Fuzzy K-Nearest Neighbor for Restaurants Business Sentiment Analysis on TripAdvisor

Baiq Billyan
Department of Information Technology Management
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
fbillyan@gmail.com

Riyanarto Sarno
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

Kelly Rossa Sungkono
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
kelly@its.ac.id

Irene R.H.T.Tangkawarow
Department of Informatics
Institut Teknologi Sepuluh Nopember, Surabaya &
Universitas Negeri Manado, Indonesia
irene.tangkawarow@unima.ac.id

Abstract*— Social media has grown so rapidly, so people easily to share their opinions, moments, etc. There are several types of research about social media, one of which is Sentiment Analysis (SA) that can also be referred to as opinions meaning (OM). Sentiment Analysis focuses on the classification of patterns that are derived from words that are positive words, negative words, and neutral words. In this paper, the researcher uses sentiment analysis with a machine learning approach and uses Fuzzy K-Nearest Neighbor (FK-NN) as the classification method. The dataset uses English text classification, to predicted sentiment of customer reviews about the positive or negative review. The predicted results show that Sentiment Analysis FK-NN is slightly close to the results of the previous research method, namely Probabilistic Latent Semantic Analysis (PLSA), which FK-NN is 72.05% and PLSA is 76%.

Keywords*— Data Analysis, Fuzzy K-Nearest Neighbor, Social Media, Sentiment Analysis, TripAdvisor

I. INTRODUCTION

Technology is growing rapidly; internet users are increasing at this time. Based on digital data in 2018, internet users reached 4.021 Billion and still continued to increase by 7% from year to year, while social media users also reached 3.196 Billion and increased by 13% every year, and also supported by the sophistication of mobile phones now that users can reach until 5.135 Billion which is it continues to increase 4% every year (smartinsights.com).

This digital world makes it easier for everyone to access data on social media, as the author did for this research, which is about public opinion in several restaurant businesses which continuously reaches the top 10 on TripAdvisor social media. TripAdvisor is the largest travel site in the world that has 661 Million reviews and opinions, 456 Million monthly average unique visitors, and 7.7 Million accommodations, airlines, experiences, and restaurants (TripAdvisor).

On social media, many visitors want to share their opinions about the restaurants that they visited based on the service, the foods, the value, and the atmosphere with some opinions containing positive and negative words and sentences. There are several reviews whose negative patterns become positive or conversely make their meanings neutral then sometimes become ambiguous, so that all depends on the context.

This research is conducted to reduce the ambiguity of meaning contained in customer reviews at several restaurants located in the city of Surabaya using the "Sentiment Analysis" method. Sentiment analysis or "Opinion Mining" focuses on specific applications that could do review classifications to see the patterns of words or sentences that are positive or negative [1].

This paper proposes the combination of sentiment analysis with machine learning and FK-NN to predict customer sentiment of positive and negative reviews. The dataset is reviews taking from the chosen website, which is TripAdvisor. Firstly, this research does pre-processing steps, such as stemmer, stopwords, and tokenizer. Then, sentiment analysis is conducted by classifying the data with Fuzzy K-Nearest Neighbor (FK-NN) method. The results of this research are reviews that are classified as positive and negative.

II. LITERATURE REVIEW

Sentiment Analysis or "opinion mining" focuses on specific applications that classify reviews by considering the patterns of words or sentences that are positive or negative [1]. Sentiment analysis depends on the context. The context is a condition that allows the loss of the characteristics of his opinion or can change its opposite character. For example, "good" expresses a positive emotion with "good" from a freight train cargo [2].

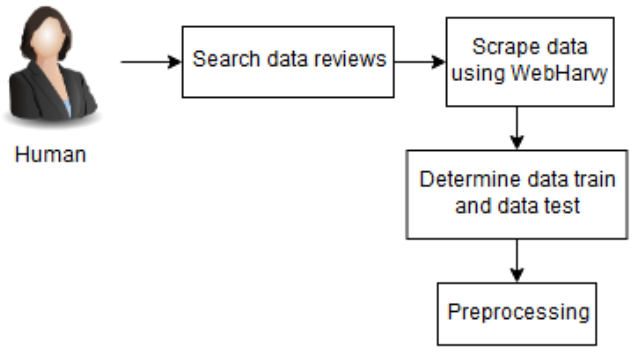

Fig. 1. Data collection

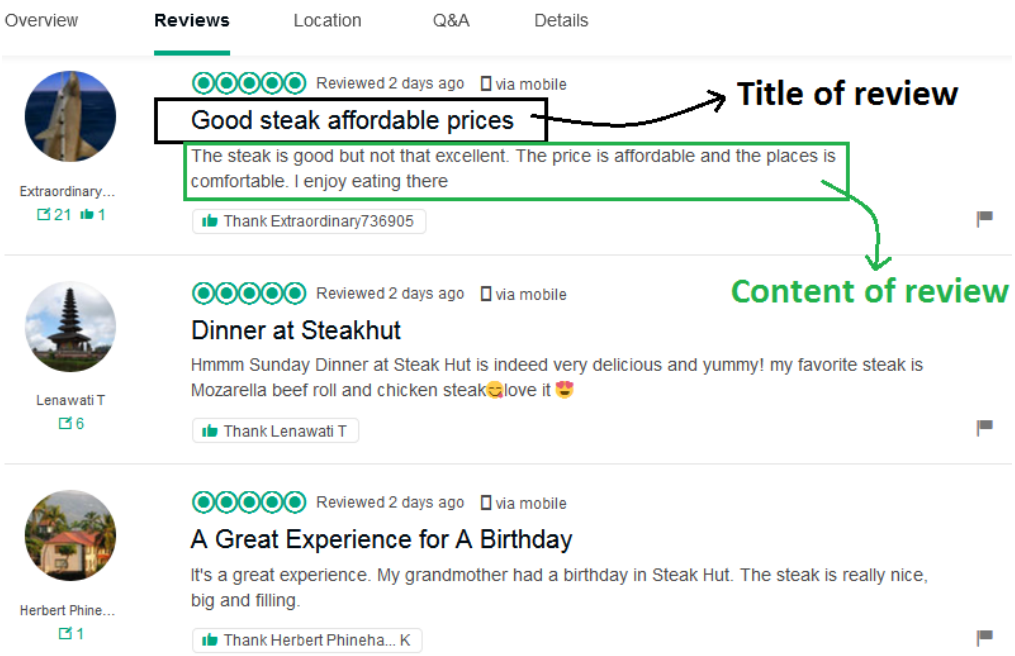

Fig. 2. Display of reviews from several customers of one restaurant

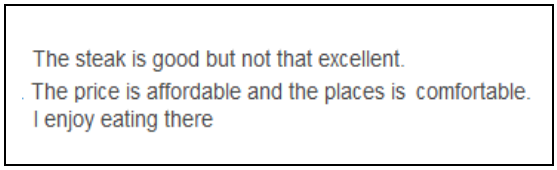

Fig. 3. Display of reviews from several customers of one restaurant

Sentiment analysis uses Natural Language Processing (NLP), text analysis and computational techniques to automate extraction or sentiment classification from sentiment review [3]. Sentiment analysis is used in the economic field, marketing strategies such as marketing where the success of a product launch can be judged determines which product and service versions are popular and also identifies demographics such as a particular features [4].

In the previous research, positive and negative words were predicted based on the customer review of a hotel website using the Probabilistic Latent Semantic Analysis (PLSA) method. The accuracy reaches 76%, needed improvisation to improve data accuracy better in English libraries such as SentiWordNet [5].

In this study, the researcher will predict positive and negative words based on the contents of several customer reviews from the TripAdvisor website in the restaurant review section in Surabaya using the Fuzzy K-Nearest Neighbor (FK-NN) method. This method can be used to compare the accuracy of the data like the meaning of a word in a sentence that is initially positive when the word is combined with another sentence, the meaning changes to negative [6]. For this reason, the FK-NN method can classify the data according to the training data and fuzzy data information taken from the method. Users can indirectly control the level of defuzzification (k value) to determine the percentage of wrong decisions, which are "having value" for the process. In many cases, setting the level of defuzzification can provide many advantages over not setting a value because when more data defects are categorized as something unknown, it is far better than classifying something that is indeed wrong. This is very accurate in many cases where classifying a faulty data defect can result in the effect of increasing or decreasing costs [6].

III. PROPOSED METHOD

In this proposed method we describe Data collection, Preprocessing, and Fuzzy K-Nearest Neighbor.

A. Data Collection

First, in Fig. 1, the author searches data reviews and chooses data for collection on the TripAdvisor website. Second, the author scrapes data using WebHarvy tools. The data used in the form of several reviews which are in the ten highest ranking restaurants in the city of Surabaya. After data collected, it is ready to determine for data train and data test. Last, data can be used to preprocessing data. Next, in Fig. 2, there are some views of the reviews of several customers who are on the TripAdvisor web.

Based on data from several customer reviews in Fig. 2, the data used is the content of the customer review, not the title of the review. As from the first customer review titled “Good steak affordable prices" (can be seen in Fig. 2) then the contents of the review data are using, then the captured data is saved in the form of a CSV file from the WebHarvy data retrieval application that is written in Fig. 3.

The data taken total is 337 customer review data. Then the data is divided into 2 data file, namely training data and testing data. Training data has more data than the testing data which is 80% of the total data is 269 data that contains positive and negative data, while testing data is 20% which contains 68 data which also contains positive data and negative data. After the data is split into training data and testing data then preprocessing data is done so that the data can be processed as sentiment analysis.

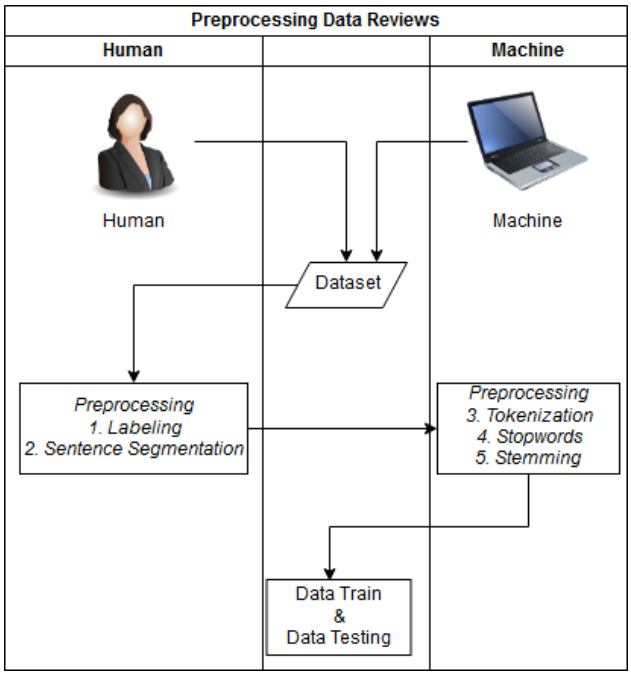

Fig. 4. Preprocessing Data

B. Preprocessing

Meanwhile, preprocessing data have some process as follows in Fig. 4.

1. Labelling

Labelling data is dividing data into positive and negative classes. Data labelling is used as a data class to facilitate the prediction of data accuracy [7]. To find out the machine works well in predicting the review as a positive or negative review . Then look at the accuracy of the prediction of the label of the machine compared to the label that has been given by humans.

2. Sentence Segmentation

Sentence segmentation is the process of determining a processing unit that is longer consisting of one or more words [8]. This involves identifying boundaries between words in different sentences [9]. Because generally written languages have punctuation marks that are at the boundary of the sentence, this sentence segmentation is usually referred to as the sentence limit detector [10]. This process has the goal of determining how a text should be divided into sentences for further processing[11].

3. Tokenization

Remove punctuation , such as:

“she’s” become “she”, “is”

“Fuzzy K-NN” become “fuzzy”, “k”, “nn”

“go away!” become “go”, “away”

Separate phrases using whitespace:

“gas station” become “gas” “station”

“flight attendant” become “flight” “attendant”

“office boy” become “office” “boy”

4. Stopwords

Stopwords is a common word in a number that is not small and sometimes considered meaningless [8]. The example of remove stopwords can be seen in Fig. 5. Based on Fig. 5, the sentence “I like reading, so I read” become “like”,”reading”,”read”.

“I like reading, so I read” become “like”, “reading”, “read”
“a stalker likes stalking, thus she stalks” become “stalker”, “likes”, “stalking”, “stalks”
“I love food, so I eat” become “love”, “food”, “eat”

Fig. 5. Example of Stopwords

5. Stemming

Stemming that is mapping the token to its basic form[12]. For examples:

“eat” can be “eating”, “eats”, “eaten”,

“week” can be “weekly”, “weeks”, “weekend”

“close” can be “close”, “closed”, “closing”, “closely”

After that, the restaurant reviews data was taken then manual data preprocessing was carried out by researchers, namely in step 1 is the labeling process and step 2 is sentence segmentation. After that pre-processing on a machine like step 3 is tokenization, step 4 is stopwords, and step 5 is stemming. This pre-processing engine is done in Weka software, training data and testing on CSV files are changed first to the ARFF file (Weka software file) to facilitate the process of further classification of data on the machine. After the preprocessing process is done, then it can be used to predict sentiment analysis using the Fuzzy K-Nearest Neighbor classification method [13].

C. Fuzzy K-Nearest Neighbor (FK-NN)

The Fuzzy K-Nearest Neighbor (FK-NN) is a combination of Fuzzy methods and K-Nearest Neighbor methods [6]. The FK-NN method is one of the classification methods that will be used as a method in the process of sentiment analysis in this study. Based on Fig. 6, this is a description of the process of sentiment analysis with the FK-NN method. After preprocessing data is completed, then the obtained data results ready to use for classification, we can use FK-NN for the method. The machine will compute data with FK-NN method and will give two choices of the results; there is Positive > Negative will be -1 value or negative sentiment value and Negative > Positive will be one value or positive value [9].

This FK-NN method is a method that, in its calculation, uses membership values[14]. There is a membership value in each class, that is, a data can be owned by different classes but with a degree of membership value at the interval [-1, 1][15]. Next, (1) is an equation that is used as the grantor of membership value in a data testing that is using the Euclidian equation [16]. After calculating all the data using the distance equation, then do the calculation to find the value of weight (w) like (3).

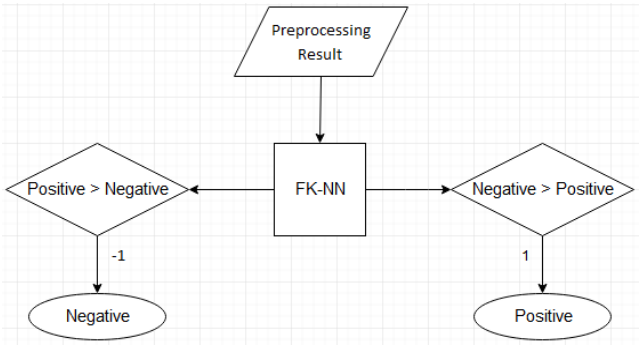

Fig. 6. Sentiment analysis process using FK-NN method

$$u_i(x) = \frac{\sum_{j=1}^{K} u_{ij}\left(\frac{1}{\|x-x_j\|^{\frac{2}{(m-1)}}}\right)}{\sum_{j=1}^{K}\left(\frac{1}{\|x-x_j\|^{\frac{2}{(m-1)}}}\right)} \quad (1)$$

where:

$u_i(x)$: the value of membership x data to u_i class,
K : the value that used to determine the number of closest neighbors,
u_{ij} : the value of membership data at K nearest neighbor in j class,
$x - x_j$: the difference in the distance of x data to x_j data in the K nearest neighbor, and
m : weight exponent whis is $m > 1$.

Then, normalization calculations are carried out as found in the data normalization in (2).

$$V' = \frac{V - min_A}{max_A - min_A} \quad (2)$$

where:

V' : normalization results whose value range from 0 and 1,
V : value of attribute A that will be normalized,
min_A : the minimum value of an attribute A,
max_A : the maximum value of an attribute A.

$$weight(w) = \frac{1}{distance(y,a)^2} \quad (3)$$

The function of $distance(y,a)$ is distance value from training data to testing data [17]. The final step is to initialize (4). Based on the calculation stages, then the results will be obtained from the analysis in this study.

$$u_{ij}(x) = \begin{cases} 0.51 + \left(\frac{n_j}{k}\right) * 0.49 & j = i \\ \left(\frac{n_j}{k}\right) * 0.49 & j \neq i \end{cases} \quad (4)$$

For every function of the initialization equation is:

u_{ij}: membership value of i class in j vector,
n_j: number of members of j class in a K dataset,
k: number of closest neighbors,
j: target class.

IV. Result and Analysis

Researchers collected data on the 10 highest rankings of restaurants in the Surabaya city, from each restaurant, the authors took 269 for training data. Then for testing data using several 68 data, the total data used is 337 data. Next, in the table below TABLE, I is a description of some of the data used. We used 34 positive sentiment data and 34 negative sentiments for test data, and the total is 68 test data. ID_Review mean is number of reviews, Review is the content of review from TripAdvisor website, and Label there is a positive and negative label.

TABLE I. Result of Processing Data

ID_Review	Review	Label
Rev 1	about the steak well the taste was nice	Positive
Rev 2	about the place felt comfortable	Positive
Rev 3	overall worth to try this restaurant	Positive
Rev 4	best luck	Positive
Rev 5	affordable price good ambience recommended for family event	Positive
Rev 6	so many promotion from any credit card	Positive
Rev 7	great service affordable price for steak	Positive
Rev 8	usually i visited this restaurant once in a month	Positive
Rev 9	recommended steak restaurant in surabaya	Positive
Rev 10	not too crowded	Positive
Rev 11	cosy place and decent food	Positive
Rev 12	nice environment and atmosphere	Positive
Rev 13	nice meal but if it is a larger portion will be perfect	Positive
Rev 35	i think they need more innovation for their menu	Negative
Rev 36	but to crowded if visit in weekend	Negative
Rev 37	area non smoking smoking small space	Negative
Rev 38	when i come here i thought it is only a small restaurant but i am totally wrong	Negative
Rev 39	have not been here for a long time and tried it for dinner the fish is a bit small *tempe* also small portion *nasi goreng merah* amazed us with the super tiny bit portion a small bowl	Negative
Rev 40	food price is expensive quality is average portion is small	Negative
Rev 41	sorry we are not coming back here again	Negative
Rev 42	but disappointed with their female staff	Negative
Rev 43	i just want to ask for spoon they serve their food without spoon to take up the food and no one coming	Negative
Rev 44	and no response	Negative
Rev 45	then she look back n yell at me with louder voice wait	Negative
Rev 46	at the end she is not coming to my table at all	Negative

In the next table, TABLE II is the best result in the approach for positive and negative sentiment values. For the positive sentiment, the value is one while for the negative sentiment, the value is -1. For the k value of the classification using FK-NN, the k value with the best results is used, k = 6.

Based on the table above Table 2, actual data is data that the label is given by human, 2:1 is for positive sentiment data and 1:-1 is for negative sentiment data. Predicted is the result of label prediction that is given by the machine, the labels are same like actual data, 2:1 is for positive sentiment prediction and 1:-1 is for negative sentiment prediction. Then, an error is the result of incorrect class data (not by the actual class), and it is given by the machine. Meanwhile, the prediction is the value of sentiment approach.

TABLE II. RESULT OF SENTIMENT CLASSIFY

ID_Review	Actual	Predicted	Error	Prediction
Rev 1	2:1	1:-1	+	0.5
Rev 2	2:1	2:1		0.666
Rev 3	2:1	2:1		0.666
Rev 4	2:1	2:1		0.666
Rev 5	2:1	2:1		0.666
Rev 6	2:1	2:1		0.5
Rev 7	2:1	1:-1	+	0.666
Rev 8	2:1	2:1		0.5
Rev 9	2:1	2:1		0.666
Rev 10	2:1	1:-1	+	0.5
Rev 11	2:1	2:1		0.5
Rev 12	2:1	2:1		0.5
Rev 13	2:1	2:1		0.5
Rev 34	2:1	2:1		0.833
Rev 35	1:-1	2:1	+	0.5
Rev 36	1:-1	1:-1		0.5
Rev 37	1:-1	2:1	+	0.5
Rev 38	1:-1	1:-1		0.5
Rev 39	1:-1	2:1	+	0.5
Rev 40	1:-1	1:-1		0.5
Rev 41	1:-1	1:-1		0.5
Rev 42	1:-1	2:1	+	0.5
Rev 43	1:-1	1:-1		0.5
Rev 44	1:-1	2:1	+	0.5
Rev 45	1:-1	2:1	+	0.5
Rev 46	1:-1	1:-1		0.5

```
=== Summary ===

Correctly Classified Instances          49               72.0588 %
Incorrectly Classified Instances        19               27.9412 %
Total Number of Instances               68
```

Fig. 7. Summary Accuracy

Fig. 7 is the final result for determining reviews to be the positive and negative sentiment from reviews that have been predicting by machine learning. The best k value used in the classification using this FK-NN is k = 6. From 68 testing data (total number of instances) used, there were some data that predicted errors. The correctly classified instances are 49 data and got accuracy 72,0588 %.

$$Accuracy = \frac{TS_1 + TS_2}{(Total\ Data)} \tag{5}$$

```
=== Detailed Accuracy By Class ===

                TP Rate   FP Rate   Precision   Recall
                0,559     0,118     0,826       0,559
                0,882     0,441     0,667       0,882
Weighted Avg.   0,721     0,279     0,746       0,721

=== Detailed Accuracy By Class ===

                F-Measure   Class
                0,667       -1
                0,759       1
Weighted Avg.   0,713
```

Fig. 8. Detailed Accuracy by Class

Meanwhile, the incorrectly classified instances are 19 data and got the accuracy of 27,9412 %. Based on (5), the equation for calculating the accuracy of all data as follows [18].

The function of TS_1 is for knowing the true prediction of the first meaning of the phrase, TS_2 is for the true prediction of the second of the phrase, and total data here means the amount of data that you want to calculate the accuracy of the data. The next figure, Fig. 8, is the calculating result of detailed accuracy by class positive and negative from customer reviews that have been predicting by machine learning. There is TP Rate (True Positive) value, FP Rate (False Positive) value, Precision result, Recall result, F-Measure result, and Weighted Avg.

The meaning of True Positive Rate is about the sensitivity of the positive sentiment that is correctly identified. The False Positive Rate is the sensitivity of the positive sentiment that is incorrectly identified. The Precision result means the ratio of the prediction from positive sentiment value. Recall Result means the result of sensitivity data, recall result is about true positive rate. The F-Measure result means for accuracy test to defined the weighted mean of precision result and recall result, and the Weighted Avg is weighted average from TP Rate, FP Rate, Precision, Recall, and F-measure. Class is 1 for positive class and -1 for negative class. In the calculation using (6) as below.

$$Precision = \frac{\left(\frac{TS_1}{(TS_1 + FS_1)} + \frac{TS_2}{(TS_2 + FS_2)}\right)}{2}$$

$$Recall = \frac{\left(\frac{TS_1}{(TS_1 + FS_2)} + \frac{TS_2}{(TS_2 + FS_1)}\right)}{2} \tag{6}$$

$$F - measure = \frac{(Precision \times Recall)}{(Precision + Recall)} \times 2$$

The function of TS_1 is for knowing the true prediction of the first meaning of the phrase, TS_2 is for the true prediction of the second of the phrase, FS_1 is for knowing the false prediction of the first meaning of the phrase, and FS_2 is for the false prediction of the second of the phrase.

```
=== Confusion Matrix ===

  a  b   <-- classified as
 19 15 |  a = -1
  4 30 |  b = 1
```

Fig. 9. Confusion Matrix

The last result, based on Fig. 8, is about the confusion matrix. That we could see in the confusion matrix, the meaning of a = -1 is a negative sentiment, b = 1 is a positive sentiment, we have 19 true negatives (TN) data, 30 true positives (TP) data, four false positives (FP) data, and 15 false negatives (FN) data.

V. CONCLUSION

In this research, the sentiment are predicted based on the content of review in TripAdvisor website. The predicted sentiment are positive and negative, and it explains the customer reviews about happy or satisfaction and not happy or dissatisfaction of the restaurants. If the words detected positive, so the customers are happy, and if the words detected negative it is mean the customers are not happy.

This research can be analysed by applying it in Weka software. The Fuzzy K-Nearest Neighbor (FK-NN) classification can be used to calculate the data, and we can find the best k value for the best results. The best k value means it is the best accuracy rate (high accuracy) that machine can predict.

In this research, we found an accuracy rate is 72.05%. To further research, may use another classification method to compare and could achieve better results by calculating this data reviews.

ACKNOWLEDGMENT

Authors give a deep thank to Institut Teknologi Sepuluh Nopember, The Ministry of Research, Technology and Higher Education of Indonesia, *Direktorat Riset dan Pengabdian Masyarakat*, and *Direktorat Jenderal Penguatan Riset dan Pengembangan Kementerian Riset, Teknologi, dan Pendidikan Tinggi Republik Indonesia* for supporting the research.

REFERENCES

[1] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Found. Trends Inf. Retr.*, vol. 2, no. 1–2, 2018.

[2] A. Weichselbraun, S. Gindl, and A. Scharl, "Enriching semantic knowledge bases for opinion mining in big data applications," *Knowledge-Based Syst.*, vol. 69, pp. 78–85, 2014.

[3] B. Agarwal, N. Mittal, P. Bansal, and S. Garg, "Sentiment Analysis Using Common-Sense and Context Information," , Vol. 2015, Article ID , 9 pages, 2015.," *Comput. Intell. Neurosci.*, vol. 2015, no. 715730, p. 9, 2015.

[4] J. Jotheeswaran and S. Koteeswaran, "Sentiment Analysis: A Survey of Current Research and Technique," *Int. J. Innov. Res. Comput. Commun. Eng. An ISO 32972007 Certif. Organ.*, vol. 3, no. 5, 2015.

[5] D. A. . Khotimah and R. Sarno, "Sentiment Detection of Comment Titles in Booking.com Using Probabilistic Latent Semantic Analysis," in *International Conference on Information and Communication Technology (ICOICT)*, 2018.

[6] M. E. Bakry, S. Safwat, and O. Hegazy, "Big Data Classification using Fuzzy K-Nearest Neighbor," *Int. J. Comput. Appl. (0975 – 8887)*, vol. 132, no. 10, 2015.

[7] A. Suhariyanto, Firmanto, and R. Sarno, "Prediction Movie Sentiment based on Reviews and Score on Rotten Tomatoes using SentiWordNet," in *International Seminar on Application for Technology of Information and Communication (iSemantic), 2018. Universitas Dian Nuswantoro, Indonesia*, 2018.

[8] U. W. Wijayanto and R. Sarno, "An Experimental Study of Supervised Sentiment Analysis Using Gaussian Naive Bayes," in *International Seminar on Application for Technology of Information and Communication (iSemantic), 2018. Universitas Dian Nuswantoro, Indonesia.*, 2018.

[9] E. Faisal, F. Nurifan, and R. Sarno, "Word Sense Disambiguation in Bahasa Indonesia Using SVM," in *International Seminar on Application for Technology of Information and Communication (iSemantic), 2018. Universitas Dian Nuswantoro, Indonesia.*, 2018.

[10] N. M. Elfajr and R. Sarno, "Sentiment Analysis using Weighted Emoticons and SentiWordNet for Indonesian Language," in *International Seminar on Application for Technology of Information and Communication (iSemantic)*, 2018.

[11] M. Fikri and R. Sarno, "A Comparative Study of Sentiment Analysis using SVM and SentiWordNet," *Indones. J. Electr. Eng. Comput. Sci. (IJEECS).*, 2017.

[12] B. S. Rintyarna, R. Sarno, and C. Fatichah, "Enhancing the performance of sentiment analysis task on product reviews by handling both local and global context," *Int. J. Inf. Decis. Sci.*, vol. 11, 2018.

[13] F. H. Wattiheluw and R. Sarno, "Developing Word Sense Disambiguation Corpuses using Word2vec and Wu Palmer for Disambiguation," in *International Seminar on Application for Technology of Information and Communication (iSemantic), 2018. Universitas Dian Nuswantoro,* 2018.

[14] T. S. Utomo and R. Sarno, "Emotion Label from ANEW dataset for Searching Best Definition from WordNet," in *International Seminar on Application for Technology of Information and Communication (iSemantic), 2018. Universitas Dian Nuswantoro, Indonesia.*, 2018.

[15] P. Pambudi and R. Sarno, "Searching Word Definitions in WordNet Based on ANEW Emotion Labels," in *International Seminar on Application for Technology of Information and Communication (iSemantic), 2018. Universitas Dian Nuswantoro, Indonesia.*, 2018.

[16] B. Rintyarna and R. Sarno, "Adapted Weighted Graph for Word Sense Disambiguation," .," in *The 4th International Conference on Information and Communication Technology (ICoICT)*, 2016.

[17] F. H. Rachman, R. Sarno, and C. Fatichah, "CBE : Corpus-Based off Emotion for Emotion Detection in Text Document," in *The 3rd International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE)*, 2016, pp. 331 – 335.

[18] F. Nurifan, R. Sarno, and C. S. Wahyuni, "Developing Corpora using Word2vec and Wikipedia for Word Sense Disambiguation," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 12, no. 3, pp. 1239–1246, 2018.