolos

Sobat Hijau mau mobilnya prima dan laik jalan, mulai dari sekarang lakukan perawatan kendaraan secara berkala dan pakai bahan bakar yang sesuai spesifikasi kendaraan, oktan tinggi dan sulfur rendah.

8:50 PM · May 30, 2021 · Twitter for Android

Figure 1. Tweet from the Ministry of Environment and Forestry regarding Emission Tests

The government makes a policy for testing exhaust emissions followed by regular and orderly maintenance of motor vehicles which is carried out correctly, effectively has the potential to minimize motor vehicle exhaust gases, and provides many benefits, namely improving air quality and maintaining public health. One of them is contained in DKI Jakarta Governor Regulation (Pergub) Number 66 of 2020. The DKI Jakarta Government has also created a system that can monitor the process of implementing this emission test through the Emission Test Information System (https://ujiemisi.jakarta.go.id /).

Information on social media can be used to evaluate emission test policies and explore public opinion about both policies and their implementation. So, the information obtained is able to describe the situation in society. In addition, information in public opinion data obtained through social media in real-time, can assist the government in evaluating policies and improving the quality of currently implemented policies, especially evaluating the application of emission tests on motorized vehicles. With ever-growing levels of data, the methods used will be updated regularly. So the best method is needed to conduct stance analysis and seek public opinion regarding emission tests policies. By raising this problem, the researcher formulates 2 questions, that is "How is the performance comparison of various algorithms to implement social media listening on Twitter social media based on accuracy, training time, and prediction time?" and "What is the public's response to the emissions test policy?".

Social Media

Social media, users can interact and convey various ideas, content, opinions, and information through social media [12]. There are various types of social media with their respective advantages. Examples of social media are Facebook, which is used to share daily activities, Twitter, which is used to quickly share textual information, Youtube, which is used to share videos, and Instagram, which is used to share photos.

The development of social media platforms is directly proportional to the increasing amount of data, which is called social media big data. The data comes from activities carried out by users, such as expressing opinions or complaining about products or services. Users are not restricted from creating content or sharing information [13]. Data available on social media is analyzed for various purposes and with various methods.

Twitter

Twitter is also a social media that has the most active users in Indonesia. According to [3] approximately 108 million Indonesians are active Twitter users. Twitter generates an average of over 500 million tweets per day [14]. Twitter can be used as a source to find out what's going on right now because news travels quickly worldwide. So that currently many individuals, companies, product vendors, and organizations have used public opinion on social media as a basis for decision-making. Corporate communication facilities, customer service, and product campaigns are conducted via Twitter. Research topics are also very broad, such as politics [5], education [6], and film [7].

Stance Analysis

Stance analysis is a technique to automatically determine the author's views on a text, whether they agree (pros), neutral, or disagree (cons). In general, the stance analysis process can be divided into two stages, the formation of word vectors and stance classification. The popularity of this model is due to its ability to determine the similarity of meaning between words. This similarity information is obtained by observing the similarity of the words around the target word. Stance analysis is a type of task within the science of Natural Language Processing, which is a scientific focus to analyze and understand the meaning contained in a text. The general method used for stance analysis, i.e. Decision Tree, Support Vector Machine (SVM), Regression, and Ensemble.

METHOD

The stages of the research were conducted to answer the research objective, namely to compare five classification algorithms and five text representation algorithms with the aim of stance analysis on Twitter's opinion about the emission test policy. This methodology shown in Figure 2.

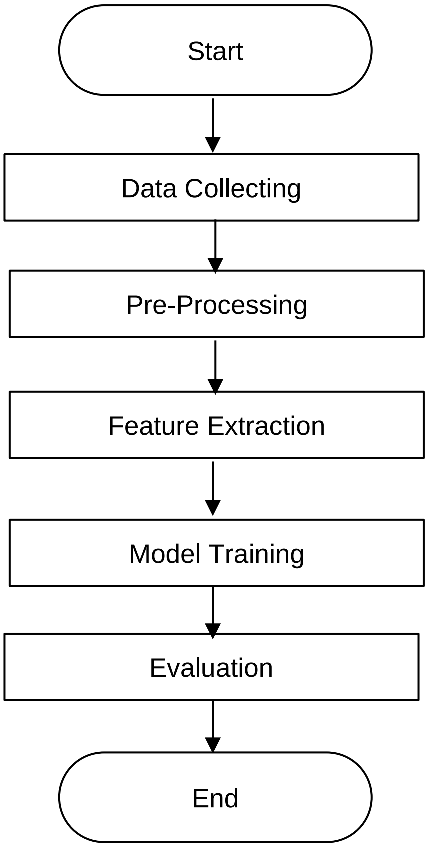


Figure 2. Methodology

Data Collecting

In collecting data, the scrapping method is used, which is a technical process of collecting data on a site through an information extraction process using Hypertext Transfer Protocol (HTTP). In this process, data was collected from October 1, 2020 to October 1, 2022. Key words are “uji emisi”, “emisi ribet”. The amount of data that could be collected were 4,500 tweets. Tweet data is then labelled into 3 classes, namely positive, negative, and neutral manually by 3 people. To maintain consistency and quality, a cross-check is carried out, in which each person will label 1,500 tweets and will then be re-checked by another person.

Pre-Processing

Pre-processing aims to reduce to a minimum the slang vocabulary or terms used in the text [15]. Social media users tend to replace formal words with slang or terms. Examples include using numbers to replace the alphabet, repeated vowel characters, and using non-standard words. Case folding changes all the letters to lowercase. Tokens are punctuation marks, terms, numbers, etc. [16], while tokenization is dividing the text into certain parts [17]. Word filtering is needed to improve and make the text of Twitter (tweets) normal by removing URLs, mentions, and hashtags. The process of removing stop-words is used to clear the text of conjunctions, which research has shown do not carry useful information.

The data generated from the labelling process must go through cleaning beforehand, because the raw data is the free text that can be randomly generated by Twitter users. There are lots of unnecessary words, then lots of characters or emojis, to meaningless punctuation in the following process. The steps taken include case folding to convert all characters into lowercase, tokenizing text into words, filtering meaningless characters, and removing stop-words. In addition to cleaning content, information data from Twitter users will also be deleted to maintain social media ethics.

Feature Extraction

An important part of text processing is converting text values into numeric values, either in vectors or other representations [18]. The determination of features for the machine learning approach (machine learning) is adjusted to text categorization. According to [19], the feature extraction method reduces original features by removing irrelevant features, which aims to increase accuracy and reduce machine learning processing time.

The first method is the bag of words (BOW). The BOW approach is a popular feature extraction method for sentences and documents. The text representation will be based on a histogram of words, i.e., considering each word count [20].

Also, the method for extracting features using the Word2vec model. Word2vec is a word augmentation technique introduced by Mikolov for word expressions that contain the meaning and context of words in a document and include two learning algorithms, a namely continuous bag of words (CBOW) and skip algorithms [21][22]. The similarity between words is calculated via the cosine similarity of the word vectors in word2vec, which includes the meanings of the words in the document [23]. Several studies regarding sentiment analysis or classification of emotions can use word2vec [24] [25].

Besides word2vec, another feature extraction method is Fasttext. Fasttext was developed by Facebook's research team to learn how to represent words and categorize text efficiently [26]. Fasttext model’s most significant contribution is that it considers the internal structure of words by examining word representations. Fasttext is especially useful for languages with a wide variety of morphologies [27]. Word representation's approach in fasttext embedding model is quite different from other word representations, such as word2vec [28]. Although fasttext assumes a word is arranged by n-gram characters where length could change from one to another, the smallest unit from each term used in word2vec. The advantage of this method is that since it stores word vectors as ngrams of characters, it can find vector representations for words that are not directly in the dictionary [28].

A significant advance in neural network models was made by models using the Transformer architecture based on the self-recognition mechanism [29], such advantages have led to the development of new models, but Bidirectional Encoder Representations from Transformers (BERT), which use context and word embeddings to overcome the limitations of RNNs and LSTMs, significantly improve the performance of sentiment analysis [30].

The biggest advantage of BERT would be an unneeded big corpus of text to train models. BERT is pre-trained, so users only need to fine-tune the BERT model based on specific training data (manually annotated). In addition to classification, BERT can also be used to obtain vector representations. Modification of a trained BERT network that uses triplet and siamese network structures to earn semantically meaningful representations called Sentence-BERT (SBERT) [31].

Data from the cleaning process will then be converted into numeric data. The numerical representation of this result is in the form of a vector or an equal number of numerical values. The methods used to generate this vector are bag of word, tf-idf, word2vec, fasttext and BERT. All algorithms are obtained using the Python instrument and the gensim library.

The word2vec model utilizes the gensim library and is trained with Indonesian Wikipedia data. Like-wise with the fasttext model is used through the gensim library and is trained using data on Indonesian language sites and Wikipedia data. This model is available on site fasttext. For BERT, utilize the sentence-bert library and the Indobert hugging face model.

To get the vector representation of the word models (word2ved and fasttext) at the sentence level, the vectors for each word will be combined, and the average will be calculated. While the sentence model (BERT) then, the representation is made directly from the sentence level. This vector will represent the features of the text which will be the input of the classification model. Word2vec vector size is 1000, fasttext is 300, while BERT is 768.

Model Training

Sentiment classification is done by classifying text into two classes, positive or negative class [32]. Neutral classes can be used, but most studies do not use them. Sentiment classification is a form of text classification in general, namely giving a label to the text where this label has been previously defined. Examples include labelling news into class categories such as sports or politics. In classification, related topic words are the main features. But in its derivatives at the sentiment level, the label defined is positive or negative, according to the subjectivity of the text writer.

According to [33]machine learning is part of research in intelligent computing that aims to create programs that can mimic human intelligence without having to write code explicitly. The approach is to analyze the data and other data around it to find patterns. There are two types of machine learning methods, namely supervised and unsupervised. In the supervised learning method, the data to be used and trained already has a label or group. In contrast, the data for training on unsupervised learning has no labels or clusters. The result of this process is the ability to assign labels or group values to unlabelled data based on patterns found from the surrounding data. There are several machine learning methods, but the methods used in this research are Decision Tree, Support Vector Machine (SVM), Logistic Regression, and Ridge.

Decision Tree is a well-known data mining and machine learning technology that takes a series of attribute values as input and outputs a

Boolean conclusion [34]. In practice, every path of a decision tree represents a decision rule that is easily translatable into either a programming language or human language. Considering all paths (rules), the complete tree corresponds to a compound Boolean expression utilizing disjunction and conjunction to produce a Boolean judgment. Decision Tree may be preferred since it is straightforward and simple to interpret [34].

SVM is a suitable method for text analysis, as Naives Bayes requires data training to get the right results [35]. On the other hand, SVM requires more computation than Naive Bayes, but the results are faster and more accurate [36].

Logistic regression is a popular linear regression analysis model, widely used in data mining, automated disease diagnosis, economic forecasting, etc.[37]. The goal of this method is to select variables with more information to identify the type of sample estimation and build a model with the lowest probability of error [38].

Ridge Regression approach for analyzing multicollinear regression statistics when statistics are generated with multicollinearity, the least squares estimate is an independent estimate, which has a big variance and consequently has a tendency to go long way from the real value. Ridge regression produces a more reliable estimate by reducing standard errors and adding some bias to the regression estimate. Ridge regression is a very flexible and subjective regression assessment, but it is an analytical approach combining qualitative and quantitative assessments. It is specific to fixing multicollinearity troubles and is regularly utilized in widespread research [39].

The model will be formed to classify tweets into negative (disagree), positive (agree), and neutral attitudes. The model architectures built are Decision Tree, SVM, Logistic Regression, Ridge Classifier, and combined methods (ensemble).

Before conducting experiments and comparing models, it is necessary to divide the data into training, test, and validation data. Training data is data used to train the model to do its job correctly and validation data to validate performance during training. Test data is used to test whether the model after the training process can predict correctly and is evaluated with a matrix such as an accuracy.

Evaluation

Five models of text representation (feature extraction models) and five classification models will each be combined into 25 models shown in Table1.

Table 1. 25 Models for Evaluation

Classification Model	**Feature Extraction Model**
Decision Tree	BOW
Decision Tree	TF-IDF
Decision Tree	Fasttext Wiki-ID
Decision Tree	Word2Vec-ID
Decision Tree	BERT
SVM	BOW
SVM	TF-IDF
SVM	Fasttext Wiki-ID
SVM	Word2Vec-ID
SVM	BERT
Logistic Regression	BOW
Logistic Regression	TF-IDF
Logistic Regression	Fasttext Wiki-ID
Logistic Regression	Word2Vec-ID
Logistic Regression	BERT
Ridge Classifier	BOW
Ridge Classifier	TF-IDF
Ridge Classifier	Fasttext Wiki-ID
Ridge Classifier	Word2Vec-ID
Ridge Classifier	BERT
Ensemble	BOW
Ensemble	TF-IDF
Ensemble	Fasttext Wiki-ID
Ensemble	Word2Vec-ID
Ensemble	BERT

From each combination, the following are the metrics compared to answer the research questions:

1.Model accuracy
2.Training time
3.Time prediction

RESULT AND DISCUSSION

Based on testing of several models with several feature extraction, the results are as follows:

Model Accuracy

According to table 2, 2 models achieve the highest evaluation with an F1 value of 66%, namely SVM and Ensemble, both of which use BERT representation vectors. Based on the average results of all text representation models, the best classification model is the Decision Tree with an average F1 value of 56%. Even though it is the best, this classification model is not significantly different from other models, which are only 1% different. Meanwhile, based on the average results of all classification models, the best text representation model is BERT, with an average F1 value of 63%. This representational model shows the best performance and has a fairly high difference 3% compared to TF-IDF.

Table 2. Comparison of model accuracy

	BOW	TF-IDF	Fasttext Wiki-ID	Word2Vec-ID	BERT
Decision Tree	0.61	0.52	0.56	0.55	0.56
SVM	0.54	0.64	0.44	0.47	**0.66**
Logistic Regression	0.53	0.59	0.44	0.54	0.65
Ridge Classifier	0.49	0.63	0.43	0.43	0.63
Ensemble	0.56	0.63	0.44	0.48	**0.66**

Table 3. Comparison of Training Time

	BOW	TF-IDF	Fasttext Wiki-ID	Word2Vec-ID	BERT
Decision Tree	4,28	3,98	1,71	1,57	6,23
SVM	0,56	0,40	0,59	6,93	11,45
Logistic Regression	206,60	20,15	0,59	8,24	47,34
Ridge Classifier	2,60	2,60	**0,03**	54,00	101,00
Ensemble	214,10	28,09	2,93	16,97	65,41

Table 4. Comparison of Time Prediction

	BOW	TF-IDF	Fasttext Wiki-ID	Word2Vec-ID	BERT
Decision Tree	0,050	0,061	0,035	**0,030**	1,921
SVM	0,057	0,067	0,035	**0,030**	1,962
Logistic Regression	0,061	0,068	**0,030**	0,031	1,992
Ridge Classifier	0,122	0,137	0,080	0,131	2,043
Ensemble	0,142	0,155	0,138	0,132	2,053

Table 5. Comparison of classification models based on average text representation models

	Average F1	***Average Train***	***Average Prediction***
Decision Tree	**<u>0.56</u>**	**<u>3.55</u>**	**<u>0.419</u>**
SVM	0.55	3.99	0.430
Logistic Regression	0.55	56.58	0.436
Ridge Classifier	0.52	32.05	0.503
Ensemble	0.55	65.50	0.524

Table 6. Comparison of text representation models based on average classification model

	Average F1	***Average Train***	***Average Prediction***
BOW	0,55	85,63	0,086
TF-IDF	0,60	11,04	0,098
Fasttext Wiki-ID	0,46	**<u>1,17</u>**	**<u>0,064</u>**
Word2Vec-ID	0,49	17,54	0,071
BERT	**<u>0,63</u>**	46,29	1,994

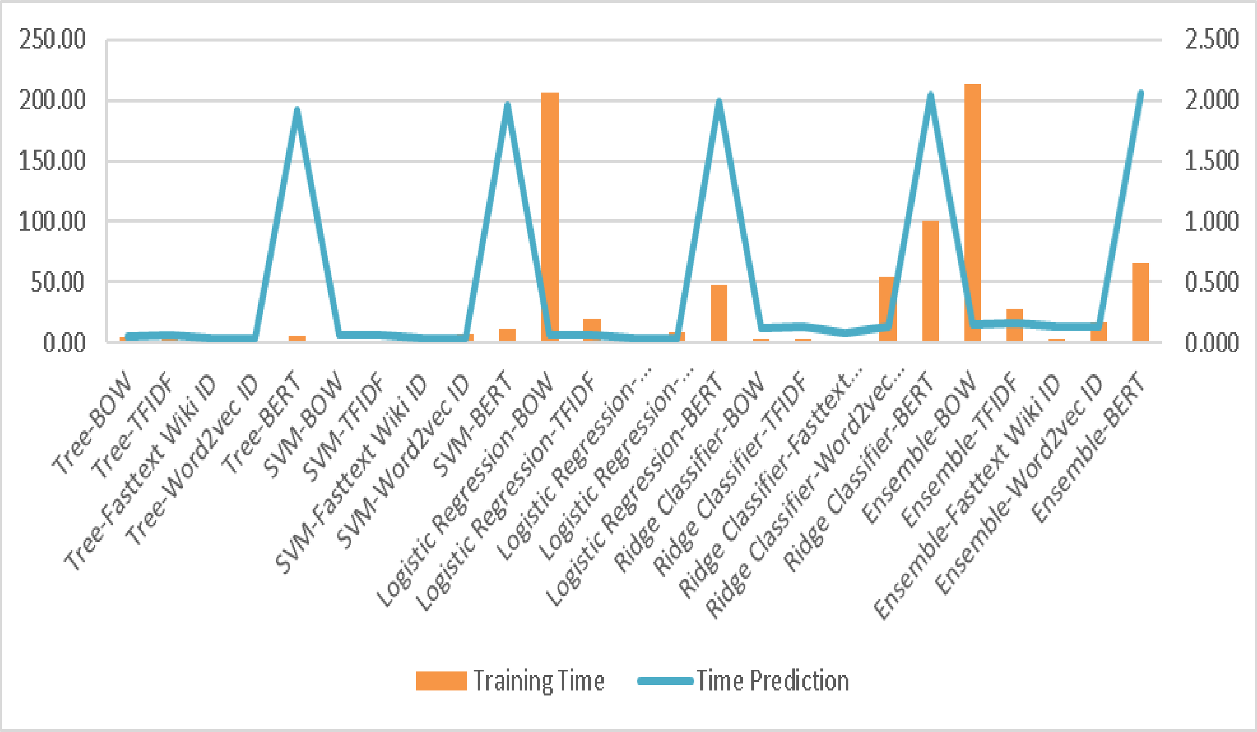


Figure. 3 Comparison of training time and time Prediction

Training Time

According to Table 3 and Figure 3, the model with an advantage in training speed is the Ridge Classifier with fasttext text representation. This combination of models can train a sentiment model with 4,320 data in 0.03 seconds.

Based on the average results of all text representation models, the best classification model based on training time is the Decision Tree, with an average time of 3.55 seconds. Based on the average results of all classification models, the best text representation model based on training time is fasttext, with an average time of 1.17 seconds.

BOW takes the most time since it needs to build the vocabulary and complexity, to calculate the frequency. BERT with its architecture complexity takes the second longest time in training. At the same time, fasttext is the fastest due to its vector size, which is only 300.

Time Prediction

According to Table 4 and Figure 3, 3 models are able to make the fastest predictions (0.03 seconds), namely Decision Tree and SVM, both with Word2Vec text representation, and Logistic Regression with fasttext text representation. Based on the average results of all text representation models, the best classification model based on prediction time is the Decision Tree with an average time of 0.430 seconds. Based on the average results of all classification models, the best text representation model based on training time is fasttext, with an average time of 0.064 seconds. For the same reason with training time, Fasttext outperformed the others due to its number of smaller vectors.

Public’s Response To The Emission Test Policy

Based on the best accuracy model, which is SVM-BERT, the following is the result of the classification of all community tweets about emission tests:

1.246 (5%) tweets pro with government
2.2778 (62%) tweets neutral
3.1475 (33%) tweets contra with government

CONCLUSION

Experiments were conducted to answer research questions regarding the best combination of text representation and classification models. It can be concluded that:

The best model based on the level of accuracy is a combination of Decision Tree and BERT, which reaches a value of 66%. Meanwhile, based on training time, the model that has the advantage is the Ridge Classifier with fasttext text representation. If based on prediction time, there are 3 combination models, namely Decision Tree with word2vec, SVM with Word2Vec, and Logistic Regression with fasttext text representation.

On average, the model representation in Table 5, Decision Tree is the best classification model because this model achieves the best performance of all components (accuracy, training time, and prediction time). Meanwhile, based on the average classification model in Table 6, FastText is the best text representation model. Even though it doesn't have a good level of accuracy, FastText has an advantage in speed in training and prediction.

In addition, the public response to the emission test policy is dominated by neutral opinions 66%, followed by opinions that contain cons to emission tests 33%, and tweets that agree with the government 5%.

The model produced in this study is expected to be a reference for the government to build a system that evaluates emission test policies and explores public opinion regarding both policies and their implementation. So from the information obtained, it can describe the situation in society. This system can be implemented in real-time and updated regularly to get good model quality.

The suggestion for further research is to explore the correlation analysis between people's attitudes and an incident. In addition, topic modeling can also be applied to find out the reasons for each community's attitude. Finally, future research can explore such as lexicon-based, neural network-based models and BERT-based classification models

REFERENCES

[1] J. Cruz-Cárdenas, E. Zabelina, O. Deyneka, J. Guadalupe-Lanas, and M. Velín-Fárez, “Role of demographic factors, attitudes toward technology, and cultural values in the prediction of technology-based consumer behaviors: A study in developing and emerging countries,” *Technol Forecast Soc Change*, vol. 149, Dec. 2019, doi: 10.1016/j.techfore.2019.119768.

[2] L. R. Men and S. Muralidharan, “Understanding Social Media Peer Communication and Organization-Public Relationships: Evidence from China and the United States,” *Journalism and Mass Communication Quarterly*, vol. 94, no. 1. SAGE Publications Inc., pp. 81–101, Mar. 01, 2017. doi: 10.1177/1077699016674187.

[3] S. Kemp, “Digital 2020: Global Digital Overview,” 2020.

[4] A. Punel and A. Ermagun, “Using Twitter network to detect market segments in the airline industry,” *J Air Transp Manag*, vol. 73, pp. 67–76, Oct. 2018, doi: 10.1016/j.jairtraman.2018.08.004.

[5] W. Budiharto and M. Meiliana, “Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis,” *J Big Data*, vol. 5, no. 1, Dec. 2018, doi: 10.1186/s40537-018-0164-1.

[6] E. J. Dommett, “Understanding student use of twitter and online forums in higher education,” *Educ Inf Technol (Dordr)*, vol. 24, no. 1, pp. 325–343, Jan. 2019, doi: 10.1007/s10639-018-9776-5.

[7] A. Benlahbib and E. H. Nfaoui, “MTVRep: A movie and TV show reputation system based on fine-grained sentiment and semantic analysis,” *International Journal of Electrical and Computer Engineering*, vol. 11, no. 2, pp. 1613–1626, Apr. 2021, doi: 10.11591/ijece.v11i2.pp1613-1626.

[8] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S. F. Chang, and M. Pantic, “A survey of multimodal sentiment analysis,” *Image Vis Comput*, vol. 65, pp. 3–14, Sep. 2017, doi: 10.1016/j.imavis.2017.08.003.

[9] E. Lim, E. I. Setiawan, and J. Santoso, “Stance Classification Post Kesehatan di Media Sosial Dengan FastText Embeddingdan Deep Learning,” *Journal of Intelligent System and Computation*, 2020.

[10] R. Jannati, R. Mahendra, C. W. Wardhana, and M. Adriani, “Stance Classification towards Political Figures on Blog Writing,” in *Proceedings of the 2018 International Conference on Asian Language Processing (IALP)* , 2018.

[11] K. Shalini, M. Anand Kumar, and K. Soman, “Deep-Learning-Based Stance Detection for Indian Social Media Text,” in *Emerging Research in Electronics, Computer Science and Technology*, V. Sridhar, M. C. Padma, and K. A. R. Rao, Eds., Singapore: Springer Singapore, 2019, pp. 57–67.

[12] N. I. M. Dawot and R. Ibrahim, “A review of features and functional building blocks of social media,” in *2014 8th. Malaysian Software Engineering Conference (MySEC)*, 2014, pp. 177–182. doi: 10.1109/MySec.2014.6986010.

[13] A. Jain and V. Jain, “Sentiment classification of twitter data belonging to renewable energy using machine learning,” *Journal of Information and Optimization Sciences*, vol. 40, no. 2, pp. 521–533, Feb. 2019, doi: 10.1080/02522667.2019.1582873.

[14] G. Sand, L. Tsitouras, G. Dimitrakopoulos, and V. Chatzigiannakis, “A big data aggregation, analysis and exploitation integrated platform for increasing social management intelligence,” in *2014 IEEE International Conference on Big Data (Big Data)*, 2014, pp. 40–47. doi: 10.1109/BigData.2014.7004411.

[15] E. Lunando and A. Purwarianti, *Indonesian Social Media Sentiment Analysis With Sarcasm Detection*. 2013. doi: 10.1109/ICACSIS.2013.6761575.

[16] T. Mangasi, A. Erwin, and H. P. Ipung, “Defined entity extraction based on Indonesian text document,” in *2014 International Conference on ICT For Smart Society (ICISS)*, 2014, pp. 61–65. doi: 10.1109/ICTSS.2014.7013152.

[17] I. Klampanos, “Introduction to information retrieval,” *Inf. Retr.*, vol. 12, pp. 609–612, Oct. 2009, doi: 10.1007/s10791-009-9096-x.

[18] B. Pang and L. Lee, “Opinion Mining and Sentiment Analysis,” *Foundations and Trends in Information Retrieval*, vol. 2, pp. 1–135, Jan. 2008, doi: 10.1561/1500000011.

[19] A. Sharma and S. Dey, “Performance Investigation of Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis,” *Special Issue of International Journal of Computer Applications*, pp. 975–8887, 2012.

[20] Y. Goldberg, “Neural Network Methods for Natural Language Processing,” *Synthesis Lectures on Human Language Technologies*, vol. 10, no. 1, pp. 1–311, 2017, doi: 10.2200/S00762ED1V01Y201703HLT037.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” Jan. 2013, [Online]. Available: http://arxiv.org/abs/1301.3781

[22] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” Oct. 2013, [Online]. Available: http://arxiv.org/abs/1310.4546

[23] S. Lai, K. Liu, S. He, and J. Zhao, “How to generate a good word embedding,” *IEEE*

Intell Syst, vol. 31, no. 6, pp. 5–14, Nov. 2016, doi: 10.1109/MIS.2016.45.

[24] Y. Zhu, E. Yan, and F. Wang, "Semantic relatedness and similarity of biomedical terms: Examining the effects of recency, size, and section of biomedical publications on the performance of word2vec," *BMC Med Inform Decis Mak*, vol. 17, no. 1, Jul. 2017, doi: 10.1186/s12911-017-0498-1.

[25] L. Ma and Y. Zhang, "Using Word2Vec to process big text data," in *Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015*, Institute of Electrical and Electronics Engineers Inc., Dec. 2015, pp. 2895–2897. doi: 10.1109/BigData.2015.7364114.

[26] C. T. Chao, W. H. Chu, C. L. Lee, J. K. Lee, M. Y. Hung, and H. W. Sung, "Devise Sparse Compression Schedulers to Enhance FastText Methods," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Aug. 2020. doi: 10.1145/3409390.3409394.

[27] P. Mojumder, M. Hasan, M. F. Hossain, and K. M. A. Hasan, "A study of fasttext word embedding effects in document classification in bangla language," in *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, Springer, 2020, pp. 441–453. doi: 10.1007/978-3-030-52856-0_35.

[28] B. Kuyumcu, C. Aksakalli, and S. Delil, "An automated new approach in fast text classification (fastText): A case study for Turkish text classification without pre-processing," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Jun. 2019, pp. 1–4. doi: 10.1145/3342827.3342828.

[29] A. Vaswani *et al.*, "Attention Is All You Need," Jun. 2017, [Online]. Available: http://arxiv.org/abs/1706.03762

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *ArXiv*, vol. abs/1810.04805, 2019.

[31] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," Aug. 2019, [Online]. Available: http://arxiv.org/abs/1908.10084

[32] W. Al-Ghaith, "Developing Lexicon-based Algorithms and Sentiment Lexicon for Sentiment Analysis of Saudi Dialect Tweets," 2019. [Online]. Available: www.ijacsa.thesai.org

[33] W. Xiu-Shen, J. Wu, and Q. Cui, "Deep Learning for Fine-Grained Image Analysis: A Survey," *arXiv.org*, Jul. 2019, [Online]. Available: https://www.proquest.com/working-papers/deep-learning-fine-grained-image-analysis-survey/docview/2254222037/se-2?accountid=17242

[34] F. J. Yang, "An extended idea about decision trees," in *Proceedings - 6th Annual Conference on Computational Science and Computational Intelligence, CSCI 2019*, Institute of Electrical and Electronics Engineers Inc., Dec. 2019, pp. 349–354. doi: 10.1109/CSCI49370.2019.00068.

[35] S. A. Alquhtani and A. Muniasamy, "Analytics in Support of E-Commerce Systems Using Machine Learning," in *International Conference on Electrical, Computer, and Energy Technologies, ICECET 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICECET55527.2022.9872592.

[36] R. Burbidge and B. Buxton, "An Introduction to Support Vector Machines for Data Mining."

[37] Y. Wang, Y. Ou, X. Deng, L. Zhao, and C. Zhang, *The Ship Collision Accidents Based on Logistic Regression and Big Data*. 2019.

[38] X. Chen and R. Ye, "Identification model of logistic regression analysis on listed firms' frauds in China," in *Proceedings - 2009 2nd International Workshop on Knowledge Discovery and Data Mining, WKKD 2009*, 2009, pp. 385–388. doi: 10.1109/WKDD.2009.35.

[39] D. Li, Q. Ge, P. Zhang, Y. Xing, Z. Yang, and W. Nai, "Ridge Regression with High Order Truncated Gradient Descent Method," in *Proceedings - 2020 12th International Conference on Intelligent Human-Machine Systems and Cybernetics, IHMSC 2020*, Institute of Electrical and Electronics Engineers Inc., Aug. 2020, pp. 252–255. doi: 10.1109/IHMSC49165.2020.00063.