Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12776 p-ISSN : 2541-044X

Comparison of Sentiment Analysis Methods on Topic Haram of Music In Youtube

Rahmat Saudi Al Fathir As[1)*], Ema Utami[2)], Anggit Dwi Hartono[3)]
[1,2,3)]Magister Of Informatics Engineering, Universitas Amikom, Indonesia
[1)]alfathir111@gmail.com, [2)]ema.u@amikom.ac.id, [3)]anggit@amikom.ac.id,

Submitted : Jul 25, 2023 | **Accepted** : Aug 25, 2023 | **Published** : Oct 1, 2023

Abstract: Sentiment analysis on video lectures on YouTube that discuss the haram of music is an exciting topic to find out public opinion. This study aims to find what factors affect the model's accuracy in sentiment analysis, especially on video lecture content on YouTube. The data used is comment data on three video lectures that discuss the haram of music, which has been labelled into two categories: positive and negative. The data is divided into two categories, namely primary data, as many as 2099 data that have not been normalized, while secondary data has 1001 data that have been normalized. The experiment shows that the validity of the data, labelling the data, the amount of data, and preprocessing are essential points in forming a good sentiment analysis classification model because, from the test results, it was found that imbalance techniques such as SMOTE, word embedding word2Vec and FastText, and SVM and KNN classification algorithms do not provide maximum accuracy if the data used primary data. However, the data imbalance process, such as oversampling and SVM and KNN classification algorithms, will provide better model accuracy if used with secondary data. Based on the trial results, it is found that when using the SVM algorithm, primary data produces the highest accuracy at 58.35%, while secondary data is 72.23%. If using KNN, the primary data provides the highest model accuracy at 53.54%, while the secondary data has the highest accuracy at 72.81%. Based on these results, it was found that the validity of the data or data must be appropriate and related to the case raised and labelling the data must be done carefully because the most crucial is the inappropriate data in preprocessing the data must be done correctly, if data preprocessing is done in an inappropriate way then data imbalance techniques such as oversampling do not have enough influence on increasing accuracy, but if on the contrary then accuracy will increase. The selection of the right word embedding also affects accuracy. It is necessary to do many experiments to select the correct algorithm and follow the data owned because selecting the correct algorithm will provide maximum accuracy model results.
Keywords: Analysis Sentiment, Classification, Imbalance Data, Machine Learning, Natural Language Processing, Word Embedding.

INTRODUCTION

The need to understand public sentiment about a problem can be done using sentiment analysis. Sentiment analysis can be done on people's perceptions of social media by analyzing people's opinions on various topics (Nurdendi, 2021). Sentiment analysis is an exciting topic to discuss, which aims to find out public opinion, reputation management, market research, crisis detection, or customer service based on public sentiment on an issue, which aims to provide valuable insights into public opinions and perceptions that can help in making better decisions and knowing positive and negative comments (Suryani, 2019). Knowing about public perceptions of haram music video lectures on YouTube is

*name of corresponding author

helpful for a lecture or channel owners to find out the public opinions on the video content created, which aims to determine the condition of public understanding of the video content. To solve this problem, sentiment analysis is needed. Sentiment analysis has several stages: data cleaning, feature extraction, model building, sentiment classification, and result evaluation (Syah, 2022).

Previously, research has been conducted using feature extraction or word embedding using TF-IDF , Word2Vec, for data balancing using the oversampling method, and model building using SVM , KNN, Naive Bayes. Wisnu (2020) explains how to compare naïve Bayes and KNN models to analyze public satisfaction sentiments on digital payment services in Indonesia, namely GO-PAY, OVO, and Link Aja, where the KNN accuracy level is better than naïve Bayes. The highest accuracy results are obtained with a k-folds value of 20 with a curation rate of 83.50% for GO-PAY, 84% for KNN and 91% for Link Aja, while naïve Bayes get an accuracy of 70.71% for GO-PAY, 75.75% for OVO and 70.60% for Link Aja.

The first thing to do is to collect the dataset; in this study, the dataset was collected using Twitter API and collected 10,000 tweets that discussed aeroplane reviews. Furthermore, the data goes through text processing; what is done is stemming, removing URLs, removing @tag hashtags, and stopping word removal. Then, the data is processed using the Naïve Bayes algorithm and Support Vector Machine with 67% training data division, and the rest is testing data. The results found by the Support Vector Machine accuracy of 82.48% have better accuracy than Naive Bayes, which just produced an accuracy of 76.56%. (Rahat, 2019).

Al Fathir [16] discusses the methods used to conduct sentiment analysis based on netizen comment data on YouTube with the title Law of Music in Islam Along with its Evidence, Complete Islamic Music Law: Halal or Haram Music, and Hadith Haramnya Dhaif music. The data collected were 2114 in Indonesian, and the data were labelled with three categories: accepting, not accepting and confused. Furthermore, the data is carried out text processing or data processing with four stages, namely cleaning, tokenizing, stopwords, and stemming; after passing these stages, the data extraction process is carried out using Term Frequency / Inverse Document Frequency (TF-IDF), and classification is carried out using the K-NN algorithm. The test results show a relatively low accuracy value, which is at a level of 65% (Al-Fathir, 2021).

This research will improvise previous research and analyze in detail the stages of the sentiment analysis process to find what factors affect the results of sentiment analysis, methodology and selection of the correct algorithm in the case at hand.

LITERATURE REVIEW

The literature review is used as a comparison in research, which aims to discover the research's position and novelty and raise relevant topics from previous studies. The following is a comprehensive summary of the topics to be discussed; the literature review is as follows.

No	Title	Conclusion	Suggestion	Comparison
1	Sentiment Analysis Provider by.U on Google Play Store Reviews with TF-IDF and Support Vector Machine	Comparison of SVM and TF-IDF with 5-Fold Validation gets good accuracy results. The highest accuracy is 86.1%. The effect of TF-IDF on the model could be better but not harmful.	Can use manual labelling, other machine learning algorithms with different feature selection comparisons,	Feature selection or word embedding using word2vec and fastText with Support Vector Machine and K-Nearest Neighbor algorithms as comparisons.

*name of corresponding author

No	Title	Conclusion	Suggestion	Comparison
2	Sentiment analysis on customer satisfaction of digital payment in Indonesia: A comparative study using KNN and Naive Bayes	The best accuracy is KNN compared to Naive Bayes.	Using other classification algorithms, such as SVM, requires preprocessing data.	The algorithms to be used are SVM and K-NN and use several experiments in performing preprocessing stages such as stemming and lemmatization.
3	Natural Language Processing on Marketplace Product Review Sentiment Analysis	Word normalizer can improve accuracy by 10% for Naive Bayes, and 4% for K-NN.	Find irrelevant data in the dataset to improve sentiment analysis	Word Normalizer / Text Processing uses several variations, such as stemming and lemmatization and using SVM and K-NN algorithms.
4	Aspect Based Sentimental Analysis of Hotel Reviews : A Comparative Study	Based on the experiments that have been carried out, the results of word2vec using SVM produce good accuracy compared to other experiments; namely, accuracy is at 76%, 72%. and 79%.	Using other classification algorithms. Have not tried using another word embedding, such as fastText.	Word embedding will use word2vec and fastText as a comparison.
5	Text Mining Based on Tax Comments as Big Data Analysis Using SVM and Feature Selection	Classification using information gain and SVM produces an average value of 75%, 70%, and 72%.	Feature Selection using information gain and SVM, no comparison with other techniques.	Comparing several techniques and methods in classification, namely using word2vec and fastText using SVM and K-NN algorithms.
6	Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset	The comparison results found that SVM is superior to Naive Bayes by getting an accuracy of 82.48% and Naive Bayes 76.56.	There is no word embedding.	word embedding using word2vec and fastText
7	Sentiment Analysis on twitter tweets about COVID-19 vaccines using NLP and Supervised KNN Classification Algorithm	Twitter sentiment analysis classification of the covid-19 vaccine using KNN resulted in the classification of Pfizer, moderna, and AstraZeneca vaccines.	The analysis needs to be more precise; it does not explain the analysis output, such as accuracy, recall, precision, and the resulting F1	The analysis output is calculated by looking at accuracy, recall, precision and F1 Measure and using a comparison algorithm, namely SVM.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12776
e-ISSN : 2541-2019
p-ISSN : 2541-044X

No	Title	Conclusion	Suggestion	Comparison
			Measure. Moreover, it only uses one algorithm, namely K-NN.	
8	Sentiment analysis of twitter data related on Rinca Island development using Dov2Vec and SVM and logistic regression as classifier	The comparison shows that the Word2Vec CBOW-SVM, PV-SVM and logistic regression methods get the best results of 87% and F1 Score of around 81%.	The use of datasets that have been labelled in a balanced way	The dataset experiment will use two experiments, namely balanced data and not. Word embedding uses Word2Vec and FastText, and the algorithms used are Support Vector Machine and K-Nearest Neighbor.
9	Analisis Sentimen Haramnya Musik Secara Umum Menggunakan Metode KNN	The result of the comparative analysis of the accuracy obtained in performing sentiment haram music using TF-IDF and K-NN is 65%.	At the data processing stage, variations are needed, different word embedding and other algorithms are used, and more data is needed.	word embedding using word2vec and fastText and algorithms using SVM and K-NN.
10	Sentiment Analysis and Classification of Indian Farmers Protest using Data Twitter	The results found are that the Random Forest algorithm provides the best results compared to other machine learning.	Using other algorithm variations. The weakness is the use of unbalanced data. Did not try variations in data division in training and testing and has not used other word embedding.	In word embedding using word2vec and fastText methods, the algorithms used are SVM and K-NN.
11	An efficient approach for sentiment analysis using machine learning algorithm	In that case, the best machine learning algorithm uses the SMO and Decision Tree comparison, which results in an accuracy rate of 89%.	Using other optimization techniques to improve performance	Perform optimization by looking at the influence of preprocessing, word embedding, and algorithms.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12776 p-ISSN : 2541-044X

No	Title	Conclusion	Suggestion	Comparison
12	Sentiment Analysis On Youtube Comments Using Word2Vec and Random Forest	experiments with parameters 1,5,20 epochs and window size 3,5,10 average model accuracy between 90.1% to 91%. Epoch and window size on skip-gram affect accuracy; the higher the epoch value and window size affects the increase in model accuracy.	Classification using Random Forests and Word Embedding using Word2Vec. Moreover, and using word embedding and other algorithms as a comparison.	Word Embedding using Word2vec and FastText. As well as algorithms for classification using SVM and K-NN as a comparison.
13	Analisis sentimen komentar di Youtube tentang ceramah ustadz Abdul Somad Menggunakan Algoritma Naive Bayes.	The resulting accuracy is 87% with a precision value of 91%, Recall 97%, and F1 Measure 93%. Naive Bayes is good at analysing sentiment on YouTube comments about Ustadz Abdul Somad.	Can use other methods in doing sentiment analysis.	The algorithms used are SVM and K-NN.
14	Analisis Sentimen Masyarakat terhadap Kasus Covid-19 pada Media Sosial Youtube dengan Metode Naive bayes	The performance of naive Bayes in sentiment analysis is quite good, producing an accuracy rate of 74%.	No variation of word embedding, and it has not used other algorithms.	Using word2vec and fastText word embedding and classification algorithm using SVM and K-NN. Using classification algorithms using SVM and K-NN

Table 1. Literature Review

METHOD

Data Collections

The data is taken from three videos on YouTube that discuss "Haram music" using web scraping technique, which collected as many as 2110 used by previous researcher. There are two types of data used to conduct sentiment analysis. First is primary data and secondary data for comparison data. Previous researchers used primary data, which amounted to 2110 (Al-Fathir, 2021). The secondary data is the primary data that has made some adjustments, such as word corrections, data labels, etc. The secondary data totals 1001. The description of the feature data is as follows.

Table 2. Description of Features

Feature	Description
Sentence	Comment from people
Label	Descript comment positive or negative

Table 3. Amounted Data

Feature	Description	Amounted
Primary Data	Data from previous researcher	2110
Secondary Data	Data with normalized	1001

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12776 p-ISSN : 2541-044X

Preprocessing Data

Data preprocessing is used to validate data to improve data quality, efficiency, and consistency. The stage carried out in data preprocessing is lower conservation, which is converting vocabulary in text data into lowercase letters to make the data consistent. Non-alphabetic character removal removes all numbers, punctuation marks, and special characters from text data. Spelling correction is to correct typo words in the data. Stopword removal serves to remove common words that are considered meaningless. Stemming is used to convert the affixed words into essential words. Tokenized is used to convert data into an array so that feature extraction can be done. (Al-Fathir, 2021)

Table 4. Preprocessing Data Flow

Preprocessing Step	Result
Initial	Tentang musik ada dua
Lower Conservation	tentang musik ada dua !
Non Alphabetic Character Removal	tentang musik ada dua
Tokenized	[tentang, musik, ada, dua]
Stemming	[tentang, musik, dua]
Stopword Removal	[tentang, musik, dua]

Word Embedding

Word Embedding is used to convert data that was initially a word into a vector, and this aims to allow the model to perform the classification process. Word embedding is used using Word2Vec and FastText. This method is already used with (Thavarasan, 2019). Word2Vec is a method for creating word embedding. It can be done using two methods, Skip Gram and Common Bag of Words (CBOW) Karani (2018). Word2Vec works using words that are to the left and right of the target word and delimited by a window to predict the target word (binus.ac.id).

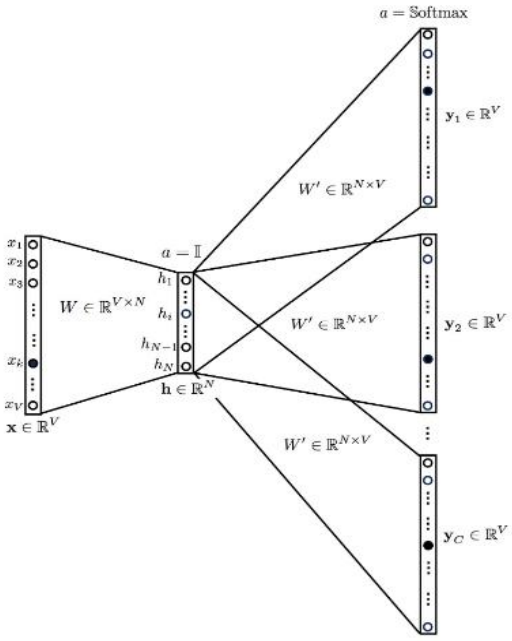

Image 1. Word2Vec Skip-n-gram

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12776
e-ISSN : 2541-2019
p-ISSN : 2541-044X

fastText is a Facebook-owned library used to generate efficient word representations and provide support for text classification. FastText is used to vectorize data based on subwords, and then the value of the vector is averaged. Khattak (2019) FastText is built on the constraints of specific methods such as word2vec and GloVe. In particular, it can handle vocabulary not in the dictionary by extending the word2vec skip-gram model with internal sub-word information. So FastText breaks words into syllables, which are then put into a vector space.

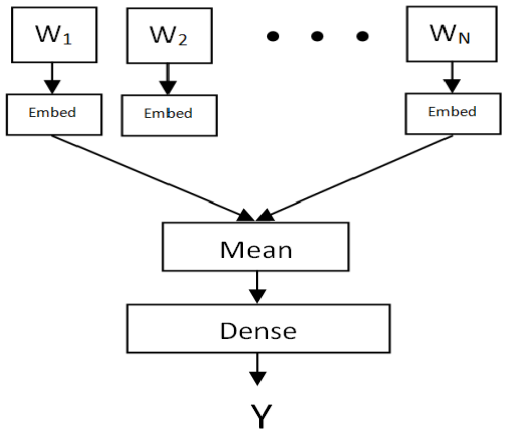

Image 2. FastText Architecture (researchgate.net)

Data imbalance and SMOTE

Data imbalance is common in various NLP tasks, such as tagging and machine reading comprehension (Li, Xiaoya, 2019). Data resampling has proven very effective for handling data imbalance (Verbiest, 2014). Techniques used for data imbalance include NCL and SMOTE techniques (Junsonboon, 2017). SMOTE is a technique for generating data through bootstrapping and the K-nearest neighbour methods. For specific data belonging to a minority class, K nearest neighbours of the same minority class are found, and new data are generated between them by creating a linear connection structure with the neighbours (Hu, Feng, 2013).

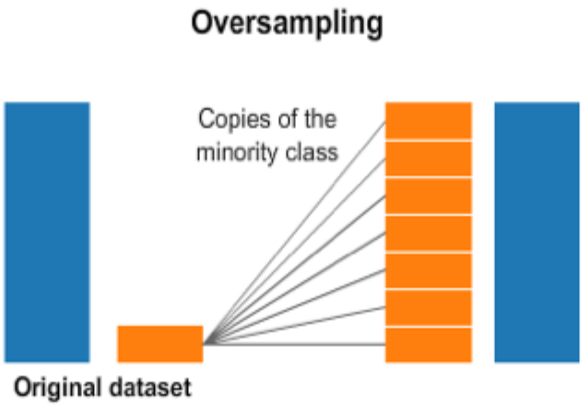

Figure 3. Smote Oversampling (oralytics.com)

Model

Support Vector Machine is a discriminative classifier with a separating hyperplane as its formal definition. In other words, the algorithm generates the optimum value of the hyperplane that categorizes new instances given labelled training data. (Tripathi, 2021)

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12776
e-ISSN : 2541-2019
p-ISSN : 2541-044X

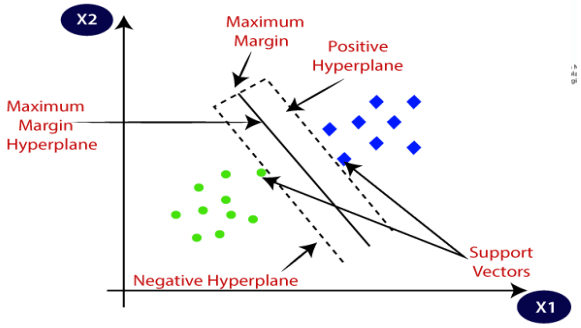

Image 3. Support Vector Machine

K-Nearest Neighbor, generally, KNN calculates the distance of one test data with all existing train data using the Euclidean distance.

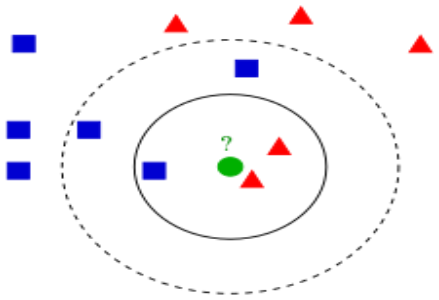

Image 4. K Nearest Neighbor

Evaluation

Evaluation is calculated by looking at the accuracy, recall, precision and F1 Measure level. Moreover, testing is done using a confusion matrix.

Confusion Matrix :

$$Accuracy = \left(\frac{TP+TN}{P+N}\right) x100\%$$

$$Recall = \left(\frac{TP}{TP+FN}\right) x100\%$$

$$Precision = \left(\frac{TP}{TP+FP}\right) x100\%$$

$$F1\ Score = 2x\left(\frac{Precisiom\ x\ Recall}{Precission + Recall}\right)$$

Figure 2. Confusion Matrix

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12776 p-ISSN : 2541-044X

RESULT

Experiments tested on two data, namely primary data and secondary data, using the same method. This session explains what factors affect the accuracy of the model. The results of the discussion are as follows.

Data Collection and Label Data

Two types of data are used in the testing process, namely primary data and secondary data. First is primary data and secondary data for comparison data. Previous researchers used primary data, which amounted to 2110 (Al-Fathir, 2021). The secondary data is the primary data that has made some adjustments, such as word corrections, data labels, etc. The secondary data totals 1001. All above data have two type labelled data: positive and negative.

Table 3. Total Data

Feature	Positive	Negative	Total
Primary Data	55.36%	44.64%	2099
Secondary Data	61.84%	38.16%	1001

Imbalance Data

Based on the amount of data that has been presented. If it is found that the data is imbalance or unbalanced, then the data needs to be normalized using oversampling techniques using SMOTE. As done by SMOTE dramatically increases accuracy (Obiedat, 2022).

Table 3. Data With Oversampling

Feature	Positive	Negative	Total
Primary Data	50%	50%	3662
Secondary Data	50%	50%	1238

Word Embedding

Word2Vec and FastText have not previously been used on this data (Al-Fathir, 2022) previously, several different studies were using TF IDF (Fransiska, 2020) and word2vec (Abro, 2020).

Support Vector Machine

Table 5. Accuracy With Support Vector Machine With Word2Vec

Methods	Test/ Pred (%)	Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)
Primary Data + Word2Vec	60/40	55.77	84.95	55.26	66.96
	70/30	55.64	85.87	54.99	67.05
	80/20	55.37	100.0	55.35	71.26
	90/10	56.93	88.24	55.90	68.44
Primary Data + Word2Vec + Smote (Oversampling)	60/40	55.54	89.97	55.87	68.33
	70/30	55.80	90.78	54.98	68.48

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12776 p-ISSN : 2541-044X

	80/20	55.37	100.0	55.35	71.26
	90/10	57.88	88.04	55.81	68.32
Secondary Data + Word2Vec	60/40	67.25	81.98	67.77	74.20
	70/30	68.60	91.45	65.56	76.37
	80/20	64.00	99.59	62.24	76.61
	90/10	63.00	75.04	66.66	70.60
Secondary Data + Word2Vec + Smote (Oversampling)	60/40	72.32	66.03	73.35	69.50
	70/30	69.54	64.43	73.61	68.71
	80/20	66.75	68.28	68.00	68.14
	90/10	65.80	63.55	70.94	67.04

Referring to Table 5 primary data + Word2vec, the resulting accuracy is not good enough, with the highest accuracy at 56.93%. However, when added using oversampling techniques, the accuracy increases with the highest point at 72.32%. Secondary Data + Word2vec has better accuracy than before at the highest point of 68.60%; when combined with oversampling, the accuracy again increases by more than 3% at 72.32%.
This proves that even though secondary data has less data than primary data, secondary data can provide better accuracy, and if added with oversampling, the accuracy will increase even more. Using Word2Vec, FastText, SVM, and KNN provides better results than previous research (Al-Fathir, 2021).

Table 6. Accuracy With Support Vector Machine With FastText

Methods	Test/ Pred(%)	Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)
Primary Data + FastText	60/40	55.30	91.26	55.10	68.71
	70/30	55.32	86.73	54.68	67.08
	80/20	55.37	83.22	54.20	65.64
	90/10	58.35	76.76	55.0	64.0
Primary Data + FastText + Smote (Oversampling)	60/40	50.81	45.62	52.04	48.62
	70/30	50.93	52.02	51.83	51.93
	80/20	51.50	43.44	52.74	48.64
	90/10	46.53	39.19	51.05	44.34

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12776
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Secondary Data + FastText	60/40	66.49	83.60	75.66	79.43
	70/30	69.00	85.68	72.74	78.68
	80/20	66.50	83.43	70.96	76.69
	90/10	70.00	80.07	74.20	77.02
Secondary Data + FastText + Smote (Oversampling)	60/40	72.23	57.14	74.12	64.53
	70/30	71.70	55.19	79.93	65.30
	80/20	70.03	51.51	79.93	62.65
	90/10	68.50	53.85	76.72	63.29

Refer to Table 6. Based on the trials that have been carried out using the support vector machine algorithm, the data's validity is very influential in modelling. In primary data, it is found that even though the data has been processed in imbalanced data using oversampling techniques, the accuracy of the model obtained has little effect. Even the accuracy tends to decrease, in contrast to secondary data + oversampling affects the accuracy level of the model, which means that data validity is essential in forming a model or caused the model could be more optimal in classifying based on the data owned. The oversampling technique can work in the SVM algorithm if the data used is valid, in agreement with the research that SMOTE oversampling can improve accuracy (Obiedat, 2022), but with a record of the validity of the data used is correct and appropriate. Word2Vec, FastText, SVM, and KNN provide better results than previous research (Al-Fathir, 2021), which only uses TF-IDF and KNN.

K-Nearest Neighbor

Table 7 Accuracy K-Nearest Neighbor With Word2Vec

Methods	Test/ Pred(%)	Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)
Primary Data + Word2Vec	60/40	49.60	58.02	54.21	56.05
	70/30	48.50	55.52	53.36	54.42
	80/20	51.30	60.10	55.56	57.74
	90/10	49.73	51.24	54.91	53.01
Primary Data + Word2Vec + Smote (Oversampling)	60/40	47.85	55.73	52.78	54.21
	70/30	49.52	57.86	54.13	55.93
	80/20	51.72	60.64	55.89	58.17
	90/10	50.79	52.77	55.87	54.27

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12776 p-ISSN : 2541-044X

Secondary Data + Word2Vec	60/40	66.72	73.92	72.75	73.33
	70/30	63.62	74.82	68.93	71.76
	80/20	63.29	76.56	68.04	72.05
	90/10	64.59	69.12	72.36	70.70
Secondary Data + Word2Vec + Smote (Oversampling)	60/40	69.85	63.88	72.47	67.90
	70/30	69.55	62.12	72.89	67.08
	80/20	67.81	59.20	71.46	64.75
	90/10	64.84	55.29	68.29	61.11

They are referring to Table 7. Primary Data + Word2vec, the resulting accuracy needs to be better because the highest accuracy is only at point 51.30%. However, if added using the data imbalance technique, namely oversampling SMOTE, the accuracy increases by more than 15% at point 69.55. Secondary Data + Word2vec, the resulting accuracy is better than before at the highest accuracy of 66.72, but if added with oversampling, the accuracy increases by more than 2% to 69.85%. This proves that the oversampling technique can improve the model's accuracy (Obiedat, 2022) and is suitable when used with KNN. Word2Vec, FastText, SVM, and KNN provide better results than previous research (Al-Fathir, 2021), which only uses TF-IDF and KNN.

Table 7. Accuracy K-Nearest Neighbour With FastText

Methods	Test/Pred(%)	Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)
Primary Data + FastText	60/40	49.60	58.02	54.21	56.05
	70/30	50.74	57.73	55.29	56.49
	80/20	49.10	58.06	53.73	55.81
	90/10	48.83	54.58	53.71	54.14
Primary Data + FastText + Smote (Oversampling)	60/40	53.54	47.05	54.03	50.30
	70/30	52.42	47.87	52.63	50.12
	80/20	52.74	46.66	53.12	49.68
	90/10	51.14	46.08	51.27	48.53
Secondary Data + FastText	60/40	72.04	80.64	75.75	78.12
	70/30	69.75	80.83	73.06	76.75
	80/20	70.66	86.26	71.88	78.42
	90/10	66.92	78.63	70.98	74.61
Secondary Data + FastText + Smote (Oversampling)	60/40	72.81	64.69	77.17	70.38
	70/30	70.81	61.20	75.71	67.68
	80/20	71.54	63.03	75.91	68.87
	90/10	67.71	53.68	74.56	62.42

Referring to Table 7, primary data + FastText has an accuracy that is not good enough, with the highest accuracy of 50.74%, but when added with oversampling, the highest accuracy point is 53.54%. Secondary data + FastText has relatively good accuracy at 72.04%. When added using oversampling, accuracy increases quite a bit at the highest accuracy point of 72.81%. Oversampling is quite effective in improving accuracy (Junsonboon, 2017). Word2Vec, FastText, SVM, and KNN provide better results than previous research (Al-Fathir, 2021), which only uses TF-IDF and KNN.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12776 p-ISSN : 2541-044X

DISCUSSIONS

Based on the results of trials that have been carried out using two different data, namely primary data and secondary data, it shows that data collection, data cleaning, and data imbalance namely SMOTE, have an essential role in providing better classification results, the use of FastText and Word2Vec word embedding and SVM and KNN classification algorithms provide pretty good results on secondary data because secondary data provides the best accuracy at 72.81% even though it provides fewer data, namely 1001 data compared to primary data, which has a total of 2099 data and gives the highest accuracy at 58.35% points after researching this is happens because the primary data is not correctly cleaned, making it difficult for the algorithm to classify.

This research contributes to previous research that has been researched by (Al-Fathir, 2021), which shows that data cleaning must be done correctly, word embedding FastText and Word2Vec provide better vectorization results than TF-IDF which has an impact on the classification results of the SVM and KNN algorithms can provide reasonably good results in the case of using secondary data.

CONCLUSION

Based on the test results that have been presented, it is found that the factors that affect the accuracy of the model on sentiment analysis using video data of lectures on the haram of music are that the validity of the data or data must be appropriate and related to the case raised and labelling the data must be done carefully because the most crucial is the inappropriate data as well as in preprocessing the data must be done correctly, if data preprocessing is done in an inappropriate way then data imbalance techniques such as oversampling do not have enough influence on increasing accuracy, but if on the contrary then accuracy will increase. The selection of the right word embedding also affects accuracy. It is necessary to do many experiments to select the correct algorithm and follow the data owned because selecting the correct algorithm will provide maximum accuracy model results. In this case, the SVM and known algorithms have almost the same accuracy when combined using oversampling and word2vec and FastText techniques, with the highest accuracy of 72.81% and 72.32%.

REFERENCES

Abro, S., Shaikh, S., Abro, R. A., Soomro, S. F., & Malik, H. M. (2020). Aspect based sentimental analysis of hotel reviews: a comparative study. *Sukkur IBA Journal of Computing and Mathematical Sciences*, *4*(1), 11-20.

Ahmadi, M. I., Gustian, D., & Sembiring, F. (2021). Analisis Sentiment Masyarakat terhadap Kasus Covid-19 pada Media Sosial Youtube dengan Metode Naive bayes. *J-SAKTI (Jurnal Sains Komputer dan Informatika)*, *5*(2), 807-814.

Al Fathir, R. S., Agus, T. R., Suyono, A. A., & Ibrahim, F. (2021). Analisis Sentimen Haramnya Musik Secara Umum Menggunakan Metode KNN. *METIK JURNAL*, *5*(2), 66-70.

Asgarnezhad, R., Monadjemi, S. A., & Aghaei, M. S. (2022). A new hierarchy framework for feature engineering through multi-objective evolutionary algorithm in text classification. *Concurrency and Computation: Practice and Experience*, *34*(3), e6594.

Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2022). Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*, *25*(1), 114-146.

Fransiska, S., Rianto, R., & Gufroni, A. I. (2020). Sentiment Analysis Provider by. U on Google Play Store Reviews with TF-IDF and Support Vector Machine (SVM) Method. *Scientific Journal of Informatics*, *7*(2), 203-212.

Harpizon, H. A. R., Kurniawan, R., Iskandar, I., Salambue, R., Budianita, E., & Syafria, F. (2022). Analisis Sentimen Komentar Di YouTube Tentang Ceramah Ustadz Abdul Somad Menggunakan Algoritma Naïve Bayes. *Analisis Sentimen Komentar Di YouTube Tentang Ceramah Ustadz Abdul Somad Menggunakan Algoritma Naïve Bayes*, *5*(1), 131-140.

Hidayat, T. H. J., Ruldeviyani, Y., Aditama, A. R., Madya, G. R., Nugraha, A. W., & Adisaputra, M. W. (2022). Sentiment analysis of twitter data related to Rinca Island development using Doc2Vec and SVM and logistic regression as classifier. *Procedia Computer Science*, *197*, 660-667.

*name of corresponding author

Karani, D. (2018). Introduction to word embedding and word2vec. *Towards Data Science*, *1*.

Khomsah, S. (2021). Sentiment Analysis On YouTube Comments Using Word2Vec and Random Forest. *Telematika: Jurnal Informatika dan Teknologi Informasi*, *18*(1), 61-72.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, *10*(4), 150.

Naresh, A., & Venkata Krishna, P. (2021). An efficient approach for sentiment analysis using machine learning algorithm. *Evolutionary Intelligence*, *14*(2), 725-731.

Neogi, A. S., Garg, K. A., Mishra, R. K., & Dwivedi, Y. K. (2021). Sentiment analysis and classification of Indian farmers' protest using twitter data. *International Journal of Information Management Data Insights*, *1*(2), 100019.

Nurdeni, D. A., Budi, I., & Santoso, A. B. (2021). *Sentiment Analysis on Covid19 Vaccines in Indonesia: From The Perspective of Sinovac and Pfizer*. 122–127. https://doi.org/10.1109/eiconcit50028.2021.9431852

Nurkholis, A., Alita, D., & Munandar, A. (2022). Comparison of Kernel Support Vector Machine Multi-Class in PPKM Sentiment Analysis on Twitter. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *6*(2), 227-233.

Obiedat, R., Qaddoura, R., Ala'M, A. Z., Al-Qaisi, L., Harfoushi, O., Alrefai, M. A., & Faris, H. (2022). Sentiment analysis of customers' reviews using a hybrid evolutionary svm-based approach in an imbalanced data distribution. *IEEE Access*, *10*, 22260-22273.

Rahat, A. M., Kahir, A., & Masum, A. K. M. (2019, November). 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)* (pp. 266-270). IEEE.

Raisa, J. F., Ulfat, M., Al Mueed, A., & Reza, S. S. (2021, February). A review on Twitter sentiment analysis approaches. In *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)* (pp. 375-379). IEEE.

Rini, R., Utami, E., & Hartanto, A. D. (2020, October). Systematic Literature Review Of Hate Speech Detection With Text Mining. In *2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS)* (pp. 1-6). IEEE.

Rohman, A. N., Musyarofah, R. L., Utami, E., & Raharjo, S. (2020, October). Natural Language Processing on Marketplace Product Review Sentiment Analysis. In *2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS)* (pp. 1-5). IEEE.

Shamrat, F. M. J. M., Chakraborty, S., Imran, M. M., Muna, J. N., Billah, M. M., Das, P., & Rahman, O. M. (2021). Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm. *Indones. J. Electr. Eng. Comput. Sci*, *23*(1).

Thavarasan, S., & Mahesan, S. (2020, July). Sentiment lexicon expansion using Word2vec and fastText for sentiment prediction in Tamil texts. In *2020 Moratuwa engineering research conference (MERCon)* (pp. 272-276). IEEE.

Tripathi, M. (2021). Sentiment Analysis of Nepali COVID19 Tweets Using NB SVM and LSTM. *Journal of Artificial Intelligence*, *3*(03), 151-168.

Utami, E., & Luthfi, E. T. (2018, March). Text mining based on tax comments as big data analysis using SVM and feature selection. In *2018 International Conference on Information and Communications Technology (ICOIACT)* (pp. 537-542). IEEE.

Wisnu, H., Afif, M., & Ruldevyani, Y. (2020). Sentiment analysis on customer satisfaction of digital payment in Indonesia: A comparative study using KNN and Naïve Bayes. In *Journal of Physics: Conference Series* (Vol. 1444, No. 1, p. 012034). IOP Publishing.

*name of corresponding author