Journal of Applied Informatics and Computing (JAIC)
Vol.8, No.2, December 2024, pp. 550~556
e-ISSN: 2548-6861

Detecting Fake Reviews Using BERT and Sublinear_TF Methods on Hotel Reviews in the Lombok Tourism Area

Zulpan Hadi [1]*, M. Zulpahmi [2], Zaenudin [3]**, Akmaludin Asrory[4]****
* Rekayasa Sistem Komputer, Universitas Teknologi Mataram, Indonesia
** Teknik Informatika, Universitas Teknologi Mataram, Indonesia
** Komputerisasi Akuntansi, Universitas Teknologi Mataram, Indonesia
** Teknik Informatika, Universitas Teknologi Mataram, Indonesia
zlpnhadi@gmail.com[1], pahmijorge04@gmail.com [2], zen3d.itb@gmail.com [3] akmaludin.asrory@gmail.com [4]

Article Info

Article history:

Received 2024-10-21
Revised 2024-11-24
Accepted 2024-11-25

Keyword:

Fake Reviews,
BERT,
Random Forest,
SVM,
Sublinear_TF.

ABSTRACT

The number of visitors to Lombok, one of the famous tourist destinations in Indonesia, increased from 400,595 in 2020 to 1,376,295 in 2022. Although the government supports the hotel industry, fake reviews are a significant problem that can damage hotel reputations and mislead tourists. This study uses BERT and Sublinear_TF feature extraction techniques to analyze fake reviews from three main areas: Gili Trawangan, Senggigi, and Kuta. BERT detects fake reviews by understanding the context of words, while Sublinear_TF emphasizes more informative words by reducing the weight of irrelevant common words. The results showed that the more extensive and diverse dataset from Gili Trawangan had the best classification results. The combination of BERT and Random Forest achieved the highest accuracy of 0.84. Overall, BERT excels in Gili Trawangan with an accuracy of 0.79 for SVM and 0.84 for Random Forest. In contrast, smaller and more homogeneous datasets such as Senggigi and Kuta have lower accuracy. In addition, Sublinear_TF performed well on Gili Trawangan with an accuracy of 0.82 using SVM and 0.83 using Random Forest; however, its performance declined in Senggigi and Kuta. BERT and Sublinear_TF techniques are more effective on large and diverse datasets such as Gili Trawangan. Sublinear_TF is better for varied data but less effective on more homogeneous datasets, while BERT with Random Forest showed the highest accuracy due to its ability to capture broader language context. This suggests that the size and variety of the dataset highly influence the success of fake review classification techniques.

This is an open access article under the CC–BY-SA license.

I. INTRODUCTION

Lombok is one of the islands famous for its natural beauty, and many local and foreign tourists come there. In the last three years, the number of tourists has increased rapidly. The number of tourists was recorded at 400,595 in 2020. It increased to 964,036 in 2021 and 1,376,295 in 2022 [1]. This positive trend has encouraged Lombok's tourism and hospitality sector to continue to develop and add facilities to receive tourists. However, as the number of tourists increases, the hospitality industry faces a new problem: fake reviews on online review platforms..

The hospitality industry faces a significant problem with fake reviews. Hotel managers and sellers often work with reviewers to create fake reviews to increase their hotel ratings[2]. In addition, fake reviews can also worsen business competition between hotels because they provide false information about the quality of service [3]. As a result, hotel managers who provide quality service are disadvantaged, losing the reputation they have worked so hard to build. Fake reviews can mislead tourists and make their experience not meet expectations. This can reduce tourists' interest in visiting Lombok.

Several studies have been conducted to find fake reviews. In some studies, Support Vector Machine (SVM) and Random Forest algorithms were used; SVM showed 98% accuracy in detecting fake reviews [2]. Another study combined n-gram and TF-IDF methods with chi-square feature selection, which showed 92.19% accuracy [3]. These methods show that understanding the context of words and selecting relevant features can improve the accuracy of identifying fake reviews.

A study also tried to find fake reviews with post tags. Using word tagging and distribution functions [4]. This study analyzed 4,478 genuine reviews and 856 fake reviews. Six hundred twenty-eight fake reviews aimed to increase product sales or the store owner's brand name, while 228 fake reviews aimed to bring down competitors. The Support Vector Machine (SVM) and Random Forest algorithms were used during the testing process, giving the best results with 98% accuracy. Although the data was imbalanced, this model showed no bias towards the majority class.

The study used a method to detect fake reviews using n-gram and TF-IDF. Then, feature selection was performed using chi-square to find the most relevant features. After that, classification was performed using logistic regression algorithms, Random Forest, and SVM [5]. This study successfully identified a thousand of the best features by understanding the context of words and using chi-square to select relevant features. Combining these features with the SVM algorithm produced the highest accuracy of 92.19%. These results indicate that, even with the same algorithm, the proper method to understand the context of words and carefully select features can increase the accuracy of detecting fake reviews by 3.49%.

Many researchers have used various feature extraction methods in this study. These include C-LSTM, HAN, Convolutional HAN, Char-level C-LSTM, BERT, DistilBERT, and RoBERTa [6]. Experimental results on two benchmark datasets show that RoBERTa outperforms other state-of-the-art methods by about 7% in the mixed domain. RoBERTa has the highest accuracy of 91.2%, indicating that it has excellent potential to be used as a basis for future studies on fake review detection.

This study used three datasets, YelpChi, YelpNYC, and YelpZip, to detect fake reviews. The results showed that the Random Forest algorithm performed the best out of the four algorithms tested: Random Forest, Logistic Regression, Naïve Bayes, and Decision Tree [7]. The results showed that the Random Forest algorithm best detected fake reviews across all datasets, including YelpChi, YelpNYC, and YelpZip. This algorithm successfully outperformed other algorithms, such as Logistic Regression, Naïve Bayes, and Decision Tree, by consistently showing high accuracy across datasets.

The purpose of combining two feature extraction methods, Bag-of-Words (BOW) and Term Frequency-Inverse Document Frequency (TF-IDF), is to improve the performance of fake review detection [8]. This combination is intended to create more relevant and accurate features for finding fake reviews. To show how effective the combination of these two techniques is in improving detection accuracy, the Support Vector Machine algorithm produces the best results.

This study suggests using the BERT model to extract word features. The results show that the classifier using the Support Vector Machine (SVM) algorithm outperforms other methods regarding accuracy and F1 score. With an accuracy of 87.81%, this is an improvement of 7.6% compared to the accuracy of the classifier used in the previous study [9].

Sublinear_TF and BERT approaches will be used in this study to find fake reviews on online review platforms. We will also use Random Forest and SVM algorithms that have been proven effective in classification. While BERT can extract more profound meaning and context from reviews, improving the accuracy of identifying fake reviews, Sublinear_TF is a weighting technique that addresses word frequency without causing excessive inflation. We will conduct experiments to validate each of the Sublinear_TF and BERT models and evaluate how both models detect fake reviews. By using these methods, we hope to improve the accuracy of detecting fake reviews, thereby helping to maintain the reputation of the hospitality sector in Lombok and supporting the growth of the tourism industry. In addition, more accurate and reliable information will be presented to tourists, allowing them to make more informed decisions in choosing accommodation.

II. METHOD

Several stages are carried out in this research, one of which is the first data collection. The hotel review dataset in tourist areas will be collected from the TripAdvisor site at this stage. Furthermore, the case folding, tokenizing, stopword, and stemming processes will be carried out on the data collected for preprocessing. After that, the feature extraction process will be carried out using sublinear_TF and BERT. These two techniques will be compared with two algorithms, SVM and Random Forest, and the last stage will be evaluated using a chaos matrix. This is the research route.

The following is an explanation of the stages in the research flowchart. The first stage is collecting data from hotel reviews from the TripAdvisor platform. This platform was chosen as the primary data source because it has often been used as a reference in hotel review sentiment analysis research [10][11][12]. The data collected includes hotel reviews in the Lombok tourist area, especially in Gili Trawangan, Kuta Mandalika, and Senggigi. This data is used to conduct further analysis, which includes reviews that are considered false and authentic. Pre-processing is done after the data is collected. This includes data cleaning, such as removing punctuation, changing the text format to be consistent, and possibly normalizing the word. Pre-processing is done to make the data more accessible to process by machine learning models. Sublinear_TF and BERT are two methods used to complete the next stage. Sublinear_TF is a

way to scale the term frequency (TF) value using a non-linear function, usually a logarithmic function. Sublinear_TF calculates frequency in a smoother way than using the frequency of word occurrence directly. This helps reduce the effects of words that appear too often in a document. TF represents how often a word appears in a document in the conventional TF-IDF (Term Frequency—Inverse Document Frequency) method [13]. A word with a very high TF value does not always indicate its importance in classification.

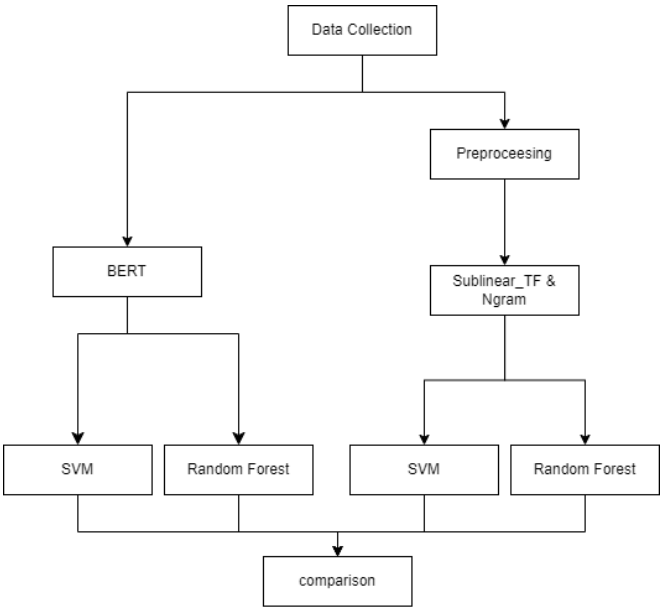

Figure 1. Research Flow

Sublinear scaling, typically used with logarithmic functions (e.g., 1 + log(TF)), reduces the effect of words that occur too frequently to prevent them from dominating the feature vector. This helps:

1) Addressing the influence of common words: TF subliners value frequently occurring words less, so they do not dominate the computation as much.
2) Improving model stability: The computation scale can become too large if many words are used. Sublinear scaling helps stabilize the contribution of these words in the overall representation.
3) Reducing data imbalance: Some words in a document may appear many times, but not significantly. TF subliners help reduce this imbalance and give fairer weight to other words.

This study also uses the N-gram technique, which identifies sequences of related words in reviews. This N-gram method can find more complex patterns and contexts in text, which often improves classification accuracy. Many algorithms such as SVM, Random Forest, and BERT have been used to classify fake reviews in recent years [9]. automatically. Combining Sublinear_TF and N-gram is expected to increase the effectiveness of detecting fake reviews. This will support the study's success in providing more accurate and reliable information.

However, BERT, known as Bidirectional Encoder Representations from Transformers, is a neural network-based technique to leverage previous learning in natural language processing (NLP). This model is beneficial for Google in helping them better understand the words involved in search queries [14]. Unlike traditional methods that only use one-way word sequences (left-to-right or a combination of left-to-right or left-to-left), BERT can train a language model by considering the entire set of words in a sentence or query through bidirectional training. BERT allows language models to learn the context of words based on surrounding words, not just those that precede or follow them. Because the contextual representation of words starts "from the very bottom of the neural network," Google calls BERT "highly bidirectional." [15]. Categorizing text fragments based on their context is one of the critical tasks in fake review detection; it allows developers to perform text classification in the natural language processing process [16].

Testing both methods BERT and Sublinear_TF was done using the Random Forest and SVM (Support Vector Machine) classification algorithms. Using a supervised machine learning method, the SVM classification algorithm developed by Vladimir Vapnik can predict class patterns [17]. Simply put, the function of SVM is to find the best hyperplane that can distinguish two classes in the input space. The ideal hyperplane can be identified by measuring the margin and finding its maximum value [18]. The kernel function in the SVM algorithm is used to handle non-linear problems by converting them into linear separables. Polynomial, linear, sigmoid, and radial basis functions (RBF) are some kernel functions often used in classification. Each kernel function has different parameter values, and the Grid Search method can be used to find the best parameter values for the selected kernel function. However, the Random Forest algorithm, developed by J. Ross Quinlan and derived from the ID3 approach to building decision trees, is considered very effective for classification problems in data mining and machine learning. To predict previously unseen data, Random Forest maps class attributes. To analyze problems with a set of independent data depicted in the form of a tree diagram, the "divide and conquer" approach to decision trees is used [19]. In a decision tree, a series of questions are systematically arranged, with each question being given a branch based on the attribute value, and the process stops at a leaf of the tree indicating the predicted class of the variable [20].

Here are the steps in making a decision tree using the Random Forest algorithm:

1) Check Data Labels. If all labels in the data are homogeneous, leaves will be formed with the overall data label value. This is the step in making a decision tree using the Random Forest algorithm.

2) Calculate Information Value. You can calculate the information value using the following formula:

$$Info(D) = -\sum_{i=1}^{m} pi \log 2(pi) \quad (1)$$

The average information required to identify a tuple D is the probability that it belongs to a particular class, also called the entropy of D [21].

A data set D is separate from the set of values A if the values of A are discrete, so the values in each branch are pure and of the same type. The number of possible subsequent branches is calculated after the first branch:

$$InfoA(D) \sum_{j}^{v} \frac{|Dj|}{|D|} xInfoA(Dj) \quad (2)$$

3) Calculating the information value with the formula.

4) Pay attention to the data content of each attribute. where |Dj| |D| is the partition's weight, and the information needed to classify tuple D in partition A is InfoA(D). The result of this equation is proportional to the quality of the resulting partition. The value of an attribute determines how important the attribute is for constructing the decision tree. If the attribute has a continuous value, the split_point will be found by sorting all data according to the attribute from the smallest to the largest, then taking the average from one data to the next. The split_point value selected sequentially will be used to calculate the information value. (4) The advantage value for each feature will be calculated using the formula (7.8). The value with the most significant advantage will be the decision tree branch.

$$Gain(A) = info(D) - InfoA(D) \quad (3)$$

5) The calculation is repeated in stages 1 to 4 after forming the decision tree branch. However, if the branch reaches the maximum number of branches allowed, a leaf will be formed with the most significant data value.

III. Results And Discussion

The dataset used in this study comes from the Tripadvisor platform, which contains various tourist reviews about their stay experiences at hotels in the Lombok area. To analyze differences in the level of satisfaction and preferences of tourists in each tourist area, the dataset is divided into three main areas based on hotel location: Senggigi, Gili Trawangan, and Kuta. An overview of the dataset for each hotel area is given here.

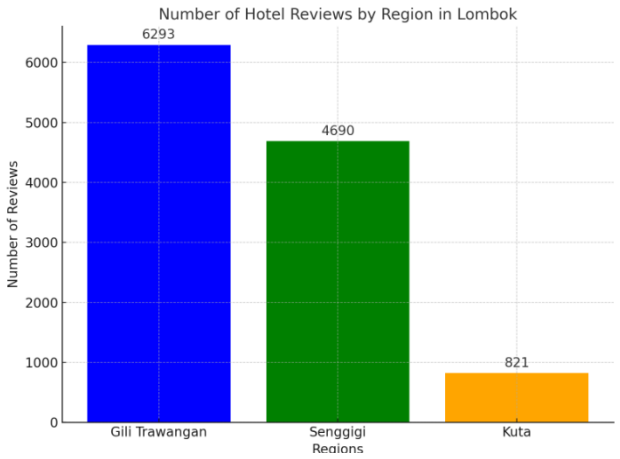

Figure 2. Dataset division

The graph above shows how hotels in Lombok are divided by region. Gili Trawangan has the highest number of reviews with 6,293, followed by Senggigi with 4,690, and Kuta with 821.

	cleaned_text	label
0	Hotel con posizione strategica vicino al porto...	1
1	Restaurant Sea Horse de lhÃtel Ã fuir Crevet...	1
2	I booked the Ombak Romantic Dinner for my girl...	0
3	Asians and women should avoid this hotel The m...	1
4	We Stay in this hotel for a family vacation th...	0

Figure 3. Example dataset

Review 1 uses a lot of nouns and verbs, such as “restaurant,” “buffet,” and “satisfied,” to show an accurate and objective experience. Review 3 also uses a lot of nouns and verbs, such as “hotel,” “manager,” “avoid,” and “yell,” to show concrete actions and specific situations, indicating a first-hand experience. In contrast, Review 2 uses many adjectives, such as “great special experience,” which are overly promotional and overly promotional, which are typical of fake reviews. Review 4 also uses adjectives such as “very welcoming” and overly adverbs such as “really,” which make the review feel less objective and overly complimentary, indicating a fake review.

We classify fake hotel reviews based on this data using various data processing techniques and classification algorithms. The following table shows the results of this classification, which shows how the model performs with various feature extraction techniques on three different hotel datasets: Senggigi, Gili Trawangan, and Kuta, Lombok. The results of the performance evaluation of the methods used are as follows.

TABLE I
PERFORMANCE EVALUATION RESULTS

Feature Exstraction	Dataset	Classification	
		SVM	Random Forest
BERT	Sengigi	0.75	0.74
	Gili Trawangan	0.79	0.84
	Kuta, Lombok	0.65	0.6
Sublinear_TF	Sengigi	0.73	0.72
	Gili Trawangan	0.82	0.83
	Kuta, Lombok	0.61	0.61

The results show that the BERT feature extraction method performs best on the Gili Trawangan dataset, with an accuracy of 0.79 for SVM and 0.84 for Random Forest. On the other hand, the Kuta Lombok dataset has the lowest accuracy, with 0.65 for SVM and 0.60 for Random Forest. This difference in performance is related to the size of the dataset and the essential differences in the operations of BERT and Sublinear_TF, the two feature extraction techniques used.

BERT is a transformer-based model that takes the context of words from the previous and following words. This approach allows BERT to produce richer feature representations because it further considers the semantic relationships between words in a sentence. In the Gili Trawangan dataset, which has 6,293 review data, BERT was able to identify more complex patterns and find variations in the language used, both in words and sentence structure. As a result, BERT performs better, especially for more extensive and varied datasets. This understanding is crucial in places where understanding semantic context is critical to improving classification accuracy.

In contrast, Sublinear_TF uses a more straightforward method, converting text into a numeric representation based on the frequency of words appearing in the reviews. However, Sublinear_TF reduces the weight of these frequencies logarithmically to avoid frequent words dominating the representation. Frequency-based methods can produce pretty good results on smaller datasets, such as the Kuta dataset, which has 821 reviews. This is because the words used are more uniform in smaller datasets. However, because Sublinear_TF does not consider the semantic context between words, this technique may have difficulty identifying complex language patterns. Ultimately, this may lead to less accurate classification of datasets with more significant language variation.

In addition, the classification model used affects the performance difference. Random Forest tends to be superior in handling more extensive and more varied datasets, such as Gili Trawangan, compared to SVM, as shown by its better performance than SVM on this dataset. Meanwhile, Random Forest still has difficulty achieving its desired goals on smaller datasets, such as Kuta.

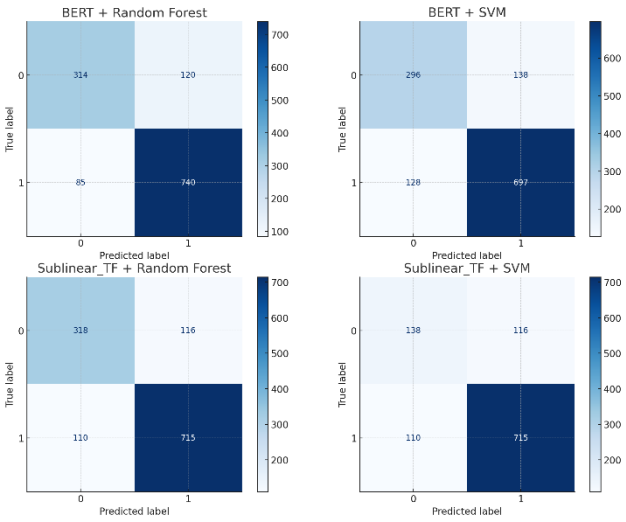

Figure 4. Confusion Matrix Gili Trawangan Dataset

The BERT + Random Forest model performed best with 740 True Positives and 314 True Negatives, low False Negatives (85) and False Positives (120). Sublinear_TF + Random Forest was also competitive, with 318 True Negatives, although the False Negatives (110) were higher. In contrast, the BERT + SVM and Sublinear_TF + SVM models produced more errors, with higher rates of False Negatives and False Positives. Overall, BERT + Random Forest and Sublinear_TF + Random Forest were the best choices to minimize prediction errors on the Gili Trawangan dataset, while the SVM-based models were less effective.

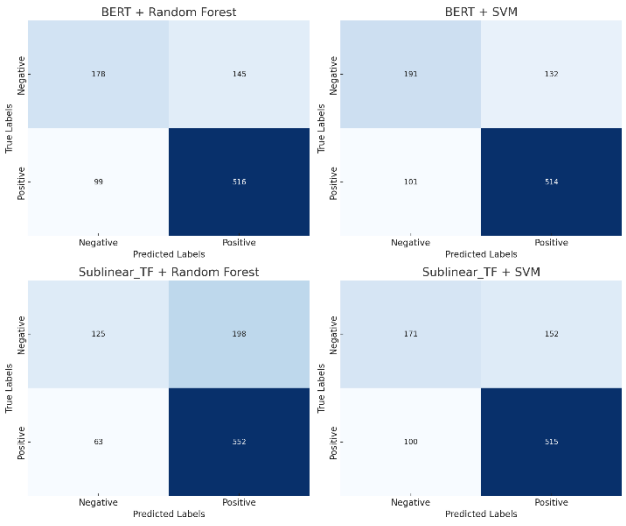

Figure 5. Confusion Matrix Sengigi Dataset

Of the four models tested, BERT + Random Forest performed the best with 516 True Positives (TP) and 178 True Negatives (TN). This model had 145 False Positives (FP) and 99 False Negatives (FN), so the data classification was quite good. BERT + SVM also gave competitive results, with 514 TP and 191 TN, although it had slightly more FP (132) and FN (101). Sublinear_TF + Random Forest recorded the

highest TP (552) but also had high FP (198) and FN (63), indicating that some of its optimistic predictions were less accurate. On the other hand, Sublinear_TF + SVM produced 515 TP, 171 TN, 152 FP, and 100 FN, indicating that the SVM-based model was less effective than Random Forest. Overall, BERT + Random Forest and BERT + SVM are the best choices for classification on the Senggigi dataset.

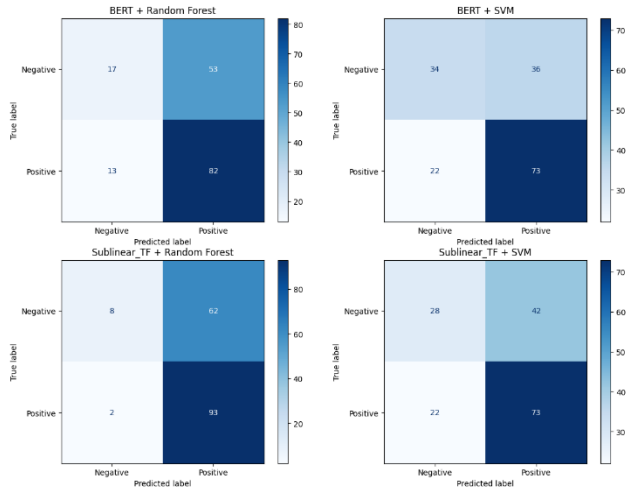

Figure 6. Confusion Matrix Kuta Lombok Dataset

The confusion matrix results of the four models on the Kuta dataset show differences in performance in classifying data. The BERT + Random Forest model performed exceptionally well identifying positive data with 82 TP and 17 TN but still made errors with 53 FP and 13 FN. On the other hand, BERT + SVM produced a slightly lower performance with 73 TP, 34 TN, and more errors, namely 36 FP and 22 FN. The Sublinear_TF + Random Forest model recorded the highest TP (93), but the error rate in classifying harmful data was relatively high, with only 8 TN and 62 FP. Sublinear_TF + SVM performed similarly to BERT + SVM, resulting in more errors in classifying harmful data.

Then matrix results show the show + Random Forest model consistently provides the best performance in all three datasets (Gili Trawangan, Senggigi, and Kuta) with lower prediction errors compared to other Sublinear_TF + Random Forest is also quite competitive. In comparison, the SVM-based model produces more errors. Overall, the Random Forest-based model is more effective for classification on this dataset than the SVM model.

IV. CONCLUSION

The results show that the BERT + Random Forest model provides the best results in classifying fake hotel reviews in Lombok, especially on the largest dataset, Gili Trawangan, with an accuracy of 0.84. Combining Sublinear_TF + Random Forest and BERT + SVM is a practical choice for classifying fake reviews, as it provides high accuracy and lower prediction errors. The Random Forest-based model outperforms SVM, while Sublinear_TF remains effective for small datasets, although it still needs improvement in reducing positive prediction errors.

REFERENCES

[1] Dinas Pariwisata NTB, "Jumlah Kunjungan Wisatawan ke Provinsi Nusa Tenggara Barat (NTB) | Satu Data NTB," Ntbprov.Go.Id. [Online]. Available: file:///E:/POLTEKPAR/PROYEK AKHIR/Jumlah Kunjungan Wisatawan ke Provinsi Nusa Tenggara Barat (NTB) _ Satu Data NTB.html

[2] G. S. Budhi, R. Chiong, and Z. Wang, "Resampling imbalanced data to detect fake reviews using machine learning classifiers and textual-based features," *Multimed. Tools Appl.*, vol. 80, pp. 13079–13097, 2021.

[3] R. Barbado, O. Araque, and C. A. Iglesias, "A framework for fake review detection in online consumer electronics retailers," *Inf. Process. Manag.*, vol. 56, no. 4, pp. 1234–1244, 2019, doi: 10.1016/j.ipm.2019.03.002.

[4] Z. Hadi, E. Utami, and D. Ariatmanto, "Detect Fake Reviews Using Random Forest and Support Vector Machine," *SinkrOn*, vol. 8, no. 2, pp. 623–630, 2023, doi: 10.33395/sinkron.v8i2.12090.

[5] Z. Hadi and S. Andi, "Detecting Fake Reviews Using N-gram Model and Chi-Square," *2023 6th Int. Conf. Inf. Commun. Technol.*, 2023, doi: 10.1109/ICOIACT59844.2023.10455895.

[6] R. Mohawesh *et al.*, "Fake Reviews Detection: A Survey," *IEEE Access*, vol. 9, pp. 65771–65802, 2021, doi: 10.1109/ACCESS.2021.3075573.

[7] M. Abdulqader, A. Namoun, and Y. Alsaawy, "Fake Online Reviews: A Unified Detection Model Using Deception Theories," *IEEE*, vol. 10, pp. 128622–128655, 2022, doi: 10.1109/ACCESS.2022.3227631.

[8] A. Ahmed, I. Bacho, and S. Talpur, "Identification of Real and Fake Reviews Written in Roman Urdu," vol. 5, no. 4, pp. 787–797, 2023.

[9] A. Q. Mir, F. Y. Khan, and M. A. Chishti, "Online Fake Review Detection Using Supervised Machine Learning And BERT Model," *Comput. Lang.*, 2023.

[10] M. Ott, C. Cardie, and J. T. Hancock, "Negative deceptive opinion spam," *NAACL HLT 2013 - 2013 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Main Conf.*, no. June, pp. 497–501, 2013.

[11] J. K. Rout, A. Dalmia, K. K. R. Choo, S. Bakshi, and S. K. Jena, "Revisiting semi-supervised learning for online deceptive review detection," *IEEE Access*, vol. 5, pp. 1319–1327, 2017, doi: 10.1109/ACCESS.2017.2655032.

[12] R. Hassan and M. R. Islam, "Detection of fake online reviews using semi-supervised and supervised learning," *2nd Int. Conf. Electr. Comput.* Commun. *Eng. ECCE 2019*, pp. 1–5, 2019, doi: 10.1109/ECACE.2019.8679186.

[13] J. Piskorski and G. Jacquet, "TF-IDF Character N-grams versus Word Embedding-based Models for Fine-grained Event Classification: A Preliminary Study," *Proc. Work. Autom. Extr. Socio-political Events from News 2020*, no. May, pp. 26–34, 2020.

[14] M. S. Isa, "Penerapan Algoritma BERT dalam Search Engine Google," Master of Computer Science. Accessed: Sep. 17, 2024. [Online]. Available: https://mti.binus.ac.id/2020/09/03/penerapan-algoritma-bert-dalam-search-engine-google/

[15] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media," *Int. Conf. Complex Networks Their Appl.*, vol. 881, 2019, doi: https://doi.org/10.1007/978-3-030-36687-2_77 .

[16] K. Florio, V. Basile, M. Polignano, P. Basile, and V. Patti, "Time of your hate: The challenge of time in hate speech detection on social media," *Appl. Sci.*, vol. 10, no. 12, 2020, doi: 10.3390/APP10124180.

[17] G. R. Ditami, E. F. Ripanti, and H. Sujaini, "Implementasi Support Vector Machine untuk Analisis Sentimen Terhadap Pengaruh Program Promosi Event Belanja pada Marketplace," *J. Edukasi dan Penelit. Inform.*, vol. 8, no. 3, p. 508, 2022, doi: 10.26418/jp.v8i3.56478.

[18] Y. X. Chu, X. G. Liu, and C. H. Gao, "Multiscale models on time series of silicon content in blast furnace hot metal based on Hilbert-Huang transform," *Proc. 2011 Chinese Control Decis. Conf. CCDC 2011*, pp. 842–847, 2011, doi: 10.1109/CCDC.2011.5968300.

[19] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical* Machine *Learning Tools and Techniques*. 2011. doi: https://doi.org/10.1016/C2009-0-19715-5.

[20] K. Dinas *et al.*, "Prediksi Jumlah Penggunaan BBM Perbulan Menggunakan Algoritma Decition Tree (C4.5) Pada," *J. Inform. dan Teknol.*, vol. 1, no. 1, pp. 56–63, 2018.

[21] L. T. E. . Kusrini, *Algoritma Data Mining. Buku Algoritma Data Mining*, I. Yogyakarta: C.V ANDI, 2009. [Online]. Available: https://books.google.co.id/books?id=-Ojclag73O8C&printsec=frontcover&hl=id#v=onepage&q&f=false