Journal of Science and Applied Engineering (JSAE)
Vol. 7, No. 2, October 2024
ISSN Print : 2621-3745
ISSN Online : 2621-3753
DOI : 10.31328/jsae.v7i2.6438
(Page.18-26)

ANALYSIS OF GOOGLE MAPS REVIEW SENTIMENT ON BLIMBING MARKET FACILITIES WITH SVM

Muhaemin[1,*], R. P. Putra[2], Istiadi[3]

[1, 2, 3]*Department of Informatics Engineering, Universitas Widya Gama Malang, Malang, Indonesia*
**Email:muhaemin1998@gmail.com*

Submitted : 28 August 2024; *Revision :* 21 October 2024; *Accepted :* 24 October 2024

ABSTRACT

Traditional markets play a crucial role in the local economy, especially in developing countries like Indonesia. Although markets and traders have officially developed, the quality of facilities in traditional markets remains a challenge, with various issues affecting visitor comfort and experience. This study proposes a sentiment analysis of public opinions on the facilities of Blimbing Traditional Market using the Support Vector Machine (SVM) method. The aim is to understand the implementation and results of sentiment classification using SVM. The data used is sourced from user reviews on Google Reviews. The analysis process includes data collection, automatic sentiment labeling using TextBlob, and data preprocessing, including case folding, text cleaning, tokenization, stemming, and stopword removal. SVM with various kernels (linear, polynomial, RBF, and sigmoid) was evaluated with train-test data split ratios ranging from 90/10 to 10/90. The results show that the linear kernel achieved an accuracy of 98.07% with an 80/20 ratio, poly 96.15% with a 90/10 ratio, RBF 97.69% with a 90/10 ratio, and sigmoid 98.46% with a 90/10 ratio. Suggestions for future research include exploring more advanced preprocessing techniques and other algorithms such as Random Forest, Naive Bayes, or Neural Networks to improve sentiment analysis performance.

Keywords Sentiment Analysis; Market; Vector Machine Support;
Paper type Research paper

INTRODUCTION

Traditional markets have an important role in the local economy, especially in developing countries such as Indonesia. Data from the Ministry of Trade shows that in 2020, there are more than 13,000 traditional markets spread throughout Indonesia. The development of traditional markets is regulated in Law Number 12 of 2008 concerning Regional Government which authorizes regions to develop and manage the potential basic needs of the community. Traditional markets not only offer more affordable prices but also allow for the bargaining process [1]. In addition, traditional markets also support the diversity of local products and the preservation of traditional economic culture.

Previous research by [2] shows that traditional markets other than as a place for sales of local products, also provide jobs. However, hygiene issues, such as poor sanitation, waste management, and lack of facilities, are still the main concerns [3], Traditional markets are generally still not good, especially in terms of facilities, such as lack of water, poor waste management, and messy parking areas. Blimbing Market, established in 1970 With ± 2,250 traders, it has not undergone significant updates, making its facilities less Adequate. The government's revitalization efforts in the last decade have gone unheard, which has finally caused uncertainty for traders, even though revitalization is important to face global competition and overcome degradation [4].

In this digital age, online reviews from consumers have become a valuable source of data to assess the quality of products and services [5]. Review data from platforms like Google Maps can provide an accurate picture of the public's perception of traditional market facilities [6]. Data mining and sentiment analysis are effective techniques to uncover patterns, evaluate online reviews, and help identify public opinion automatically [7]-[8] according to research conducted by A. Gupita, P. Tyagi, one of the techniques that is often used in sentiment analysis is the Support Vector Machine (SVM) method, which has a good performance in text classification [9].

This research was conducted to analyze public sentiment towards the Blimbing Traditional Market facility using the Support Vector Machine (SVM) method based on reviews on Google Maps. This research is important because until now there has been no study that specifically analyzes public sentiment related to the traditional market facilities, this research aims to provide recommendations for improving facilities and offer useful insights for market managers and local governments in an effort to improve the quality of facilities and shopping experience in the market.

METHOD

In this study, the data collection process uses Google Maps as the main source because this platform contains many user reviews that directly reflect their experience at the Blimbing Traditional Market [10]. Google Maps users often provide reviews that cover a variety of aspects, including amenities. These reviews are particularly valuable for sentiment analysis as they often express genuine user opinions. The data is collected using web scraping techniques with the Instant Data Scraper tool, which focuses on the reviews provided by users. The flow of the research framework can be seen in the following figure.

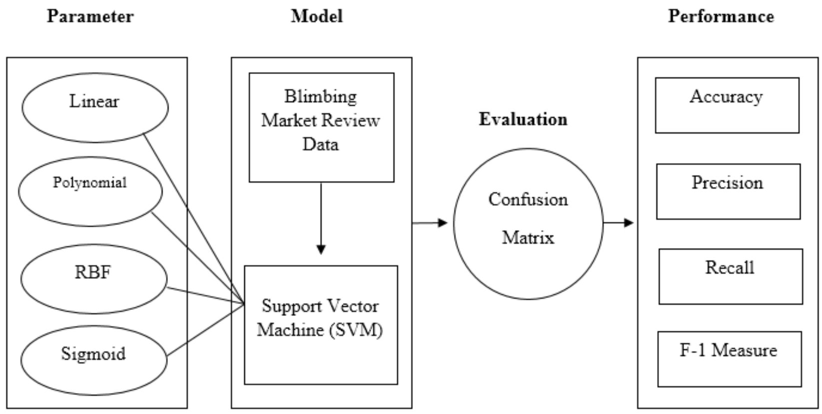

Figure 1. Conceptual Framework

Dataset

Data is collected through web scraping techniques using the Instant Data Scraper extension tool, which generates .xlsx-formatted documents. This dataset includes 1,299 reviews from Google Maps visitors related to Blimbing Traditional Market. The collected data will then be automatically labeled using the Python TextBlob library [11], which not only automates the sentiment labeling process, but also allows for clearer visualization of sentiment distribution in the dataset [12].

Data Labeling

Once the data is successfully obtained, the next step is to label the data. This labeling is done automatically using a set of Python and Google Colab code that has been developed. This process results in two categories of sentiment, namely 54 positive sentiments, and 1245 negative sentiments, as detailed in Table I.

TABLE I. LABELING RESULTS DATA

Reviews	Sentiment
It needs renovation, needs additional lighting, lack of cleanliness, and the distance of toilets. So that in some alleys the smell of pee, traders want to go to the toilet far away, afraid that customers will rush to leave	Negative

Table I shows the data from the labeling of reviews about the blimbing market, including visitor experiences related to market facilities such as lack of cleanliness, smell, and distance from toilets. This data provides valuable insights into areas that need attention and improvement strategies.

ISSN Print : 2621-3745
ISSN Online : 2621-3753
DOI : 10.31328/jsae.v7i2.6438
(Page.18-26)

Preprocessing

Data preprocessing is essential to improve the accuracy of sentiment analysis with Deep Belief Networks and classify review datasets. It involves cleaning, normalizing, and transforming the data into a format suitable for analysis. This process is to ensure that the data is free of distractions and inconsistencies before the analysis is carried out [13]. The steps include case folding up to TF-IDF, which will be discussed further in the next section. Proper preprocessing improves data quality, allowing the model to produce more accurate analysis [14].

Case Folding

This process aims to convert the entire dataset to lowercase [8], eliminating the distinction between uppercase and lowercase letters. For example, the word 'Market' was changed to 'market,' which helps maintain data consistency. This step is very important to ensure that the analysis is not affected by variations in the use of letters, as shown in table 1 below, so that each word is normalized to ensure the accuracy of the subsequent analysis.

TABLE II. CASE FOLDING PROCESS

Before	Case Folding
The location of this market is strategic and clean, so it's comfortable to shop here -_-	the location of this market is strategic and clean, so it's comfortable to shop here -_-

Cleaning

This process removes symbols such as periods (.), commas (,), exclamation marks (!), percent (%), numbers (0-9), and more (^), ensuring cleaner and more relevant data for analysis. This step focuses on truly informative text, improving data consistency and accuracy. In addition, this cleaning process also reduces noise in the data that can interfere with the results of the analysis.

TABLE III. CLEANING PROCESS

Case Folding	Cleaning
the location of this market is strategic and clean, so it's comfortable to shop here -_-	the location of this market is strategic and clean, so it is convenient to shop here

Tokenization

Tokenization divides the text into small parts or tokens, by separating them based on spaces or punctuation. This process also normalizes compound words like 'no' and 'eat' for consistency of the text. By converting text into manageable units, tokenization allows for more detailed analysis. Tokenization facilitates analysis by processing each token separately, reducing data complexity, and improving the efficiency of the model in recognizing important patterns.

TABLE IV. TOKENIZATION PROCESS

Cleaning	Tokenization
the location of this market is strategic and clean, so it is convenient to shop here	[location, market, this, strategic, and, clean, so, comfortable, shopping, here]

Stemming

This stage removes inflection or affixes from the word to abandon its basic form, such as changing 'run' and ‘running around' to 'run' [15]. Stemming reduces the complexity of the data and improves accuracy, such as search words, even if it results in non-standard words. This technique helps in achieving uniformity in word representation. This process simplifies word variations into consistent forms, improving the model's ability to recognize and group words with similar meanings.

ISSN Print : 2621-3745 Journal of Science and Applied Engineering (JSAE)
ISSN Online : 2621-3753 Vol. 7, No. 2, October 2024
(Page.18-26) DOI : 10.31328/jsae.v7i2.6438

TABLE V. STEMMING PROCESS

Tokenization	Stemming
[location, market, this, strategic, and, clean, so, comfortable, shopping, here]	['location', 'market', 'this', 'strategy', 'and', 'clean', 'so', 'comfortable', 'shopping', 'here']

Normalization

Normalize the process for converting text into a standard form by removing punctuation, normalizing numbers, removing the conjunction 'and' or 'or', and standardizing spelling. It also involves correcting typos and inconsistencies to ensure a uniform text representation. This process prepares the text for tokenization and stemming, reduces unnecessary diversity, and ensures formatting consistency. Normalization also improves model accuracy by simplifying text and improving the efficiency of model analysis and training.

TABLE VI. NORMALIZATION PROCESS

Stemming	Normalization
['location', 'market', 'this', 'strategy', 'and', 'clean', 'so', 'comfortable', 'shopping', 'here']	['location', 'market', 'strategy', 'clean', 'comfortable', 'shopping']

Stopword Removal

After tokenization, the next step is to remove stop words like 'and,' 'in,' and 'that.' This involves specifying a list of stop words, then splitting the text into tokens. Removing unimportant words makes the analysis more effective and reduces the computational load of the model.

TABLE VII. STOPWORD REMOVAL PROCESS

Normalization	Stopword Removal
['location', 'market', 'strategy', 'clean', 'comfortable', 'shopping']	['location', 'market', 'strategy', 'clean', 'comfortable', 'shopping']

TF-IDF (Term Frequency-Inverse Document Frequency)

After the review text has successfully undergone the pre-filing stages such as case folding, cleaning, tokenization, and deletion of the word stop as described earlier, the next step is to apply the TF-IDF process or word weighing to improve the vector representation of the review text [9], using the following equation.

$$idf(t, D) = log \frac{|D|}{df(t,D)} \qquad (1)$$

Where |D| is the total number of documents in corpus D and df (t,D) is how many t are contained in the document [16]. The following is the training data that has passed the data preprocessing stage and can be seen in table 7 below.

TABLE VIII. RESULTS OF THE PRE PROCESSING PROCESS

Review Data	Sentiment
['location', 'market', 'strategy', 'clean', 'comfortable', 'shopping']	Positive

In the sentence "location" will be made as an example, there are many documents N=6, and term "location" appears in 1 document. This demonstrates how frequently the term is present in the dataset relative to the total number of documents.

Journal of Science and Applied Engineering (JSAE)
Vol. 7, No. 2, October 2024
DOI : 10.31328/jsae.v7i2.6438
ISSN Print : 2621-3745
ISSN Online : 2621-3753
(Page.18-26)

$$IDF(location) = log\frac{6}{1} = \log(6) \approx 0.778 \qquad (2)$$

Thus, the term 'location' has a weight of 0.778, which is the result of TF-IDF calculations. These results provide an indication of the importance of the term in the context of the dataset used. Full details of the results of the TF-IDF calculation can be seen in Table IX below.

TABLE IX. TF-IDF WEIGHTING RESULTS

Term	TF (Doc 1)	DF	IDF	TF (Doc 1)
location	1	1	0.778	0.778
market	1	1	0.778	0.778
strategy	1	1	0.778	0.778
clean	1	1	0.778	0.778
comfortable	1	1	0.778	0.778
shopping	1	1	0.778	0.778

DISCUSSION

SVM implementation begins with the data capture stage using the Instant Data Scraper tool, followed by labeling and pre-processing stages such as case folding, cleaning, and TF-IDF. This ensures that the data is well-prepared for accurate model training. The pre-processed data is then used to train the SVM model effectively. Furthermore, Support Vector Machine is implemented, including model training and evaluation. The following sections will discuss each stage in detail, including the methodology and outcomes, to provide a comprehensive understanding of the process.

Support Vector Machine Classification

This Support Vector Machine (SVM) method was chosen because of its ability to use kernel techniques to map data to higher dimensions[17]-[18], find hyperplane lines that maximize margins between classes, and handle complex data. according to S. Salcedo-Sanz, SVM is also effective in handling data with many features and high variability, making it suitable for a variety of practical applications [19]. In addition, this algorithm is also known for its ability to handle data with a large number of features and high variability. This capability makes it particularly useful in a variety of practical applications in the real world, where data often has varying structures and complexities.

Model Training

The data that has gone through preprocessing and converted into TF-IDF representations is used to train the SVM model. First, the data is divided into two parts: 90% for training and 10% for testing, and this ratio is tested to 10% of training data and 90% of testing data. The total number of data that has gone through preprocessing is 1,299 samples. Furthermore, the kernels mentioned above are used with parameters C = 1 and gamma = 0.1. The C parameter determines the balance between the correct classification and the decision margin, while the gamma parameter controls the extent to which a single training instance affects the model [20]. Additionally, this process involves tuning parameters to achieve an optimal balance between overfitting and underfitting, so that the model can perform at its best on data that has never been seen before.

Classification Results

The results of SVM testing with linear, polynomial (poly), RBF, and sigmoid kernels show that the sigmoid kernel provides the best performance in data classification. This indicates its superior ability to capture complex patterns in the data. Evaluation involves confusion matrix, accuracy, precision, recall, and F1-score with various ratios of training and testing data. Figure 2, showing the confusion matrix for the sigmoid kernel, illustrates the number of true and false predictions for each class, helping to assess the model's strengths and weaknesses and identify areas for improvement.

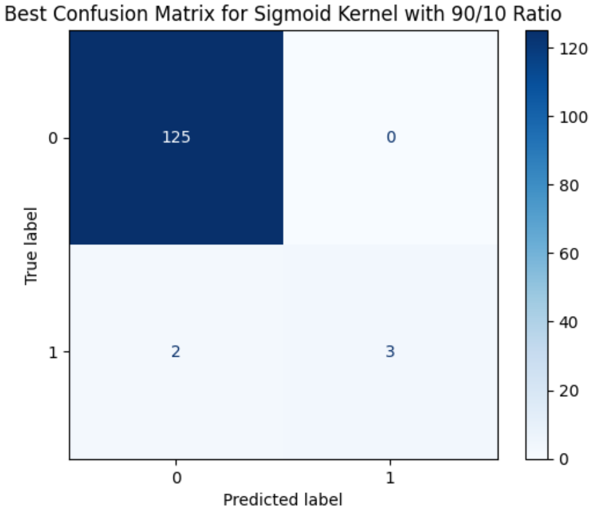

Figure 2. Highest Accuracy Kernel

Figure 2 above shows the best Confusion Matrix of the Sigmoid Kernel with a 90/10 data ratio: 125 true-negative, 0 false-positive, 2 false-negative, and 3 true-positive. The following are the results of manual calculations using the Confusion Matrix.

$$Accuracy = \frac{125+3}{125+3+2+0} = \frac{128}{130} = 98.46\%$$

$$Precision = \frac{125}{125+2} = \frac{125}{127} = 98.46\%$$

$$Recall = \frac{125}{125+0} = 1\ atau\ 100\%$$

$$f1 - scrore = 2 \times \frac{0.9843 \times 1}{0.9843 \times 1} 2 \times \frac{0.9843}{1.9843} = 99\%$$

The following table presents the overall results of testing and training for each kernel type, including Linear, Poly, RBF, and Sigmoid, with varying ratios of training and test data. This table provides an overview of the performance of each Kernel under various conditions.

TABLE X. LINEAR KERNEL TEST RESULTS

	Ratio	Accuracy	Precision	Recall	F1 Score
LINEAR KERNEL	90/10	97.69%	100.00%	40.00%	57.14%
	80/20	98.08%	100.00%	54.55%	70.59%
	70/30	97.69%	100.00%	43.75%	60.87%
	60/40	97.12%	81.82%	40.91%	54.55%
	50/50	96.62%	85.71%	22.22%	35.29%
	40/60	96.28%	80.00%	12.50%	21.62%
	30/70	95.82%	0.00%	0.00%	0.00%
	20/80	95.87%	0.00%	0.00%	0.00%
	10/90	95.81%	0.00%	0.00%	0.00%

In table 8, describes the SVM linear kernel that shows varying performance with training and testing ratios. Highest accuracy and precision are achieved at 90/10, 80/20, and 70/30 ratios, but lower recall performance.

Journal of Science and Applied Engineering (JSAE)
Vol. 7, No. 2, October 2024
ISSN Print : 2621-3745
ISSN Online : 2621-3753
DOI : 10.31328/jsae.v7i2.6438
(Page.18-26)

TABLE XI. POLYNOMIAL KERNEL TEST RESULTS

	Ratio	Accuracy	Precision	Recall	F1 Score
. POLYNOMIAL KERNEL	90/10	96.15%	0.00%	0.00%	0.00%
	80/20	95.77%	0.00%	0.00%	0.00%
	70/30	95.90%	0.00%	0.00%	0.00%
	60/40	95.77%	0.00%	0.00%	0.00%
	50/50	95.85%	0.00%	0.00%	0.00%
	40/60	95.90%	0.00%	0.00%	0.00%
	30/70	95.82%	0.00%	0.00%	0.00%
	20/80	95.87%	0.00%	0.00%	0.00%
	10/90	95.81%	0.00%	0.00%	0.00%

Based on Table XI, the polynomial kernels in the SVM model show 95% accuracy across all training/testing ratios and consistency in other metrics such as precision, recall, and F1-score. This signifies that the polynomial kernel is highly effective and stable in data classification, suitable for applications that require high accuracy and stability.

TABLE XII. RBF KERNEL TEST RESULTS

	Ratio	Accuracy	Precision	Recall	F1 Score
RBF KERNEL	90/10	97.69%	100.00%	40.00%	57.14%
	80/20	97.31%	100.00%	36.36%	53.33%
	70/30	97.18%	100.00%	31.25%	47.62%
	60/40	96.73%	100.00%	22.73%	37.04%
	50/50	95.85%	0.00%	0.00%	0.00%
	40/60	95.90%	0.00%	0.00%	0.00%
	30/70	95.82%	0.00%	0.00%	0.00%
	20/80	95.87%	0.00%	0.00%	0.00%
	10/90	95.81%	0.00%	0.00%	0.00%

In Table XII. It shows high accuracy and precision at 90/10, 80/20, and 70/30 ratios, signaling the model's good ability in classification and avoiding false positives. However, recalls and F1-scores vary and tend to be lower, suggesting some positive cases are missed. Overall, the RBF kernel is consistent in accuracy and precision, but less optimal in recall and F1-score at certain ratios.

TABLE XIII. SIGMOID KERNEL TEST RESULTS

	Ratio	Accuracy	Precision	Recall	F1 Score
SIDGMOID KERNEL	90/10	97.69%	100.00%	40.00%	57.14%
	80/20	97.31%	100.00%	36.36%	53.33%
	70/30	97.18%	100.00%	31.25%	47.62%
	60/40	96.73%	100.00%	22.73%	37.04%
	50/50	95.85%	0.00%	0.00%	0.00%
	40/60	95.90%	0.00%	0.00%	0.00%
	30/70	95.82%	0.00%	0.00%	0.00%
	20/80	95.87%	0.00%	0.00%	0.00%
	10/90	95.81%	0.00%	0.00%	0.00%

ISSN Print : 2621-3745
ISSN Online : 2621-3753
(Page.18-26)
Journal of Science and Applied Engineering (JSAE)
Vol. 7, No. 2, October 2024
DOI : 10.31328/jsae.v7i2.6438

Table XIII, Support Vector Machine model with Sigmoid kernel shows very high accuracy at most data ratios, close to 1. However, precision decreased at 20/80 ratios, and recalls were lower at 90/10 and 20/80 ratios, indicating difficulty in identifying all positive cases. Recall and F1-score declined more with smaller data ratios, exposing challenges with less training data.

This study analyzes sentiment about the facilities of the Blimbing Traditional Market using the Support Vector Machine (SVM) method, which was chosen because of its ability to classify effectively by maximizing class margins. The SVM's performance will be evaluated in terms of its precision and accuracy in capturing user sentiment. The results will be compared with previous studies to assess the strengths and limitations of the method in sentiment analysis.

TABLE XIV. COMPARISON OF PREVIOUS RESEARCH AUTHOR'S STUDY

Researcher	Object	Metode		Testing	Accuracy
(Satria, Suarjaya, and Pratama 2022)	Waste	Support Vector Machine 2.339 dataset		4 X	91.67%
		Naïve Bayes 2.339 dataset		4 X	63.89%
Our Proposed	Waste Cleanliness	Support Vector Machine 1.299 Dataset	Linear	9 X	98.07%
			Polynomial	9 X	96.15%
			Rbf	9 X	97.69%
			Sigmoid	9 X	98.46%

A comparison of previous research with current research can be seen in Table XIV. From the table, it is known that the current research uses the Support Vector Machine (SVM) method with various kernels and a 90/10 to 10/90 training and test data-sharing ratio in 9 tests. This approach allows for a more comprehensive evaluation of kernel performance across different data splits. In addition, this study also considers various variations in kernel selection to get optimal results. The accuracy obtained is higher than that of previous research, with the best accuracy of each kernel reaching 98.46%.

Meanwhile, previous research used two methods, namely SVM and Naive Bayes. The data used amounted to 2,339 with 4 tests. Previous research has not only compared these two methods but also explored the limitations of each in the context of different datasets. This comparison highlights the effectiveness of SVM in various settings. The accuracy obtained in SVM reached 91.67%, while in the Naive Bayes method reached 63.89%. This difference suggests that the use of SVMs with multiple kernels in the current study provides a significant improvement in the accuracy of sentiment analysis.

CONCLUSION

In the tests carried out, the researcher found that the Support Vector Machine (SVM) was effective for the analysis of the cleanliness sentiment of the traditional Blimbing market. Four kernel types were tested: Linear, Polynomial (Poly), Rbf, and Sigmoid, with a ratio of training and test data from 90/10 to 10/90. The use of these multiple kernels allows for a thorough evaluation and identification of the best performance of each model. The results show that Linear Kernel achieves an accuracy of 98.07% at 80/20 ratio, poly 96.15% at 90/10 ratio, Rbf 97.69% at 90/10 ratio, and Sigmoid 98.46% at 90/10 ratio.

This indicates that each Kernel has a high potential for data classification at a given ratio, providing insight into the performance of each Kernel under various data sharing conditions. The varied performance of each Kernel also highlights the importance of selecting the appropriate Kernel for specific datasets. This SVM model surpasses the accuracy of previous studies, which reached 78.76% for similar topics. This proves that SVM with Kernel variations and test ratios can perform better sentiment analysis of traditional market facilities compared to the methods in previous studies, showing its potential in improving the quality of sentiment analysis in this field.

REFERENCES

[1] M. Purnomo, F. Otten, and H. Faust, "Indonesian traditional market flexibility amidst state promoted market competition," *Soc. Sci.*, vol. 7, no. 11, 2018, doi: 10.3390/socsci7110238.
[2] M. H. Shuman, *The Small-Mart Revolution: How Local Businesses Are Beating the Global*

ISSN Print : 2621-3745
ISSN Online : 2621-3753
DOI : 10.31328/jsae.v7i2.6438
(Page.18-26)

Competition. San Francisco, California, Amerika Serikat.: Berrett-Koehler Publishers, 2006.

[3] L. G. Palamuleni, "Effect of sanitation facilities, domestic solid waste disposal and hygiene practices on water quality in Malawi's urban poor areas: a case study of South Lunzu Township in the city of Blantyre," *Phys. Chem. Earth*, no. January 2002, pp. 845–850, 2002, doi: 10.1016/S1474-7065(02)00079-7.

[4] W. Fahmi and K. Sutton, "Cairo's contested garbage: Sustainable solid waste management and the Zabaleen's right to the city," *Sustainability*, vol. 2, no. 6, pp. 1765–1783, 2010, doi: 10.3390/su2061765.

[5] M. Trenz and B. Berger, "Analyzing online customer reviews - An interdisciplinary literature review and research agenda," in *ECIS 2013 - Proceedings of the 21st European Conference on Information Systems*, 2013.

[6] J. R. Taylor and S. T. Lovell, "Mapping public and private spaces of urban agriculture in Chicago through the analysis of high-resolution aerial images in Google Earth," *Landsc. Urban Plan.*, vol. 108, no. 1, pp. 57–70, 2012, doi: 10.1016/j.landurbplan.2012.08.001.

[7] F. Hemmatian and M. Karim Sohrabi, "A survey on classification techniques for opinion mining and sentiment analysis," *Springer*, 2017.

[8] K. Ravi and V. Ravi, *A survey on opinion mining and sentiment analysis: Tasks, approaches and applications*, vol. 89, no. November. 2015. doi: 10.1016/j.knosys.2015.06.015.

[9] A. Gupta, P. Tyagi, T. Choudhury, and M. Shamoon, "Sentiment Analysis Using Support Vector Machine," *Proc. 4th Int. Conf. Contemp. Comput. Informatics, IC3I 2019*, pp. 49–53, 2019, doi: 10.1109/IC3I46837.2019.9055645.

[10] M. Ahmad, K. Jawad, M. B. Alvi, and M. Alvi, "Google Maps Data Analysis of Clothing Brands in South Punjab, Pakistan," *EAI Endorsed Trans. Scalable Inf. Syst.*, vol. 10, no. 3, 2023, doi: 10.4108/eetsis.v10i3.2677.

[11] B. S. Shekhawat, "Sentiment Classification of Current Public Opinion on BREXIT: Naïve Bayes Classifier Model vs Python's TextBlob Approach," *Natl. Coll. Irel.* , 2019.

[12] K. Kucher, C. Paradis, and A. Kerren, "The state of the art in sentiment visualization," *Comput. Graph. Forum*, vol. 37, no. 1, pp. 71–96, 2018, doi: 10.1111/cgf.13217.

[13] S. Zhou, Q. Chen, X. Wang, and X. Li, "Hybrid deep belief networks for semi-supervised sentiment classification," *COLING 2014 - 25th Int. Conf. Comput. Linguist. Proc. COLING 2014 Tech. Pap.*, pp. 1341–1349, 2014.

[14] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, "Big data preprocessing: methods and prospects," *Big Data Anal.*, vol. 1, no. 1, pp. 1–22, 2016, doi: 10.1186/s41044-016-0014-0.

[15] J. Singh and V. Gupta, *A systematic review of text stemming techniques*, vol. 48, no. 2. 2017. doi: 10.1007/s10462-016-9498-2.

[16] H. D. Abubakar and M. Umar, "Sentiment Classification: Review of Text Vectorization Methods: Bag of Words, Tf-Idf, Word2vec and Doc2vec," *SLU J. Sci. Technol.*, vol. 4, no. 1&2, pp. 27–33, 2022, doi: 10.56471/slujst.v4i.266.

[17] R. Zuo and E. J. M. Carranza, "Support vector machine: A tool for mapping mineral prospectivity," *Comput. Geosci.*, vol. 37, no. 12, pp. 1967–1975, 2011, doi: 10.1016/j.cageo.2010.09.014.

[18] N. Cristianini and B. Schölkopf, "Support vector machines and kernel methods: The new generation of learning machines," *AI Mag.*, vol. 23, no. 3, pp. 31–41, 2002.

[19] S. Salcedo-Sanz, J. L. Rojo-Álvarez, M. Martínez-Ramón, and G. Camps-Valls, "Support vector machines in engineering: An overview," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 4, no. 3, pp. 234–267, 2014, doi: 10.1002/widm.1125.

[20] I. S. Al-Mejibli, J. K. Alwan, and D. H. Abd, "The effect of gamma value on support vector machine performance with different kernels," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 5, pp. 5497–5506, 2020, doi: 10.11591/IJECE.V10I5.PP5497-5506.