

Enhance sentiment analysis in big data tourism using hybrid lexicon and active learning support vector machine

Ni Wayan Sumartini Saraswati¹, I Ketut Gede Darma Putra², Made Sudarma¹, I Made Sukarsa²

¹Department of Engineering Science, Faculty of Engineering, Udayana University, Bali, Indonesia

²Department of Information Technology, Faculty of Engineering, Udayana University, Bali, Indonesia

Article Info

Article history:

Received Feb 15, 2024

Revised Mar 12, 2024

Accepted Mar 20, 2024

Keywords:

Active learning

Big data

Lexicon

Sentiment analysis

Support vector machine

ABSTRACT

Sentiment analysis is a review analysis process used to determine whether an opinion is neutral, negative, or positive. Sentiment analysis can be done using lexicon-based or machine learning-based approaches. Lexicon can perform sentiment analysis without training data because it is dictionary-based but performs worse than machine learning. Machine learning can perform well in completing sentiment analysis but requires training data so that the model does not experience underfitting. In the case of sentiment analysis on big data, manual labeling of training data is an inefficient job. Support vector machine (SVM) has the opportunity to be used together with the active learning (AL) method to make small training data but still have good performance. This research proposed a hybrid lexicon and AL-SVM method to complete sentiment analysis on big data tourism. This research used polarity from the valence aware dictionary and sentiment reasoner (VADER) lexicon as a reference for the query by user process from the AL-SVM to automate the sentiment analysis process on big data. The experimental results showed that using the hybrid lexicon and AL-SVM increased the sentiment analysis performance compared to the VADER lexicon, SVM, and lexicon SVM, which run separately.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Ni Wayan Sumartini Saraswati

Departement of Engineering Science, Faculty of Engineering, Udayana University

Denpasar, Bali, Indonesia

Email: sumartini.saraswati@instiki.ac.id

1. INTRODUCTION

As information technology develops, the tourism industry evolves into an e-tourism industry. E-tourism digitalizes the tourism sector, both in hospitality, travel, and other subsectors. E-tourism is an approach to building business relationships in selling the tourism sector via the internet. With e-tourism, tourism managers can promote travel, services, and tourism products more efficiently than conventional methods.

This phenomenon's impact is the fast growth of tourist track record data on the internet, one of which is review data [1]. Visitors tend first to analyze reviews about the hotel, restaurant, or tourist attraction they want to visit to get a travel experience that suits their preferences and minimizes the risk of failure. Review data is growing very rapidly and has become a source of big data tourism that has the potential to provide valuable insight for tourism actors if it can be managed well.

Sentiment analysis is a form of data analytics in data review. Sentiment analysis aims to automatically identify whether an opinion is neutral, negative, or positive. Sentiment analysis was built using two approaches: lexicon-based and machine learning-based methods, as shown in Figure 1.

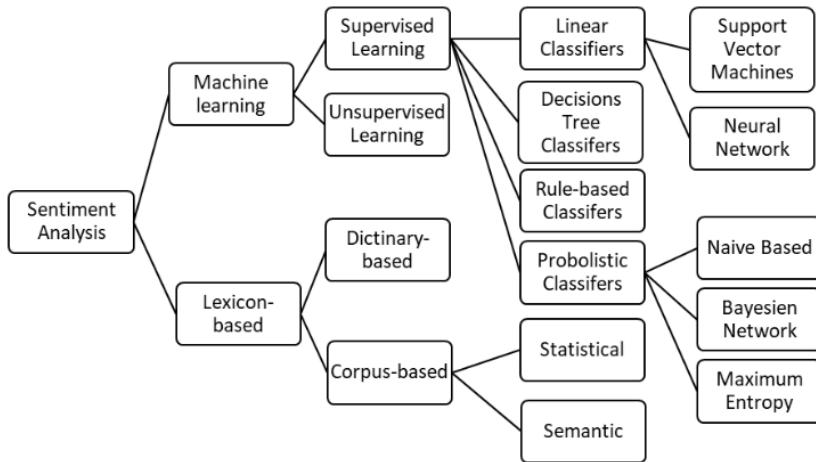


Figure 1. Two approaches to conducting sentiment analysis [2]

Valence aware dictionary and sentiment reasoner (VADER), is a lexicon-based sentiment analysis tool developed by Hutto and Gilbert to overcome problems related to analyzing language, symbols, and text styles on social media. Depending on their semantic orientation, phrases are generally categorized as either positive or negative in this rule-based sentiment analyzer. Among several lexicon-based sentiment analysis methods, the VADER lexicon shows the best performance [3], [4]. It is because the VADER lexicon is designed to be context-aware, assigning valence scores to individual words, allowing for a more granular analysis of sentiment, effective in handling negations and sentiment boosters, which are crucial for understanding the true sentiment of a text and recognizing emoticons and assigns sentiment scores to them.

Lexicons can correlate sentiment scores with specific words and expressions. Due to its lack of sensitivity to the quantity and quality of the training dataset and the minimal effort required in human-labeled documents, lexicon-based methods are also very competitive. Its main advantages are the simplicity of use and interpretability of lexicon-based sentiment analysis. Since the lexicon is directly accessible, anyone can easily examine, comprehend, expand, and modify it. Since it is simple to identify the words that affect the final score and update each word score as needed, lexicon-based sentiment analysis is easier to interpret than machine learning-based sentiment analysis [5]. On the other hand, machine learning methods are more complex to interpret but promise higher accuracy, i.e., fewer false classifications [6]. In sentiment analysis based on machine learning, the scoring algorithm is kept in an inaccessible black box to the user, making it impractical to make a rapid algorithmic adjustment.

One of the difficulties in training machine learning-based sentiment analysis models is the requirement for a large amount of training data. A large portion of the total samples for each sentiment class—positive, neutral, and negative—must be prepared to train a model from scratch. Building datasets of such size takes a significant amount of time. However, machine learning-based approaches perform much better than the lexicon method in solving sentiment analysis problems [3], [4], [7], [8]. One reason for this is that machine learning models capture the contextual nuances of language. They learn to understand the relationships between words, their order, and the overall context, allowing them to discern sentiment in a more nuanced way than lexicon-based methods. Machine learning models also can adapt to different domains and datasets. They learn from the specific features and patterns presented in the training data, making them more versatile across diverse types of text. Lexicon-based methods, on the other hand, might struggle when faced with domain-specific or evolving language. Machine learning models can be fine-tuned and updated with new data, allowing them to improve their performance over time continuously. Lexicon-based methods, in contrast, may require manual adjustments to accommodate changes in language or sentiment expressions.

Several studies have been carried out to overcome the weaknesses and the excess of lexicon-based and machine learning-based sentiment analysis by combining the use of both in sentiment analysis work. Research by Bhalerao [9] examined the joint use of the lexicon SentiWordNet and WordNet methods and several other machine learning methods, such as support vector machine (SVM), Naïve Bayes, and logistic regression. The results showed that the SVM method with term frequency-inverse document frequency (TF-IDF) feature extraction provided the best results. In addition, SVM is a machine learning method that provides the best performance compared to other machine learning methods when combined with the lexicon method, as reported by [10]–[13]. Several advantages of SVM compared to other machine learning methods are that SVM works well in high-dimensional spaces, making them suitable for sentiment analysis where the

feature space can be complex and extensive. SVM tends to be less prone to overfitting, which is crucial when working with sentiment analysis datasets that may have a limited number of examples for certain sentiment classes. SVM prioritizes balancing fitting the training data well and generalizing it to new, unseen data. SVM also performs global optimization during the training process and is suitable for small datasets.

Research on the lexicon and machine learning hybrid methods was also carried out by [14]–[21] for different variations of the lexicon and machine learning methods. All of these studies reported that using the hybrid method provided better performance for sentiment analysis than using each method independently. While machine learning models benefit from large labeled datasets during training, lexicon-based features could reduce the dependency on extensive labeled data. It was beneficial in scenarios where obtaining labeled data is challenging or expensive. Lexicon-based approaches excel at capturing specific sentiment words and their polarity. By integrating lexicon-based features into a machine learning model, the model could benefit from these explicit sentiment indicators while leveraging its ability to understand contextual nuances, making the analysis more comprehensive. Combining lexicon-based features with machine learning models often improved accuracy and robustness. Lexicon features provide a strong foundation, and machine learning models refine predictions based on the overall context, contributing to a more reliable sentiment analysis system. Sentiment analysis tasks often involve ambiguous language, sarcasm, or subtle nuances that may be challenging for either lexicon-based methods or machine learning models alone. The hybrid approach leverages the rule-based nature of lexicons for explicit sentiment signals and the learning ability of machine-learning models for understanding complex linguistic structures.

Lexicon-based methods might perform well in a specific domain but struggle when applied to a different domain. Machine learning models can adapt to different domains, especially when trained on diverse datasets. Integrating lexicon features into the machine learning model allows for better domain adaptation, combining the specificity of lexicons with the adaptability of machine learning. In specific applications or industries, a hybrid approach may outperform individual methods. For example, in financial sentiment analysis, where specific terms and phrases have pronounced sentiment implications, combining lexicon-based features and machine learning models can offer superior performance.

This research by Khan and Lee [22] improved this hybrid model by using a hybrid of semi-supervised learning and lexicon, named the lineament extraction and stripe statistical analysis (LeSSA) method. The methods used in semi-supervised learning were active learning (AL), self-training, co-training, and their combination for sentiment classification. LeSSA consists of three layers: feature engineering, multi-model sentiment learning, and sentiment classification. It used this lexicon method, such as AFFIN, GI, OL, SentiWordNet, semantic orientation calculator (SO-CAL), subjectivity lexicons, WordNet-Affect, NRC Hashtag sentiment lexicon, SenticNet5, and SentiSense. For the semi-supervised learning method, it used self-training and co-training methods. The results showed that LeSSA outperforms the semi-supervised approach and the stand-alone lexicon.

As explained previously regarding the advantages of the SVM method, this research [23]–[25] used SVM to perform sentiment analysis. Sentiment analysis using SVM was a binary classification approach that would classify reviews as positive or negative. In general, SVM models were built using several training data and tested on several test data. For some cases, such as big data, preparing training data with a proportional ratio of the overall data to be classified using a SVM was impossible. Manual labeling was a time-consuming and error-prone task. For this reason, several studies have developed AL on SVM to overcome this problem. With the idea that AL is carried out on some data closest to the hyperplane, it is possible to obtain better classification results than without AL. The main goal of AL is to reduce labeling effort without reducing classification accuracy by determining which samples require manual labeling with intelligence, called query samples. Numerous experiment results indicated that the proposed AL approach significantly improved classification accuracy while demanding less labeling work. However, this AL method still has weaknesses because it still involves manual labeling by humans on sample queries.

Based on previous research studies regarding methods combining lexicon and machine-learning usages in text processing and the weaknesses of the AL method, we propose the hybrid lexicon and AL-SVM (LeALSVM) method. This research filled the research gap regarding the absence of research that used those methods to improve sentiment analysis performance. The method proposed in this research was a novel method that previous researchers had never used. This research proposed using the VADER lexicon labeling to answer query samples generated by AL-SVM. It aimed to automate the sentiment analysis process on big data. In addition, this research also tested LeALSVM performance on big data in a limited scope.

2. METHOD

VADER lexicon is a lexicon-based sentiment analysis method developed by Hutto and Gilbert to overcome problems related to sentiment analysis, symbols, and text styles on social media. VADER could detect the polarity of sentences, whether the sentences had positive, negative, or neutral polarity. VADER

could detect abbreviations, symbols, or slang often found on social media, such as the abbreviation laugh out loud (LOL) and emoticon symbols: for example, those emoticons representing a smile could be detected as having a positive polarity [26].

In VADER, sentences were classified based on the positive, negative, and neutral polarity value range. The value range was that a sentence was said to have positive polarity if the compound value was greater than or equal to 0.05, a sentence was said to have negative polarity if the compound value was less than or equal to -0.05, and a sentence was said to have neutral polarity if the compound value was from -0.05 to 0.05. The intended compound value was calculated by adding the equivalent value to each word in the lexicon and adjusting or normalizing according to the rules to -1 (most negative) and +1 (most positive) [26].

SVM is a supervised learning method that is quite popular for classifying high-dimensional data, such as text data [27], [28]. SVM has advantages because its computation is lighter than deep learning but has an excellent classification ability. The SVM method is relatively easy to implement because determining the support vector can be formulated in a quadratic programming (QP) problem [29]. As for data that wants to be processed using the SVM method must first be transformed into vector space.

In the case of linear classification, SVM separated the data into two parts by building a hyperplane from the support vector. Classified data was based on vector position in hyperplane space. There were many possible hyperplanes for classifying data. One of the normal choices for the best hyperplane was the one that represented the separation with the largest margin between the two classes. Thus, a hyperplane was chosen to maximize the distance to and from the nearest data points on each side. If such a hyperplane exists, it is known as a maximum margin hyperplane, and the linear classifier it defines is known as a maximum margin classifier. The SVM's maximum margin and margin hyperplanes were trained with samples from two classes. The samples at the margin were called support vectors.

Suppose some data was randomly selected and classified manually by an expert. In that case, that dataset could be used to reproduce labels for all the data held through a supervised learning algorithm, such as SVM. Since the work was done manually, labeling is expected to be carried out on a relatively small number of documents.

It motivated researchers to develop an approach where instead of randomly selecting documents, researchers could guide the selection of such documents in such a way. Therefore, it was only required to label a minimum number of documents to get a certain classification accuracy level. AL aimed to address this specific problem. The classification model was retrained using the labeled samples obtained through the manual labeling of the sample with the highest classification uncertainty at each iteration in the AL scenario to achieve an adequate classification of the unlabeled data.

AL-SVM is a pool-based AL method that allows the model to use a small amount of training data but obtains good classification accuracy results. This method divided unlabeled data into several groups and executed sequentially. In each execution, the selected unlabeled data was labeled by an expert and added to the training dataset while discarding the unlabeled data. The process continued until all data groups have been executed. The selection of unlabeled data was determined based on the distance of the data from the separating hyperplane found by SVM during the training process. There were several AL-SVM methods, including both binary classification and multiclass classification. In research Tong and Koller [30], the AL-SVM method consisted of a simple margin method, which was developed again using the maxmin margin method and the maxratio margin method, as shown in Figures 2(a) to (c).

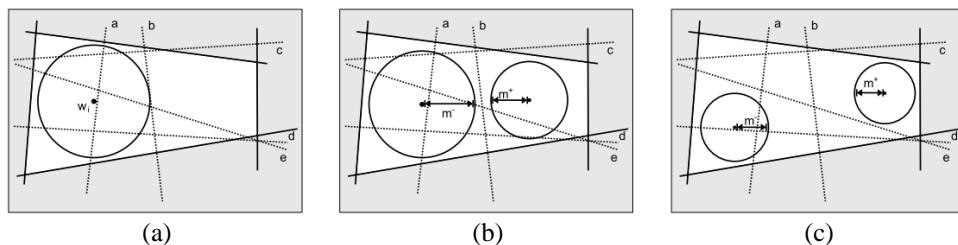


Figure 2. AL-SVM method variation; (a) simple margin will choose to query data a, (b) maxmin will choose data b, and (c) maxratio will choose data e [30]

This research set up three experimental scenarios to understand LeALSVM performance. This research's first experiment aimed to compare the AL-SVM model and a model without using the AL method. It was called passive learning SVM (PLSVM). The first data source was labeled English-language hotel

reviews for several hotels in Europe. Data taken from Booking.com [31]. This dataset contained 515,000 customer reviews and ratings for 1,493 European luxury hotels. The labels on this review data were given directly by the user, considering that when filling in a review, the user would be asked to fill in a positive and negative review separately. After cleaning the data, this research had 19,000 positive and 19,000 negative reviews selected from the dataset to support this AL experiment. This research only used part of the data to limit computing time. This research used 1,000 positive and 1,000 negative review data. For the AL-SVM testing experiments, this research used Monte Carlo simulation. The following are the Monte Carlo simulation stages for AL-SVM learning, as shown in Figure 3, namely:

- Divided the data randomly into training data and testing data. This experiment divided 1,500 records for training data and 500 for testing data.
- For training data, conducted SVM training (building a classification model), then selected the ten data points closest to the hyperplane.
- For the closest data, delete it from the training data, then add it to the data pool.
- Performed SVM training for the data pool.
- Tested the testing data at point 1 for the model from the data pool.
- Repeated steps b to e 20 times so the research had 20 models and 20 error rates.

The experiment was repeated ten times to obtain a more stable error rate value by taking the average error rate of the experimental results.

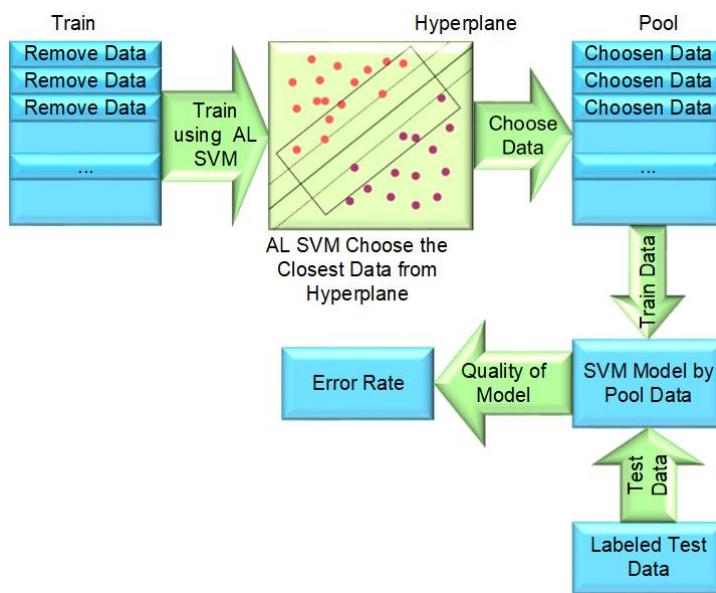


Figure 3. Monte Carlo simulation for AL-SVM

As a comparison, this research conducted PLSVM learning with the following simulation stages. Just like the AL experiment, the experiment was repeated ten times to obtain a more stable error rate value by taking the average error rate of the experimental results. The following are the PLSVM Monte Carlo simulation steps.

- Divided the data randomly into training data and testing data. This experiment divided 1,500 records for training data and 500 for testing data.
- Selected 10 data points randomly.
- For this data, delete it from the training data, then add it to the new training data.
- Performed SVM training for the data pool.
- Tested the testing data at point 1 for the model from the new train data.
- Repeated steps b to e 20 times so the research had 20 models and 20 error rates.

Thus far, this research has described how the AL-SVM method works. This research proposed how to develop the AL-SVM method into the LeALSM method. The following illustration in Figure 4 shows the LeALSVM primary process and where AL-SVM was in this new method. AL-SVM would select the closest vector data from the hyperplane as query samples and manually ask the user for the label. Thus, to overcome this case, this research proposed to use the VADER lexicon method to provide labels to query samples so that

labeling query samples runs automatically. The labeling results would compile training data used to classify big data.

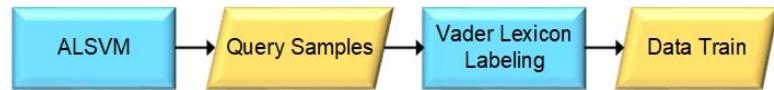


Figure 4. LeALSV primary process

In the second experiment, this research still used the labeled dataset [31]. This experiment aimed to compare the model performance of LeALSV, SVM, and VADER lexicon. This research used 2% of the data as training data, and the iteration in LeALSV stopped when the data pool reached 20% of the total data.

The experimental flow to measure the performance of LeALSV for the data in this second experiment is shown in Figure 5. It was to use 2% big data as training data in a data pool where this data was labeled for polarity using VADER lexicon. This experiment selected records for the data pool, which consisted of reviews that had a rating score of 1 for extreme negative data and a rating of 5 for extreme positive data, in the hope that the data pool does not provide biased values. The research carried out the LeALSV process on the remaining 98% of the big data that had been labeled with the VADER lexicon, which was named train data.

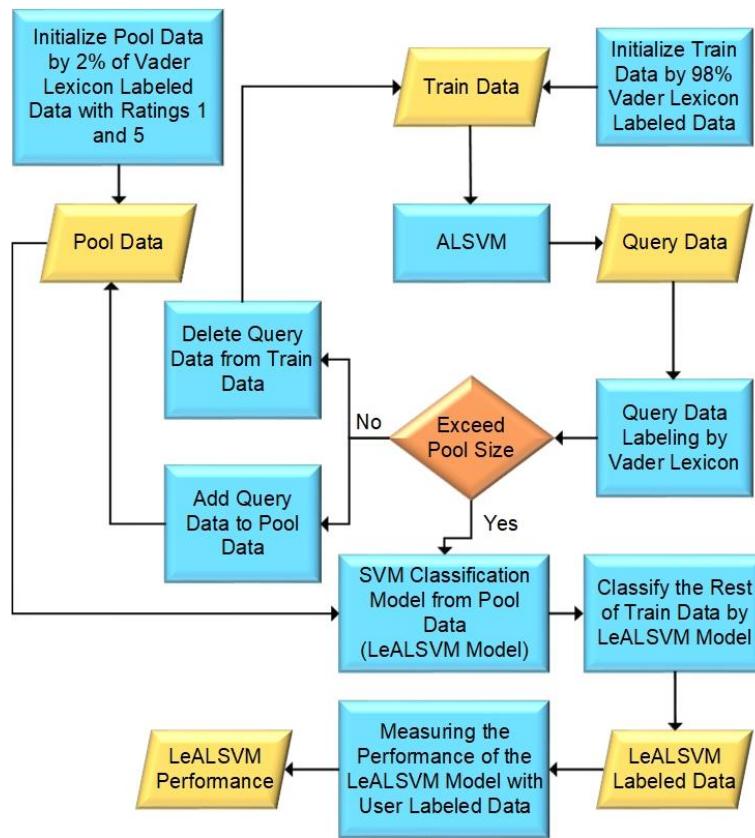


Figure 5. LeALSV process by user labeled data from dataset [31]

AL-SVM modeling was carried out by searching for the data closest to the hyperplane as the selected query data. In AL-SVM theory, this data query would be labeled manually by the user, but in LeALSV, the labeling was done by VADER lexicon to automate the query labeling process. Furthermore, this research added query data that had been labeled with the VADER lexicon to the data pool and removed it

from the train data. The process repeated until the data pool size reached 20% of the total data train size. For this experiment, this research measured the performance of LeALSVIM by comparing the predicted labels with the review labels on the data given by the user.

The third experiment used unlabeled big data from Bali's tourist attraction reviews on TripAdvisor. The data consisted of 46,796 records for 159 tourist attractions in Bali. The experiment measured LeALSVIM model performance based on the model's accuracy concerning ground truth polarity (gt-polarity). Gt-polarity was built from user ratings in reviews. With the consideration that ratings could be used to represent customer satisfaction. Ratings 3, 4, and 5 would be given positive polarity, while ratings 1 and 2 would be given negative polarity. In addition, this research measured the VADER lexicon model performance against gt-polarity and the lexicon SVM model against gt-polarity separately.

The LeALSVIM flow in this experiment was the same as the LeALSVIM flow in the second experiment, with slight differences in testing model performance, as depicted in Figure 6. This experiment used the SVM model formed from the pool data to classify the remaining data in the train data and measured its performance against gt-polarity (polarity, which was built from ratings). This research compared LeALSVIM performance against the VADER lexicon performance, lexicon-SVM with 2% of big data as training data, and lexicon-SVM with 20% of big data as training data. Lexicon-SVM was an SVM model where the training data was labeled using the VADER lexicon method.

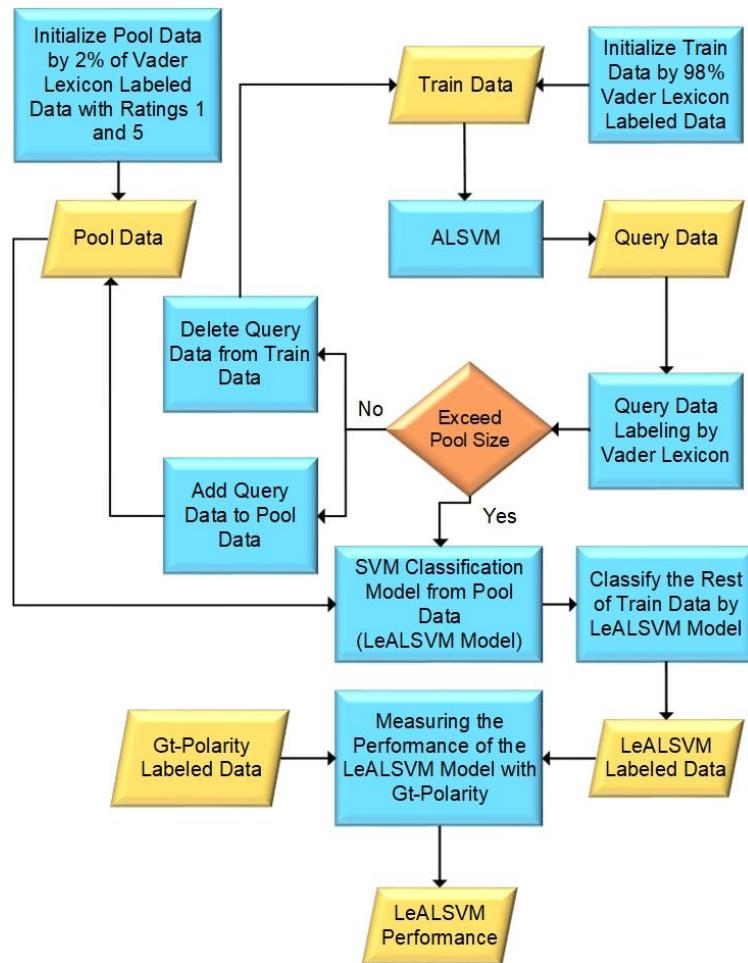


Figure 6. LeALSVIM process from TripAdvisor unlabeled data

Lexicon-SVM in this experiment was a model where VADER lexicon sentiment analysis prepared the training data for the SVM model, as shown in Figure 7. The data was split into train data and test data. Then, the SVM model was built from the train data and used to polarize the test data. Model performance was measured by comparing the predicted label with the gt-polarity label. Considering that there was a gray area where a rating of 3 in a review could mean both positive and negative polarity, this research analyzed a

sample of 50 records with a rating of 3 where the polarity of the LeALSV model differed from the polarity of the VADER lexicon and compared them with manual labeling. The test flow is shown in Figure 8.

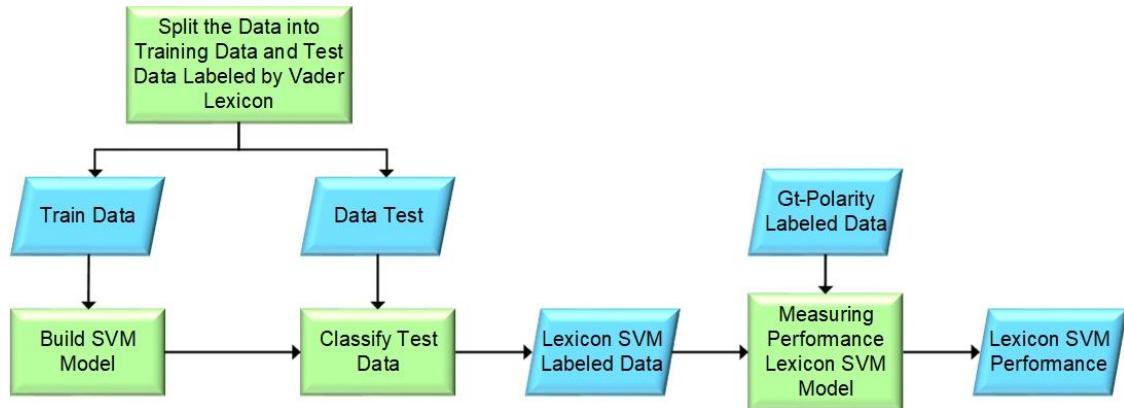


Figure 7. Lexicon SVM model flowchart



Figure 8. Additional testing for LeALSV with expert labeling

3. RESULTS AND DISCUSSION

The first experiment results using Monte Carlo simulation are shown in Table 1. The longer the training process runs, the smaller the model's error rate. It is clearly shown by the graph in Figure 9. The performance of the AL-SVM model or PLSVM model improved as training progressed. The research could obtain the information from the first model until the 20th model.

Table 1. Monte Carlo simulation experiment results for datasets [31]

SVM model	PLSVM error rate	AL-SVM error rate	SVM model	PLSVM error rate	AL-SVM error rate
SVM model 1	0.4812	0.4776	SVM model 11	0.1846	0.1702
SVM model 2	0.4226	0.4008	SVM model 12	0.1772	0.1484
SVM model 3	0.365	0.3286	SVM model 13	0.1666	0.154
SVM model 4	0.308	0.2754	SVM model 14	0.152	0.1388
SVM model 5	0.2634	0.2652	SVM model 15	0.1534	0.1338
SVM model 6	0.2436	0.22	SVM model 16	0.149	0.1282
SVM model 7	0.2302	0.2104	SVM model 17	0.1496	0.13
SVM model 8	0.2018	0.2036	SVM model 18	0.145	0.122
SVM model 9	0.1932	0.1794	SVM model 19	0.1364	0.125
SVM model 10	0.1844	0.184	SVM model 20	0.1384	0.1122

In general, learning using AL-SVM provided better results than PLSVM learning, as can be seen in the graph in Figure 9. The AL-SVM error rate was lower than PLSVM, showing fewer classification errors occurred. Based on Table 1, it was found that in the 20th model, the error rate of AL-SVM was 11.22% compared to 13.84% by PLSVM.

AL supported labeling sample data with the highest classification uncertainty represented by the closest vector of the hyperplane. When data with the highest uncertainty, which was likely to cause the model to misclassify, was formed, manual labeling for that data would generally be able to represent the classification class space constructed from the hyperplane after manual labeling. In this case, the AL-SVM model would be better than the PLSVM model, which chose document labeling randomly.

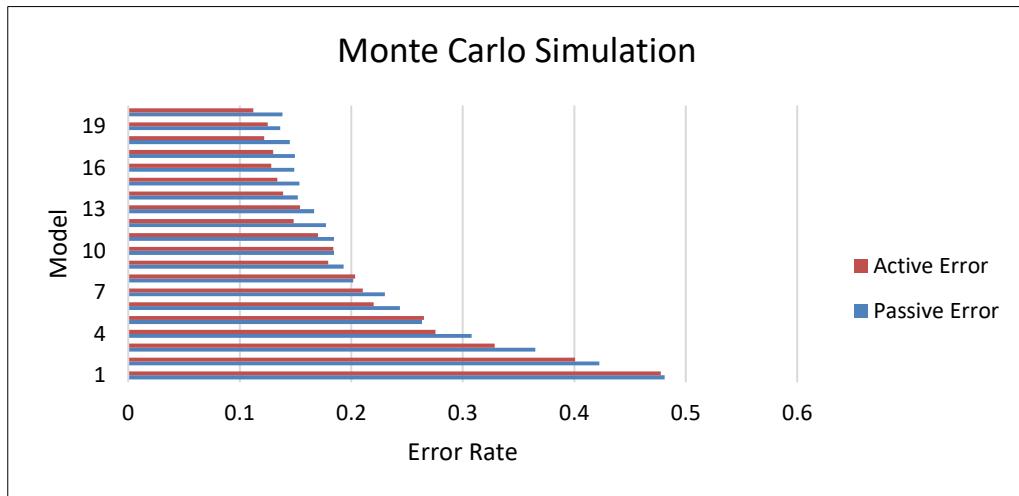


Figure 9. Experimental results using Monte Carlo simulation

The second experiment aimed to measure LeALSV model performance compared to the VADER lexicon and SVM methods, which run independently on user labeled data [31]. This research used a confusion matrix to measure the classification model performance. A confusion matrix is a performance evaluation tool in machine learning, representing the accuracy of a classification model. It displayed the number of true positives, true negatives, false positives, and false negatives. From the confusion matrix, this research measured the model's accuracy, precision, recall, and F1 score.

The experimental results are shown in Table 2. The experiment was carried out on 2000 data records, divided equally between positive and negative reviews. SVM and LeALSV used 2% of the data as randomly selected training data. The experimental results showed that the LeALSV model provided the best performance compared to the VADER lexicon or SVM methods, which run independently. Measurements were carried out in accuracy, precision, recall, and F1 score units, whereas in the LeALSV model, the magnitudes of all measurement units were relatively balanced. Thus, it could be said that the LeALSV model could recognize both positive and negative reviews well.

Table 2. Performance results of VADER lexicon, SVM, and LeALSV for datasets [31]

	VADER lexicon performance	SVM performance	LeALSV performance
Accuracy	0.679500	0.730000	0.823000
Precision	0.610462	0.738095	0.767828
Recall	0.992000	0.713000	0.926000
F1 score	0.755810	0.725331	0.839529

VADER used a combination of sentiment lexicon as a list of lexical features (e.g., words) generally labeled according to their semantic orientation as positive or negative. Lexical words were adjectives that described everything related to words, vocabulary, or even language more generally. Since it was built from an adjective dictionary, it was very likely that the VADER lexicon would fail to classify documents that lacked adjectives but had a negative sentimental meaning. For example, the sentence "There is a lot of rubbish on the beach". For this condition, machine-learning classifier models, such as SVM had advantages because the TF-IDF feature calculated the weight of the occurrence of all words against the weight of that word in the document class without limiting it to adjectives.

Table 3 shows the third experiment results, which compared the performance of VADER lexicon, lexicon SVM, and LeALSV against gt-polarity for unlabeled tourist attraction big data. Lexicon SVM was an SVM experiment where training data labeling was done using VADER lexicon. The experimental results showed that LeALSV provided the best performance compared to the performance of VADER lexicon or lexicon SVM for the same data, measured against gt-polarity. LeALSV also showed a fairly good model with a balance of accuracy, precision, recall, and F1 score in this experiment.

Since the LeALSV model was run on unlabeled data and polarity performance was measured from ratings, where a rating of 3 provided an ambiguity value. Thus, this research analyzed the LeALSV results for review data with different polarities between VADER lexicon and LeALSV. Table 4 shows the

LeALSVM performance compared to VADER lexicon in recognizing review polarity and measuring its suitability to the polarity of the user manual labeling. Experimental results showed that LeALSVM was better at recognizing review polarity than the VADER lexicon method. Based on the analysis results, LeALSVM classifies much closer to the actual polarity represented by the expert's manual labeling.

Table 3. Performance results of VADER lexicon, lexicon SVM, and LeALSVM for big data tourism

VADER lexicon performance	Lexicon SVM performance with 2% train data	Lexicon SVM performance with 20% train data	LeALSVM performance
Accuracy	0.887683	0.904407	0.958136
Precision	0.933176	0.917179	0.961123
Recall	0.943116	0.982877	0.996339
F1 score	0.938120	0.948893	0.978414

Table 4. VADER lexicon and LeALSVM performance against polarity by the expert

VADER lexicon performance	LeALSVM performance
Accuracy	0.200000
Precision	0.363636
Recall	0.108108
F1 score	0.166667

Thus far, LeALSVM has shown excellent performance in the sentiment analysis work domain. When compared with research [25] which also worked on a sentiment analysis model using a hybrid method of semi-supervised learning and lexicon, named the LeSSA method, the LeALSVM method provides better accuracy of 95.8% compared to LeSSA which provides an accuracy of 85.1%. This research contributed a new and effective method with good performance to solve sentiment analysis problems on unlabeled data without having to manually label the training data. In this research, it was found that the LeALSVM method improved the sentiment analysis performance compared to using the VADER lexicon, lexicon SVM, and SVM methods, which run independently.

4. CONCLUSION

AL-SVM allowed classification using the SVM method to be carried out by labeling a small amount of training data while still obtaining optimal results. Selecting query data closest to the hyperplane (AL-SVM) improved the classifier model performance compared to selecting query data randomly (PLSVM). Considering the capabilities of AL-SVM and to fully automate the classification processing of AL-SVM, the labeling process on query data could be supported by the VADER lexicon method. This research named the novel method as LeALSVM. The VADER lexicon method did not require training data for the sentiment analysis process but still had quite good capabilities. Based on the experimental results, the LeALSVM method was proven to improve the sentiment analysis model performance in big data tourism compared to sentiment analysis carried out using the VADER lexicon, lexicon SVM, and SVM methods, which run separately. Experimental results also showed that LeALSVM was better at recognizing review polarity than the VADER lexicon method, which LeALSVM classifies much closer to the actual polarity represented by the expert's manual labeling.

ACKNOWLEDGEMENTS

We would like to thank Udayana University for their support so that this research can run optimally.

REFERENCES

- [1] N. D. Zaki, N. Y. Hashim, Y. M. Mohialden, M. A. Mohammed, T. Sutikno, and A. H. Ali, "A real-time big data sentiment analysis for iraqi tweets using spark streaming," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 4, pp. 1411–1419, 2020, doi: 10.11591/eei.v9i4.1897.
- [2] I. A. Ozen and Nevsehir, "Tourism Products and Sentiment Analysis," in *Advances in Hospitality and Tourism Information Technology*, Florida: USF M3 Publishing, LLC, 2021.
- [3] A. Veluchamy, H. Nguyen, M. L. Diop, and R. Iqbal, "Comparative Study of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches," *SMU Data Science Review*, vol. 1, no. 4, pp. 1–22, 2018.
- [4] R. Srivastava, P. K. Bharti, and P. Verma, "Comparative Analysis of Lexicon and Machine Learning Approach for Sentiment Analysis," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 3, pp. 71–77, 2022, doi: 10.14569/IJACSA.2022.0130312.

[5] D. H. Abd, A. R. Abbas, and A. T. Sadiq, "Analyzing sentiment system to specify polarity by lexicon-based," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 1, pp. 283–289, 2021, doi: 10.11591/eei.v10i1.2471.

[6] C. Siebert, J. Hartmann, M. Heitmann, and C. Schamp, "Accuracy of Automated Sentiment Analysis," *SSRN Electron. Journal*, 2019, doi: 10.2139/ssrn.3489963.

[7] I. Rustanto and N. A. Rakhmawati, "Media Sentiment Analysis of East Java Province: Lexicon-Based vs Machine Learning," *IPTEK Journal of Proceedings Series*, vol. 0, no. 6, p. 203, 2021, doi: 10.12962/j23546026.y2020i6.11094.

[8] K. Machova, M. Mach, and M. Vasilko, "Comparison of machine learning and sentiment analysis in detection of suspicious online reviewers on different type of data," *Sensors*, vol. 22, no. 1, 2022, doi: 10.3390/s22010155.

[9] K. V. Bhalerao, "Customer Reviews Sentiment Analysis: A hybrid technique of Lexicon and Machine Learning based Classification model (SVM,NB, Logistic Regression)," Aug. 2021. [Online]. Available: <https://norma.ncirl.ie/5135/>. (Accessed: Jul. 04, 2023).

[10] M. E. Moussa, E. H. Mohamed, and M. H. Haggag, "Opinion mining: a hybrid framework based on lexicon and machine learning approaches," *International Journal of Computers and Applications*, vol. 43, no. 8, pp. 786–794, 2021, doi: 10.1080/1206212X.2019.1615250.

[11] A. I. Saad, "Opinion Mining on US Airline Twitter Data Using Machine Learning Techniques," in *16th International Computer Engineering Conference, ICENCO 2020*, Dec. 2020, pp. 59–63, doi: 10.1109/ICENCO49778.2020.9357390.

[12] M. Lepe-Faúndez, A. Segura-Navarrete, C. Vidal-Castro, C. Martínez-Araneda, and C. Rubio-Manzano, "Detecting aggressiveness in tweets: A hybrid model for detecting cyberbullying in the Spanish language," *Applied Sciences*, vol. 11, no. 22, Nov. 2021, doi: 10.3390/app112210706.

[13] S. Sazzed, "A Hybrid Approach of Opinion Mining and Comparative Linguistic Analysis of Restaurant Reviews," in *International Conference Recent Advances in Natural Language Processing, RANLP*, 2021, pp. 1281–1288, doi: 10.26615/978-954-452-072-4_144.

[14] Murni, T. Handhika, A. Fahrurrozi, I. Sari, D. P. Lestari, and R. I. M. Zen, "Hybrid Method for Sentiment Analysis Using Homogeneous Ensemble Classifier," in *2019 2nd International Conference of Computer and Informatics Engineering (IC2IE)*, 2019, pp. 232–236, doi: 10.1109/IC2IE47452.2019.8940896.

[15] B. Erşahin, Ö. Aktaş, D. Kılıç, and M. Erşahin, "A hybrid sentiment analysis method for Turkish," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 27, no. 3, pp. 1780–1793, Jan. 2019, doi: 10.3906/elk-1808-189.

[16] S. Tammina, "A Hybrid Learning approach for Sentiment Classification in Telugu Language," in *2020 International Conference on Artificial Intelligence and Signal Processing (AISP)*, 2020, pp. 1–6, doi: 10.1109/AISP48273.2020.9073109.

[17] K. Machová, M. Mikula, X. Gao, and M. Mach, "Lexicon-based sentiment analysis using particle swarm optimization," *Electronics*, vol. 9, no. 8, pp. 1–22, Aug. 2020, doi: 10.3390/electronics9081317.

[18] A. Mardjo and C. Choksuchat, "HyVADRF: Hybrid VADER-Random Forest and GWO for Bitcoin Tweet Sentiment Analysis," *IEEE Access*, vol. 10, pp. 101889–101897, 2022, doi: 10.1109/ACCESS.2022.3209662.

[19] S. A. S. Neshan and R. Akbari, "A Combination of Machine Learning and Lexicon Based Techniques for Sentiment Analysis," *Electronics*, pp. 8–14, Apr. 2020, doi: 10.1109/ICWR49608.2020.9122298.

[20] P. Paul and R. P. Singh, "A weighted hybrid recommendation approach for user's contentment using natural language processing," *AIP Conference Proceedings*, vol. 2705, no. 1, p. 020006, Jun. 2023, doi: 10.1063/5.0148413.

[21] N. Amirah, M. Yusoff, and M. Kassim, "Hybrid Machine Learning Methods with Malay Lexicon for Public Polarity Opinion on Water Related Issue," *2022 IEEE International Conference in Power Engineering Application (ICPEA)*, Shah Alam, Malaysia, 2022, pp. 1–5, doi: 10.1109/ICPEA53519.2022.9744713.

[22] J. Khan and Y. K. Lee, "LeSSA: A unified framework based on lexicons and semi-supervised learning approaches for textual sentiment classification," *Applied Sciences*, vol. 9, no. 24, Dec. 2019, doi: 10.3390/app9245562.

[23] V. Nurcahyawati and Z. Mustaffa, "Improving sentiment reviews classification performance using support vector machine-fuzzy matching algorithm," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 3, pp. 1817–1824, 2023, doi: 10.11591/eei.v12i3.4830.

[24] A. Kumar, V. Dutt, V. García-Díaz, and S. K. Narang, "Twitter sentimental analysis from time series facts: The implementation of enhanced support vector machine," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 5, pp. 2845–2856, 2021, doi: 10.11591/eei.v10i5.3078.

[25] L. K. Ramasamy, S. Kadry, and S. Lim, "Selection of optimal hyper-parameter values of support vector machine for sentiment analysis tasks using nature-inspired optimization methods," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 1, pp. 290–298, 2021, doi: 10.11591/eei.v10i1.2098.

[26] M. Al-Shabi, "Evaluating the performance of the most important Lexicons used to Sentiment analysis and opinions Mining," *Int. IJCSNS International Journal of Computer Science and Network Security*, vol. 20, no. 1, pp. 51–57, 2020.

[27] J. Philip, B. VeeraSekharReddy, M. Harshmi, I. V. S. L. Haritha, S. Patil, and S. K. Shareef, "A Comparative Study of Text Classification using Selective Machine Learning Algorithms," in *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2023, pp. 482–484, doi: 10.1109/ICICCS56967.2023.10142474.

[28] H. M. Lee and Y. Sibaroni, "Comparison of IndoBERTweet and Support Vector Machine on Sentiment Analysis of Racing Circuit Construction in Indonesia," *Jurnal Media Informatika Budidarma*, vol. 7, no. 1, pp. 99–106, 2023, doi: 10.30865/mib.v7i1.5380.

[29] N. W. S. Saraswati, K. K. Widiartha, and L. P. A. Prapitasari, "Vector machine to predict student retention: A computerized approach," *Journal of Physics: Conference Series*, vol. 1469, no. 1, 2020, doi: 10.1088/1742-6596/1469/1/012045.

[30] S. Tong and D. Koller, "Support Vector Machine Active Learning with Applications to Text Classification," *Journal of Machine Learning Research*, no. June, pp. 45–66, 2001, doi: 10.1162/153244302760185243.

[31] J. Liu, "515K Hotel Reviews Data in Europe," *Kaggle*, 2007. <https://www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe> (accessed May 04, 2022).

BIOGRAPHIES OF AUTHORS



Ni Wayan Sumartini Saraswati is senior lecturer at Indonesian Institute of Business and Technology. Currently studying in the engineering sciences doctoral program at Udayana University. She obtained his Master of Engineering (M.T) degree from Udayana University, Indonesia, in 2011. She received Bachelor's degree in engineering from Telkom University in 2003. Her research interests are big data analytics, machine learning, and business intelligence system. She can be contacted at email: sumartini.saraswati@instiki.ac.id.



I Ketut Gede Darma Putra hold a Doctoral degree from Gajah Mada University, Indonesia, in 2007. He also obtained Master of Engineering (M.T) degree from Gajah Mada University, Indonesia, in 2000. He received his S. Kom degree in informatics engineering from the Institute of Ten November Technology Surabaya, Indonesia, in 1997 and now he is a lecturer in the Department of Electrical Engineering and Information Technology, Udayana University Bali, Indonesia. He is currently a professor of information technology science at the Faculty of Engineering, Udayana University, since 2014. His research interests are biometrics, image processing, data mining, and soft computing. He can be contacted at email: ikgdarmaputra@unud.ac.id.



Made Sudarma holds a Doctorate from Udayana University, Indonesia, in 2012. He also holds a Master of Applied Science (M.A.Sc.) from SITE-OU: School of Information Technology and Engineering, Ottawa University Canada in 2000. During his studies at SITE-OU, he was an assistant professor and a member of the research team of the built-in self-testing compaction generator field VLSI technology. He is also a professor of information technology science at the Electrical Engineering Study Program, Faculty of Engineering, Udayana University at Udayana University since 2019. His research includes internet and web applications, cloud computing, artificial intelligence, data warehousing and data mining, computer graphics and virtual reality, as the author of books and as a reviewer in international and national journals. In addition, he also completed vocational education (IPU., ASEAN Eng) and is active in academic activities, and he also work as an Information Technology consultant in local government. He can be contacted at email: msudarma@unud.ac.id.



I Made Sukarsa hold a Doctoral degree from Udayana University, Indonesia, in 2019. He also obtained his Master of Engineering (M.T) degree from Gajah Mada University, Indonesia, in 2005. He received his S.T degree in informatics engineering from the Gajah Mada University, Indonesia, in 2000 and now he is a lecturer in the Lecturer at the Department of Information Technology, Faculty of Engineering Udayana, Indonesia. Currently actively teaching and conducting research on IT governance, dialog models on chatbot engines, datawarehouses, and system integration. He can be contacted at email: sukarsa@unud.ac.id.