ISMODE
2021

#70 (1570781710): Sentiment Analysis of Madura Tourism in New Normal Era using Text Blob and KNN with Hyperparameter Tuning

DATE: **29-30 January 2022**

URL: https://ismode.unkris.ac.id/

VIRTUAL Conference

Manuscripts must not contain page numbers, headers or footers. Please use the **A4** template available at https://www.ieee.org/conferences/publishing/templates.html

#70 (1570781710): *Sentiment Analysis of Madura Tourism in New Normal Era using Text Blob and KNN with Hyperparameter Tuning*

Hide details

BibTeX

	Drag to change order	Author name	Author affiliation (edit for paper)	Author email	Email	Delete
Authors	⠿	Fika Rachman	University of Trunojoyo Madura, Indonesia	hastarita.fika@gmail.com	➤	✕
	⠿	Imamah Imamah	Institut of Technology Sepuluh Nopember, Indonesia	imamah.207022@mhs.its.ac.id	➤	✕
	⠿	Bagus Setya Rintyarna ✎	Universitas Muhammadiyah Jember, Indonesia	bagus.setya@unmuhjember.ac.id	➤	✕

⊞ ➤ ↶

Paper title	*Sentiment Analysis of Madura Tourism in New Normal Era using Text Blob and KNN with Hyperparameter Tuning* Only the chairs can edit
Conference and track	**2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)** - *Full papers*
Abstract	Only the chairs can edit Tourism during the Covid-19 pandemic has paralysis, even though tourism is a source of regional...
Keywords	Sentiment Analysis; Text Blob; TF-IDF; KNN; Tuning Parameter; Tourism Only the chairs can edit
Topics	Machine Learning ✎
Similarity	On Dec 15, 2021 01:13 America/New_York, ithenticate computed a similarity score of 5 for the review manuscript.
Personal notes	⊞

Roles	You are a reviewer for this conference. You are an author for this paper.
Status	Published ⊗
Copyright	⊞ IEEE; IEEE: Jan 25, 2022 00:37 America/New_York
Registration	Fika Rachman has registered and paid for Batch 3:Pro Reg ⊗
Visa letter	! Need to pay for registration first.
Presented	by Fika Rachman ⊞ in session 1A: *Parallel Session 1A* from Sat, January 29, 2022 07:30 WIB until 09:30 (5th paper) in Room A (15 min.)

Review manuscript	PDF
Presentation	PDF
Final (accepted) manuscript / PDF must be submitted to PDF eXpress	PDF
Stamped	PDF
Stamped-e	PDF
Auxiliary files	

Review

Actions	Technical content and scientific rigour		Novelty and originality		Quality of presentation		Relevance and timeliness		Recommendation		[] Text and mathematical formulas Is text clear and simple? Are math formulas clear and understandable? [] Conclusion/Summary Is the Summary/Conclusion section of the paper a good summary of what is presented? Thank you again for your time.">Check list		
completed	Average	3	Good	4	Average	3	Average	3	Accept.	3		6	👍

Length of pages

Nothing to remove. Just more improved.

English writing quality

Also major problems with the grammar and proofreading

Points to stress

First of all, I would like to thank the authors for considering ISMODE as a publication outlet for their paper "The research conducted in this paper is Sentiment Analysis of Madura Tourism in New Normal Era using Text Blob and KNN with Hyperparameter Tuning."
This paper has improved compared to the previous version. This paper has improved compared to the previous version. There are strengths and weaknesses

"Strength*

Actions	Technical content and scientific rigour	Novelty and originality	Quality of presentation	Relevance and timeliness	Recommendation	[] Text and mathematical formulas Is text clear and simple? Are math formulas clear and understandable? [] Conclusion/Summary Is the Summary/Conclusion section of the paper a good summary of what is presented? Thank you again for your time.">Check list	

1. Topic is interesting and will attract the attention of the audience.
2. In the discussion, results have an interesting concept because it uses questions
3. There is a discussion about the use of blockchain in companies
4. Reference is ok

"Weakness"
1. The abstract section is well-written, summarizes the literature review, includes a summary of the study's problem, purpose, methods, results, discussion, and conclusion. Concerning section "Related Works," I consider that the authors should make a more in-depth review of specific models,
2. introduction
what problems were raised so that a hybrid framework is needed
3. literature review
what is meant by a hybrid framework
what is the difference between the research conducted and the previous related research."

And too many grammar mistakes and wrong spelling in the manuscript.

The similarity results are standard. I recommended this paper be accepted with a minor revision."

completed	Average 3	Average 3	Average 3	Average 3	Possible Accept. 2		5 👍

Length of pages

The length of the paper is okay

English writing quality

Brush up your grammar.
Check your spelling before and after writing.

Points to stress

This paper is very interesting because by using sentiment analysis we can extract people's opinions, emotions, attitudes, and feelings about a topic or situation from a large amount of unstructured data.
The author should explain more about the train set, development set and test set.
Describe the best validation set accuracy for each n-gram.
Elaborate more about the precision, recall and F1 score

completed	Good 4	Good 4	Good 4	Good 4	Accept. 3		3 👍

Actions	Technical content and scientific rigour	Novelty and originality	Quality of presentation	Relevance and timeliness	Recommendation	[] Text and mathematical formulas Is text clear and simple? Are math formulas clear and understandable? [] Conclusion/Summary Is the Summary/Conclusion section of the paper a good summary of what is presented? Thank you again for your time.">Check list

Length of pages	English writing quality	Points to stress

the length of pages is appropriate	the writing is clear enough to convey the meaning	all equation should be rewrite. Please be carefully using underscore , minus, multiply etc. Examples
		TF-IDF = TF x IDF
		F-Measure = 2 x (......)

EDAS at foxtrot for 2001:448a:5122:5c64:61bc:7d32:3be7:3c4b (Sun, 18 Jun 2023 09:53:33 -0400 EDT) [User 1357592 using Win10:Chrome 0.0 0.159/2.053 s] Request help

Sentiment Analysis of Madura Tourism in New Normal Era using Text Blob and KNN with Hyperparameter Tuning

***Abstract*— Tourism during the Covid-19 pandemic has paralysis, even though tourism is a source of regional income. In the new normal period, tourism began to rise again. Madura Tourism Sentiment Analysis is needed for regional parties and tourism developers to find out public opinion about tourism places in Madura that have been vacuumed for a long time. The dataset used is opinion data on Twitter for the categories of nature, culinary and religious tourism in Madura. Data was taken during the New Normal period between April 2020 to August 2021. This research compared Manual Lexicon Based and TextBlob for labeling data. TF-IDF for term weighting. SVM, Naïve Bayes and KNN methods with Tuning Parameters are compared for classification method in sentiment analysis. Based on this research, the best accuracy value is 68% for KNN method with K-Value = 5 and based on the data distribution of Text Blob labeling results, it is known that the most positive labels are obtained sequentially for 3 tourism categories: nature tourism, culinary tourism and religious tourism.**

Keywords—Sentiment Analysis, TextBlob, TF-IDF, KNN, Tuning Parameter, Tourism

I. Introduction

Location of Madura is in East Java that has a diversity of tourist attractions. There are several categories of tourist attractions managed in Madura, including: nature tourism, historical tourism, cultural tourism, culinary tourism, religious tourism, and artificial tourism. The Covid-19 Pandemic period, which lasted almost two years, had indirectly paralyzed the tourism sector. Whereas the tourism sector is one source of local revenue [1]. Along with the New Normal, Tourism began to rise again. People are starting to follow health protocols in their activities outside the home, especially while on vacation.

Twitter is one of the social media platforms used by the public to accommodate opinions or share information through the internet [2]. In terms of tourism, tourists also sometimes provide reviews of places that have been visited through tweets on social media Twitter [3]. This New Normal period is the initial period for the development of tourist areas after a long vacuumd due to the Covid-19 Pandemic. Sentiment analysis techniques can be used to analyze review data from tourists to determine the level of tourist satisfaction with the places visited. The use of this technique can be useful for the management of tourism places or local parties to develop the place according to tourist attractions.

Previous research has used sentiment analysis techniques to determine visitor expectations of natural attractions [4]. In addition, sentiment analysis techniques have also been applied to determine the location of halal tourism in the world which is widely reviewed by visitors on Twitter [5]. The results of sentiment analysis can also be a feature in the forecasting concept [6] and can also be applied to the case of predicting visitors to a tourist spot [7]. The application of sentiment analysis as a complementary technique in the tourism recommendation system has been carried out previously [8]. The classification method used in sentiment analysis can also affect the system's accuracy value. Research [9] shows the results that the use of the KNN method is better than the SVM method for real time-based twitter data sentiment analysis. So in this study using KNN as a method of classification. The data to be used is tweet data for each tourism category, not only nature tourism. It is hoped that the three categories of popular tourist attractions and the level of satisfaction with these tourism categories will be known.

Twitter data (tweet) was taken using a scrapper technique. From this twitter data, a Groundtruth dataset is created for training and testing data. In the dataset labeling process, humans are often used as experts to label the data. However, for large amounts of data, the labeling process in this way takes a very long time. The scrapper process can generate hundreds, thousands and even hundreds of thousands of review data that will be used as datasets. With this condition, it is hoped that there will be other labeling techniques that can help make groundtruth with good accuracy. Previous studies used different lexicon-based techniques in the dataset labeling process. Research [10] used a lexicon manual-based technique with the help of a lexicon dictionary. Research [11][12] uses a lexicon based technique using the python library, namely TextBlob. The use of TextBlob can be used for annotating tweets [13].

This study aims to analyze the level of tourist satisfaction with several categories of tourist attractions in Madura. The contribution of this study is to measure the best accuracy of the K-Nearest Neighbor (KNN) method using hypertuning parameters and compare the performance of the lexicon based manual with TextBlob in the dataset labeling process.

II. Proposed Method

A. Dataset

The process of scrapping twitter data is done using the python library: twint. In the process there are keywords that are used to produce reviews that are in accordance with the topic of Madura tourism. Some of keyword used are: “Wisata Madura” (Madura Tourism), “Wisata Bangkalan” (Bangkalan Tourism), “Wisata Sampang” (Sampang Tourism), “Wisata Pamekasan” (Pamekasan Tourism), and “Wisata Sumenep” (Sumenep Tourism), with a time period between April 2020 to August 2021. Other keywords used are in accordance with the characteristics of the tourist attractions as in Table 1.

TABLE I. Tourism Category Keyword for Scrapping Data

Tourism category	Keywords
Nature tourism	‘pantai’, ‘gunung’, ‘bukit’, ‘air terjun’, ‘gua’, ‘api alam’

	('beach', 'mountain', 'hill', 'waterfall', 'cave', 'natural fire')
Artificial tourism	‘mercusuar’, ’wisata buatan’ ('lighthouse', 'artificial tourism')
Culinary tourism	‘kuliner’, ‘soto’, ’sate’, ’rujak’, ’nasi’, ’keripik’ ('culinary', 'soto', 'sate', 'rujak', 'rice', 'chips')
Religious tourism	‘makam’, ‘sunan’, ‘masjid’, ‘wali’, ‘religi’ ('tomb', 'sunan', 'mosque', 'wali', 'religion')
History tourism	‘sejarah’, ‘museum’ ('history', 'museum')
Culture tourism	‘tari’, ‘kerapan sapi’ ‘adat’ ('dance', 'kerapan sapi' 'custom')

Then the data from scrapper is going to cleaned and duplication data removed process. The data text using Bahasa Indonesian. The amount of data that will be used is 522 data. It can be seen from the amount of data that the largest distribution is for the category of natural tourism. This shows that in New Normal Era, many people make visits to natural or outdoor attractions than other category of tourism. The distribution of the scrapper data is shown in Figure 1.

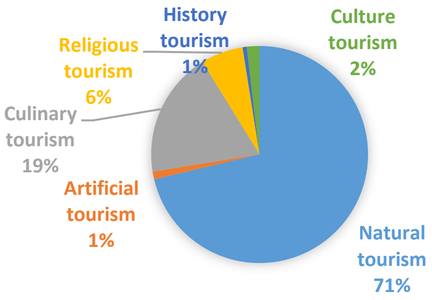

Fig 1. Distribution of tourism category tweet data

Then the data will be labeled to create Groundtruth. In the labeling process, this research compares the labeling method using the manual lexicon based and Python library Textblob. This dataset is preprocessed so as to produce terms that will later be extracted. Feature extraction is done using TF-IDF. Feature data is used in the classification process so as to produce a sentiment label.

The stages of the sentiment analysis process are shown in Figure 2. From the proposed methods, the stages carried out are preparing the dataset, feature extraction, classification process, and evaluation. In the stage of preparing the dataset, there is a twitter data scrapping process, data cleaning process, preprocessing and dataset labeling process. The label sentiments used as the target class are 'positive', 'negative', and 'neutral'. The evaluation process is carried out by using a confusion matrix to determine the accuracy value of analysis sentiment. Hypertuning parameters are performed during the classification process in the formation of the best model. The best model is used by data testing to predict data sentiment.

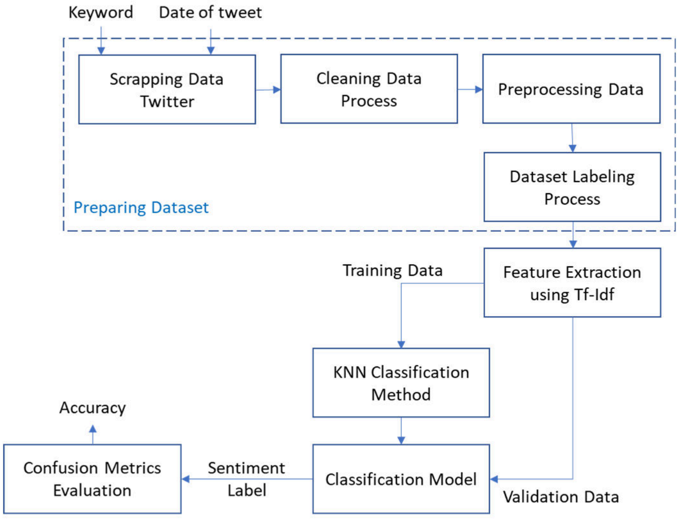

Fig 2. Proposed methods of sentiment analysis

B. Scrapping Data Twitter

The Twint Python library is used for scrapping process. The additional configuration to filter data are `search`, `since`, `until`, and `output`. In the configuration `search` is enter the keywords used.

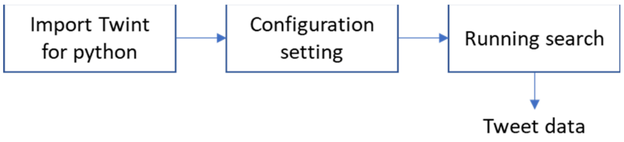

Fig. 3 Stages of Scraping data using Twin

Configuration `since` inputted time to start for scrapping data, which is April 1, 2020. Configuration `until` inputted time to finish scrapping data, which is August 30, 2021. Configuration `output` is used to save tweet data from scrapping by specific file name.

C. Preprocessing Data

Before preprocessing the data, there is a data cleaning stage that is passed. Data Cleaning Process is the process of cleaning tweet data which is not a review but in the form of information. In addition, the deleted tweet data is duplicate tweet data, this happens because users often retweet the previous tweet data. This kind of data needs to be deleted and not included in the dataset formation process.

Preprocessing is carried out on the data resulting from the cleaning process. The preprocessing stages carried out are case folding, tokenizing, stopword removal, stemming using the python library sastrawi.

D. Dataset Labeling

The process of labeling tweet data is done by comparing the concept of manual lexicon based and Text Blob. The concept of manual lexicon based is analyzing data by looking at the context of the sentiment lexicon of the words used in composing sentences. This process requires a lexicon dictionary according to the language used in the tweet data. For the Indonesian lexicon dictionary, the dictionary produced from the research [10] is used. This lexicon dictionary has 6,609 negative and 3,609 positive words with scoring between -5 to +5. So the label is seen based on the total scoring value. A negative score means negative sentiment, a score of 0 means neutral sentiment and a positive score means positive sentiment.

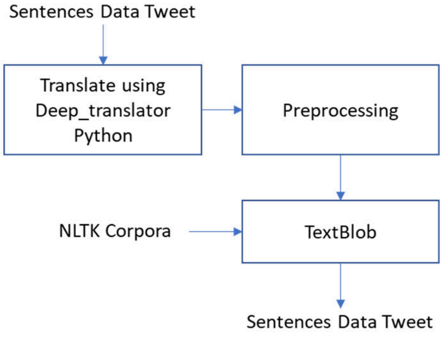

Fig 4. Diagram of Labelling Process using TextBlob

Using the Python library: Textblob is a tool for sentence-level sentiment analysis. This textblob is actually also a lexicon based, only the corpus is taken from the NLTK corpora [14]. Polarity is taken based on the maximum number of words that fall into the positive, negative and neutral categories. The polarity score is worth -1 to 1 and there is a subjectivity value that is worth 0 to 1. The problem that arises is that the corpora NLTK is a collection of English words so that for Indonesian documents a translator will be needed.

E. Feature Extraction using TF-IDF

This research uses the TF-IDF feature obtained from tweet data. This feature is expected to represent and characterize in a review that has a certain polarity of sentiment [15]. Term Frequency (TF) is the value of the occurrence of a word in the document. Document Frequency (DF) describes how many documents contain a certain word. Each document will have a TF-IDF feature that will be used in the document classification process. The TF-IDF formulation according to [16] is as follows:

$$TF_{m,k} = \frac{X_{m,k}}{\sum_n X_{n,k}} \quad (1)$$

$$DF_{m,k} = \frac{|d_k \in D : X_k \in d_k|}{|D|} \quad (2)$$

$$IDF_{m,k} = log \frac{|D|}{|d_k \in D : X_k \in d_k|} \quad (3)$$

$$TF - IDF = TF\ x\ IDF \quad (4)$$

Where :
$|D|$ = total documents
$|d_k \in D : X_k \in d_k|$ = number of documents that have term X_k
$X_{m,k}$ = number of occurences term X_k in document d_k
$\sum_n X_{n,k}$ = number of occurences all term in document d_k

F. Classification Process

There are 3 classification methods used, namely the K-Nearest Neigbor (KNN), Support Vector Machine (SVM) and Naïve Bayes methods.

In KNN, a distance between the test sample and the training data is identified by various types of calculations. Distance similarity measure is an important role for final classification results. Euclidean distance is one of the most frequently used similarity measure methods in the KNN classification [17]. In this research, comparison of accuracy with similarity measure using Eucledian Distance, Cosine and Manhattan will be carried out. The K value is known to also affect the accuracy [18], so a test scenario will also be carried out by changing the K value.

Classification in sentiment analysis functions in determining the class of sentiment document. The methods often used by previous studies are SVM [15][19] and Naïve Bayes [20]. In this study, the method will be compared with the best KNN model after the parameter tuning process is carried out.

The evaluation of the system that will be used is accuracy, recall, precision, and F-Measure. Here is the formula that will be used:

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

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

$$Precision = \frac{TN}{(TN+FP)} \quad (7)$$

$$F - Measure = 2\ x \frac{Recall\ x\ Precision}{(Recall+Precision)} \quad (8)$$

Where:

TN = True Negative
TP = True Positive
FP = False Positive
FN = False Negative

III. Result And Discussion

The data processed in this sentiment analysis process amounted to 522 tweets. There is a test scenario in the Labeling Process and Classification Method.

A. Labeling Process Testing

The results of system testing using dataset labeling from lexicon based and TextBlob are shown in Table 2. The use of TextBlob in the system produces higher accuracy than the use of Lexicon Based, which is 0.68. Although it is better than Lexicon Based, this model does not provide optimal accuracy values. This is possible because the NLTK corpora uses English, so the translator process can also affect the accuracy of the results.

TABLE II. Comparing Accuracy Result using Lexicon Based and TextBlob

Labeling model	Accuracy
Lexicon Based	0,58
TextBlob	**0,68**

From the Table 2, it is known that the use of Text Blob is better than manual lexicon based. Figure 5 is the distribution of data based on the results of labeling with Text Blob for each category of tourist attractions. From the Figure 5, it can be seen that the most positive labels were obtained sequentially for 3 tourism categories, namely: the category of nature tourism, culinary tourism and religious tourism.

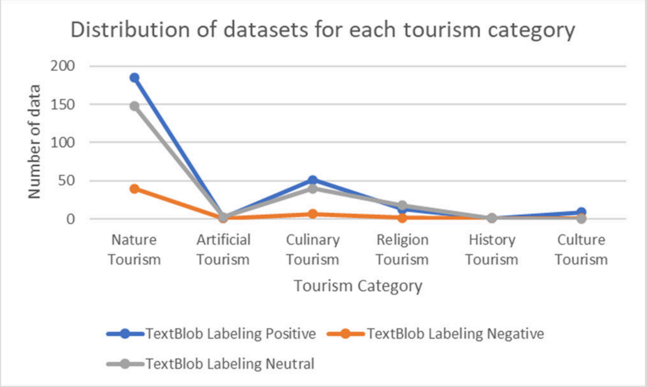

Fig 5. Distribution of dataset for each tourism category

B. *Classification Process Testing*

In the classification process testing, there are two scenarios that are carried out. The first test scenario is measuring the accuracy of the system using the KNN method by tuning the K value parameter and metric of similarity measure. The results of the comparison of accuracy are shown in Table 3.

TABLE III. COMPARING ACCURACY USING KNN CLASSIFICATION WITH TUNING PARAMETER K VALUE AND SIMILARITY MEASURE

K value	Metric Similarity Measure	Accuracy
K=1	Cosine Similarity	0.53
K=3	Cosine Similarity	0.61
K=5	**Cosine Similarity**	**0.68**
K=10	Cosine Similarity	0.62
K=1	Manhattan	0.51
K=3	Manhattan	0.52
K=5	Manhattan	0.55
K=10	Manhattan	0.51
K=1	Euclidean Distance	0.52
K=3	Euclidean Distance	0.61
K=5	Euclidean Distance	0.68
K=10	Euclidean Distance	0.61

From Table 3 it can be seen that the best K-value is 5 with a metric similarity measure using Cosine Similarity. For Fig.5 we can analyze that for the overall metric simlarity measure the value of K-value = 5 is the peak value of the system's maximum accuracy.

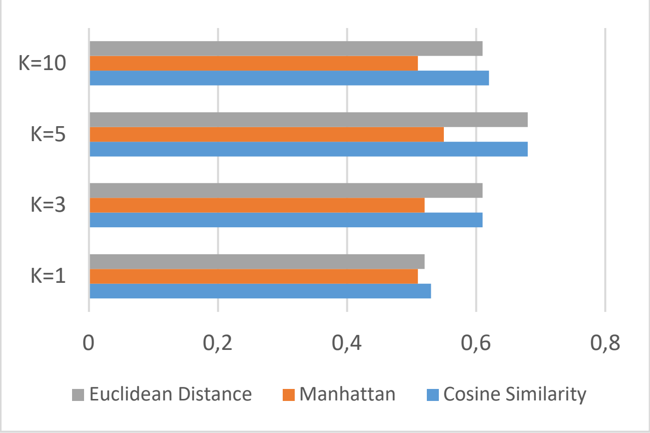

Fig 5. Graph of Kvalue - similarity measure on KNN method

The second test scenario is to compare the system accuracy values using 3 classification methods, namely: KNN (K-value=5, cosine similarity), SVM and Naïve Bayes. Table 4 is the experimental results obtained in the test. The comparison of the 3 classification methods shows that the classification with the KNN method obtains the highest accuracy than the others, which is 0.68. Even if viewed from the F-Measure value, the best value is owned by the SVM method.

TABLE IV. COMPARATION RESULT OF CLASSIFICATION METHOD

Metode	Akurasi	Recall	Precision	F-Measure
KNN	**0.68**	0.64	0.58	0.61
SVM	0.62	0.66	0.59	0.62
Naive Bayes	0.58	0.57	0.59	0.58

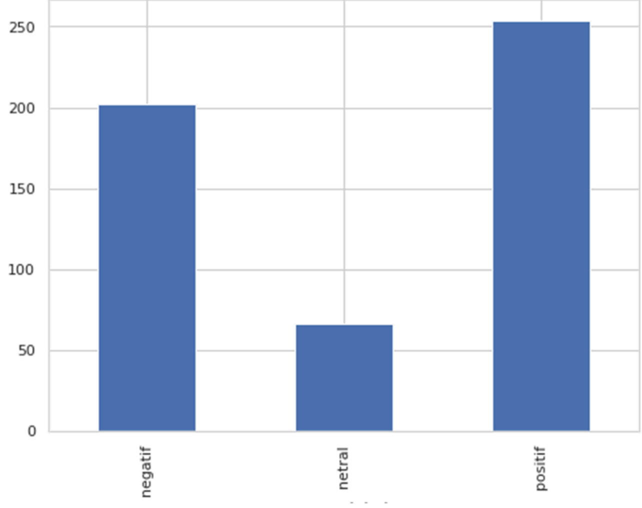

Fig 6. Graph of distribution polarity on tweet data

The selection of algorithm performance that can be used as a reference is generally seen from the amount of FN and FP data. If the value is close to or symmetric then the best reference that can be used is the accuracy value, but if it is not symmetric then the F-Measure value is the reference.

Figure 6 is the distribution of sentiment labels for classified tweet data, it can be seen that Madura Tourism is still considered good by the public with positive reviews that are still higher than negative reviews.

IV. Conclusion

Based on the experimental results that have been carried out, it can be concluded that Madura Tourism is still considered good by the community as evidenced by the high polarity value of tweet review data for positive sentiment compared to negative sentiment, which is 48.7%.

This research found that for own dataset, the labeling process using Text Blob produces better accuracy than manual lexicon based. Based on the data distribution of Text Blob labeling results, it is known that the most positive labels are obtained sequentially for 3 tourism categories, namely: the category of nature tourism, culinary tourism and religious tourism.

From the test scenario, it is found that sentiment analysis uses the KNN method with a K value of 5 and the metric used is Cosine Similarity which has the best accuracy value, which is 68%. However, this value is considered to be still low, this is probably due to the less than optimal translation results and the need for other processes such as word correction, the use of POS Tagging in tweet reviews. The use of the word refinement method helps in changing the shape of the data to be structured. While the use of POS Tagging leads to the introduction of semantics in sentiment analysis.

Acknowledgment

This research was made for Independent Research 2021 in LPPM Universitas Trunojoyo Madura.

References

[1] M. T. Jaenuddin and P. Independen, "Upaya Peningkatan Pendapatan Asli Daerah melalui Pengembangan Pariwisata di Kabupaten Mamuju," *Goverment: Jurnal Ilmu Pemerintahan*, vol. 12, no. 2, pp. 67–71, 2019.

[2] W. A. S. Hootsuite, "DIGITAL 2021: The Latest Insights into The 'State of Digital,'" 2021.

[3] G. W. Tan, V. Lee, J. Hew, and K. Ooi, "Telematics and Informatics The interactive mobile social media advertising : An imminent approach to advertise tourism products and services ?," *Telematics and Informatics*, vol. 35, no. 8, pp. 2270–2288, 2018.

[4] F. Mirzaalian and E. Halpenny, "Exploring destination loyalty : Application of social media analytics in a nature-based tourism setting," *Journal of Destination Marketing & Management*, vol. 20, no. March, p. 100598, 2021.

[5] S. Ainin, A. Feizollah, N. B. Anuar, and N. A. Abdullah, "Sentiment analyses of multilingual tweets on halal tourism," *Tourism Management Perspectives*, vol. 34, no. January 2019, p. 100658, 2020.

[6] S. Yoo, J. Song, and O. Jeong, "Social media contents based sentiment analysis and prediction system," vol. 105, pp. 102–111, 2018.

[7] X. Li, R. Law, G. Xie, and S. Wang, "Review of tourism forecasting research with internet data," *Tourism Management*, vol. 83, no. October 2020, 2021.

[8] Z. Abbasi-moud, H. Vahdat-nejad, and J. Sadri, "Tourism recommendation system based on semantic clustering and sentiment analysis," *Expert Systems With Applications*, vol. 167, no. May 2020, p. 114324, 2021.

[9] A. Ali, "Sentiment Analysis on Twitter Data using KNN and SVM," vol. 8, no. 6, pp. 19–25, 2017.

[10] F. Koto, "InSet Lexicon : Evaluation of a Word List for Indonesian Sentiment Analysis in Microblogs InSet Lexicon : Evaluation of a Word List for Indonesian Sentiment Analysis in Microblogs," no. December, 2017.

[11] S. Kunal, A. Saha, A. Varma, and V. Tiwari, "Textual Dissection Of Live Twitter Reviews Using Naive Bayes," *Procedia Computer Science*, vol. 132, no. Iccids, pp. 307–313, 2018.

[12] F. Rustam, R. Khan, K. Kanwal, R. Y. Khan, and G. S. Choi, "US Based COVID-19 Tweets Sentiment Analysis Using TextBlob and Supervised Machine Learning," pp. 1–8, 2021.

[13] R. Guzman-cabrera, "Exploring the use of lexical and psycho-linguistic resources for sentiment analysis," in *Mexican International Conference on Artificial Intelligent, MICAI*, 2020, no. December, pp. 109–121.

[14] V. Bonta, N. Kumaresh, and N. Janardhan, "A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis," no. March, 2019.

[15] S. S. and P. K.V., "Sentiment analysis of malayalam tweets using machine learning techniques," *ICT Express*, no. xxxx, pp. 2–7, 2020.

[16] S. W. Kim and J. M. Gil, "Research paper classification systems based on TF - IDF and LDA schemes," *Human-centric Computing and Information Sciences*, pp. 9–30, 2019.

[17] M. Sudarma and I. G. Harsemadi, "Design and Analysis System of KNN and ID3 Algorithm for Music Classification based on Mood Feature Extraction," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 7, no. 1, pp. 486–495, 2017.

[18] K. N. Classifier, V. B. S. Prasath, H. Arafat, A. Alfeilat, and O. Lasassmeh, "Distance and Similarity Measures Effect on the," pp. 1–50.

[19] A. Borg and M. Boldt, "Expert Systems with Applications Using VADER sentiment and SVM for predicting customer response sentiment q," *Expert Systems With Applications*, vol. 162, p. 113746, 2020.

[20] Y. Tan and P. P. Shenoy, "A bias-variance based heuristic for constructing a hybrid logistic regression-naïve Bayes model for classification," *International Journal of Approximate Reasoning*, vol. 117, pp. 15–28, 2020.

Sentiment Analysis of Madura Tourism in New Normal Era using Text Blob and KNN with Hyperparameter Tuning

Fika Hastarita Rachman [1*] **Imamah**[2] **Bagus Setya Rintyarna**[3]

[1,2]Departement of Informatics, University of Trunojoyo Madura, Indonesia
[3]Departement of Electro, University Muhammadiyah Jember, Indonesia

* Corresponding author’s Email: hastarita.fika@gmail.com

Abstract— **Tourism during the Covid-19 pandemic has paralysis, even though tourism is a source of regional income. In the new normal period, tourism began to rise again. Madura Tourism Sentiment Analysis is needed for regional parties and tourism developers to find a public opinion about tourism places in Madura that have been vacuumed for a long time. The dataset used is opinion data on Twitter for nature, culinary and religious tourism in Madura. Data was taken during the New Normal period between April 2020 to August 2021. This research compared Manual Lexicon Based and TextBlob for labeling data. TF-IDF for term weighting. SVM, Naïve Bayes, and KNN methods with Tuning Parameters are compared for classification methods in sentiment analysis. Based on this research, the best Accuracy value is 94% for SVM Method or KNN Method using Manhattan measure and K-Value = 1. The most positive labels are obtained for three tourism categories: nature, culinary, and religious.**

Keywords—Sentiment Analysis, TextBlob, TF-IDF, KNN, Tuning Parameter, SVM, Tourism

I. Introduction

The location of Madura is in East Java that has a diversity of tourist attractions. Several categories of tourist attractions managed in Madura include nature tourism, historical tourism, cultural tourism, culinary tourism, religious tourism, and artificial tourism. The Covid-19 Pandemic period had indirectly paralyzed the tourism sector, which lasted almost two years. In contrast, the tourism sector is one source of local revenue [1]. Along with the New Normal, Tourism began to rise again. People are starting to follow health protocols in their activities outside the home, especially while on vacation.

Twitter is one of the social media platforms used by the public to accommodate opinions or share information through the internet [2]. In terms of tourism, tourists also sometimes provide reviews of places visited through tweets on social media Twitter [3]. This New Normal period is the initial period for developing tourist areas after a long vacuum due to the Covid-19 Pandemic. Sentiment analysis techniques can be used to analyze review data from tourists to determine tourist satisfaction with the places visited. This technique can be helpful for the management of tourism places or local parties to develop the place according to tourist attractions.

Previous research has used sentiment analysis techniques to determine visitor expectations of natural attractions [4]. In addition, sentiment analysis techniques have also been applied to determine the location of halal tourism globally, which visitors widely review on Twitter [5]. The sentiment analysis results can also be a feature in the forecasting concept [6] and can also be applied to the case of predicting visitors to a tourist spot [7]. The application of sentiment analysis as a complementary technique in the tourism recommendation system has been carried out previously [8]. The classification method used in sentiment analysis can also affect the system's accuracy value. Research [9] shows results that the use of the KNN method is better than the SVM method for real-time-based twitter data sentiment analysis. So in this study uses KNN as a method of classification. The data to be used is tweet data for each tourism category, not only nature tourism. It is hoped that the three categories of popular tourist attractions and the level of satisfaction with these tourism categories will be known.

Twitter data (tweet) was taken using a scrapper technique. A Groundtruth dataset is created for training and testing data from this Twitter data. Humans are often used as experts in the dataset labeling process to label the data. However, for large amounts of data, the labeling process in this way takes a very long time. The scrapper process can generate hundreds, thousands, and even hundreds of thousands of review data used as datasets. With this condition, it is hoped that there will be other labeling techniques that can help make ground truth with good accuracy. Previous studies used different lexicon-based techniques in the dataset labeling process. Research [10] used a lexicon manual-based technique with the help of a lexicon dictionary. Research [11][12] uses a lexicon-based technique using the python library, namely TextBlob. The use of TextBlob can be used for annotating tweets [13].

This study aims to analyze tourist satisfaction with several categories of tourist attractions in Madura. The contribution of this study is to measure the best accuracy of the K-Nearest Neighbor (KNN) method using hyper tuning parameters and compare the performance of the Lexicon-based manual with TextBlob in the dataset labeling process.

II. Proposed Method

A. Dataset

Scrapping Twitter data is done using the python library: twint. In the process, there are keywords used to produce reviews by Madura tourism. Some of keyword used are: "Wisata Madura" (Madura Tourism), "Wisata Bangkalan" (Bangkalan Tourism), "Wisata Sampang" (Sampang

Tourism), "Wisata Pamekasan" (Pamekasan Tourism), and "Wisata Sumenep" (Sumenep Tourism), with a time period between April 2020 to August 2021. Other keywords used are by the characteristics of the tourist attractions as in Table 1.

TABLE I. TOURISM CATEGORY KEYWORD FOR SCRAPPING DATA

Tourism category	Keywords
Nature tourism	‘pantai’, ‘gunung’, ‘bukit’, ‘air terjun’, ‘gua’, ‘api alam’ ('beach', 'mountain', 'hill', 'waterfall', 'cave', 'natural fire')
Artificial tourism	‘mercusuar’, ’wisata buatan’ ('lighthouse', 'artificial tourism')
Culinary tourism	‘kuliner’, ‘soto’, ’sate’, ’rujak’, ’nasi’, ’keripik’ ('culinary', 'soto', 'sate', 'rujak', 'rice', 'chips')
Religious tourism	‘makam’, ‘sunan’, ‘masjid’, ‘wali’, ‘religi’ ('tomb', 'sunan', 'mosque', 'wali', 'religion')
History tourism	‘sejarah’, ‘museum’ ('history', 'museum')
Culture tourism	‘tari’, ‘kerapan sapi’ ‘adat’ ('dance', 'kerapan sapi' 'custom')

Then the data from the scrapper will be cleaned, and duplication data removed process. The data text uses Bahasa Indonesian. The amount of data that will be used is 522 data. It can be seen from the amount of data that the most significant distribution is for the category of natural tourism. This shows that in the New Normal Era, many people visit natural or outdoor attractions than another category of tourism. The distribution of the scrapper data is shown in Figure 1.

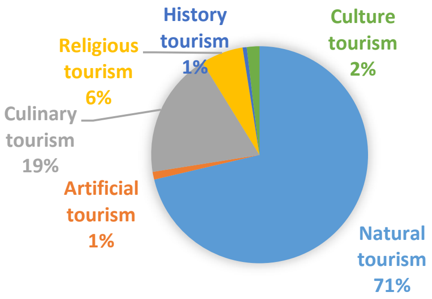

Fig 1. Distribution of tourism category tweet data

Then the data will be labeled to create Groundtruth. In the labeling process, this research compares the labeling method using the manual Lexicon-based and Python library Textblob. This dataset is preprocessed so as to produce terms that will later be extracted. Feature extraction is done using TF-IDF. Feature data is used in the classification process so as to produce a sentiment label.

The stages of the sentiment analysis process are shown in Figure 2. The stages carried out from the proposed methods are preparing the dataset, feature extraction, classification process, and evaluation. In preparing the dataset, there is a Twitter data scraping process, data cleaning process, preprocessing, and dataset labeling process. The label sentiments used as the target class are 'positive', 'negative', and 'neutral'. The evaluation process is carried out by using a confusion matrix to determine the evaluation value of analysis sentiment. Hypertuning parameters are performed during the classification process to form the best model. The best model is used by data testing to predict data sentiment.

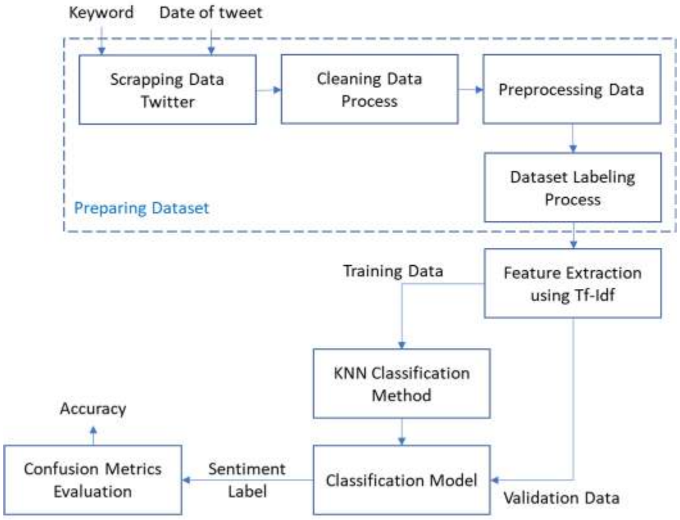

Fig 2. Proposed methods of sentiment analysis

B. *Scrapping Data Twitter*

The Twint Python library is used for the scrapping process. The additional configuration to filter data search, since, until, and output. In the configuration, the search is to enter the keywords used.

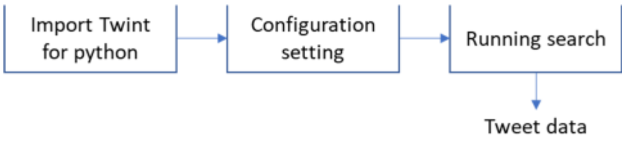

Fig. 3 Stages of Scraping data using Twint

Configuration since inputted time to start scrapping data is April 1, 2020. Configuration until inputted time to finish scrapping data is August 30, 2021. Configuration output is used to save tweet data from scrapping by a specific file name.

C. *Preprocessing Data*

Before preprocessing the data, a data cleaning stage is passed. Data Cleaning Process is cleaning tweet data that is not a review but in the form of information. In addition, the deleted tweet data is duplicate tweet data, and this happens because users often retweet the last tweet data. This kind of data needs to be deleted and not included in the dataset formation process.

Preprocessing is carried out on the data resulting from the cleaning process. The preprocessing stages carried out are case folding, tokenizing, stopword removal, stemming using the python library sastrawi.

D. Dataset Labeling

The process of labeling tweet data is done by comparing the concept of manual Lexicon-based and Text Blob. The manual lexicon-based concept analyzes data by looking at the context of the sentiment lexicon of the words used in composing sentences. This process requires a lexicon dictionary according to the language used in the tweet data. The dictionary produced from the research [10] is used for the Indonesian lexicon dictionary. This lexicon dictionary has 6,609 negative and 3,609 positive words with scoring between -5 to +5. So the label is seen based on the total scoring value. A negative score means negative sentiment, a score of 0 means neutral sentiment, and an upbeat score means positive sentiment.

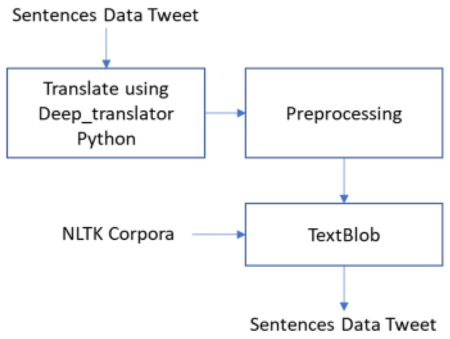

Fig 4. Diagram of Labelling Process using TextBlob

Using the Python library: Textblob is a tool for sentence-level sentiment analysis. This textblob is also lexicon-based; only the corpus is taken from the NLTK corpora [14]. Polarity is taken based on the maximum number of words in the positive, negative, and neutral categories. The polarity score is worth -1 to 1, and there is a subjectivity value worth 0 to 1. The problem is that the corpora NLTK is a collection of English words, so that a translator will be needed for Indonesian documents.

E. Feature Extraction using TF-IDF

This research uses the TF-IDF feature obtained from tweet data. This feature is expected to represent and characterize in a review that has a specific polarity of sentiment [15]. Term Frequency (TF) is the value of the occurrence of a word in the document. Document Frequency (DF) describes how many documents contain a certain word. Each document will have a TF-IDF feature used in the document classification process. The TF-IDF formulation, according to [16], is as follows:

$$TF_{m,k} = \frac{X_{m,k}}{\sum_n X_{n,k}} \quad (1)$$

$$DF_{m,k} = \frac{|d_k \in D: X_k \in d_k|}{|D|} \quad (2)$$

$$IDF_{m,k} = log \frac{|D|}{|d_k \in D: X_k \in d_k|} \quad (3)$$

$$TF - IDF = TF \ x \ IDF \quad (4)$$

Where :
$|D|$ = total documents
$|d_k \in D: X_k \in d_k|$ = number of documents that have term X_k
$X_{m,k}$ = number of occurences term X_k in document d_k
$\sum_n X_{n,k}$ = number of occurences all term in document d_k

F. Classification Process

Three classification methods are used, namely the K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Naïve Bayes methods.

In KNN, various types of calculations identify a distance between the test sample and the training data. The distance similarity measure is an important role for final classification results. Euclidean distance is one of the most frequently used similarity measure methods in the KNN classification [17]. In this research, a comparison of accuracy with similarity measures using Euclidean Distance, Cosine, and Manhattan will be carried out. The K value also affects the accuracy [18], so a test scenario will also be carried out by changing the K value.

Classification techniques is used to determine the class of sentiment document. The methods often used by previous studies are SVM [15][19] and Naïve Bayes [20]. This study will compare the method with the best KNN model after the hypertuning parameter process is carried out.

The evaluation of the system that will be used is accuracy, recall, precision, and F-Measure. Here is the formula that will be used:

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

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

$$Precision = \frac{TN}{(TN+FP)} \quad (7)$$

$$F - Measure = 2 \ x \frac{Recall \ x \ Precision}{(Recall+Precision)} \quad (8)$$

Where:

TN = True Negative
TP = True Positive
FP = False Positive
FN = False Negative

III. Result And Discussion

The data processed in this sentiment analysis process amounted to 522 tweets. There is a test scenario in the Labeling Process and Classification Method.

A. Labeling Process Testing

The system testing results using dataset labeling from Lexicon-based and TextBlob are shown in Table 2. The use of

TextBlob using KNN classifiar in the system produces higher accuracy than Lexicon Based, which is 0.94. Although it is better than Lexicon Based, this model does not provide optimal accuracy values. It is possible because the NLTK corpora use English, so the translator process can also affect the accuracy of the results.

TABLE II. COMPARING ACCURACY RESULT USING LEXICON BASED AND TEXTBLOB

Labeling model	Accuracy
Lexicon Based	0,58
TextBlob	**0,94**

From Table 2, it is known that the use of Text Blob is better than manual Lexicon based. Figure 5 is the data distribution based on the results of labeling with Text Blob for each category of tourist attractions. From Figure 5, it can be seen that the most positive labels were obtained sequentially for three tourism categories: category of nature tourism, culinary tourism, and religious tourism.

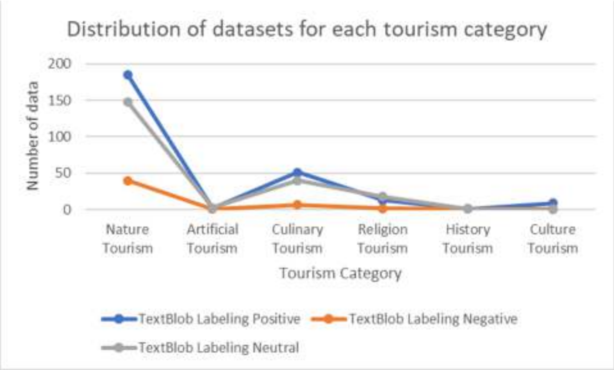

Fig 5. Distribution of dataset for each tourism category

B. *Classification Process Testing*

In the classification process testing, two scenarios are carried out. The scenario use a split dataset, 80% training and 20% testing. The first test scenario measures the system's accuracy using the KNN method by tuning the K-value parameter and metric of the similarity measure. The results of the comparison of accuracy are shown in Table 3.

TABLE III. COMPARING ACCURACY USING KNN CLASSIFICATION WITH TUNING PARAMETER K VALUE AND SIMILARITY MEASURE

K value	Metric Similarity Measure	Accuracy
K=1	Cosine Similarity	0.93
K=3	Cosine Similarity	0.90
K=5	Cosine Similarity	0.89
K=10	Cosine Similarity	0.84
K=1	**Manhattan**	**0.94**
K=3	Manhattan	0.88
K=5	Manhattan	0.89
K=10	Manhattan	0.84
K=1	Euclidean Distance	0.93
K=3	Euclidean Distance	0.87
K=5	Euclidean Distance	0.91
K=10	Euclidean Distance	0.82

From Table 3, it can be seen that the best K-value is 5 with a metric similarity measure using Cosine Similarity. For Fig.6, we can analyze that the value of K-value = 1 is the peak value of the system's maximum accuracy for the overall metric similarity measure.

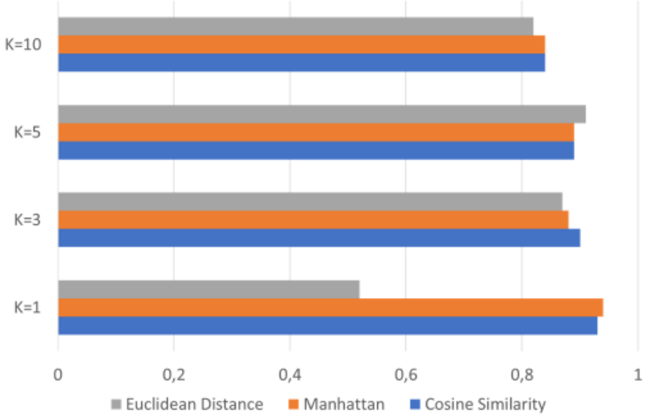

Fig 6. Graph of Kvalue - similarity measure on KNN method

The second test scenario compares the system accuracy values using three classification methods: KNN (K-value=1, Manhattan similarity), SVM, and Naïve Bayes. Table 4 shows the experimental results obtained in the test. The comparison of the three classification methods shows that the classification with the SVM method or KNN obtains the highest accuracy than the others, which is 0.94.

TABLE IV. COMPARATION RESULT OF CLASSIFICATION METHOD

Method	Accuracy	Recall	Precision	F-Measure
KNN	**0.94**	0.92	0.94	0.94
SVM	**0.94**	0.92	0.94	0.94
Naive Bayes	0.93	0.92	0.93	0.93

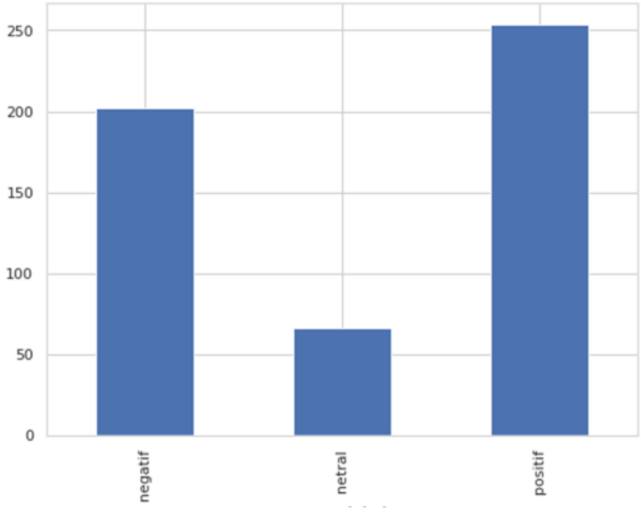

Fig 7. Graph of distribution polarity on tweet data

The selection of algorithm performance that can be used as a reference is generally seen from the amount of FN and FP data. If the value is close to or symmetric, then the best reference that can be used is the accuracy value, but the F-Measure value is the reference if it is not symmetric.

Figure 6 shows the distribution of sentiment labels for classified tweet data; it can be seen that Madura Tourism is still considered reasonable by the public, with positive reviews that are still higher than negative reviews.

IV. Conclusion

Based on the experimental results that have been carried out, it can be concluded that Madura Tourism is still considered reasonable by the community, as evidenced by the high polarity value of tweet review data for positive sentiment compared to negative sentiment, which is 48.7%.

This research found that the labeling process using Text Blob produces better accuracy for the own dataset than manual lexicon-based. Based on the data distribution of Text Blob labeling results, it is known that the most positive labels are obtained sequentially for three tourism categories, namely: category of nature tourism, culinary tourism, and religious tourism.

From the test scenario, it is found that sentiment analysis uses the SVM Method or KNN method with a K-value of 1, and the metric used is Manhattan which has the best accuracy value, which is 94%. The translation results for tweet before labelling process using TextBlob, word correction method and POS Tagging in tweet reviews can affect the evaluation result.

Acknowledgment

This research is part of the Independent Research 2021 with the research topic of Madura tourism trend prediction in LPPM Universitas Trunojoyo Madura.

References

[1] M. T. Jaenuddin and P. Independen, "Upaya Peningkatan Pendapatan Asli Daerah melalui Pengembangan Pariwisata di Kabupaten Mamuju," *Goverment: Jurnal Ilmu Pemerintahan*, vol. 12, no. 2, pp. 67–71, 2019.

[2] W. A. S. Hootsuite, "DIGITAL 2021: The Latest Insights into The 'State of Digital,'" 2021.

[3] G. W. Tan, V. Lee, J. Hew, and K. Ooi, "Telematics and Informatics The interactive mobile social media advertising : An imminent approach to advertise tourism products and services ?," *Telematics and Informatics*, vol. 35, no. 8, pp. 2270–2288, 2018.

[4] F. Mirzaalian and E. Halpenny, "Exploring destination loyalty : Application of social media analytics in a nature-based tourism setting," *Journal of Destination Marketing & Management*, vol. 20, no. March, p. 100598, 2021.

[5] S. Ainin, A. Feizollah, N. B. Anuar, and N. A. Abdullah, "Sentiment analyses of multilingual tweets on halal tourism," *Tourism Management Perspectives*, vol. 34, no. January 2019, p. 100658, 2020.

[6] S. Yoo, J. Song, and O. Jeong, "Social media contents based sentiment analysis and prediction system," vol. 105, pp. 102–111, 2018.

[7] X. Li, R. Law, G. Xie, and S. Wang, "Review of tourism forecasting research with internet data," *Tourism Management*, vol. 83, no. October 2020, 2021.

[8] Z. Abbasi-moud, H. Vahdat-nejad, and J. Sadri, "Tourism recommendation system based on semantic clustering and sentiment analysis," *Expert Systems With Applications*, vol. 167, no. May 2020, p. 114324, 2021.

[9] A. Ali, "Sentiment Analysis on Twitter Data using KNN and SVM," vol. 8, no. 6, pp. 19–25, 2017.

[10] F. Koto, "InSet Lexicon : Evaluation of a Word List for Indonesian Sentiment Analysis in Microblogs InSet Lexicon : Evaluation of a Word List for Indonesian Sentiment Analysis in Microblogs," no. December, 2017.

[11] S. Kunal, A. Saha, A. Varma, and V. Tiwari, "Textual Dissection Of Live Twitter Reviews Using Naive Bayes," *Procedia Computer Science*, vol. 132, no. Iccids, pp. 307–313, 2018.

[12] F. Rustam, R. Khan, K. Kanwal, R. Y. Khan, and G. S. Choi, "US Based COVID-19 Tweets Sentiment Analysis Using TextBlob and Supervised Machine Learning," pp. 1–8, 2021.

[13] R. Guzman-cabrera, "Exploring the use of lexical and psycho-linguistic resources for sentiment analysis," in *Mexican International Conference on Artificial Intelligent, MICAI*, 2020, no. December, pp. 109–121.

[14] V. Bonta, N. Kumaresh, and N. Janardhan, "A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis," no. March, 2019.

[15] S. S. and P. K.V., "Sentiment analysis of malayalam tweets using machine learning techniques," *ICT Express*, no. xxxx, pp. 2–7, 2020.

[16] S. W. Kim and J. M. Gil, "Research paper classification systems based on TF - IDF and LDA schemes," *Human-centric Computing and Information Sciences*, pp. 9–30, 2019.

[17] M. Sudarma and I. G. Harsemadi, "Design and Analysis System of KNN and ID3 Algorithm for Music Classification based on Mood Feature Extraction," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 7, no. 1, pp. 486–495, 2017.

[18] K. N. Classifier, V. B. S. Prasath, H. Arafat, A. Alfeilat, and O. Lasassmeh, "Distance and Similarity Measures Effect on the," pp. 1–50.

[19] A. Borg and M. Boldt, "Expert Systems with Applications Using VADER sentiment and SVM for predicting customer response sentiment q," *Expert Systems With Applications*, vol. 162, p. 113746, 2020.

[20] Y. Tan and P. P. Shenoy, "A bias-variance based heuristic for constructing a hybrid logistic regression-naïve Bayes model for classification," *International Journal of Approximate Reasoning*, vol. 117, pp. 15–28, 2020.