SEMARAK ILMU PUBLISHING

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Clustering on Sentiment Analysis: Effect of Twitter Dataset

Sri Redjeki[2], Satria Abadi[1,*], Deborah Kurniati[2], Sri Rezeki Candra Nursari[3], Ariesta Damayanti[2], Edi Iskandar[2]

[1] Faculty of Computing and Meta Technology, Universiti Pendidikan Sultan Idris, Perak, Malaysia
[2] Indonesia Digital Technology University, Yogyakarta, Indonesia
[3] Universitas Pancasila Jakarta, Indonesia

ARTICLE INFO

Article history:
Received 23 October 2023
Received in revised form 29 December 2023
Accepted 6 August 2024
Available online 2 September 2024

Keywords:
Auto Labeling; Clustering; Deep Learning; LSTM; Sentiment Analysis

ABSTRACT

The process of labeling text datasets presents a challenge in sentiment analysis, especially those done manually. This is because it takes time, effort, and skill which is taxing in Twitter data labeling. This study aims to auto-label Twitter dataset using a clustering approach to classify tourism twitter sentiment using one of the LSTM (Long Short Term Memory) deep learning algorithms. The clustering used for the auto labeling process is K-means, while the deep learning sentiment classification used is LSTM. The research datasets consist of 10,228 tweets about Yogyakarta tourism in Indonesia. The Twitter data language used in this study is Indonesian. The classification process using LSTM is carried out twice, the first process uses a manual label dataset, and the second process uses an auto-labeling dataset. The sentiment class is divided into 3, namely negative, positive and neutral. The results indicates that the classification of tourism twitter sentiment using the auto-labeling dataset provide better accuracy results than the manual-labeling dataset. LSTM classification model with auto-labeling dataset produces optimum graphs with an average accuracy of 99% while manual-labeling datasets produce overfitting charts with an average accuracy of 40%. The results showed that the auto-labeling process of the class dataset using K-Means clustering can improve the accuracy of the classification results of Yogyakarta tourism Twitter sentiment. The model produced in this study can help in solving class labeling problems in sentiment classification.

1. Introduction

In recent decades, information technology's rapid growth has made the distribution of information becomes particularly relevant [1]. The introduction of social media has given Internet users the ability to express and share their opinions and views on various issues and events [2]. Twitter is one of the social media that is overgrowing in processing public opinion. In particular, Twitter provides a platform on which discussions on various topics can be detected more rapidly than other standard information channels. In the scientific literature, Twitter can offer a more fertile

* *Corresponding author.*
E-mail address: satriaabadi@meta.upsi.edu.my

https://doi.org/10.37934/araset.51.1.3951

ground for sentiment analysis than Facebook. The main reason is that emoticons icons can easily extract twitter data and tweets.

A very large number of very short messages created by the users of this microblogging site are included in Twitter. From personal opinions to official statements, the contents of the messages differ. Data from these sources can be used in opinion mining and sentiment analysis tasks as the audience of microblogging sites and services expand daily [3].

Analysis of social media data that mostly used to identify public opinion is sentiment analysis. This method has been widely used to predict an organization's early identification in capturing social media views [4]. Sentiment analysis is one of computer science's fastest-growing research fields, making it difficult to track all field activities [5]. One of the problems that arise in sentiment analysis is the exceedingly large number of datasets assigned labels or classes. The raw information used to train these classifiers is plentiful and easy to gather, but labels are missing. Labels are important to train supervised classifiers; however, considerable human effort is needed for the labeling process [6].

Also, because of the unstructured nature of reviews, labeling is a challenging job, with many of them featuring slang, lack of punctuation, and inaccurate grammar [7]. It is costly to procure training labels for specific text classification issues, although vast amounts of unlabeled documents are readily accessible [8]. Several studies have developed auto labeling techniques to classify dataset labels through grouping or classifying datasets before using a classifier algorithm [6,8-10]. Text datasets with clear class labels in sentiment analysis will increase the classifier algorithm in developing a successful classification. A large volume of Twitter slang poses a challenge for researchers to collect tweet data accurately.

The research objective is to develop an auto labeling model on the Twitter dataset using the K-Means clustering approach to classify Twitter sentiments regarding tourism in Yogyakarta-Indonesia. Yogyakarta is one of the tourist destinations for local and foreign tourists in Indonesia, where the number of tourist visits has increased annually. This is because Yogyakarta has many types of tourism, namely nature, art, culture, and culinary tourism [10-12].

The sentiment classification used in this study is one of deep learning algorithms, namely LSTM. Several studies on sentiment classification texts using LSTM [10-13] can be a classifier algorithm on the topic of text mining or sentiment analysis. Sentiment analysis regarding tourist opinion [15,16] is very useful for decision-makers because sentiment analysis results show tourist satisfaction on tourist attraction’s management in an area. This study is related to tourism located in one of famous tourist destinations in Indonesia, namely Yogyakarta. This paper's structure consists of section 1 for introduction, section 2 for literature review, section 3 for research framework, section 4 for results and discussion, and the last, section 5 for conclusion.

2. Literature Review

Sentiment analysis, which is a part of text mining, is a collection of methods, techniques, and tools for detecting and extracting personal information from a sentence in a language, such as opinions and attitudes [17,18]. Sentiment Analysis or Opinion Mining is a way to compare people's emotions, attitudes, and opinions about a particular entity. Entities can represent individuals, events, or topics [19]. The target of sentiment analysis is to find opinions, feedback, or reviews, and then identify the sentiments they want to express and then classify their polarity as positive, negative, or neutral [20]. Sentiment analysis has become a leading framework for scientific and commercial market research. It examines people's thoughts, views, actions, attitudes, and emotions towards

individuals, organizations, goods, services, problems, and their qualities in written or spoken language [21,22,35,36].

Online data extraction and interpretation, polarity and subjectivity, feature collection, sentiment analysis of comparative sentences, opinion quest, and retrieval discovery are the majority of sentiment analysis studies. One of the service sectors that uses public opinion to improve its services is tourism. Tourists are now able to access numerous sources of knowledge and create their content and share their thoughts and experiences. In terms of both reputation and results, tourism content exchanged via social media has become a significant source of knowledge that impacts tourism [23-25]. Public opinion regarding tourism in an area becomes the content of social media accounts that will be an input for tourism managers or the government [25]. Sentiment analysis can manage opinions using the classification method.

Long Short Term Memory (LSTM) is a deep learning algorithm that can classify opinions or sentiments well. Deep learning has made significant breakthroughs in text categorization tasks in recent years, and the network model based on deep learning technology has also produced better classification results in aspect-based affective analysis tasks than conventional machine learning approaches. Sentiment analysis can manage opinions using the classification method [26]. Several studies on sentiment analysis related to tourism (tourist spots, travel, trips, tourism facilities) using the LSTM deep learning algorithm show classification results with good accuracy [27-29]. LSTM which is part of deep learning recurrent neural network has the form of a series of recurring neural network modules. LSTM has the same structure but has an additional feature in the form of a gate on the cell [30,31,34].

Text mining research provides maximum results in the classification or identification of opinions if many datasets support it. The problem arises when the necessary information used to train this classifier is large and easy to collect, but the label is missing [7]. The process of labeling class to very large datasets is a challenging task as it is time-consuming, expensive, and has the potential for manual mislabelling.

The automatic labeling process uses the K-Means grouping method. The data grouping method is one of the unsupervised learning techniques, which means that the clustering method can group data into several groups. One of the most widely used classification methods is the K-Means Method [31,32].

2.1 Long Short Term Memory (LSTM)

Long Short Term Memory or commonly abbreviated as LSTM, is a special form of RNN that can carry out learning on long-term dependencies. This model was introduced by Hochreiter and Schmidhuber in 1997. LSTM, which is part of a deep learning recurrent neural network, has the form of a series of recurring neural network modules. LSTM has the same structure but has an additional feature, such as a gate on the cell. Figure 1 shows the details of the LSTM structure.

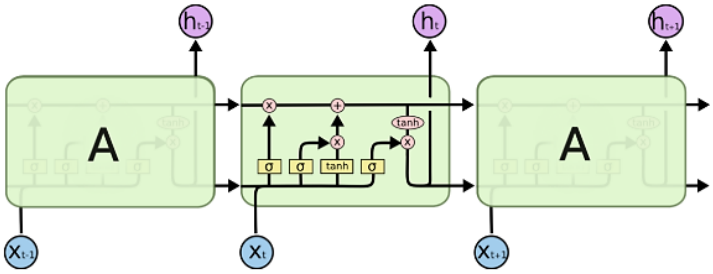

Fig. 1. LSTM structure

The LSTM will determine what information will be removed from the cell. This decision is made by the forget gate layer (see Figure 2). This layer will pay attention to h_{t-1} and x_t so that it will produce output between 0 and 1. The mathematical equation can be seen in Eq. [(1). Output 0 represents that information will be forgotten, while output 1 represents that information will not be forgotten.

$$f_t = \sigma(W_f . [h_{t-1}, x_t] + b_f) \quad (1)$$

where f_t is a function to produce outputs 0 and 1.

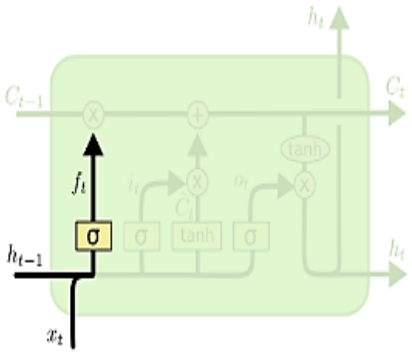

Fig. 2. Forget structure layer

A sigmoid layer called the input gate layer determines which values to update can be seen in Eq. (2). Next, a tanh layer creates a vector of the new candidate value, C_t, which can be added to the state. The Equation can be seen in Eq. (3). These two layers will be combined to update the state (see in Figure 3).

$$i_t = \sigma(W_i . [h_{t-1}, x_t] + b_i) \quad (2)$$

$$C_t = tanh(W_c . [h_{t+1}, x_t] + b_C) \quad (3)$$

Where i_t is an activation function that functions as an input gate layer.

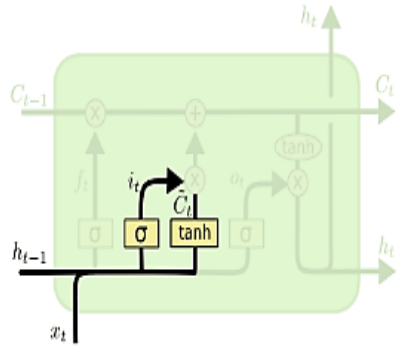

Fig. 3. Remember gate structure

The old state will be updated, C_{t-1} to the new cell state C_t (see Figure 4). Then, f_t will be multiplied by the old state, ignoring the previously forgotten information. Then, i_t is added with C_t, the equation can be seen in Eq. (4).

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t) \quad (4)$$

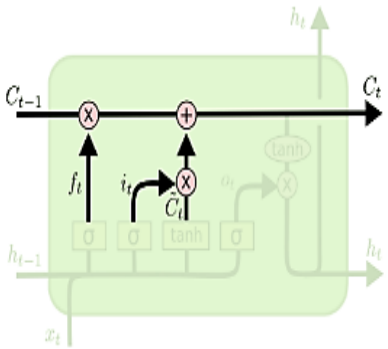

Fig. 4. Update layer structure

The final step is to determine what the output is O_t. The sigmoid layer will determine the part of the cell that will be removed, the equation can be seen in Eq. (5). Then, the cell is passed to the tanh layer (to force the output value between -1 and 1) and multiplied by the output of the sigmoid gate, this equation can be seen in Eq. (6).

$$O_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = O_t * tanh(C_t) \tag{6}$$

2.2 K-Means Clustering

The data grouping method is an unsupervised learning technique where the clustering method can group data into several groups. One of the clustering methods that is widely used is the K-Means Method [33]. K-Means is able to minimize the average distance of each data to the cluster. Mac Queen developed this method in 1967. This method partitioned data into groups so that data that had the same characteristics were grouped into the same group, and data that had different characteristics were grouped into other groups.

The K-Means algorithm has several stages, namely

i. Input: *n* points, distance function *d*(), number *k* of clusters to find.
ii. Start with *k* centers.
iii. Compute *d* (each point *x*, each center *c*).
iv. For each *x*, find closest center *c*(*x*).
v. If no point has changed "owner" *c*(*x*), stop.
vi. Each *c* ← mean of points owned by it.
vii. Repeat from ii.

Three user-defined parameters are required for the K-means algorithm: number of K clusters, cluster initialization, and distance metric. K is the most critical choice. The data mean it is defined below in Eq. (7).

$$C_i = \frac{1}{M}\sum_{j=1}^{M} x_j \tag{7}$$

Calculating the distance d can use several distance formulas, including Euclidean, Manhattan/City Block, and Minkowsky. The allocation of owner membership uses equation in Eq. (8).

$$a_{il} = \begin{cases} 1, & d = min\{D(x_i, C_l)\} \\ 0, & lainnya \end{cases} \quad (8)$$

The K-means algorithm finds the partitions in such a way that the squared error between the empirical mean of a cluster and the points in the cluster is minimized.

3. Research Methodology

The research object used to develop an auto labeling model on the Twitter dataset using the K-Means clustering approach in the classification of opinions related to tourism in Yogyakarta-Indonesia as many as 10228 tweet datasets using Indonesian. The data is obtained through a crawling process using keywords related to some of the most visited tourist objects in Yogyakarta. The stages of the research process carried out are shown in Figure 5. This study will compare the LSTM classification results' accuracy using a dataset that is labeled automatically and manually.

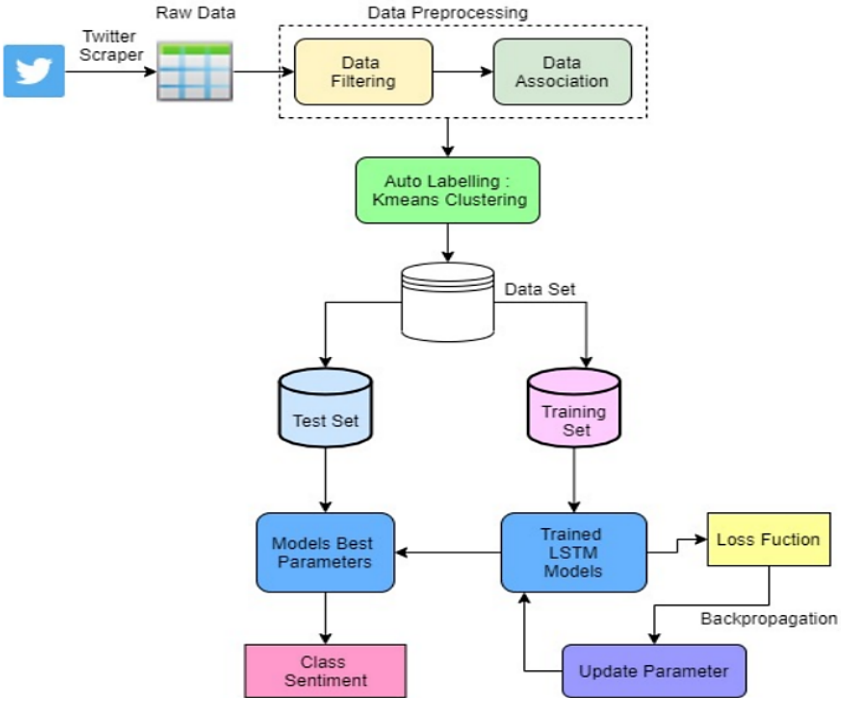

Fig. 5. Research framework

The data source comes from Twitter with the Yogyakarta tourism data item. Primary research data in the form of Twitter data requires special steps before the data is processed into sentiment analysis. The details of the stages carried out in this study are Figure 5, namely

I. Collecting papers on tourism in Yogyakarta.
II. Create a word cloud to find the number of words that frequently appear from the collected paper.
III. Perform data crawling using keywords from several words from word cloud processing.
IV. Data crawling results that change Twitter's raw data are stored in CSV form, and after that, the preprocessing stages are carried out, namely filtering and association.
V. Data filtering has the following stages.
   a. Converting to Lowercase - Steps are taken to convert tweets to lowercase. This prevents the same word from having multiple copies.

b. Removes punctuation marks because when handling text data, it doesn't add any extra details. Removing all these examples will also help us reduce the size of the training results.
c. Stop Words Removal should be removed from text data with words that appear frequently.
d. Deletion of unusual words: Remove rare words from the text. Because they are very rare, noise dominates the relationship between them and other words. For more general forms, you can substitute unfamiliar words and these will have a larger number.
e. Spelling correction - We've all seen tweets with lots of misspellings. Our timeline is sometimes filled with hastily sent tweets that are often almost unreadable. We'll be using the TEXT BLOB python library for this.
f. Tokenization refers to breaking the text into a series of words or phrases.
g. Lemmatization, instead of simply eliminating enough, turns the word into the root word. To get the roots, use vocabulary and perform morphological analysis. After the data filtering stage continues to look for associations between words in each tweet.

Manually marking large amounts of training data is the most difficult method of sentiment analysis. The automatic labeling of datasets in research using the K-Means algorithm is grouped into 3, namely positive, neutral and negative. Furthermore, the dataset that has been labeled using the K-Means clustering approach is divided into 2 parts, namely the 80% dataset for training and 20% dataset for testing. The algorithm used for sentiment classification in this study is LSTM. The software used in this study is based on python, namely Google Colab by using several existing libraries in python [34,37].

4. Result and Discussion

This section will explain the results and discussion related to the research process starting from the input, process, and output. A display of a portion of the dataset that has passed the preprocessing stage is shown in Table 1.

Table 1
Overview dataset

	Text	Times tamp
0	Malioboro area looking for top	1/13/18 23:34
1	Some people think they study in an old school city…	1/31/18 22:06
2	When I was a fan of Tatu, I bought the CD…..	1/31/18 18:47
3	You've never been in a relationship, what do you do?...	1/31/18 18:15
4	Yes, it's raining in Malioboro	1/31/18 17:09

This study compares the ability of LSTM to classify Twitter sentiments using datasets that are labeled automatically and manually. The dataset that has been labeled manually will then be analyzed using the sentiment analyzer in Python software. The sentiment analyzer produces output scores for three classes (negative, positive, neutral) called compound_score and result_sentiment. The results of automatic labeling through a sentiment analyzer are then tested by comparing the results of labeling using K-means clustering. The K-means results show a change in the labeling of the dataset so that there is a change in the amount of data for each class (Figure 6). This Figure shows the number of each class labeling results using the K-Means clustering algorithm between manual labeling and the sentiment analyzer.

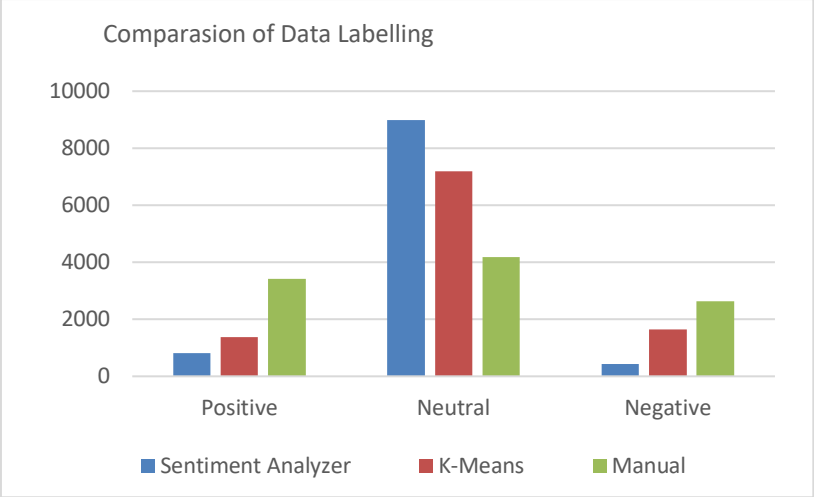

Fig. 6. Comparison of labeling methods

Furthermore, the dataset with manual labeling data and the K-means algorithm will be classified using one of the deep learning algorithms, namely LSTM. The training data uses 80% of the dataset, and for testing, data uses 20% of the dataset. The LSTM model parameters are shown in Figure 7.

```
model.summary()

Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, None, 64)          1118528
_________________________________________________________________
spatial_dropout1d (SpatialDr (None, None, 64)          0
_________________________________________________________________
bidirectional (Bidirectional (None, 128)               66048
_________________________________________________________________
dense (Dense)                (None, 64)                8256
_________________________________________________________________
dropout (Dropout)            (None, 64)                0
_________________________________________________________________
dense_1 (Dense)              (None, 3)                 195
_________________________________________________________________
activation (Activation)      (None, 3)                 0
=================================================================
Total params: 1,193,027
Trainable params: 1,193,027
Non-trainable params: 0
```

Fig. 7. LSTM model parameters

The summary of the LSTM model displayed by python software in Figure 8 shows there are seven layers in the LSTM model, namely the embedding for text input consisting of 64 neurons, spatial_dropoutid to reduce overfitting consisting of 64 neurons, bidirectional consisting of 128 neurons, a dense layer consisting of 64 neurons, the dropout layer consists of 64 neurons with a dropout value of = 0.3, the dense_1 layer consists of 3 neurons, and the activation layer consists of 3 neurons that use the softmax function and the adam optimizer. The above LSTM model parameters are used to classify manually labeled datasets using K-means. The results of the LSTM classification accuracy of the two types of datasets are shown in Table 2 and Table 3.

Table 2
LSTM performance using manual labeling

Epoch	Final Accuracy	Final ValAccuracy	Final Loss	Final Val_Loss
20	0.6547	0.4435	0.7799	1.3936
30	0.7582	0.4341	0.6012	2.3793
40	0.8042	0.4292	0.5122	3.4888
50	0.8297	0.4440	0.4586	4.3008
60	0.7769	0.4194	0.5613	3.0731
70	0.8707	0.4248	0.3841	4.4927
80	0.8110	0.4351	0.5029	3.9519
90	0.8689	0.4410	0.3785	4.9013
100	0.8364	0.4400	0.4523	4.3072

This study conducted LSTM training experiments from epoch 20 to 100 for manual-labeling and auto-labeling datasets. Table 2 shows the LSTM performance for the manual datasets in each epoch. In the manual-labeling dataset, LSTM performance shows an overfitting curve, which means that the accuracy and val_accuracy values show significantly different values on the epoch to 20 to 100 with an average accuracy of 40% (see Figure 8).

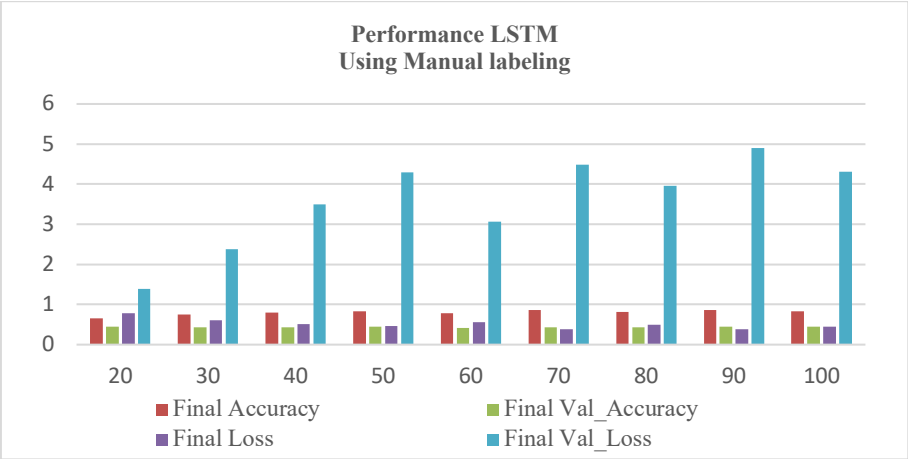

Fig. 8. Comparation of manual labeling result

Table 3 shows the performance results of the LSTM model using auto labelling. This is different from the auto-labeling dataset which produces a good fit or optimum curve, which means that the accuracy and val_accuracy values are almost the same at the 20th to 100th epoch with an average final accuracy of 99.6% and an average val_accuracy of 99.2%.

Table 3
LSTM performance using auto-label

Epoch	Final Accuracy	Final ValAccuracy	Final Loss	Final Val_Loss
20	0.9864	0.9932	0.057	0.0242
30	0.9945	0.9912	0.0232	0.0678
40	0.9954	0.9888	0.0202	0.2249
50	0.9939	0.9946	0.027	0.0348
60	0.9962	0.9956	0.018	0.1206
70	0.9936	0.9951	0.0261	0.0273
80	1.0000	0.9927	0.000015	0.0564
90	1.0000	0.9907	0.0000016	0.0725
100	1.0000	0.9922	0.0000001	0.0526

The graph in Figure 9 shows the values for final accuracy, final Val_Accuray, Final Loss, and Final Val_loss for each epoch for a dataset labeled using K-Means clustering. The data shows that the Final Accuracy and Final Val_Acurracy values are almost the same, especially at the 50th and 70th epoch at 99.5%.

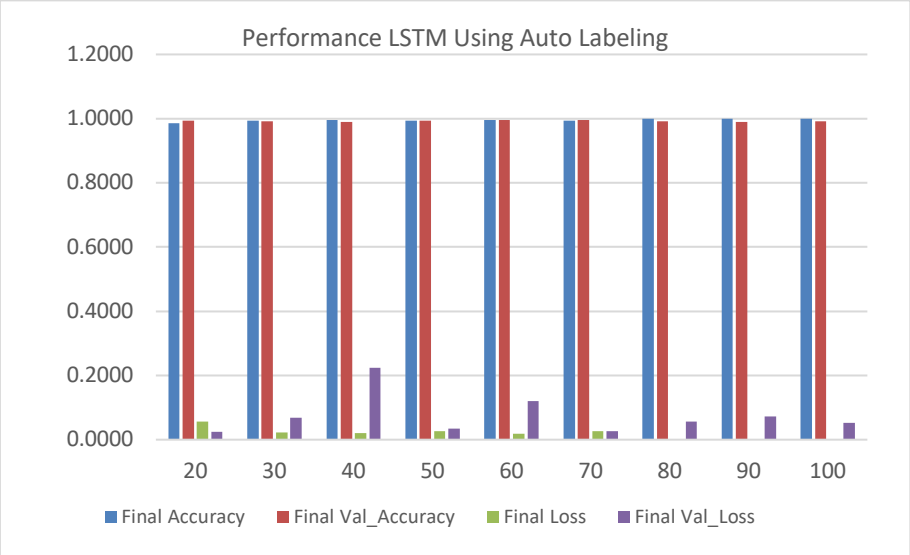

Fig. 9. Comparation of auto labeling result

Detailed data on the LSTM performance of the auto-labeling dataset for each epoch 20 to 100 in Table 3 shows that epoch 50 and epoch 70 have almost the same final accuracy and final val_accuracy values, namely 99.5%.

Meanwhile, the final loss and val_loss values at epoch 50 (see in Figure 10) and epoch 70 show the same value, namely 3%. The highest final accuracy value is shown by epoch 80 to 100, which is 100%, but this value has a higher difference to the final val_accuracy value compared to epoch 50 and epoch 70. Besides, the final_loss and final val_loss epoch values 80 to 100 have quite a difference. take effect.

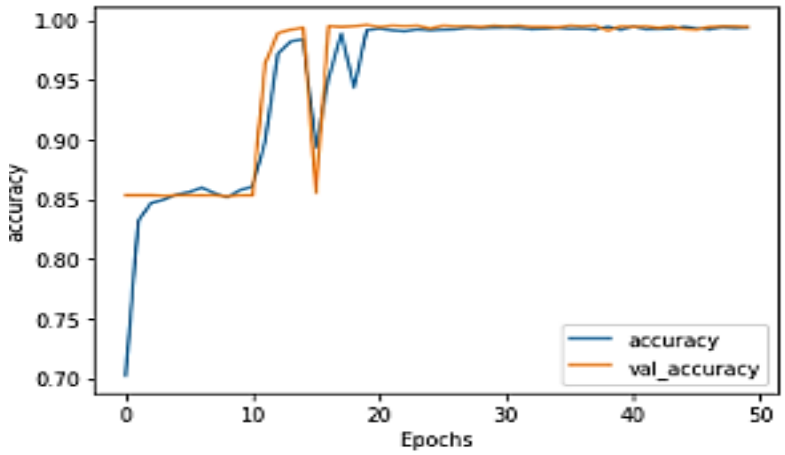

Fig. 10. Training performance of epoch 50

5. Conclusions

The dataset in big data research is very important, especially sentiment analysis research in the scope of text mining; this is because the data collected must be large and have the correct class label.

The process of labeling a dataset in the text requires a lot of energy and time when done manually. This study succeeded in comparing the use of manually labeled datasets and using the K-Means clustering process for auto-labeling on the tourism dataset in classifying sentiments regarding tourist opinion in Yogyakarta-Indonesia. The classification algorithm of tourist opinion sentiment via Twitter using LSTM shows that the auto-labeling class's classification results provide optimum results with average accuracy and validation accuracy of 99% at epoch 50 and 70. While the classification using a dataset with manual class labeling gives overfitting results for all epochs trained in the model. The results of the study note that the process of class labeling on large text datasets can be done through auto-labeling using the K-Means clustering approach. This process can reduce the inconsistency of class labels and increase the accuracy of the model results.

References

[1] Troussas, Christos, Maria Virvou, and Spyridon Mesaretzidis. "Comparative analysis of algorithms for student characteristics classification using a methodological framework." In *2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pp. 1-5. IEEE, 2015. https://doi.org/10.1109/IISA.2015.7388038

[2] Krouska, Akrivi, Christos Troussas, and Maria Virvou. "The effect of preprocessing techniques on Twitter sentiment analysis." In *2016 7th international conference on information, intelligence, systems & applications (IISA)*, pp. 1-5. IEEE, 2016. https://doi.org/10.1109/IISA.2016.7785373

[3] Pak, Alexander, and Patrick Paroubek. "Twitter as a corpus for sentiment analysis and opinion mining." In *LREc*, vol. 10, no. 2010, pp. 1320-1326. 2010.

[4] Hui, June Ling Ong, Gan Keng Hoon, and Wan Mohd Nazmee Wan Zainon. "Effects of word class and text position in sentiment-based news classification." *Procedia Computer Science* 124 (2017): 77-85. https://doi.org/10.1016/j.procs.2017.12.132

[5] Mäntylä, Mika V., Daniel Graziotin, and Miikka Kuutila. "The evolution of sentiment analysis—A review of research topics, venues, and top cited papers." *Computer Science Review* 27 (2018): 16-32. https://doi.org/10.1016/j.cosrev.2017.10.002

[6] Wigness, Maggie, Bruce A. Draper, and J. Ross Beveridge. "Efficient label collection for image datasets via hierarchical clustering." *International Journal of Computer Vision* 126 (2018): 59-85. https://doi.org/10.1007/s11263-017-1039-1

[7] McIlroy, Stuart, Nasir Ali, Hammad Khalid, and Ahmed E. Hassan. "Analyzing and automatically labelling the types of user issues that are raised in mobile app reviews." *Empirical Software Engineering* 21 (2016): 1067-1106. https://doi.org/10.1007/s10664-015-9375-7

[8] Nigam, Kamal, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. "Text classification from labeled and unlabeled documents using EM." *Machine learning* 39 (2000): 103-134. https://doi.org/10.1023/A:1007692713085

[9] Färber, Ines, Stephan Günnemann, Hans-Peter Kriegel, Peer Kröger, Emmanuel Müller, Erich Schubert, Thomas Seidl, and Arthur Zimek. "On using class-labels in evaluation of clusterings." In *MultiClust: 1st international workshop on discovering, summarizing and using multiple clusterings held in conjunction with KDD*, p. 1. 2010.

[10] Rif'an, Achmad Andi. "Tourism Components and Tourists Characteristic of Prambanan Temple as The World Culture Heritage Site in Yogyakarta, Indonesia." *International Journal of Tourism and Hospitality Study* 1, no. 1 (2016): 1-10.

[11] Amanda, Rima, PInsap Santosa, and M. Nur Rizal. "Analysis of tourists preferences on smart tourism in Yogyakarta (Case: Vredeburg Fort Museum)." In *Journal of Physics: Conference Series*, vol. 1007, no. 1, p. 012040. IOP Publishing, 2018. https://doi.org/10.1088/1742-6596/1007/1/012040

[12] Wachyuni, Suci Sandi, and Yudha Dwi Saputro. "Local culinary development strategy as a tourist attraction in Yogyakarta." In *Proceedings the 5th International Conferences on Cultural Studies, Udayana University* (2019): 313–322.

[13] Sundermeyer, Martin, Ralf Schlüter, and Hermann Ney. "LSTM neural networks for language modeling." In *Thirteenth annual conference of the international speech communication association*. pp 194–197, 2012.

[14] Jin, Zhigang, Yang Yang, and Yuhong Liu. "Stock closing price prediction based on sentiment analysis and LSTM." *Neural Computing and Applications* 32 (2020): 9713-9729. https://doi.org/10.1007/s00521-019-04504-2

[15] Lee, Minwoo, Linda L. Lowry, and John D. Delconte. "Social media in tourism research: A literature review." (2015).

[16] Nezakati, Hossein, Asra Amidi, Yusmadi Yah Jusoh, Shayesteh Moghadas, Yuhanis Abdul Aziz, and Roghayeh Sohrabinezhadtalemi. "Review of social media potential on knowledge sharing and collaboration in tourism industry." *Procedia-social and behavioral sciences* 172 (2015): 120-125. https://doi.org/10.1016/j.sbspro.2015.01.344
[17] Araque, Oscar, Ignacio Corcuera-Platas, J. Fernando Sánchez-Rada, and Carlos A. Iglesias. "Enhancing deep learning sentiment analysis with ensemble techniques in social applications." *Expert Systems with Applications* 77 (2017): 236-246. https://doi.org/10.1016/j.eswa.2017.02.002
[18] Ain, Qurat Tul, Mubashir Ali, Amna Riaz, Amna Noureen, Muhammad Kamran, Babar Hayat, and A. Rehman. "Sentiment analysis using deep learning techniques: a review." *International Journal of Advanced Computer Science and Applications* 8, no. 6 (2017). https://doi.org/10.14569/IJACSA.2017.080657
[19] Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. "Sentiment analysis algorithms and applications: A survey." *Ain Shams engineering journal* 5, no. 4 (2014): 1093-1113. https://doi.org/10.1016/j.asej.2014.04.011
[20] Osimo, David, and Francesco Mureddu. "Research challenge on opinion mining and sentiment analysis." *Universite de Paris-Sud, Laboratoire LIMSI-CNRS, Bâtiment* 508 (2012).
[21] Anjaria, Malhar, and Ram Mohana Reddy Guddeti. "A novel sentiment analysis of social networks using supervised learning." *Social Network Analysis and Mining* 4 (2014): 1-15. https://doi.org/10.1007/s13278-014-0181-9
[22] Tuhin, Rashedul Amin, Bechitra Kumar Paul, Faria Nawrine, Mahbuba Akter, and Amit Kumar Das. "An automated system of sentiment analysis from Bangla text using supervised learning techniques." In *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, pp. 360-364. IEEE, 2019. https://doi.org/10.1109/CCOMS.2019.8821658
[23] Fuchs, Matthias, Wolfram Höpken, and Maria Lexhagen. "Big data analytics for knowledge generation in tourism destinations–A case from Sweden." *Journal of destination marketing & management* 3, no. 4 (2014): 198-209. https://doi.org/10.1016/j.jdmm.2014.08.002
[24] Murnawan, Murnawan. "Pemanfaatan Analisis Sentimen Untuk Pemeringkatan Popularitas Tujuan Wisata." *Jurnal Penelitian Pos Dan Informatika* 7, no. 2 (2017): 109-120. https://doi.org/10.17933/jppi.2017.070203
[25] Alaei, Ali Reza, Susanne Becken, and Bela Stantic. "Sentiment analysis in tourism: capitalizing on big data." *Journal of travel research* 58, no. 2 (2019): 175-191. https://doi.org/10.1177/0047287517747753
[26] Liu, Pengfei, Duxian Nie, Xiaxu He, Weifeng Zhang, Zhirui Huang, and Kejing He. "Sentiment analysis of chinese tourism review based on boosting and LSTM." In *2019 International Conference on Communications, Information System and Computer Engineering (CISCE)*, pp. 664-668. IEEE, 2019. https://doi.org/10.1109/CISCE.2019.00154
[27] An, Hyeon-woo, and Nammee Moon. "Design of recommendation system for tourist spot using sentiment analysis based on CNN-LSTM." *Journal of Ambient Intelligence and Humanized Computing* (2022): 1-11. https://doi.org/10.1007/s12652-019-01521-w
[28] Maw, Soe Yu, and May Aye Khine. "Aspect based Sentiment Analysis for travel and tourism in Myanmar Language using LSTM." PhD diss., MERAL Portal, 2019.
[29] Taib, Abidah Mat, and Nurul Nabila Khairu Azman Azman. "Experimental Analysis of Trojan Horse and Worm Attacks in Windows Environment.", *Journal of Advanced Research in Computing and Applications* 13 (1), 1-9. 2018.
[30] Martín, Carlos Alberto, Jesús M. Torres, Rosa María Aguilar, and Sergio Diaz. "Using deep learning to predict sentiments: case study in tourism." *Complexity* 2018 (2018). https://doi.org/10.1155/2018/7408431
[31] Wang, Yequan, Minlie Huang, Xiaoyan Zhu, and Li Zhao. "Attention-based LSTM for aspect-level sentiment classification." In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 606-615. 2016. https://doi.org/10.18653/v1/D16-1058
[32] Gowda, Harsha S., Mahamad Suhil, D. S. Guru, and Lavanya Narayana Raju. "Semi-supervised text categorization using recursive K-means clustering." In *Recent Trends in Image Processing and Pattern Recognition: First International Conference, RTIP2R 2016, Bidar, India, December 16–17, 2016, Revised Selected Papers 1*, pp. 217-227. Springer Singapore, 2017. https://doi.org/10.1007/978-981-10-4859-3
[33] Jain, Anil K. "Data clustering: 50 years beyond K-means." *Pattern recognition letters* 31, no. 8 (2010): 651-666. https://doi.org/10.1016/j.patrec.2009.09.011
[34] Lei, Yuan, Shir Li Wang, Minghui Zhong, Meixia Wang, and Theam Foo Ng. "A Federated Learning Framework Based on Incremental Weighting and Diversity Selection for Internet of Vehicles." *Electronics* 11, no. 22 (2022): 3668. ttps://doi.org/10.3390/electronics11223668
[35] Abadi, Satria, Muhamad Hariz Muhamad Adnan, Sri Redjeki, and Citrawati Jatiningrum. "Using Analytical Hierarchy Process for Double Auction to Optimize Financial Performance of Private Higher Education Institutions." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 31, no. 3 (2023): 13-24. https://doi.org/10.37934/araset.31.3.1324
[36] Adnan, Muhamad Hariz Muhamad, Siti Fatimah Mohamed, Nurul Fazilah Ahmad, Nurul Naqibah Binti Annual, Satria Abadi, and Nor Masharah Husain. "AI Meets Entrepreneurship: A Framework of Web Platform for Enhancing Skills,

Streamlining Finance and Identifying Multiple Intelligence." In *2023 International Conference on Disruptive Technologies (ICDT)*, pp. 318-324. IEEE, 2023.

[37] Zulkifli, C. Z., and N. N. Noor. "Wireless Sensor Network and Internet of Things (IoT) Solution in Agriculture." *Pertanika Journal of Science & Technology* 25, no. 1 (2017).