Building of Informatics, Technology and Science (BITS)
Volume 4, No 4, Maret 2023 Page: 1714−1720
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i4.2644

Method comparison of Naïve Bayes, Logistic Regression, and SVM for Analyzing Movie Reviews

Muhammad Maulidan Aziz*, Mahendra Dwifebri Purbolaksono, Adiwijaya

Informatics, School of Computing, Telkom University, Bandung, Indonesia
Email: [1,*]maulidanaziz@student.telkomuniversity.ac.id, [2]mahendradp@telkomuniversity.ac.id,
[3]adiwijaya@telkomuniversity.ac.id
Correspondence Author Email: maulidanaziz@student.telkomuniversity.ac.id
Submitted: **07/12/2022**; Accepted: **27/03/2023**; Published**: 31/03/2023**

Abstract−A film can be categorized as a successful film based on the reviews given by the critics. The reviews can range from professional critics to public reviews from the general audience. Due to a large number of reviews and opinions on a film, this study aims to create a sentiment analysis model and compare the methods used to analyze datasets from a movie review to determine which methods suit best for this research. Sentiment Analysis is a method for studying and analyzing opinions, then classifying these opinions into several classes. This research will use the Naïve Bayes method, Logistic Regression, and Support Vector Machine (SVM) to analyze film review data. The film review dataset used is a collection of film reviews taken from the Rotten Tomatoes website and will be pre-processed before implementing the Naïve Bayes, Logistic Regression, and SVM methods. This research is our experimental work to develop an analysis system for a public review, we choose movie reviews to test the performance of the system since movie reviews are more complex than other types of other reviews. The result from this research is the SVM classifier with 80:20 data splitting has the best performance, with a result of 99.4% accuracy score and 93.5% F1 score.

Keywords: Movie Reviews; Sentiment Analysis; Naïve Bayes; Logistic Regression; Support Vector Machine

1. INTRODUCTION

The entertainment industry is a huge industry and has a significant impact on our daily lives, one of them was the filming industry because filming industry is one of the oldest institutions and has been successful in gaining public attraction. A film can be called a success if that film won public attention or got a positive review from professional critics [1]. As technology developed, the amount of movie streaming viewers also increased and indirectly increases the reviews of the movies. Due to the massive amount of movie reviews and critics, this research decided to use a big review website to easily collect and compile data of the public perception of that movie or shows. One of the big review websites is Rotten Tomatoes, Rotten Tomatoes or RT are a website that reviewed a movie and tv shows from multiple critics. RT was initially launched in 1998 by a UC Berkley student name Senh Duong, Senh started RT as a website where users can access a movie and tv reviews from multiple professional critics [2]. In 2021 RT has a total of 84 million visitors and one of the largest movie review websites. RT has a scoring system that scales from 1 to 100, and it also has two types of certifications for the main evaluation. The first one is the "Certified Fresh Seal", if a movie or tv has a scoring of 60 to 100 then it will be labeled as "fresh". If the movie or tv has a score below 60 then it will be labeled as "Rotten" and will be labeled as a bad movie. The rotten tomatoes scoring system is very suitable for this research because the scoring and the opinion from the critics can be utilized as a decider for the film if that movie is good or bad movie. To analyze the opinion of the film we can use the sentiment analysis method, Sentiment analysis is a method that can be used to study people's opinions, sentiments, emotions, appraisals, and attitudes towards other entities such as products, services, organizations, individuals, issues, event, and topics [3]. This research decides to use sentiment analysis on movie reviews because movie reviews are more complex than other sentiment reviews such as product, service, or organization reviews. That's because a movie review can be interpreted in many different ways, for example if an acting or a visual effect was given a bad review does not necessarily mean the overall movie is bad, but if a product has a bad review then it will automatically mean that the product is bad [4].

The first research is about sentiment analysis on a Commercial airline review in 2019 by Rahat et al. [5]. The research has concluded that the method using Naïve Bayes has an accuracy of 76.56%, while the Support Vector Machine (SVM) has an accuracy of 82.42%. In this study, the writer hopes that the Naïve Bayes and SVM can be applied to many more review systems such as music, news, politics, and movies.

In 2018 Bayhaqy et al. did research on sentiment analysis for E-commerce using crawling data from Twitter [6]. The methods used in that research are Decision Tree, K-Nearest Neighbor (KNN), and Naïve Bayes. The research focus on customer experience and satisfaction with using the e-commerce platform, the research crawl multiple tweets about e-commerce platform such as Tokopedia, Bukalapak, Shopee and will be classified if that tweet has positive or negative sentiment. The result for precision is Decision tree has 79.96%, KNN has 85.67%, and Naïve bayes has 88.50% meaning that naïve bayes is the best classifier for analyzing social media datasets. For future reference, the writer hopes to be able to use a larger and more complex dataset with an increased number of labels and more e-commerce reach and can include not only Indonesian sites but also sites like Amazon, eBay, Alibaba, etc.

Then in 2019 Hasanli et al. did research on Sentiment analysis for Azeri language tweets [7]. The research uses Twitter API automatically for tweets that are in Azeri language and other Proto-Turkic languages and classified them using Logistic Regression, Naïve Bayes, and Support Vector Machine (SVM). This research automatically searches for the tweets and then pre-processed them, after pre-processing the tweets are classified into two class sentiments

Building of Informatics, Technology and Science (BITS)
Volume 4, No 4, Maret 2023 Page: 1714−1720
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i4.2644

that are Positive and Negative. The result for the classifier is the Naïve bayes has an accuracy of 94%, SVM of 93%, and Logistic Regression of 93%. The naïve bayes method gives the best results on the vector and TF-IDF calculation compared to the other two methods. For future reference, the writers hope that this research can be developed for more languages, not just Azeri and Proto-Turkic languages.

In 2020 Santoso et al. did sentiment analysis research for hoaxes and misinformation in Indonesian news by using the Naïve bayes classifier [8]. This research aims to classify every viral news in Indonesia and analyze the news if it's valid or a hoax. The methods that are used in this research are the Naïve bayes and for the feature selection Particle Swarm Optimization (PSO), Informatic Gain (IG), and Genetic Algorithm (GA). The data use 30 fake news samples that went viral and social media, that sample will be classified based on Cosine Similarity (CS) and accuracy on every method. The Naïve bayes + PSO has a CS of 91.6%, while the average was 77% it was the highest of the other method and have 19 correct out of 30 samples.

In 2021 Dashtipour et al. did research on sentiment analysis for Persian language movies using Deep learning [9]. The data that are used are multiple reviews that are collected manually from Persian movie websites such as www.caffecinema.com and www.cinematicket.org, that data will be divided into multiple subsets. About 60% of the data will be for the training dataset, 30% will be used to validate the performance of the trained neural network testing set, and 10% will be used for the validation set. Classification will be divided into Positive and Negative class, and for the stemming process it will use the HAZM application. The result of this research concluded that Deep Learning is more effective than Shallow Machine Learning, especially the stacked-bidirectional-LSTM achieved the highest accuracy of up to 95.61%. The writers hope that this research can be developed to be more multilingual in terms of movie reviews, not just Farsi language movie reviews.

This research aims to improve the analysis of movie reviews by testing the Naïve Bayes, Logistic Regression, and SVM classifier and to decide which classifiers suits best for determining the movie quality. This research will determine the best classifier based on the accuracy and the F1-score. This research also aims to make it easier for users to determine movie quality and contributing to the development theories that are related to sentiment analysis of product or service review.

2. RESEARCH METHODOLOGY

2.1 System Design

In this research, we will analyze multiple reviews from rotten tomatoes. First, we will gather raw data from multiple movie reviews on Rotten tomatoes. Then those data will be pre-processed before being split into training and testing set. The data will be split into 80:20 and 90:10 scenarios and will be tested with Naïve bayes, Logistic Regression, and SVM classifier. Then there will be a method comparison and after that there will be a hyperparameter tuning stage to find out does tuning affect the performance or not. The stages of the research are shown in Figure 1.

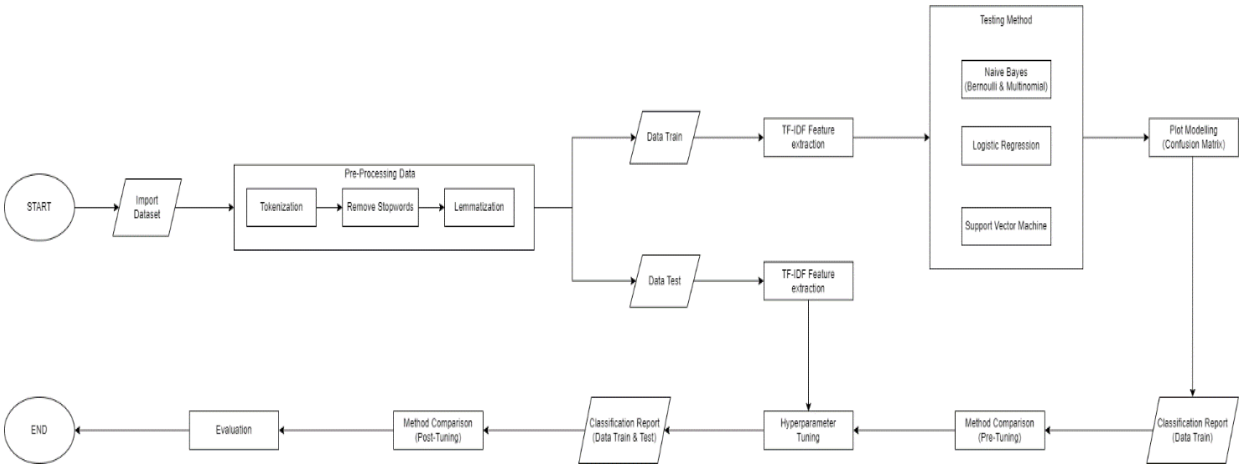

Figure 1. The stages of the research

2.2 Dataset

In this research we use collected movie reviews from the Rotten Tomatoes website, The data is downloaded from the Kaggle website in the tab-separated values (tsv) format. The data are divided into two classes that are Training class and the Testing class, The data has a total of 156060 rows of data which this research will split into two scenarios which are 80:20 and 90:10, and will be evaluate which scenario are the best.

The dataset has 4 columns, the first one is the Phrase_id columns that function as a unique Phrase identifier per phrase. Multiple phrases originate from the same sentence and its data type is "numeric". Then the Sentence_id column is a unique sentence that functions as an index for the next column. The Phrase column is a type of 'string' and it stems from the Sentence that is referenced by Sentence_id. And then the Sentiment column is a numerical type

Building of Informatics, Technology and Science (BITS)
Volume 4, No 4, Maret 2023 Page: 1714−1720
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i4.2644

column that has 5 unique values (0,1,2,3,4) and each of the values is unbalanced. Table 1 is the Sentiment distribution on the dataset.

Table 1. Sentiments on the rotten tomatoes database

Sentiment	Amount
0 (Negative)	7072
1 (Semi-negative)	27273
2 (Neutral)	79582
3 (Semi-positive)	32927
4 (Positive)	9206

2.3 Preprocessing Data

Preprocessing is a necessary data preparation step for classification. For this research we will use Tokenization, then Remove Stop words to remove a phrase that aren't correlated to movie reviews, and Lemmatization to simplify the phrase. Both data training and data testing will use the TF-IDF vectorization before testing the classifier methods. Figure 2 is the steps of preprocessing stages.

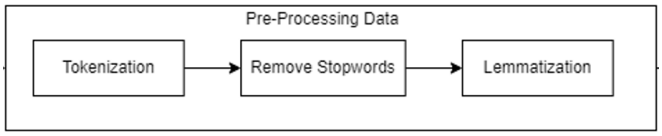

Figure 2. Preprocessing Stage

2.3 Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF is a weighting calculation that is used to weigh each word (term), TF-IDF is a combination of two methods that is Term Frequency and Inverse Document Frequency. Term Frequency is used to measure how many times a term is present in a document, while Inverse Document Frequency is used to assign lower weight to more frequent words and assign a greater weight for the words that are infrequent [10]. In short, TF-IDF is a method to weigh each term by calculating the frequency of occurrence of words in each document and the frequency of occurrence of words in all documents [11]. Equation 1 is the base formula for TF-IDF.

$$W_{t,d} = tf_{t,d} \times \log \frac{N}{df} \qquad (1)$$

2.4 Naïve Bayes

Naïve bayes classifier is a probabilistic classifier that applies the Bayesian theorem, The Bayesian theorem is used to calculate the posterior probability [12]. This research will use two types of Bayesian method that is the Bernoulli Naïve Bayes and the Multinomial Naïve bayes and will be decided which method are better for hyperparameter tuning phase. The base formula for Naïve bayes is shown in equation 2.

$$\mathrm{P(c|x)} = \frac{P(X|C) * P(c)}{P(x)} \qquad (2)$$

2.5 Logistic Regression

Logistic regression is a statistical model that is widely used for dependent variables and dependent continuous binary models [13]. For this research, we use Logistic regression because the data variable is a dependent continuous type and want to compare it with other methods. The base formula of logistic regression is shown in equation 3.

$$P = \frac{e^{b0+b1x}}{1+ e^{b0+b1x}} \qquad (3)$$

2.6 Support Vector Machine

SVM is a method where the main objective is to separate the training dataset into several classes in a hope to maximize the generalization ability [14]. The SVM model main principle is Structural Risk Minimization (SRM), where the main goal is to minimize the bound of a generalization. This research will use the Linear kernel SVM since the dataset format tab-separated values (tsv) are a linearly separated type. The base formula for linear kernel SVM is shown in equation 4.

$$Linear\ Kernel = x^T x \qquad (4)$$

2.5 Performance Evaluation

To evaluate the performance of every classification method, this research will calculate the Accuracy, Precision, Recall, and F1-Score for the data train. The method will also record the time duration for every classification method

Building of Informatics, Technology and Science (BITS)
Volume 4, No 4, Maret 2023 Page: 1714−1720
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i4.2644

to compare which one is the fastest and which one is the slowest. The following table 2 is the formula table of confusion matrix [15].

Table 2. Confusion Matrix

	False Positive (FP)	False Negative (FN)
True Positive (TP)	TP	FN
True Negative (TN)	FP	TN

Confusion matrix is the base component to find the accuracy, precision, and recall. The accuracy is a comparison of the correct prediction that was classified by the system. The base formula for accuracy is in equation 5.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

The precision is to compare the positive true prediction to the overall positive prediction result. The following equation 6 is the formula for Precision.

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

A recall is calculated to find the between the true positive prediction and all the true positive data. Equation 7 is the base formula for Recall.

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

F1-Score is a weighted comparison of Precision and Recall. Equation 8 is the formula for F1-Score.

$$F1\ Score = \frac{2*Precision*Recall}{Precision + Recall} \quad (8)$$

3. RESULTS AND DISCUSSION

3.1 Preprocessing Results

a. Tokenization

Tokenization is a process of text separation into sentences and words by specifying the basic linguistic units [16]. The first step for tokenization is to convert all words into a lowercase, then remove all abbreviations and punctuation in the words. Then the next step is a segmentation of hyphenated words which answers the question of one or two words. Table 3 is an example of the tokenization process.

Table 3. Tokenization process on movie review

Before Tokenization	After Tokenization
I remember this film,it was the first film i had watched at the cinema the picture was dark in places i was very nervous it was back in 74/75 my Dad took me my brother & sister to Newbury cinema in Newbury Berkshire England. I recall the tigers and the lots of snow in the film also the appearance of Grizzly Adams actor Dan Haggery i think one of the tigers gets shot and dies. If anyone knows where to find this on DVD etc please let me know.The cinema now has been turned in a fitness club which is a very big shame as the nearest cinema now is 20 miles away, would love to hear from others who have seen this film or any other like it.	['I', 'remember', 'this', 'film', 'it', 'was', 'the', 'first', 'film', 'i', 'had', 'watched', 'at', 'the', 'cinema', 'the', 'picture', 'was', 'dark', 'in', 'places', 'i', 'was', 'very', 'nervous', 'it', 'was', 'back', 'in', '74', '75', 'my', 'Dad', 'took', 'me', 'my', 'brother', 'sister', 'to', 'Newbury', 'cinema', 'in', 'Newbury', 'Berkshire', 'England', 'I', 'recall', 'the', 'tigers', 'and', 'the', 'lots', 'of', 'snow', 'in', 'the', 'film', 'also', 'the', 'appearance', 'of', 'Grizzly', 'Adams', 'actor', 'Dan', 'Haggery', 'i', 'think', 'one', 'of', 'the', 'tigers', 'gets', 'shot', 'and', 'dies', 'If', 'anyone', 'knows', 'where', 'to', 'find', 'this', 'on', 'DVD', 'etc', 'please', 'let', 'me', 'know', 'The', 'cinema', 'now', 'has', 'been', 'turned', 'in', 'a', 'fitness', 'club', 'which', 'is', 'a', 'very', 'big', 'shame', 'as', 'the', 'nearest', 'cinema', 'now', 'is', '20', 'miles', 'away', 'would', 'love', 'to', 'hear', 'from', 'others', 'who', 'have', 'seen', 'this', 'film', 'or', 'any', 'other', 'like', 'it', '']

b. Remove Stop words

Remove stopwords is a method to remove all unused and uncorrelated words from all tokenized words. Remove stopwords is a natural language processing that may contain a variety of stop-lists, and one per language [17]. Some of the more commonly used stopwords are like “a”, “of”, “the”, “it”, “my”, and “was”. Table 4 is an example of the remove stopwords process.

Table 4. Remove stopwords process on movie review

Before Remove Stopwords	After Remove Stopwords

Building of Informatics, Technology and Science (BITS)
Volume 4, No 4, Maret 2023 Page: 1714−1720
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i4.2644

['I', 'remember', 'this', 'film', 'it', 'was', 'the', 'first', 'film', 'i', 'had', 'watched', 'at', 'the', 'cinema', 'the', 'picture', 'was', 'dark', 'in', 'places', 'i', 'was', 'very', 'nervous', 'it', 'was', 'back', 'in', '74', '75', 'my', 'Dad', 'took', 'me', 'my', 'brother', 'sister', 'to', 'Newbury', 'cinema', 'in', 'Newbury', 'Berkshire', 'England', 'I', 'recall', 'the', 'tigers', 'and', 'the', 'lots', 'of', 'snow', 'in', 'the', 'film', 'also', 'the', 'appearance', 'of', 'Grizzly', 'Adams', 'actor', 'Dan', 'Haggery', 'i', 'think', 'one', 'of', 'the', 'tigers', 'gets', 'shot', 'and', 'dies', 'If', 'anyone', 'knows', 'where', 'to', 'find', 'this', 'on', 'DVD', 'etc', 'please', 'let', 'me', 'know', 'The', 'cinema', 'now', 'has', 'been', 'turned', 'in', 'a', 'fitness', 'club', 'which', 'is', 'a', 'very', 'big', 'shame', 'as', 'the', 'nearest', 'cinema', 'now', 'is', '20', 'miles', 'away', 'would', 'love', 'to', 'hear', 'from', 'others', 'who', 'have', 'seen', 'this', 'film', 'or', 'any', 'other', 'like', 'it', '']	['I', 'remember', 'film', 'first', 'film', 'watched', 'cinema', 'picture', 'dark', 'places', 'nervous', 'back', '74', '75', 'Dad', 'took', 'brother', 'sister', 'Newbury', 'cinema', 'Newbury', 'Berkshire', 'England', 'I', 'recall', 'tigers', 'lots', 'snow', 'film', 'also', 'appearance', 'Grizzly', 'Adams', 'actor', 'Dan', 'Haggery', 'think', 'one', 'tigers', 'gets', 'shot', 'dies', 'If', 'anyone', 'knows', 'find', 'DVD', 'etc', 'please', 'let', 'know', 'The', 'cinema', 'turned', 'fitness', 'club', 'big', 'shame', 'nearest', 'cinema', '20', 'miles', 'away', 'would', 'love', 'hear', 'others', 'seen', 'film', 'like', '']

c. Lemmatization

The main goal of Lemmatization is to reduce the inflectional form to a more common base form, although Stemming and Lemmatization have the same goal, they have different implementations. The difference between Lemmatization and Stemming is lemmatization are more focused on using the right vocabulary and morphological analysis of words, with the aim to remove inflectional endings only and to return the base form of a word. Meanwhile Stemming is more focused on removing the ends of the word and often includes the removal of derivational affixes [18]. Table 5 is an example of the lemmatization process.

Table 5. Lemmatization process on movie review

Before Lemmatizing	After Lemmatizing
['I', 'remember', 'this', 'film', 'it', 'was', 'the', 'first', 'film', 'i', 'had', 'watched', 'at', 'the', 'cinema', 'the', 'picture', 'was', 'dark', 'in', 'places', 'i', 'was', 'very', 'nervous', 'it', 'was', 'back', 'in', '74', '75', 'my', 'Dad', 'took', 'me', 'my', 'brother', 'sister', 'to', 'Newbury', 'cinema', 'in', 'Newbury', 'Berkshire', 'England', 'I', 'recall', 'the', 'tigers', 'and', 'the', 'lots', 'of', 'snow', 'in', 'the', 'film', 'also', 'the', 'appearance', 'of', 'Grizzly', 'Adams', 'actor', 'Dan', 'Haggery', 'i', 'think', 'one', 'of', 'the', 'tigers', 'gets', 'shot', 'and', 'dies', 'If', 'anyone', 'knows', 'where', 'to', 'find', 'this', 'on', 'DVD', 'etc', 'please', 'let', 'me', 'know', 'The', 'cinema', 'now', 'has', 'been', 'turned', 'in', 'a', 'fitness', 'club', 'which', 'is', 'a', 'very', 'big', 'shame', 'as', 'the', 'nearest', 'cinema', 'now', 'is', '20', 'miles', 'away', 'would', 'love', 'to', 'hear', 'from', 'others', 'who', 'have', 'seen', 'this', 'film', 'or', 'any', 'other', 'like', 'it', '']	['I', 'remember', 'this', 'film', 'it', 'wa', 'the', 'first', 'film', 'i', 'had', 'watched', 'at', 'the', 'cinema', 'the', 'picture', 'wa', 'dark', 'in', 'place', 'i', 'wa', 'very', 'nervous', 'it', 'wa', 'back', 'in', '74', '75', 'my', 'Dad', 'took', 'me', 'my', 'brother', 'sister', 'to', 'Newbury', 'cinema', 'in', 'Newbury', 'Berkshire', 'England', 'I', 'recall', 'the', 'tiger', 'and', 'the', 'lot', 'of', 'snow', 'in', 'the', 'film', 'also', 'the', 'appearance', 'of', 'Grizzly', 'Adams', 'actor', 'Dan', 'Haggery', 'i', 'think', 'one', 'of', 'the', 'tiger', 'get', 'shot', 'and', 'dy', 'If', 'anyone', 'know', 'where', 'to', 'find', 'this', 'on', 'DVD', 'etc', 'please', 'let', 'me', 'know', 'The', 'cinema', 'now', 'ha', 'been', 'turned', 'in', 'a', 'fitness', 'club', 'which', 'is', 'a', 'very', 'big', 'shame', 'a', 'the', 'nearest', 'cinema', 'now', 'is', '20', 'mile', 'away', 'would', 'love', 'to', 'hear', 'from', 'others', 'who', 'have', 'seen', 'this', 'film', 'or', 'any', 'other', 'like', 'it', '']

3.2 Results

For all the classification results, this research will compare by using a classification report on the training and testing dataset, after that the data will conclude a hyperparameter tuning to try to improve the performance. The following is the classification report on Naïve bayes (Bernoulli & Multinomial), Logistic Regression, and Support Vector Machine. The following table 6 is the classification report of the 80:20 scenario

Table 6. Classification report 80:20 scenario

	Accuracy	Precision	Recall	F1-Score	Training Duration
Linear SVM	65.4%	64.2%	65.4%	64.5%	5.5 second
Logistic Regression	63.1%	62.5%	63.5%	63.1%	10.5 second
Multinomial Naïve Bayes	60.6%	61.2%	60.7%	55%	0.2 second
Bernoulli Naïve Bayes	60.2%	58.6%	60.7%	56.8%	0.3 second

Based on table 6, the SVM classification has the highest accuracy and f1 score than the others. Theoretically the data splitting can affect the performance score as well, as we can see on table 7 for the scenario of 90:10 data splitting.

Table 7. Classification report 90:10 scenario

	Accuracy	Precision	Recall	F1-Score	Training Duration
Linear SVM	66.2%	65.1%	66.2%	65.4%	8.1 second

Building of Informatics, Technology and Science (BITS)
Volume 4, No 4, Maret 2023 Page: 1714−1720
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i4.2644

Logistic Regression	64.5%	63.6%	64.5%	61.3%	13.7 second
Multinomial Naïve Bayes	61.4%	61.3%	6.14%	56.1%	0.2 second
Bernoulli Naïve Bayes	61.3%	59.2%	61.3%	58.9%	0.3 second

Table 7 proved that data splitting can affect the performance score, although not significant there are still a little difference between the 80:20 and 90:10 scenarios. From the results above we can conclude that the overall average score is around about 60%, we think that the results are not optimal enough for this research. This research will try to improve the overall score by using hyperparameter tuning on all classifier methods.

3.3 Hyperparameter Tuning Results

Hyperparameter tuning is a method of choosing a set of optimal hyperparameters to reach a robust performance result [19]. This research aims to optimize every classifier method and to improve the overall score, especially the accuracy and F1-Score. For Naïve bayes the hyperparameter tuning will only use the multinomial naïve bayes because of the higher performance score than the Bernoulli. For Logistic regression this research will fine-tune the key parameters like C, random state, and solver. And for SVM this research will search the best scores on C and min_df. This research will also use Grid search and the pipeline framework, because grid search will find the combination of the parameter values that will give a better result in terms of Accuracy and F1-Score [20]. The following table 8 and table 9 are the results and accuracy comparison after the hyperparameter tuning.

Table 8. Classification report post tuning on 80:20 scenario

	Accuracy	F1-Score	Training Duration
Linear SVM	96.4%	93.5%	18.2 second
Logistic Regression	94.1%	91.9%	22.5 second
Multinomial Naïve Bayes	92%	90.4%	0.5 second

Table 9. Classification report post tuning on 90:10 scenario

	Accuracy	F1-Score	Training Duration
Multinomial Naïve Bayes	77.4%	74.6%	0.7 second
Linear SVM	78.9%	74.2%	17.4 second
Logistic Regression	75.6%	75.4%	25 second

Based on table 8 and 9, the score is significantly higher for the 80:20 scenario than the 90:10 scenario. The reason is because testing proportion is higher for the 80:20. It also stated that the SVM is still the best performing classifier, although in the 90:10 the Multinomial naïve bayes is slightly better than the logistic regression. Figure 3 represent the chart of performance comparison of before and after hyperparameter tuning on each classifier.

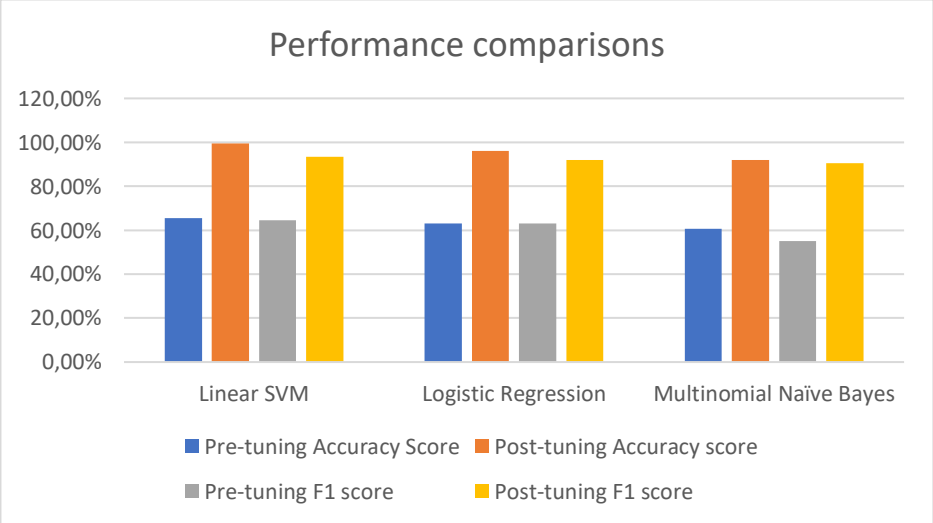

Figure 3. Before and after hyperparameter tuning accuracy score on 80:20 scenarios

3.4 Discussion

Based on table 8 and figure 3 we can conclude that the best result for sentiment analysis on rotten tomatoes movie reviews with a ratio of 80% data training and 20% data testing is the Linear Support Vector Machine classifier, with a result of 99.4% accuracy score and 93.5% F1-score. This research proves that the larger proportion of data testing, the higher the overall performance score, and this research also proves that hyperparameter tuning is very crucial for the improvement of every classifier method performance.

Building of Informatics, Technology and Science (BITS)
Volume 4, No 4, Maret 2023 Page: 1714−1720
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i4.2644

4. CONCLUSION

Based on the research that has been done, it can be concluded that sentiment analysis plays a vital role in resolving the issue of the polarity of reviews. The best classification method for sentiment analysis on rotten tomatoes movie reviews is the Support vector machine method. With the hyperparameter tuning SVM has a final performance result of 96.4% accuracy, 92% precision, 96.6% recall, and 93.5% F1-score. This research also concluded that hyperparameter tuning played a big role in improving the overall performance of every classifier method. Suggestions for further researchers is to use a dataset that is simpler and more balanced in terms of sentiment, we suggest the sentiment of dataset only has Positive and Negative and balanced in proportion, also avoid a database that has semi-positive, semi-negative, or neutral sentiment. This will eliminate the possibility of uncertainty in the classification phase and the final performance results. Also, this research can be extended by using other classification methods such as Random Forrest, Decision Tree, and other classification methods.

REFERENCES

[1] M. del Vecchio, A. Kharlamov, G. Parry, and G. Pogrebna, "The Data science of Hollywood: Using emotional arcs of movies to drive business model innovation in entertainment industries," arXiv preprint arXiv:1807.02221, 2018, doi: https://doi.org/10.1080/01605682.2019.1705194.

[2] A. Vo, "The history of Rotten Tomatoes: A Uniquely Asian-American success story," May 22, 2021. https://editorial.rottentomatoes.com/article/rotten-tomatoes-asian-american

[3] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," Wiley Interdiscip Rev Data Min Knowl Discov, vol. 8, no. 4, p. e1253, 2018, doi: 10.1002/widm.1253.

[4] T. T. Thet, J.-C. Na, and C. S. G. Khoo, "Aspect-based sentiment analysis of movie reviews on discussion boards," J Inf Sci, vol. 36, no. 6, pp. 823–848, 2010, doi: 10.1177/0165551510388123.

[5] A. M. Rahat, A. Kahir, and A. K. M. Masum, "Comparison of Naive Bayes and SVM Algorithm based on sentiment analysis using review dataset," in 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART), 2019, pp. 266–270. doi: 10.1109/SMART46866.2019.9117512.

[6] A. Bayhaqy, S. Sfenrianto, K. Nainggolan, and E. R. Kaburuan, "Sentiment analysis about E-commerce from tweets using decision tree, K-nearest neighbor, and naïve bayes," in 2018 international conference on orange technologies (ICOT), 2018, pp. 1–6. doi: 10.1109/ICOT.2018.8705796.

[7] H. Hasanli and S. Rustamov, "Sentiment analysis of Azerbaijani twits using logistic regression, Naive Bayes and SVM," in 2019 IEEE 13th International Conference on Application of Information and Communication Technologies (AICT), 2019, pp. 1–7. doi: 10.1109/AICT47866.2019.8981793.

[8] H. A. Santoso, E. H. Rachmawanto, A. Nugraha, A. A. Nugroho, and R. S. Basuki, "Hoax classification and sentiment analysis of Indonesian news using Naive Bayes optimization," TELKOMNIKA (Telecommunication Computing Electronics and Control), vol. 18, no. 2, pp. 799–806, 2020, doi: 10.12928/telkomnika.v18i2.14744.

[9] K. Dashtipour, M. Gogate, A. Adeel, H. Larijani, and A. Hussain, "Sentiment analysis of persian movie reviews using deep learning," Entropy, vol. 23, no. 5, p. 596, 2021, doi: 10.3390/e23050596.

[10] S. Qaiser and R. Ali, "Text mining: use of TF-IDF to examine the relevance of words to documents," Int J Comput Appl, vol. 181, no. 1, pp. 25–29, 2018, doi: 10.5120/ijca2018917395.

[11] G. Yunanda, D. Nurjanah, and S. Meliana, "Recommendation system from microsoft news data using TF-IDF and cosine similarity methods," Building of Informatics, Technology and Science (BITS), vol. 4, no. 1, pp. 277–284, 2022, doi: 10.47065/bits.v4i1.1670.

[12] P. H. Gunawan, T. D. Alhafidh, and B. A. Wahyudi, "The Sentiment Analysis of Spider-Man: No Way Home Film Based on IMDb Reviews," Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), vol. 6, no. 1, pp. 177–182, 2022, doi: 10.29207/resti.v6i1.3851.

[13] M. E. Shipe, S. A. Deppen, F. Farjah, and E. L. Grogan, "Developing prediction models for clinical use using logistic regression: an overview," J Thorac Dis, vol. 11, no. Suppl 4, p. S574, 2019, doi: 10.21037/jtd.2019.01.25.

[14] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," Neurocomputing, vol. 408, pp. 189–215, 2020, doi: 10.1016/j.neucom.2019.10.118.

[15] C. Nanda, M. Dua, and G. Nanda, "Sentiment analysis of movie reviews in hindi language using machine learning," in 2018 International Conference on Communication and Signal Processing (ICCSP), 2018, pp. 1069–1072. doi: 10.1109/ICCSP.2018.8524223.

[16] M. Yasen and S. Tedmori, "Movies reviews sentiment analysis and classification," in 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), 2019, pp. 860–865. doi: 10.1109/JEEIT.2019.8717422.

[17] K. Mouthami, K. N. Devi, and V. M. Bhaskaran, "Sentiment analysis and classification based on textual reviews," in 2013 international conference on Information communication and embedded systems (ICICES), 2013, pp. 271–276. doi: 10.1109/ICICES.2013.6508366.

[18] C. Manning, "Introduction to Information Retrieval," Cambridge University Press, 2008. https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

[19] P. Schratz, J. Muenchow, E. Iturritxa, J. Richter, and A. Brenning, "Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data," Ecol Modell, vol. 406, pp. 109–120, 2019, doi: 10.1016/j.ecolmodel.2019.06.002.

[20] S. Ambesange, A. Vijayalaxmi, S. Sridevi, and B. S. Yashoda, "Multiple heart diseases prediction using logistic regression with ensemble and hyper parameter tuning techniques," in 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), 2020, pp. 827–832. doi: 10.1109/WorldS450073.2020.9210404.