International Journal of Computer and Information System (IJCIS)
Peer Reviewed – International Journal
Vol : Vol. 01, Issue 01, May 2020
e-ISSN : 2745-9659
https://ijcis.net/index.php/ijcis/index

STB Sentiment Analysis Classification Multiclass Modeling Using Calibrated Classifier With SGDC Tuning As Basis and Sigmoid Method

1st M. Hafidz Ariansyah, 2nd Sri Winarno, 3rd Abu Salam
[1,2]Department of System Information
[3]Department of Information Technology
[1,2,3]Dian Nuswantoro University
Semarang, Indonesia
e-mail: *[1]112201906146@mhs.dinus.ac.id, [2]sri.winarno@dsn.dinus.ac.id, [3]abu.salam@dsn.dinus.ac.id

Abstract*—A set-top box (STB) is a device that converts digital signals into the picture and sound we see on a regular analog television (TV). More and more people are looking for STB these days as the normal 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, the 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 researcher chose the SGDC algorithm as the basis because this algorithm can train and test text datasets with good performance for the classification process. The test results show that the best optimization for this model is the SGDC-based Calibrated Classifier method with an accuracy value of 80% on the test data. It shows that the Calibrated Classifier method can slightly improve test performance on the SGDC Classifier with an accuracy value of 78.4% on the test data.

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

I. INTRODUCTION

STB is the main component in the transition of television (TV) technology from analog to digital [1]. 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 processors 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. The sentiment of switching from using analog TV to using STB reaps many pros and cons from the public. In this paper, researchers will classify sentiments about this transition so that the government can predict public sentiments so that this policy can be assessed in the end.

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) [2]. 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 [3]. Twitter users can use this platform to convey information in the form of critical comments or suggestions to an agency or a policy [4]. 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 [5]. 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 [6], 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 [7], 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 which produces good performance for text classification. This research 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 best performance will be sought so that this model can provide the best prediction of sentiment so that the government can use this sentiment as an assessment of the policy of switching to Digital TV using STB.

II. RESEARCH METHODS

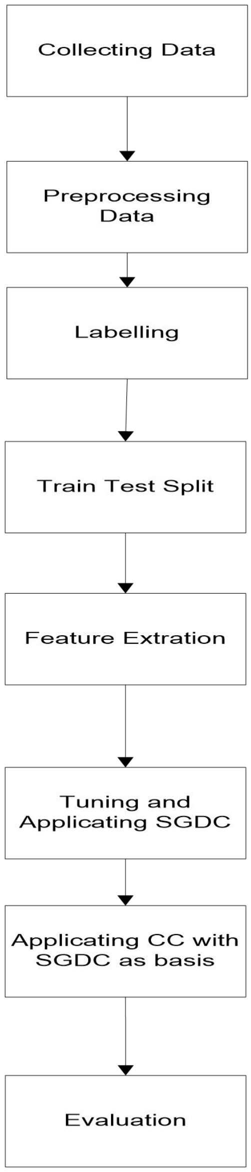

Figure 1. Research Method

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. Collecting Data

Twitter mining is the hottest topic these days as it provides substantial information to be used and applied in various fields [5]. 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 [8]. 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 data is the first step in text sentiment analysis [9]. Using a good preprocessing technique also improves the performance of classifier models [10]. 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 [9]. 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 [11]. By doing this lowercasing, the same words will merge and reduce the dimensions of the problem [9]. 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 [12]. 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 [13,14]. 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 [14,15]. 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.

2.3. TF-IDF

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 [7]. 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)[16].

$$IDF(word) = \log\frac{tf}{df} \text{ …………….. (1)}$$

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 [17,18]. 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 [12].

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

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

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 [18].

2.5. Calibrated Classifier

Calibrated classifier refers to the scale on which classifier scores are reported. The classifier ultimately assigns instances to different classes, but it is useful to decompose this assignment into an evaluation classifier which produces one or more real numbers and a decision rule which transforms these numbers into predictive classes [19]. For example, a linear classifier can generate a positive or negative score that is proportional to the distance between the instances and the decision boundary. In this case, the decision rule will be a simple threshold on that score. The advantage of fitting these values to a known domain-independent scale is that the decision rule is also in domain-independent form and does not need to be investigated.

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 [20]. 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 [21,22,23]. Accuracy (4) is the comparison between the correct prediction and the overall prediction. Precision (5) is the proportion of positive predictions that are correct from all positives. Recall (6) or true positive rate is the known fraction of positives that are predicted correctly. Figure 2 is an overview of the Confusion Matrix [21,24,25].

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

Figure 2. Confusion Matrix

$$Accuracy = \frac{(\text{TP+TN})}{(\text{TP+FP+FN+TN})}\text{…………… (4)}$$

International Journal of Computer and Information System (IJCIS)
Peer Reviewed – International Journal
Vol : Vol. 01, Issue 01, May 2020
e-ISSN : 2745-9659
https://ijcis.net/index.php/ijcis/index

$$Precision = \frac{(TP)}{(TP+FP)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(5)}$$

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

III. RESULT AND ANALYSIS

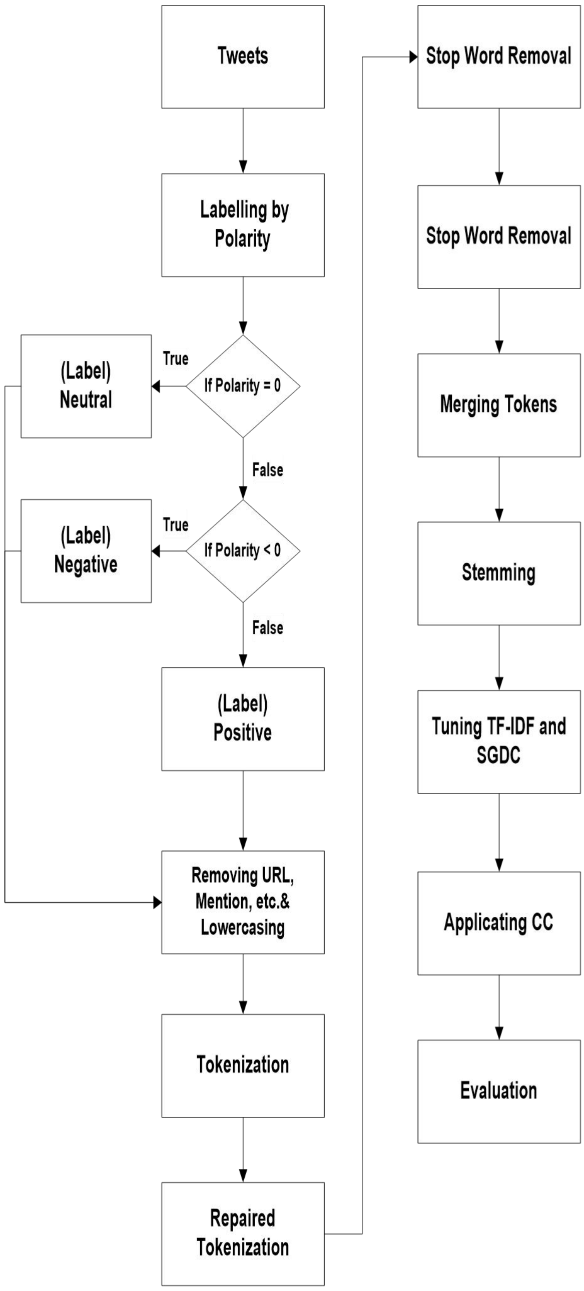

Figure 3. Classification Flow Chart

Figure 3 shows the running process and the results will be explained in this section of the paper. In Figure 3, 12 processes will be carried out starting from the labeling stage to evaluating the classification using the confusion matrix from the scikit-learn library.

3.1 Labeling and Preprocessing Data

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

Features	Values
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
Polarity	0.4
Label	Positive (2)

Tweets source = https://developer.twitter.com/en

3.2 TF-IDF and 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 4 shows the best parameter of this modeling.

International Journal of Computer and Information System (IJCIS)
Peer Reviewed – International Journal
Vol : Vol. 01, Issue 01, May 2020
e-ISSN : 2745-9659
https://ijcis.net/index.php/ijcis/index

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

Figure 4. 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 5, 6, and 7 are visualizations of TF-IDF for positive, neutral, and negative sentiments.

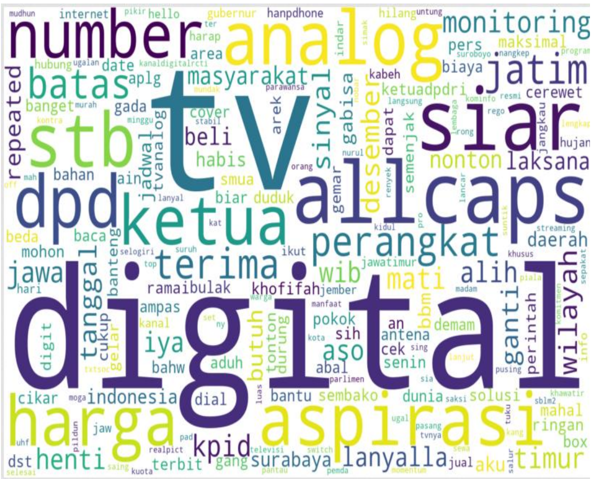

Figure 5. Neutral TF-IDF

Figure 6. Positive TF-IDF

Figure 7. 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 isotonic method to conduct training on training data. The isotonic method is used because this method is the best parameter for modeling this time.

3.5 Evaluation

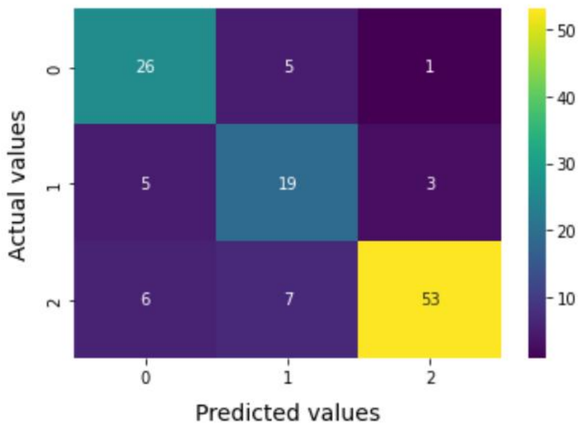

Figure 8. SGDC Evaluation

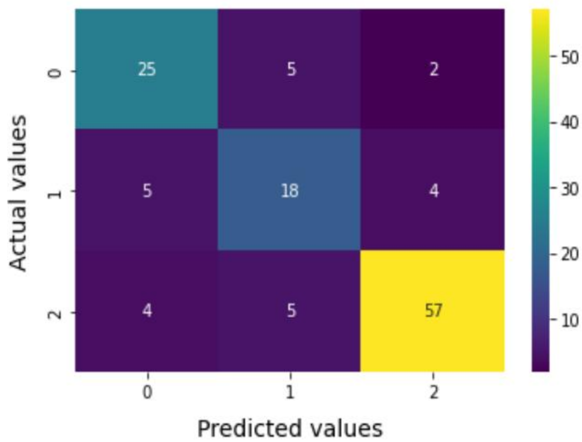

Figure 9. Calibrated Classifier Evaluation

International Journal of Computer and Information System (IJCIS)
Peer Reviewed – International Journal
Vol : Vol. 01, Issue 01, May 2020
e-ISSN : 2745-9659
https://ijcis.net/index.php/ijcis/index

Figures 8 and 9 show the resulting performance differences between the two models. Then Figures 8 and 9 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 10 and 11.

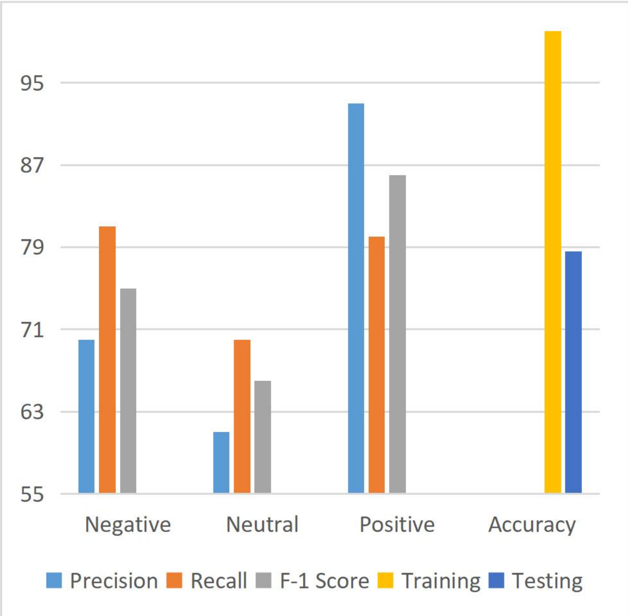

Figure 10. SGDC Report

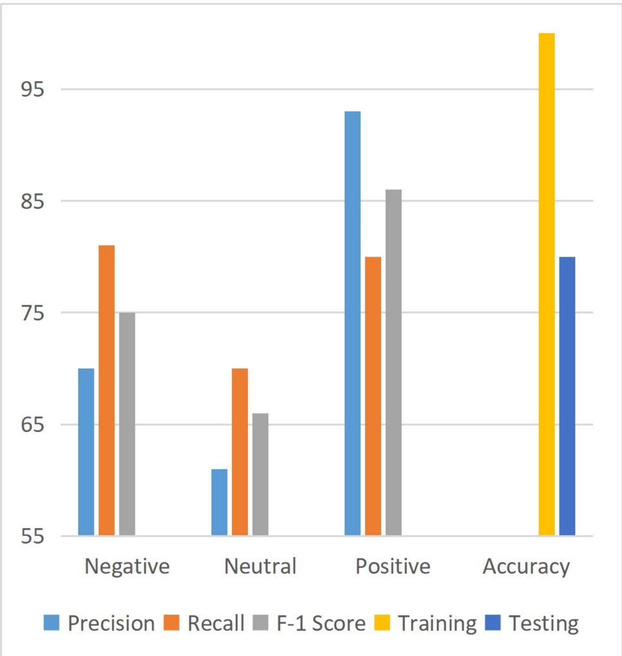

Figure 11. CC Report

The performance accuracy comparison results from Figures 10 and 11 will be presented in Figure 12. In Figure 12, it can be seen that for training accuracy, the two models can carry out the training process well. During testing, the Calibrated Classifier can make the SGDC model that has been tuned better by increasing the accuracy by 1.6% from 78,4% to 80%..

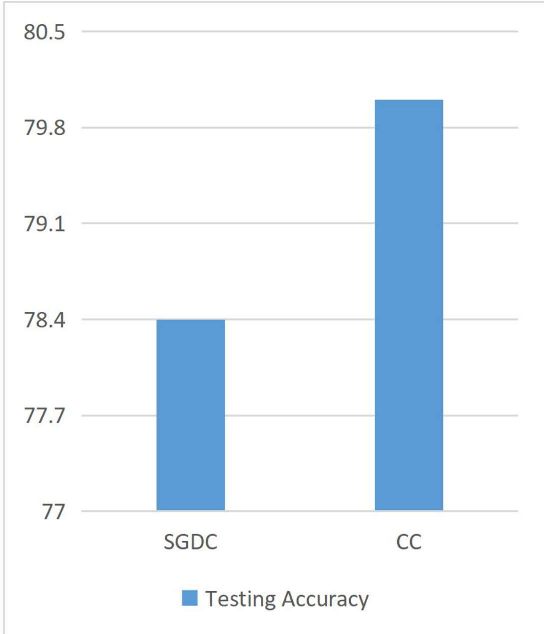

Figure 12. Accuracy Comparison Report

VI. CONCLUSION

The conclusion of this paper is the result of sentiment analysis of the STB, many of which view this as something positive. The neutral sentiment here may indicate that the public is still questioning the use of STB. The positive sentiment shows that many people are happy with STB and changes to Digital TV because it looks so good and there are more TV channels than before. Negative sentiment shows that this transition 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 any tweets that contain negative sentiments for future Digital TV policy evaluation. From the model side, the Calibrated Classifier model based on SGDC tuning has better accuracy than using only SGDC tuning. It is evidenced by the distance accuracy of 1.6% from 78.4% to 80%. The calibration module allows models to better calibrate specific model probabilities or add support for probability predictions.

REFERENCES

[1] Kominfo (Kementrian Komunikasi dan Informatika Republik Indonesia), "Beralih ke Teknologi TV Digital, Apakah TV di Rumah Perlu Diganti?", 2022, 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].
[2] Undang-undang Republik Indonesia nomor 11 tahun 2020 tentang Cipta Kerja. Jakarta: Kementerian Sekretariat Negara Republik Indonesia.
[3] D. Darwis, N. Siskawati, and Z. Abidin, "Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Review

Data Twitter Bmkg Nasional", Jurnal Tekno Kompak, vol.15, no.1, pp. 131-145, 2021, DOI : https://doi.org/10.33365/jtk.v15i1.744.

[4] G. A. Ruz, P. A. Henríquez, and A. Mascareño, "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers", Future Generation Computer Systems, vol.106, pp. 92-104, 2020, DOI : https://doi.org/10.1016/j.future.2020.01.005.

[5] A. Sharma, and U. Ghose, "Sentimental analysis of twitter data with respect to general elections in India". Procedia Computer Science, vol.173, pp. 325-334, 2020, DOI : https://doi.org/10.1016/j.procs.2020.06.038.

[6] S. M. Tambunan, Y. Nataliani, and E.S. Lestari, "Perbandingan Klasifikasi dengan Pendekatan Pembelajaran Mesin untuk Mengidentifikasi Tweet Hoaks di Media Sosial Twitter", JEPIN (Jurnal Edukasi dan Penelitian Informatika), vol.7, no.2, pp. 112-120, 2021, DOI : http://dx.doi.org/10.26418/jp.v7i2.47232.

[7] R. Dwiyansaputra, G. S. Nugraha, F. Bimantoro, and A. Aranta, "Deteksi SMS Spam Berbahasa Indonesia menggunakan TF-IDF dan Stochastic Gradient Descent Classifier", Jurnal Teknologi Informasi, Komputer, dan Aplikasinya (JTIKA), vol.3, no.2, pp. 200-207, 2021, DOI : https://doi.org/10.29303/jtika.v3i2.145.

[8] A. K. Fauziyyah, "Analisis sentimen pandemi Covid19 pada streaming Twitter dengan text mining Python", Jurnal Ilmiah SINUS, vol. 18, no.2, pp. 31-42, 2020, DOI : http://dx.doi.org/10.30646/sinus.v18i2.491

[9] S. Symeonidis, D. Effrosynidis, and A. Arampatzis, "A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis", Expert Systems with Applications, vol.110, pp. 298-310, 2018, DOI : https://doi.org/10.1016/j.eswa.2018.06.022.

[10] N. Hafidz, and D. Y. Liliana, "Klasifikasi Sentimen pada Twitter Terhadap WHO Terkait Covid-19 Menggunakan SVM, N-Gram, PSO", Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), vol.5, no.2, pp. 213-219, 2021, DOI : https://doi.org/10.29207/resti.v5i2.2960.

[11] A. Salam, J. Zeniarja, and R. S. U. Khasanah, "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, 2018.

[12] G. Mostafa, I. Ahmed, and M.S. Junayed, "Investigation of different machine learning algorithms to determine human sentiment using Twitter data", International Journal of Information Technology and Computer Science, vol.13, no.2, pp. 38-48, 2021, DOI : 10.5815/ijitcs.2021.02.04.

[13] E. R. Kalaivani, and E. R. Marivendan, "The Effect of Stop Word Removal and Stemming In Datapreprocessing", Annals of the Romanian Society for Cell Biology, vol.25, no.6, pp. 739-746, 2021.

[14] E. Mulyani, F. P. B. Muhamad, and K. A. Cahyanto, "Pengaruh N-Gram terhadap Klasifikasi Buku menggunakan Ekstraksi dan Seleksi Fitur pada Multinomial Naïve Bayes", Jurnal Media Informatika Budidarma, vol.5, no.1, pp. 264-272, 2021, DOI : http://dx.doi.org/10.30865/mib.v5i1.2672.

[15] F. Z. Ruskanda, "Study on the effect of preprocessing methods for spam email detection", Indonesia Journal on Computing (Indo-JC), vol.4, no.1, pp. 109-118, 2019, DOI : https://doi.org/10.21108/INDOJC.2019.4.1.284.

[16] A. Deolika, K. Kusrini, and E.T. Luthfi, "Analisis Pembobotan Kata Pada Klasifikasi Text Mining", (JurTI) Jurnal Teknologi Informasi, vol.3, no.2, pp. 179-184, 2019.

[17] B. J. Sowmya, C. S. Nikhil Jain, S. Seema, and S. KG, "Fake News Detection using LSTM Neural Network Augmented with SGD Classifier", Solid State Technology, vol.63, no.6, pp. 6985-9665, 2020.

[18] F. T. Admojo, and Y. I. Sulistya, "Analisis performa algoritma Stochastic Gradient Descent (SGD) dalam mengklasifikasi tahu berformalin", Indonesian Journal of Data and Science, 3(1), pp. 1-8, 2022, DOI : https://doi.org/10.56705/ijodas.v3i1.42.

[19] A. Tuomo, J. Suutala, J. Röning, and H. Koskimäki, "Better classifier calibration for small datasets", ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 14, no.3, pp. 1-19, 2020, DOI : https://doi.org/10.1145/3385656.

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

[21] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure driven view". Information sciences, vol.507, pp. 772-794, 2020, DOI : https://doi.org/10.1016/j.ins.2019.06.064.J.

[22] D. Chicco, N. Tötsch, and G. Jurman, "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation", *BioData mining*, vol.14, no.1, pp. 1-22, 2021, DOI : https://doi.org/10.1186/s13040-021-00244-z.

[23] A. Vanacore, M. S. Pellegrino, and A. Ciardiello, "Fair evaluation of classifier predictive performance based on binary confusion matrix", Computational Statistics, 1-21, 2022, DOI : https://doi.org/10.1007/s00180-022-01301-9.

[24] G. Zeng, "On the confusion matrix in credit scoring and its analytical properties", *Communications in Statistics-Theory and Methods*, vol. 49, no.9, pp. 2080-2093, 2020, DOI : https://doi.org/10.1080/03610926.2019.1568485.

[25] A. Luque, A. Carrasco, A. Martín, and A. de Las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix", *Pattern Recognition*, vol.91, pp. 216-231, 2019, DOI : https://doi.org/10.1016/j.patcog.2019.02.023.