JUSTINDO
Jurnal Sistem dan Teknologi Informasi
http://jurnal.unmuhjember.ac.id/index.php/JUSTINDO

E-ISSN: 25541-5735, P-ISSN: 2502-5724
Volume 6, No. 2, Agustus 2021

STB Sentiment Analysis Classification Multiclass Modeling Ratio With SGDC Tuning And Calibrated Classifier With SGDC Tuning As Basis

M. Hafidz Ariansyah[1], Sri Winarno[2], Abu Salam[3]

[1,2]Department of Information System, Dian Nuswantoro University
[3]Department of Information Technology, Dian Nuswantoro University
Email: [1]112201906146@mhs.dinus.ac.id, [2]sri.winarno@dsn.dinus.ac.id, [3]abu.salam@dsn.dinus.ac.id

(Naskah masuk: dd mmm yyyy, diterima untuk diterbitkan: dd mmm yyyy)

ABSTRAK

Set-top box (STB) adalah perangkat yang mengubah sinyal digital menjadi gambar dan suara yang dapat kita lihat di televisi (TV) analog biasa. Semakin banyak orang mencari STB akhir-akhir ini, karena penggunaan televisi (TV) analog biasa akan berakhir pada bulan Desember di tahun 2022. Penghentian penggunaan normal televisi (TV) analog telah menimbulkan banyak sentimen positif, netral dan negatif. Data sentimen ini diperoleh dari media sosial Twitter dengan menggunakan teknik crawling data. Proses ekstraksi fitur pada penelitian ini menggunakan metode TF-IDF. Pada penelitian ini *Stochastic Gradient Descent Classifier* (SGDC) digunakan sebagai dasar penentuan metode optimal dan membandingkan metode *tuning* SGDC dengan metode *Calibrated Classifier*. Hasil pengujian menunjukkan bahwa optimasi terbaik untuk model ini adalah metode *Calibrated Classifier* dengan SGDC sebagai basisnya dengan nilai akurasi 80% pada data uji dan 100% pada data pelatihan. Hal ini menunjukkan bahwa metode *Calibrated Classifier* dapat meningkatkan sedikit performa pengujian dan pelatihan pada SGDC *Classifier* yang hanya memiliki nilai akurasi sebesar 78,4% pada data uji dan 100% pada data pelatihan.

Kata kunci: *analisis sentimen, calibrated classifier, set-top box, SGDC, classification*

ABSTRACT

A set-top box (STB) is a device that converts digital signals into images and sounds that we can see on a regular analog television (TV). More and more people are looking for STB these days because the ordinary use of analog television (TV) will end in December 2022. The discontinuation of the normal use of analog television (TV) has generated a lot of positive, neutral, and negative sentiments. This sentiment data was obtained from social media Twitter using crawling data techniques. The feature extraction process in this study uses the TF-IDF method. In this study Stochastic Gradient Descent Classifier (SGDC) is used as a basis for determining the optimal method and comparing the SGDC tuning method with the Calibrated Classifier method. The test results show that the best optimization for this model is the Calibrated Classifier method with SGDC as its basis with an accuracy value of 80% on the test data and 100% on the training data. It shows that the Calibrated Classifier method can slightly improve the performance of testing and training on the SGDC Classifier has an accuracy value of 78.4% on the test data and 100% on the training data.

Keywords: *calibrated classifier, classification, set-top box, SGDC, sentiment analysis*

1. INTRODUCTION

STB is the main component in the transition of television (TV) technology from analog to digital (Kominfo, 2022). STB allows ordinary people to watch digital TV with better quality than analog. In other words, STB is an information technology device whose main components are processor and memory chips. Its main job is to convert digital signals into analog signals. The STB allows an analog TV to read the digital signal picked up by the antenna.

Statute (UU) Number 11 of 2020 concerning Job Creation Contracts to eliminate Analog Broadcasting and switch to Digital Broadcasting (Article 72 Paragraph 8 of the Broadcasting Law, insert Article 60A) (UU RI, 2020). For this reason, the role of STB is important. As a device, the STB can also provide additional features such as disaster information. Transmission can also be adjusted according to the age of the audience, so parents can control what their children see.

The growth of social media, especially Twitter, continues to increase every time (Darwis et al, 2021). Twitter users can use this platform to convey information in the form of critical comments or suggestions to an agency or a policy (Ruz et al, 2020). Twitter users will provide the latest news or comments about things that are currently the main topic in the world. If many users comment on something, it will cause a problem or a trending topic on Twitter.

The scope of this research includes community comments about the switch from using ordinary analog TV to digital TV on social media Twitter. Stopping the use of regular analog TVs has drawn many comments from the public, ranging from compliments, and suggestions, to criticism of this policy. The researcher collects this comment data using the crawling data method using the API available on the Twitter Developer platform (Sharma & Ghose, 2020). Then, data will be processed using a document preprocessing process. This process can generate positive, negative, and neutral class sentiments. Several techniques in text mining, among others, handle classification, clustering, information extraction, and information retrieval problems. In the next stage, the researcher will use the classification method and process the data using the SGDC tuning method as the basis for determining the optimal method and compare it with the SGDC tuning method with the Calibrated Classifier.

From the results of research conducted by Shanto Moyrano Tambunan et al. (2021), SGDC has the best accuracy, recall, and F-1, namely 86.41%, 89.36%, and 84.39%, while the best results on precision are obtained from the Random Forest algorithm, which is 86.95%. Meanwhile, for the comparison of average accuracy, recall, and the best F-1 are found in the SGD algorithm with successive results of 84.92%, 85.05%, and 82.42% while the highest precision average comparison is obtained by the Naive Bayes algorithm of 82.95%. The research shows that SGDC is the best algorithm for text classification. In research conducted by Ramaditia et al (2021), the TF-IDF and SGDC methods were used with two loss function methods, namely logistic (log) and smooth hinge lost to detect SMS spam in Indonesian. Based on the research results, the TF-IDF and Stochastic Gradient Descent Classifier methods produce a quite good performance, namely the highest in terms of accuracy of 97% with a Recall value of 97.2% and a Precision of 96.9% for recognizing SMS spam compared to the method used in previous studies namely SVM and NBC.

Based on several previous studies regarding the SGDC method that produces good performance for text classification, this study will use the SGDC method and combine it with the TF-IDF method to carry out text classification modeling. Then the model will be calibrated using the Calibrated Classifier to get better performance. The results of the two algorithms will be compared, and the best performance will be sought so that this model can make the best predictions on sentiment.

2. RESEARCH METHOD

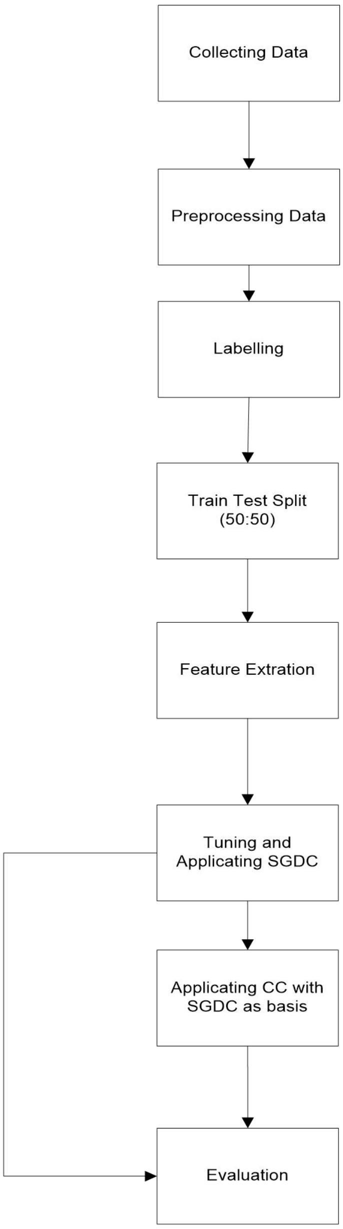


Figure 1. Research Method

To uncover patterns or opportunities in various datasets, this research uses data mining approaches, feature extraction, and classification. So this is a research model for predicting sentiment labels. The best performance is chosen by comparing SGDC and calibrated classifiers. As shown in Figure 1, the phases of this research are data acquisition, data preprocessing, labeling, splitting the dataset into training and testing, feature extraction, implementing the SGDC algorithm, implementing the calibrated classifier algorithm, and evaluation.

2.1. Data Collection

Twitter mining is the hottest topic these days as it provides substantial information to be used and applied in various fields (Sharma & Ghose, 2020). It is one of the major research areas. Various tweets can be collected using several public APIs and analyzed for research purposes. An authenticated request establishes the Twitter API.

The dataset used in this study is in the form of data tweets containing the phrases 'STB' and 'Digital TV' amounted to 294 tweets. This tweet data consists of 136 tweets with positive sentiments, 79 with neutral, and 78 with negative. Labeling of successfully withdrawn tweets is done by utilizing the Text Blob library. TextBlob is a library that is used to process textual data. This library provides a simple API for diving into Natural Language Processing(NLP) tasks (Fauziyyah, 2020). Examples include tagging words, extracting nouns, sentiment analysis, classification, translation, and so on. In this research, TextBlob is used for sentiment analysis. The sentiment analysis model in TextBlob is only available in English, so users must translate it into English if they want to use TextBlob.

2.2. Preprocessing Data

Preprocessing is the first step in text sentiment analysis (Hafidz & Liliana, 2021). Using a good preprocessing technique also improves the performance of classifier models (Symeonidis et al, 2018). Preprocessing techniques used in this study included URLs, mentions, hashtag and punctuation removal, case sensitivity, tokenization, stop word filtering, and stemming.

2.2.1. Removing URL, Mention, Hashtag, and Punctuation

In this process, URLs, mentions, hashtags and punctuation in the tweet text are removed (Hafidz & Liliana, 2021). Example of a tweet and its processed text:
Tweet :
RT @lanyallacenter_: Ketua DPD RI Terima Aspirasi Keterbatasan Perangkat Monitoring Siaran TV Digital dari KPID Jatim #LaNyalla #jawatimur.
Result :
Ketua DPD RI Terima Aspirasi Keterbatasan Perangkat Monitoring Siaran TV Digital dari KPID Jatim.

2.2.2. Lowercasing

Lowercasing is a method for converting all words in a text into lowercase words (Salam et al, 2018). By doing this lowercasing, the same words will merge and reduce the dimensions of the problem (Hafidz & Liliana, 2021). Lowercase example:
Tweet :
Ketua DPD RI Terima Aspirasi Keterbatasan Perangkat Monitoring Siaran TV Digital dari KPID Jatim.
Result :
ketua dpd ri terima aspirasi keterbatasan perangkat monitoring siaran tv digital dari kpid jatim.

2.2.3. Tokenizing

Tokenization is the process of slicing the input string according to the terms that compose it and distinguishing exclusive characters that can be treated as word separators or not (Mostafa et al, 2021). Tokenization example :
{“ketua”, “dpd”, “ri”, “terima”, “aspirasi”, “keterbatasan”, “perangkat”, “monitoring”, “siaran”, “tv”, “digital”, “dari”, “kpid”, “jatim”}

2.2.4. Stopwords Filtering

Stopwords are a list of words that are considered meaningless (Kalaivani & Marivendan, 2021; Mulyani et al, 2021). Words listed in this list are discarded and are not processed at a later stage. Words included in stopwords are generally words that often appear in every document, so these words cannot be used as identifiers. Stopwords example :
{“ketua”, “dpd”, “ri”, “terima”, “aspirasi”, “keterbatasan”, “perangkat”, “monitoring”, “siaran”, “tv”, “digital”, “kpid”, “jatim”}

2.2.5. Stemming

Stemming is the process of removing the inflection of a word into its basic form, but the basic shape does not mean the same as the root word (Ruzkanda, 2019; Kalaivani & Marivendan, 2021). In terms of efficiency stemming, it aims to reduce the number of unique words in the index thereby reducing the need for storage space for the index and speeding up the search process. Stemming example :

2.3. Feature Extration

The Term Frequency Inverse Document Frequency (TF-IDF) is a method used to determine how far a word (term) is related to a document by giving terms a weight. The TF-IDF method combines two concepts, the frequency of occurrence of a word in a document and the inverse frequency of documents containing that word [6]. In calculating the weight using TF-IDF, the TF value per word is calculated with the heft of the term being 1.

Meanwhile, the IDF value is formulated in Equation (1) (Deolika et al, 2019).

$$IDF(word) = \log \frac{tf}{df} \quad (1)$$

Explanation :
td = the total number of documents that exists
df = the number of occurrences of the word in all documents.

2.4. SGD Classifier

SGDC is a simple and efficient approach to classifying linearly using discriminatory learning. This method is an iterative (re)optimization algorithm that is useful for finding the minimum function point that can be derived (Sowmya, 2020; Tambunan et al, 2021). At the beginning of the algorithm, the process begins by making guesses. Guessing errors are corrected during repeated guessing using the gradient (derivative) rule of the function to be minimized. SGD can learn faster in conducting classification training (Admojo & Sulistya, 2022).

$$\theta j = \theta j - \alpha \frac{\partial y}{\partial x} j(\theta) \quad (2)$$

$$j\theta = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x)) + \alpha R(W) \quad (3)$$

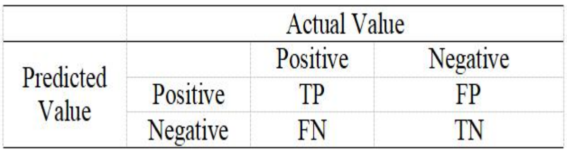
The process of the SGD algorithm is to find the value θ that can minimize the function J(θ). In determining the initial value of θ, a search algorithm is used, then in each iteration, the value of θ is continuously updated until it finds the minimum point or the minimum J value. The process of updating the value of θ at each iteration uses Equation (2). Updates are performed simultaneously for all j values = 0, ..., n. Variable α is the learning rate that regulates how much the value renewal is. The equation for the value of J(θ) can be seen in Equation (3), where L is the loss function used in the training data (x_1, y_1)...(x_n, y_x), and R is the regularization or penalty for model complexity (Admojo & Sulistya, 2022).

2.5. Calibrated Classifier

This classifier uses cross-validation to estimate classifier parameters and tune the classifier. Using the default ensemble=True, for each CV fold, copy the base estimator to fit the training subset and adjust it to the test subset. For predictions, the prediction probabilities are averaged across these calibrated individual classifiers (Pedregosa et al, 2011). Researchers use the calibration module to calibrate the probabilities of certain models or add support for probabilistic forecasting. A properly trained classifier is a probabilistic classifier whose output from the predict_proba method can be directly interpreted as a confidence level.

2.6. Confusion Matrix

The performance metrics used to evaluate the performance of the classifier model in this study are accuracy, precision, and recall. Accuracy is the comparison between the correct prediction and the overall prediction (4). Precision is the proportion of positive predictions that are correct from all positives (5). Recall or true positive rate is the known fraction of positives that are predicted correctly (6). Figure 2 is an overview of the Confusion Matrix (Xu et al, 2020).



		Actual Value	
Predicted Value		Positive	Negative
	Positive	TP	FP
	Negative	FN	TN

Figure 2. Confusion Matrix

$$Accuracy = \frac{(\mathrm{TP+TN})}{(\mathrm{TP+FP+FN+TN})} \quad (4)$$

$$Precision = \frac{(\mathrm{TP})}{(\mathrm{TP+FP})} \quad (5)$$

$$Recall = \frac{(\text{TP})}{(\text{TP+FN})} \qquad (6)$$

3. RESULT AND DISCUSSION

3.1. Preprocessing and Labelling

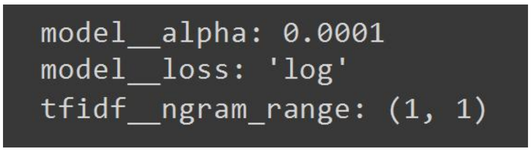
Researchers get raw data from Twitter which is still dirty. There are Twitter links, mentions, hashtags, RT posts, weird characters, emoticons, and others. These words certainly have no meaning if sentiment analysis is carried out, and will affect the performance of the accuracy of sentiment analysis. Data cleansing is required to clean up the words. Then the researcher also did some data preprocessing such as lowercasing, tokenizing, stopwords filtering, and stemming. The example of the results can be seen in Table 1.

Tabel 1. Preprocessing Result

Tweets	Mantap selamat menikmati siaran TV digital bagi masyarakat yang sudah mendapatkan STB gratis dari Kemkominfo.
Cleaning Tweets	mantap selamat menikmati siaran tv digital bagi masyarakat yang sudah mendapatkan allcaps stb allcaps gratis dari kemkominfo
Tokenization	['mantap', 'selamat', 'menikmati', 'siaran', 'tv', 'digital', 'bagi', 'masyarakat', 'yang', 'sudah', 'mendapatkan', 'allcaps', 'stb', 'allcaps', 'gratis', 'dari', 'kemkominfo']
Repaired Token	['mantap', 'selamat', 'menikmati', 'siaran', 'tv', 'digital', 'bagi', 'masyarakat', 'yang', 'sudah', 'mendapatkan', 'allcaps', 'stb', 'allcaps', 'gratis', 'dari', 'kemkominfo']
Stop Word Removal	['mantap', 'selamat', 'menikmati', 'siaran', 'tv', 'digital', 'masyarakat', 'allcaps', 'stb', 'allcaps', 'gratis', 'kemkominfo']
Merging Token	mantap selamat menikmati siaran tv digital masyarakat allcaps stb allcaps gratis kemkominfo
Stemming	mantap selamat nikmat siar tv digital masyarakat allcaps stb allcaps gratis kemkominfo
Subjectivity	0.5125
Polarity	0.4
Label	Positive (2)

3.2. TF-IDF & SGDC Tuning

The researcher does this step to determine the selected parameters that are used to find the optimal combination. The parameters used in this research are alpha, loss, and ngram_range for TF-IDF. In the alpha parameter, the higher the value, the stronger the regularization. This parameter can calculate learning rate when it is set to learning_rate is set to 'optimal'. Figure 3 is the best parameter of this modeling.



```
model__alpha: 0.0001
model__loss: 'log'
tfidf__ngram_range: (1, 1)
```

Figure 3. Model Best Parameters

3.3. TF-IDF

The TF-IDF weighting is carried out based on model adjustment in the previous stage, namely the range (1,1). At this stage, each word will be weighted according to how important the word is in the sentence. These words can represent positive, neutral, and even negative sentiments. Of course, this will be very influential when modeling with a classifier because these words have a big influence on the model. Figures 4, 5, and 6 are visualizations of TF-IDF for positive, neutral, and negative sentiments.

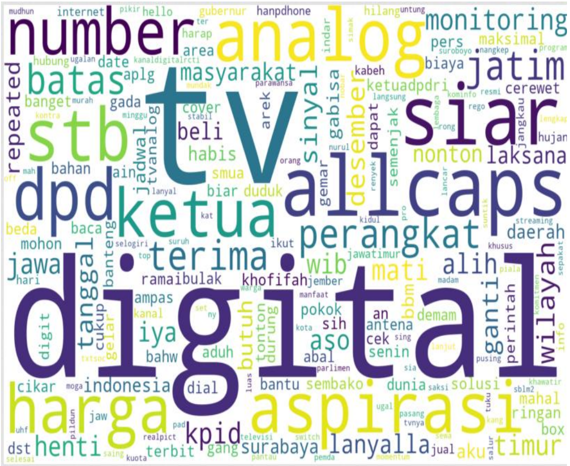


Figure 4. Neutral TF-IDF

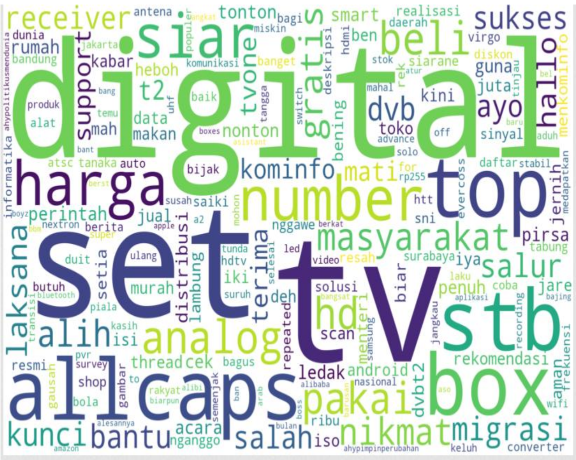


Figure 5. Positive TF-IDF

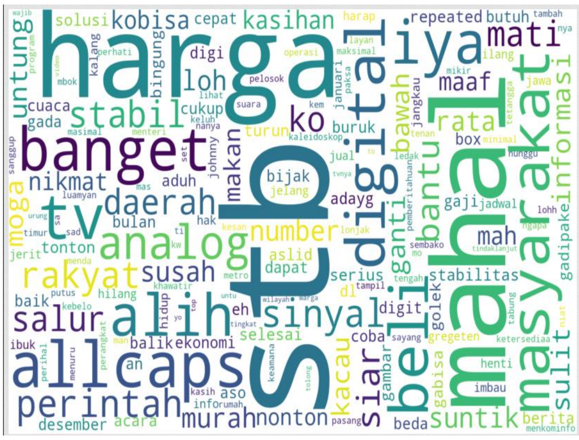


Figure 6. Negative TF-IDF

3.4. Calibrated Classifier

At this stage, the researcher uses the Calibrated Classifier with the SGDC algorithm that has been set for modeling. Researchers used cross-validation with a value of K = 10 and the sigmoid method to carry out the training process. The sigmoid method is used because this method is the best parameter for modeling this time.

3.5. Evaluation

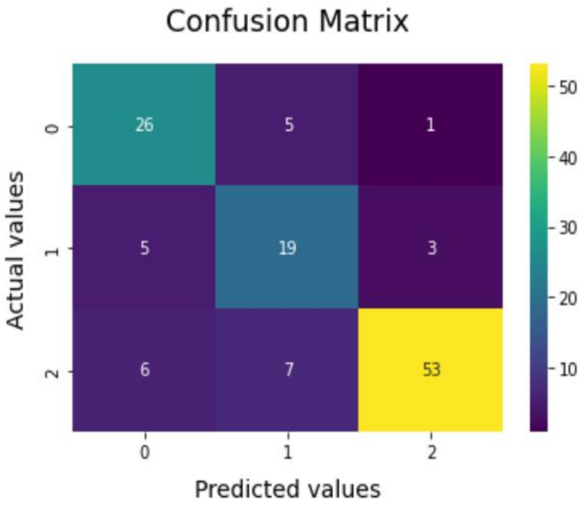


Figure 7. SGDC Evaluation

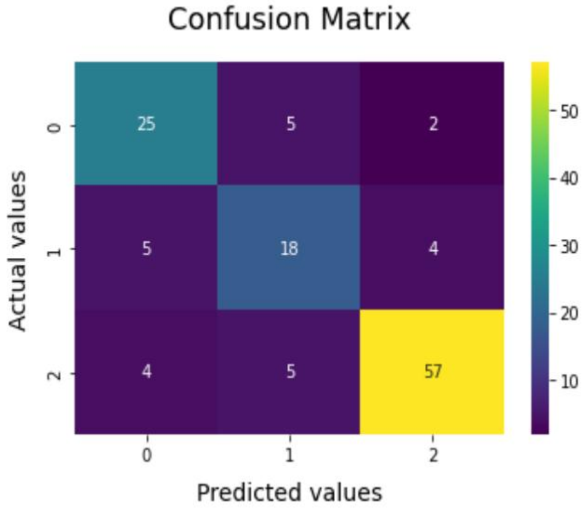
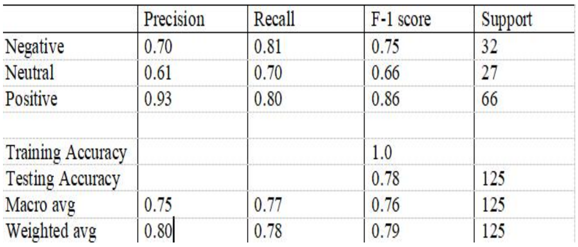


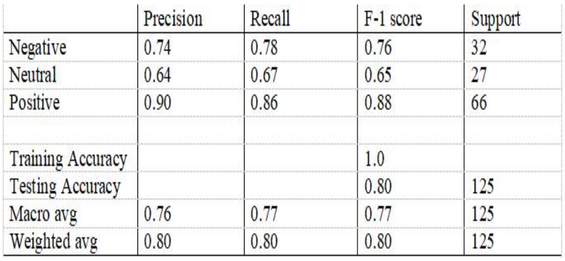
Figure 8. Calibrated Classifier Evaluation

Figures 7 and 8 show the resulting performance differences between the two models. Then Figures 7 and 8 will be analyzed more deeply with accuracy, precision, recall, and F-1 scores obtained from the scikit learn library, namely classficication_report which can be seen in Figures 9 and 10.



	Precision	Recall	F-1 score	Support
Negative	0.70	0.81	0.75	32
Neutral	0.61	0.70	0.66	27
Positive	0.93	0.80	0.86	66
Training Accuracy			1.0	
Testing Accuracy			0.78	125
Macro avg	0.75	0.77	0.76	125
Weighted avg	0.80	0.78	0.79	125

Figure 9. SGDC Report



	Precision	Recall	F-1 score	Support
Negative	0.74	0.78	0.76	32
Neutral	0.64	0.67	0.65	27
Positive	0.90	0.86	0.88	66
Training Accuracy			1.0	
Testing Accuracy			0.80	125
Macro avg	0.76	0.77	0.77	125
Weighted avg	0.80	0.80	0.80	125

Figure 10. Calibrated Classifier Report

The performance accuracy comparison results from Figures 9 and 10 will be presented in Table 2. In Table 2, it can be seen that for training accuracy, the two models can carry out the training process well. However, during testing, the Calibrated Classifier can make the SGDC model that has been tuned better by increasing the accuracy by 1.6%.

Tabel 2. Comparating Performance

Classifier	Acc Training	Acc Testing
SGDC Tuning	100%	78,6%
Calibrated Classifier	100%	80%

4. CONCLUSION

The conclusion that can be drawn from this paper is the result of sentiment analysis about the STB, many consider this as something positive. The neutral sentiment here may indicate that the use of STB is still being questioned by the public. The positive sentiment shows that many people are happy with the STB and the switch to Digital TV because it looks better and there are more and more TV channels. Negative sentiment shows that this transitional policy still has deficiencies and still needs to be improved. What can be drawn from this analysis is that the government needs to pay more attention to every tweet that contains negative sentiments for future evaluation of Digital TV policies. From the model side, the Calibrated Classifier model based on SGDC tuning has better accuracy than using only SGDC tuning. This is evidenced by the accuracy distance of 1.6%. The calibration module allows the model to calibrate the probability of a particular model better or add support for probability prediction.

DAFTAR PUSTAKA

Admojo, F. T., & Sulistya, Y. I. 2022. Analisis performa algoritma Stochastic Gradient Descent (SGD) dalam mengklasifikasi tahu berformalin. Indonesian Journal of Data and Science, 3(1), pp. 1-8.

Darwis, D., Siskawati, N., & Abidin, Z. 2021. Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Review Data Twitter Bmkg Nasional. Jurnal Tekno Kompak, 15(1), pp. 131-145.

Deolika, A., Kusrini, K., & Luthfi, E. T. 2019. Analisis Pembobotan Kata Pada Klasifikasi Text Mining. (JurTI) Jurnal Teknologi Informasi, 3(2), pp. 179-184.

Fauziyyah, A. K. 2020. Analisis sentimen pandemi Covid19 pada streaming Twitter dengan text mining Python. Jurnal Ilmiah SINUS, 18(2), pp. 31-42.

Hafidz, N., & Liliana, D. Y. 2021. Klasifikasi Sentimen pada Twitter Terhadap WHO Terkait Covid-19 Menggunakan SVM, N-Gram, PSO. Jurnal RESTI (Rekayasa Sistem

Dan Teknologi Informasi), 5(2), pp. 213-219.

Hassan, N., Gomaa, W., Khoriba, G., & Haggag, M. 2020. Credibility detection in twitter using word n-gram analysis and supervised machine learning techniques. International Journal of Intelligent Engineering and Systems, 13(1), pp. 291-300.

Kalaivani, E. R., & Marivendan, E. R. 2021. The Effect of Stop Word Removal and Stemming In Datapreprocessing. Annals of the Romanian Society for Cell Biology, 25(6), pp. 739-746.

Kominfo (Kementrian Komunikasi dan Informatika Republik Indonesia), 2022. Beralih ke Teknologi TV Digital, Apakah TV di Rumah Perlu Diganti?. https://www.kominfo.go.id/content/detail/34888/beralih-ke-teknologi-tv-digital-apakah-tv-di-rumah-perlu-diganti/0/tv_digital [Accessed in 15 Desember 2022].

Mostafa, G., Ahmed, I., & Junayed, M. S. 2021. Investigation of different machine learning algorithms to determine human sentiment using Twitter data. International Journal of Information Technology and Computer Science, 13(2), pp. 38-48.

Mulyani, E., Muhamad, F. P. B., & Cahyanto, K. A. 2021. Pengaruh N-Gram terhadap Klasifikasi Buku menggunakan Ekstraksi dan Seleksi Fitur pada Multinomial Naïve Bayes. Jurnal Media Informatika Budidarma, 5(1), pp. 264-272.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, pp. 2825-2830.

Ruskanda, F. Z. 2019. Study on the effect of preprocessing methods for spam email detection. Indonesia Journal on Computing (Indo-JC), 4(1), pp. 109-118.

Ruz, G. A., Henríquez, P. A., & Mascareño, A. 2020. Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers. Future Generation Computer Systems, 106, pp. 92-104.

Salam, A., Zeniarja, J., & Khasanah, R. S. U. 2018. Analisis Sentimen Data Komentar Sosial Media Facebook Dengan k-Nearest Neighbor (Studi Kasus Pada Akun Jasa Ekspedisi Barang J&T Ekspress Indonesia). Proceeding of SINTAK, pp. 480 - 486.

Sharma, A., & Ghose, U. (2020). Sentimental analysis of twitter data with respect to general elections in india. Procedia Computer Science, 173, pp. 325-334.

Sowmya, B. J., Nikhil Jain, C. S., Seema, S., & KG, S. 2020. Fake News Detection using LSTM Neural Network Augmented with SGD Classifier. Solid State Technology, 63(6), pp. 6985-9665.

Symeonidis, S., Effrosynidis, D., & Arampatzis, A. 2018. A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. Expert Systems with Applications, 110, pp. 298-310.

Tambunan, S. M., Nataliani, Y., & Lestari, E. S. 2021. Perbandingan Klasifikasi dengan Pendekatan Pembelajaran Mesin untuk Mengidentifikasi Tweet Hoaks di

Media Sosial Twitter. JEPIN (Jurnal Edukasi dan Penelitian Informatika), 7(2), pp. 112-120.

Undang-undang Republik Indonesia nomor 11 tahun 2020 tentang Cipta Kerja. Jakarta: Kementerian Sekretariat Negara Republik Indonesia.

Xu, J., Zhang, Y., & Miao, D. 2020. Three-way confusion matrix for classification: A measure driven view. Information sciences, 507, pp. 772-794.