Sentiment Analysis of Online Media Headlines Using Lexicon Method and K-Nearest Neighbor Algorithm

Jasmine Nabila Novel,
Information Systems Department, Trilogi University, South Jakarta, 12760, Indonesia
Email: jasmine.nabila@trilogi.ac.id

Rudi Setiawan*,
Information Systems Department, Trilogi University, South Jakarta, 12760, Indonesia
Email: rudi@trilogi.ac.id

*Corresponding author Email: rudi@trilogi.ac.id

Abstract

During his two terms as president, Joko Widodo has launched several policies that have elicited various reactions from the public and have frequently been reported by online news media. With the increasing frequency of news coverage regarding President Joko Widodo, it is important to analyze the sentiment of news headlines published by online media. Therefore, appropriate labeling methods and classification algorithms are needed to accurately categorize the sentiment in news headlines as positive, negative, or neutral. In this study, sentiment analysis is conducted using the lexicon method (Inset Lexicon), which leverages a word dictionary to determine sentiment polarity, and the K-Nearest Neighbors (KNN) method for the classification algorithm. The results of the study, conducted on three websites—kumparan.com, detik.com, and cnnindonesia.com—indicate that a k value of 8 is the most optimal k-neighbors parameter, with the highest accuracy of 0.7067 on detik.com. Additionally, the average values for the three sentiment classes recorded a precision of 0.86, a recall of 0.39, and an f1-score of 0.37. This study aims to demonstrate how the Lexicon and K-Nearest Neighbors algorithms can be used to automatically determine sentiment, reducing the need for human judgment, and producing more objective sentiment analysis.

Keywords: Sentiment Analysis; Online News; K-Nearest Neighbor; Lexicon

I. Introduction

According to Lecheler & de Vreese (2019) in their book News Framing Effects: Theory and Practice, Journalistic news frames take a starting point in journalists' discretion and autonomy; these frames help journalists and news media organizations shape their selected topics in their own particular manner and style. and journalistic news frames are used in the adaptation and modification of frames from elites. This indicates that news can be formulated and framed to produce sentiment effects on the public. Sophie Lecheler, in the same book, also states that a news framing model affects opinions, where the framing process is defined by lending additional weight to an already accessible concept. Choosing specific news headlines can potentially trigger sentiment because it serves as the starting point where readers begin to form their initial perceptions of an issue or individual discussed.

President Joko Widodo, one of the most influential politicians in Indonesia who has served as President of the Republic of Indonesia for two terms from 2014 to 2024, has generated numerous sentiment perceptions, particularly in online news media. As known, President Joko Widodo's two terms in office have naturally led to public opinions, reflected in online news coverage, whether positive, negative, or neutral. As a significant figure in Indonesian politics, President Joko Widodo attracts considerable attention in online media. The choice of news headlines can greatly impact how the public reacts to and interprets the news, regardless of its content. Emotional or controversial headlines tend to provoke strong reactions from readers, whether in support of or criticism of President Joko Widodo.

To understand sentiment analysis, this study refers to several previous research works. The first study, conducted by Musfiroh et al. (2021) titled sentiment analysis of online lectures in indonesia from twitter dataset using inset lexicon aims to identify public views on online learning during the pandemic, particularly among students. The chosen method utilizes the InSet Lexicon as an Indonesian opinion word dictionary. This study, which classified 5,811 data points, found that 63.4% were negative, 27.6% were positive, and 8.9% were neutral. With an 80:20 training-to-testing data ratio, the study achieved an accuracy of 79.2%, precision of 72.9%, recall of 62.8%, and an F-measure of 67.4%.

The objective of this study is to understand the application of the Inset Lexicon method in sentiment labeling, with the goal of identifying the sentiment tendencies of three online news websites regarding President Joko Widodo—whether they report on him with positive, negative, or neutral sentiment. The labeling process is intended to be automated without requiring human assessment. Additionally, the study seeks to analyze the Confusion Matrix results from classifying news headlines with the keyword "Joko

Widodo" using the K-Nearest Neighbor (KNN) Classifier, based on data obtained from objective scraping. This will help determine the accuracy of the KNN Classifier based on the calculations from the confusion matrix.

II. Literature Review

Sentiment analysis has become an important tool in understanding public opinion in the digital era. Over the past five years, research in this field has focused on the use of deep learning techniques, especially transformers such as BERT (Devlin et al., 2019) and GPT Wu & Lode (2020), which have shown significant improvements in accuracy in capturing context and language nuances compared to traditional approaches. According to Sun et al. (2019), transformer models effectively overcome the limitations of dictionary-based and simple machine-learning methods in handling irony and sarcasm.

Research on specific applications, such as finance and banking, shows that sentiment analysis can be used to predict stock market trends and investor behavior through news and social media analysis (Bharathi & Geetha, 2017). In addition, Obagbuwa et al. (2023) highlighted the importance of sentiment analysis in mental health applications, especially in predicting language patterns related to depression on social media. In marketing, Yan & Li (2022) found that sentiment analysis can measure consumer perceptions of products more effectively and help companies respond to consumer needs quickly. Challenges in sentiment analysis include cross-language adaptability and complex context processing. The study Kowsher et al. (2022) introduces Bangla-BERT as a monolingual model specifically developed for Bangla language, trained on the largest dataset for the language. Bangla-BERT is evaluated through transfer learning based on hybrid deep learning models such as LSTM, CNN, and CRF for NER, and shows better results compared to the latest methods. However, challenges related to bias in language models and cross-cultural adaptability are still topics that need attention (Dominic et al., 2023).

III. Methodology

The objective of this study is to understand the application of the Inset Lexicon method in sentiment labeling to determine the polarity trends of online news websites and to enable automation without requiring human assessment. During the data collection process, observations were made on the activities of online news media related to news about Joko Widodo from November 2023 to February 2024. Subsequently, data collection focused on news headlines from online sources associated with the keyword "Joko Widodo" within the same timeframe. Using scraping techniques with Python's Selenium tool, the collected data included news headlines, sources, and publication years, which were compiled into a single CSV file. The preprocessing steps included case folding, tokenizing, stop word removal, and stemming. The polarity calculation used the Inset Lexicon, an Indonesian sentiment lexicon based on the research by Koto & Rahmaningtyas (2017). Each word's polarity was calculated with weights ranging from -5 to 5, where negative values represent negative sentiment and positive values represent positive sentiment; a sum of zero indicates a neutral sentiment. The classification was performed using the KNN algorithm. Each data point, already weighted in the previous steps, was used for the classification process. The steps are illustrated in Figure 1.

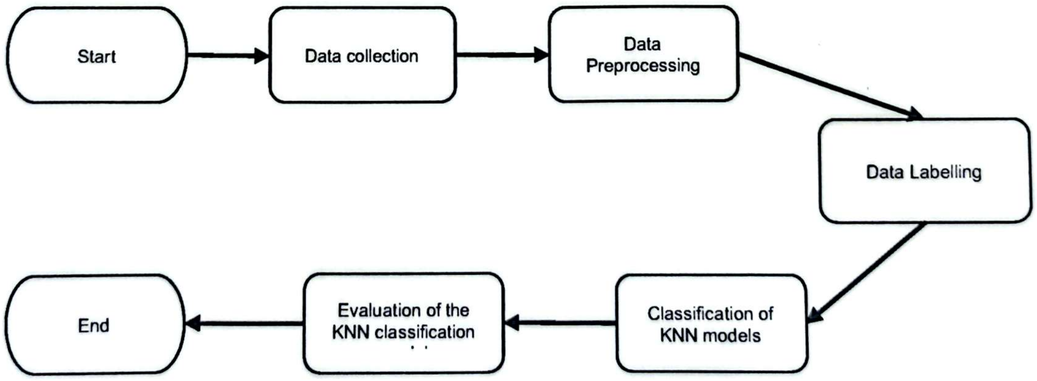

Start
Data collection
Data Preprocessing
Data Labelling
Classification of KNN models
Evaluation of the KNN classification
End

Figure 1. Flow Diagram of Research Method

III.A Data Collection

The data collection process for this research involves scraping online news media. The scraping technique will use Selenium with Python programming. The data used will be the scraped news headlines from online media, taken as an example from the Kumparan website with the keyword "Joko Widodo," collected on April 26, 2024. The reason for choosing this keyword is that it is the full name of the current President of Indonesia, which ensures the availability of many relevant news articles and facilitates the data collection process. The data to be scraped will be news headlines relevant to this subject, and only those headlines published between November 2023 and February 2024 will be collected and stored.

Table 1: Data Distribution

Data Source	Total Data Collection	Total Data After Preprocessing
Kumparan	927	683
Detik	539	500
CNN Indonesia	1072	1048

III.B Data Preprocessing

Text preprocessing involves several stages, including case folding, tokenizing, stop words removal, and stemming. These processes will transform the previously scraped text into a simpler form, making it easier to use. The results of the data preprocessing can be seen in Table 2.

Table 2: Text Preprocessing

Title	Case Folding	Tokenizing	Stopwords	Stemming
Jokowi Memihak, Kabinet Bergejolak	jokowi memihak kabinet bergejolak	['jokowi', 'memihak', 'kabinet', 'bergejolak']	["jokowi","memihak","kabinet","bergejolak"]	["jokowi","pihak","kabinet","gejolak"]
Pujian Prabowo untuk Jokowi di Debat Capres	pujian prabowo untuk jokowi di debat capres	['pujian', 'prabowo', 'untuk', 'jokowi', 'di', 'debat', 'capres']	["pujian","prabowo","jokowi","debat","capres"]	["puji","prabowo","jokowi","debat","capres"]
Joko Widodo dan Istri Nonton Konser NOAH di Jakarta	joko widodo dan istri nonton konser noah di jakarta	['joko', 'widodo', 'dan', 'istri', 'nonton', 'konser', 'noah', 'di', 'jakarta']	["joko", "widodo", "istri", "nonton", "konser", "noah", "jakarta"]	["joko", "widodo", "istri", "nonton", "konser", "noah", "jakarta"]

III.C Data Labeling

The calculation of polarity for each document or sentence is a crucial step in sentiment analysis labeling. During this phase, a lexicon containing a list of words with polarity values is used. The Inset Lexicon includes 3,609 positive words and 6,609 negative words. The Inset Lexicon consists of word-number pairs where the number represents the weight of the word. The weights in the InSet Lexicon range from -5 (very negative) to +5 (very positive). The results of the polarity calculation, leading to the assigned sentiment labels, can be seen in Table 3.

Table 3: Polarity Counting

Title	Stemmed Title	Weight Count	Result	Polarity
Jokowi Memihak, Kabinet Bergejolak	["jokowi","pihak","kabinet","gejolak"]	{"jokowi": 0, "pihak": -3, "kabinet": -3, "gejolak": 1}	-5	negatif
Pujian Prabowo untuk Jokowi di Debat Capres	["pujian","prabowo","jokowi","debat","capres"]	{"puji": 4, "prabowo": 0, "jokowi": 0, "debat": 3, "capres": 0}	7	positif
Joko Widodo dan Istri Nonton Konser NOAH di Jakarta	["joko","widodo","istri","nonton","konser","noah","jakarta"]	{"joko": 0, "widodo": 0, "istri": 0, "nonton": 0, "konser": 0, "noah": 0, "jakarta": 0}	0	neutral

III.D Classification of KNN Models

The classification process of the model is carried out after data labeling by the Inset Lexicon, KNN is a classification algorithm defined by its name: K represents the number of nearest neighbors used to determine the similarity between a new point and its neighboring points. KNN does not generalize the model and can work with any type of data distribution in the training set, making predictions by calculating similarities between training data and the test data. Various methods can be used to measure the distance between data points in the KNN algorithm, including Euclidean distance. Euclidean distance is commonly used to calculate the distance between data points and is employed to evaluate the proximity between two objects (Ismai, 2021).

$$dist = \sqrt{\sum_{i=0}^{n}(X_{i2} - X_{i1})^2} \qquad (1)$$

The classification process will involve training and evaluation stages. The training data will begin with data splitting. Data that has undergone preprocessing will be divided and trained using the KNN algorithm to obtain a classification model, with a ratio of 70% training data and 30% test data. This approach is crucial to ensure that the model is well-trained on various types of data, which in turn improves the model's ability to evaluate its performance more effectively.

IV. Results and Findings

The research utilized an Apple M1 Pro device with an Apple M1 Pro Chip (8-Core CPU, 14-Core GPU) and 16GB RAM. After Preprocessing the polarity distribution of news headlines from three online news portals is as follows: kumparan.com shows 64.42% positive, 7.91% neutral, and 27.67% negative out of 683 headlines. detik.com has 69.40% positive, 8.20% neutral, and 22.40% negative out of 500 headlines. cnnindonesia.com reports 65.36% positive, 7.06% neutral, and 27.58% negative out of 1,048 headlines.

IV.A Evaluation of KNN Classification model

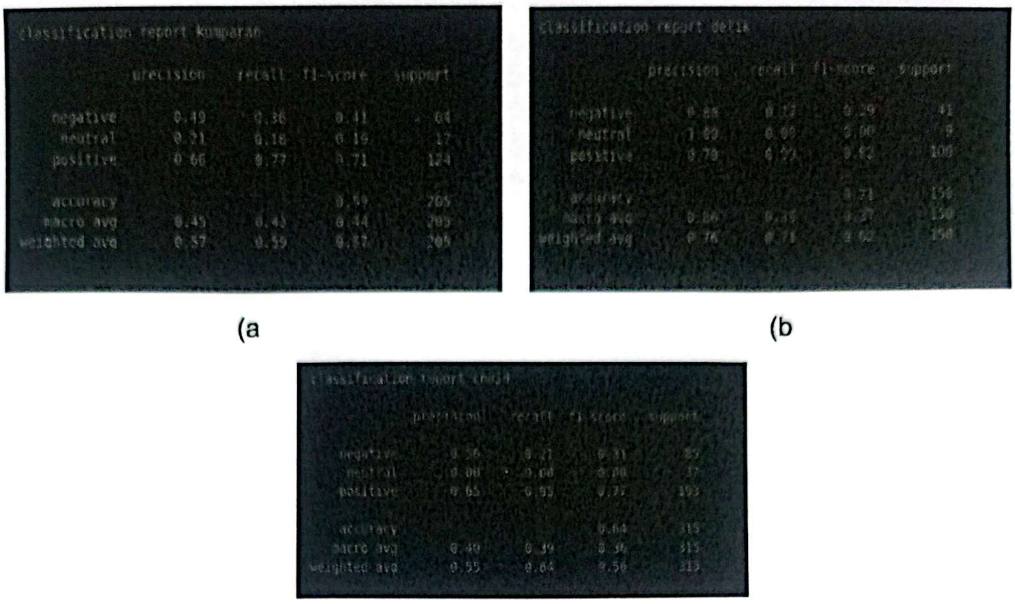
In this process, the evaluation of the KNN model's performance is conducted using data that has been previously split into 70% training data and 30% test data. The evaluation involves selecting the optimal number of neighbors (k) for the KNN algorithm by assessing the model's performance for various values of k. This evaluation includes three websites: kumparan.com, detik.com, and cnnindonesia.com. The values of k tested range from 1 to 9. The best k value selected from the evaluation is K = 8 for all websites. Specifically, kumparan.com achieved a K = 8 value of 0.590, detik.com achieved a K = 8 value of 0.707, and cnnindonesia.com achieved a K = 8 value of 0.638. Detailed results can be seen in Table 4.

Table 4: Accuracy by K Values

Website	k = 1	k = 2	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9
Kumparan	0.507	0.395	0.512	0.502	0.458	0.546	0.561	0.590	0.571
Detik	0.653	0.54	0.653	0.633	0.653	0.693	0.707	0.707	0.7
CNN Indonesia	0.533	0.406	0.530	0.559	0.594	0.622	0.619	0.638	0.635

IV.B Classification Report and Confusion Matrix

The following are the evaluation results of the KNN algorithm model for each website using K = 8, as shown in Figure 4. The results indicate that for the detik.com news website, the accuracy is 0.71. For the positive class, the metrics are precision 0.70, recall 0.99, and F1-score 0.82. For the negative class, the metrics are precision 0.88, recall 0.17, and F1-score 0.29. Finally, for the neutral class, the metrics are precision 1.00, recall 0, and F1-score 0.



```
classification report kumparan

              precision    recall  f1-score   support

    negative       0.49      0.36      0.41        64
     neutral       0.21      0.18      0.19        17
    positive       0.66      0.77      0.71       124

    accuracy                           0.59       205
   macro avg       0.45      0.43      0.44       205
weighted avg       0.57      0.59      0.57       205
```

(a

```
classification report detik

              precision    recall  f1-score   support

    negative       0.88      0.17      0.29        41
     neutral       1.00      0.00      0.00         9
    positive       0.70      0.99      0.82       100

    accuracy                           0.71       150
   macro avg       0.86      0.39      0.37       150
weighted avg       0.76      0.71      0.62       150
```

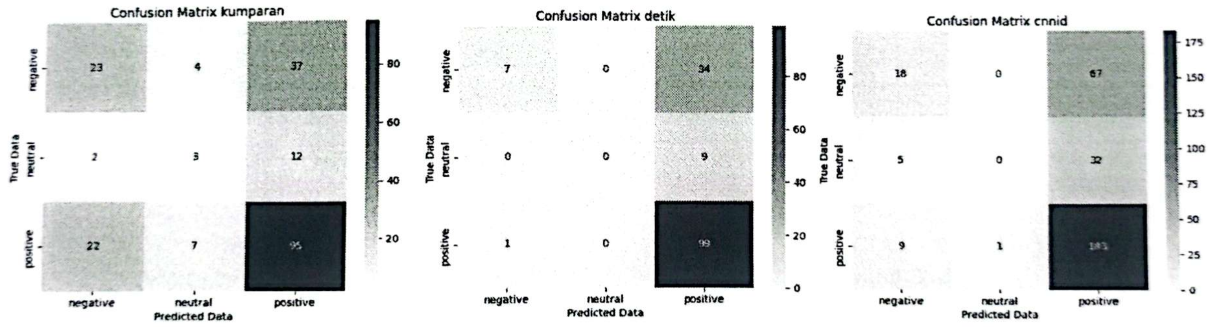
(b

```
classification report cnnid

              precision    recall  f1-score   support

    negative       0.56      0.21      0.31        85
     neutral       0.00      0.00      0.00        37
    positive       0.65      0.95      0.77       193

    accuracy                           0.64       315
   macro avg       0.40      0.39      0.36       315
weighted avg       0.55      0.64      0.56       315
```

(c)

Figure 2. Classification Report

Figure 3 shows detik.com with two parameters: True Data (actual data) and Predicted Data (predicted data). For positive data, out of 100 cases, 99 were correctly predicted as positive, 1 as negative, and none as neutral. For neutral data, all 9 cases were correctly predicted as neutral. For negative data, out of 41 cases, 7 were correctly predicted as negative, 34 were predicted as positive, and none were predicted as neutral.



(a)

(b)

(c)

Figure 3. Confusion Matrix

V. Conclusion

The sentiment analysis research on kumparan.com, detik.com, and cnnindonesia.com using InSet Lexicon labeling and the KNN method with a 70% training and 30% test data ratio shows that detik.com has the most positive news content and the highest accuracy (0.707) compared to cnnindonesia.com (0.638) and kumparan.com (0.590), with the best K value being K=8. The Confusion Matrix results confirm that detik.com has higher precision and recall values compared to the other two sites. While InSet Lexicon proves effective for automatic labeling, it still requires vocabulary expansion to improve accuracy. Overall, the implementation of KNN in this study provides reasonably good results with accuracy ranging from 60-70 percent.

References

Bharathi, S., & Geetha, A. (2017). Sentiment analysis for effective stock market prediction. *International Journal of Intelligent Engineering and Systems*, *10*(3). https://doi.org/10.22266/ijies2017.0630.16

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*.

Dominic, P., Purushothaman, N., Kumar, A. S. A., Prabagaran, A., Blessy, J. A., & John, A. (2023). Multilingual Sentiment Analysis using Deep-Learning Architectures. *Proceedings - 5th International Conference on Smart Systems and Inventive Technology, ICSSIT 2023*. https://doi.org/10.1109/ICSSIT55814.2023.10060993

Ismai. (2021). *Machine Learning: Teori, Studi Kasus, dan Implementasi Menggunakan Phyton*.

Koto, F., & Rahmaningtyas, G. Y. (2017). Inset lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs. *Proceedings of the 2017 International Conference on Asian Language Processing, IALP 2017*, *2018-January*. https://doi.org/10.1109/IALP.2017.8300625

Kowsher, M., Sami, A. A., Prottasha, N. J., Arefin, M. S., Dhar, P. K., & Koshiba, T. (2022). Bangla-BERT: Transformer-Based Efficient Model for Transfer Learning and Language Understanding. *IEEE Access*, *10*. https://doi.org/10.1109/ACCESS.2022.3197662

Lecheler, S., & de Vreese, C. H. (2019). News framing effects: Theory and practice. In *News Framing Effects*.

Musfiroh, D., Khaira, U., Utomo, P. E. P., & Suratno, T. (2021). Analisis Sentimen terhadap Perkuliahan Daring di Indonesia dari Twitter Dataset Menggunakan InSet Lexicon. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *1*(1). https://doi.org/10.57152/malcom.v1i1.20

Obagbuwa, I. C., Danster, S., & Chibaya, O. C. (2023). Supervised machine learning models for depression sentiment analysis. *Frontiers in Artificial Intelligence*, *6*. https://doi.org/10.3389/frai.2023.1230649

Sun, C., Huang, L., & Qiu, X. (2019). Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*.

Wu, X., & Lode, M. (2020). Language Models are Unsupervised Multitask Learners (Summarization). *OpenAI Blog*, *1*(May).

Yan, H. Bin, & Li, Z. (2022). Review of sentiment analysis: An emotional product development view. In *Frontiers of Engineering Management* (Vol. 9, Issue 4). https://doi.org/10.1007/s42524-022-0227-z