Sriwijaya Journal of Informatic and Applications
Vol. 4, No. 2, August 2023, pp. 69-75
ISSN 2807-2391

Sentiment Analysis Using Pseudo Nearest Neighbor and TF- IDF Text Vectorizer

Yogi Pratama [a,1,*], Abdiansah [b,2], Kanda Januar Miraswan [b,3]

[a] Department of Informatics Engineering, University Sriwijaya, Palembang, Sumatera Selatan
[1] 1998tahun@gmail.com*; [2] Abdiansah@ilkom.unsri.ac.id; [3] KandaJanuar@ilkom.unsri.ac.id
* corresponding author

A R T I C L E I N F O

Article history
Received
Revised
Accepted

Keywords
Sentiment
Analysis TF-IDF
Vectorizer
Pseudo Nearest Neighbor

ABST RACT

Twitter is one of the social media that is often used by researchers as an object of research to conduct sentiment analysis. Twitter is also a good indicator in influencing research, problems that often arise in research in the field of sentiment analysis are the many factors such as the use of colloquial or informal language and other factors that can affect sentiment results. To improve the results of sentiment classification, it is necessary to carry out a good information extraction process. One of the word weighting methods resulting from the information extraction process is the TF-IDF Vectorizer. This study examines the effect of the TF-IDF Vectorizer weighting results in sentiment analysis using the Pseudo Nearest Neighbor method. The results of the f-measure classification of sentiment using the TF-IDF Vectorizer at parameters k-2 = 89%, k-3 = 89%, k-4 = 71% and k-5 = 75% while without using the TF-IDF Vectorizer on the parameters k-2 = 90%, k-3 = 92%, k-4 = 84% and k-5 = 89%. From the results of the classification of sentiment analysis that does not use the TF-IDF Vectorizer, the f-measure value is slightly better than using it.

1. Introduction

Twitter is now a very popular communication device among internet users. At the official Chirp 2010 Twitter developer conference, the company relayed statistics regarding twitter sites and users. As of April 2010, Twitter had 106 million accounts and 180 million unique visitors each month. The number of Twitter users is said to continue to increase by 300,000 users every day (Buntoro, 2017). The problem that often arises in research in the field of Sentiment Analysis is the number of factors such as the use of non-standard language or colloquial among the public and other factors that can affect the results of sentiment.

Sentiment Analysis refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, measure, and study affective status and subjective information, in this final task will be used pseudo nearest neighbor method which will be used to perform sentiment. In 2018 there was a lot of problems about the LRT used for the ASEAN Games causing a lot of comments from the public about the phenomenon. So that the phenomenon is considered interesting to do Sentiment Analysis.

Entitled Multi-label classification of Indonesian news topics using Pseudo Nearest Neighbor Rule conducted research using PNN. The results of his research where it turns out that the Pseudo Nearest Neighbor algorithm can also be used to classify from various kinds of news texts in Indonesian. The-algorithm value for the k parameter and proximity type used affect the performance of the algorithm. Of the three types of proximity tested, Cosine's proximity provided the best performance compared to manhattan and euclid estimates.

a. Literature Study / Hypotheses Development

Sentiment analysis

Sentiment analysis is the process of extracting data automatically to obtain sentiment information contained in an opinion sentence. Sentiment analysis is a reflection of the attitude of the speaker or author regarding certain topics (Liu, 2010). In his research, the usual sentiment class consists of negative sentiments, namely sentiment for tweets that vilify or insult brands, positive sentiments for tweets that praise the brand and neutral sentiments for tweets containing question sentences, promo tweets, or news tweets.

Pseudo Nearest Neighbor

Pseudo Nearest Neighbor is the latest variant of K-NN, which is used to address the weaknesses of the K-NN method which generally provides low performance for data containing noise. P-NN works by calculating the total distance between the input pattern (unlabeled) and the number of k closest patterns in each class with proportional weighting based on Euclidean distance. Here is the Euclidean formula:

$$d = \sqrt{(x_1 - x} \qquad (2.1)$$

Description :

- d = range
- x_1= feature X data ke 1
- x_2 = feature X data ke 2
- y_1 = feature Y data ke 1
- y_2 = feature Y data ke 2

Then decide on the class with the minimum total distance as the decision class for the pattern.

TF-IDF Text Vectorizer

TF-IDF is one of the vectorizers that can be used to change text in vector form so that it can be processed for Machine Learning training (Ramadan, 2020). In search for information, TF-IDF (Term Frequency-Inverse document Frequency) is a numerical statistic intended to reflect how important a word is in a document. TF-IDF is the result of multiplication of tf vectors with idf, here is the TF-IDF formula:

$$w_{i,j} = tf_{i,j} * \log\left(\frac{N}{df_i}\right) \qquad (2.2)$$

Description:

- tfi,j = The number of i-word appearances in document j.
- N = Number of documents (sample data training).

dfi = Number of documents that have the word i

b. Methodology

Data Collection Method

The data used was obtained from Aprillia's research (2018) on LRT Palembang Sumsel. The data has been grouped into both positive and negative sentiments. The number of data obtained as many as 500 texts, 293 positive sentiments and 207 negative sentiments.

Framework

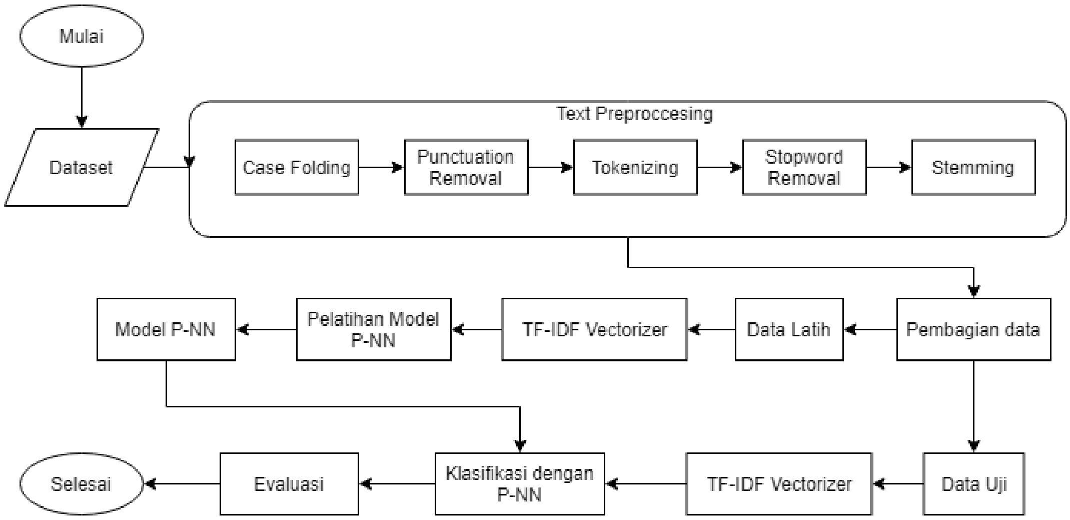

Figure 1. Framework

PNN

1. The results of the TF-IDF Vetorizer calculation that have been obtained become a comparison model that will be used as a determinant of class categories in the testing data. Examples of testing data are shown in table.

Table 1. Sample Data Testing

Data	Text	Label
T1	Suka saya naik lrt pegawainya ramah	1
T2	Lrt sepi karena Mahal banget #lrtpalembangpunyokito	0

2. The testing data above will be preprocessed as was done in the previous training data. From the results of the preprocessing, TF-IDF Vectorizer was then carried out using the previous data training model. The results of preprocessing and TF-IDF Vectorizer are shown in table

Table 2. Results of Preprocessing Data Testing

T1	T2
Suka	lrt
lrt	sepi
Pegawai	mahal
Ramah	-

3. Calculating Using PNN

Table 3. TF-IDF Vectorizer Data Testing Results

Term	W=TF*IDF	
	T1	T2
suka	0.699	0
lrt	0.079	0.079

pegawai	0.699	0
ramah	0.699	0
sepi	0	0.699
mahal	0	0.699

Then look for the Euclidean distance referring to equation 2.1 in the subchapter using the results of the previous TF-IDF Vectorizer.

1. Euclidean T1 distance with Train data a.

d(D1,T1) = $\sqrt{(d_1 - t_1)^2 + \cdots + (d_i + t_i)^2}$

$= \sqrt{(0.079 - 0.079)^2 + (0.398 - 0)^2 + \cdots + (0 - 0)^2}$

$= \sqrt{3.581}$

= 1.892

b. d(D2,T1) = $\sqrt{(0.079 - 0.079)^2 + (0 - 0)^2 + \cdots + (0 - 0)^2}$

= $\sqrt{3.912}$

= 1.977

c. d(D3,T1) = $\sqrt{(0.079 - 0.079)^2 + (0 - 0)^2 + \cdots + (0 - 0)^2}$

$= \sqrt{2.934}$

=1.712

d. d(D4,T1) $= \sqrt{(0.079 - 0.079)^2 + (0 - 0)^2 + \cdots + (0 - 0)^2}$

$= \sqrt{0.489}$

= 0.699

e. d(D5,T1) = $\sqrt{(0.158 - 0.079)^2 + (0.398 - 0)^2 + \cdots + (0 - 0)^2}$

$= \sqrt{4.076}$

= 2.018

2. Euclidean T2 distance with Train data a.

d(D1,T2) = $\sqrt{(0.079 - 0.079)^2 + (0.398 - 0)^2 + \cdots + (0 - 0)^2}$

$= \sqrt{2.154}$

= 1.467

b. d(D2,T2) = $\sqrt{(0.079 - 0.079)^2 + (0 - 0)^2 + \cdots + (0 - 0)^2}$

$= \sqrt{2.445}$

= 1.563

c. d(D3,T2) = $\sqrt{(0.079 - 0.079)^2 + (0 - 0)^2 + \cdots + (0 - 0)^2}$

$= \sqrt{2.445}$

= 1.563

d. d(D4,T2) = $\sqrt{(0.079 - 0.079)^2 + (0 - 0)^2 + \cdots + (0 - 0)^2}$

$= \sqrt{2.934}$

$= 1.712$

e. $d(D5,T2) = \sqrt{(0.158-0.079)^2 + (0.398-0)^2 + \cdots + (0-0)^2}$

$= \sqrt{3.587}$

$= 1.893$

From the results of the calculation of the Euclidean distance above, we need k parameters to perform sentiment. Suppose that the k parameter in the sentiment of this case is 2, then the 2 closest distances between the training data and testing data from each positive sentiment and also negative sentiment are added up and seen a comparison between the two.

1. Data *Testing* T1

   Negative sentiment = 1.892 + 1.977

   = 3.869

   Positive sentiment = 1.712 + 0.699

   = 2.411

   From the results of the distance calculation above, it is found that the T1 test data in the testing data is included in positive sentiment because the distance is closer to positive sentiment.

2. Data *Testing* T2

   Negative sentiment = 1.467 + 1.563

   = 3.03

   Positive sentiment = 1.563 + 1.712

   = 3.275

From the results of the distance calculation above, it is found that the T2 test data in the testing data is included in negative sentiment because the distance is closer to negative sentiment.

c. Result and Discussion

A. The results of the tests that have been carried out are summarized in the confusion matrix table shown in table V-1.

Table 4. Confusion Matrix Table of Classification Results

Parameter k-	*Pseudo Nearest Neighbor*				PNN + TF-IDF			
	TP	FP	TN	FN	TP	FP	TN	FN
2	52	0	37	11	50	0	37	13
3	56	3	34	7	53	3	34	10
4	47	2	35	16	35	0	37	28
5	57	8	29	6	39	2	35	24

Based on the results of the confusion matrix that has been carried out, it is known that for the assessment of the TP value when using PNN itself it looks better than using PNN
+ TF-IDF but the larger the parameters used, the higher the FP value. From these results, it can be seen that the results of the comparison of the performance of the classification stage are based on the value of the k parameter given to the test data which can be seen in table V-2.

Table 5. Performance Value

Parameter k-	F-Measure	
	PNN	PNN + TF-IDF
2	0.9	0.89
3	0.92	0.89
4	0.84	0.71
5	0.89	0.75

This section describes the result of the analysis process which can be delivered in table, chart, or descriptive format. This section discusses the results of the study. In this part authors are suggested to synthesize the findings, link the finding with the existing literatures, and highlight the novelties of the study.

B. Discussion

Based on the graph in Figure V-1, the f-measure comparison value of the two classification methods shows that for each k parameter, the PNN classification method is better than the PNN + TF-IDF classification method, where at the time of parameter k = 2, the comparison between the two is 90%. :89%, when using the parameter k=3 here, the classification using PNN has the highest value with 92% and the PNN + TF-IDF method is 89% when using the parameter value k=3 and 4 here they have a rather far gap value where the comparison is 84%:71% and 89%:75%.

d. Conclusion

Based on the experimental results described in the previous chapter, there are several conclusions, namely as follows:.

1. Based on the results of the research, Pseudo Nearest Neighbor can be used to classify in Indonesian-language twitter tweets.
2. Based on the results, the performance value of the classification using Pseduo Nearest Neighbor is relatively high.

Based on the results of research using Pseudo Nearest Neighbor, f-measure values are better in sentiment classification than using Pseudo Nearest Neighbor + TF-IDF Vectorizer.

References

[1] Adiwijaya, I. (2006). Text Mining dan Knowledge Discovery. *Kolokium Bersama Komunitas Datamining Indonesia & Soft-Computing Indonesia*, 1–9. http://web.ipb.ac.id/~ir-lab/pdf/tm (text summarization).pdf

[2] Aprillia (2018). "Pengaruh *Parf Of Speech Tagging* dalam Analisis Sentimen menggunakan Algoritma *Multinomial Naïve Bayes*." Skripsi Jurusan Teknik Informatika, Fak. Ilmu Komputer, UNSRI – Tidak diterbitkan.

[3] Buntoro, G. A. (2017). Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter. *Integer Journal Maret*, *1*(1), 32–41. https://www.researchgate.net/profile/Ghulam_Buntoro/publication/316617194_Analisis_Sentimen_Calon_Gubernur_DKI_Jakarta_2017_Di_Twitter/links/5907eee44585152d2e9ff992/Analisis-Sentimen-Calon-Gubernur-DKI-Jakarta-2017-Di-Twitter.pdf

[4] Kruchten, P., & Kroll, P. (2003). The Rational Unified Process Made Easy. In *Rational Unified Process Made Easy: A Practitioner's Guide to the RUP*.

[5] Langgeni, D. P. 2010. 'Komunitas Data *mining* Indonesia & *Soft-computing* Indonesia.', *Text Mining* dan *Knowledge Discovery*., 0(semnasIF), pp. 1–10. doi: 10.1007/s13398-014-0173-

7.2.

[6] Liu, B. (2010). *Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing, Second Edition,* 2, 568.

[7] Pambudi, R. A., Adiwijaya, & Mubarok, M. S. (2019). Multi-label classification of Indonesian news topics using Pseudo Nearest Neighbor Rule. *Journal of Physics: Conference Series*, *1192*(1). https://doi.org/10.1088/1742-6596/1192/1/012031

[8] Pan, W., Narasimhan, H., Protopapas, P., Kar, P., & Ramaswamy, H. G. (2017). *Optimizing the multiclass F-measure via biconcave programming*. *Proceedings - IEEE International Conference on Data Mining*, ICDM, 1101–1106. https://doi.org/10.1109/ICDM.2016.184.

[9] Patterson, J. & Gibson A. 2017. *Deep Learning A Practitioner's Approach,* O'Reilly Media, Sebastopol, USA.

[10] Pratiwi, H., Mukid, M. A., Hoyyi, A., & Widiharih, T. (2019). Credit scoring analysis using pseudo nearest neighbor. *Journal of Physics: Conference Series*, *1217*(1). https://doi.org/10.1088/1742-6596/1217/1/012100

[11] Ramadhan M. L. (2020). "*Email Spam Filtering Menggunakan Multilayer Perceptron dengan Metode Pelatihan Stochastic Gradient Descent (SGD) dan Momentum.*" Skripsi Jurusan Teknik Informatika, Fak. Ilmu Komputer, UNSRI – Tidak diterbitkan.

[12] Suyanto (2017). *Data Mining untuk Klasifikasi dan Klasterisasi Data*. Bandung, Jawa Barat, Indonesia: Informatika Bandung.