Jurnal Mantik, 6 (3) (2022), ISSN 2685-4236 (Online)

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik

Application Of N-Gram On K-Nearest Neighbor Algorithm To Sentiment Analysis Of TikTok Shop Shopping Features

Riska Dwi Ayu Lestari[1], Bagus Setya Rintyarna[2], Moh. Dasuki[3]
[1,3]Department of Informatics Engineering, University of Muhammadiyah Jember, Indonesia
[2]Department of Electrical Engineering, University of Muhammadiyah Jember, Indonesia

A R T I C L E I N F O

Article history:

Received Sep 12, 2022
Revised Sep 29, 2022
Accepted Oct 15, 2022

Keywords:

Analysist
KNN
N-Gram
Sentiment
Undersampling

ABSTRACT

This study contains sentiment analysis on Twitter data with the direction of sentiment on the TikTokShop feature. In this study, the k nearest neighbor method is implemented in which the metric distance cosine similarity is used with the value of the nearest neighbor distance k = 3, 5, 7, and 9. In the modeling, a k-fold cross-validation scenario is used with a value of k = 10 fold. This study also uses unigram, bigram, and trigram selection features to handle imbalanced data using undersampling techniques. From the modeling results, it is found that the best modeling is the model with unigram feature selection with nearest neighbor k = 3. From this model, the average accuracy value is 89.92%, the average precision is 90.54% and the recall average is 87.37%. In the test, the results showed that the unigram feature selection had the best performance with 91% accuracy, 92% precision, and 89% recall.

Corresponding Author:

Riska Dwi Ayu Lestari,
Department of Informatics Engineering,
University of Muhammadiyah Jember,
Gumuk Kerang, Karangrejo, Kec. Sumbersari, Kabupaten Jember, Jawa Timur 68124,
Email: riskadwiayu10@gmail.com

1. INTRODUCTION

Online marketing of teenagers' products can be found on various social media. One of the social media that is widely used for this is TikTok. There is even a new feature on Tiktok, namely TikTokShop which is useful for making buying and selling transactions to users who are interested in products being marketed or promoted. Tiktok is a social media platform that has been widely enjoyed in the last two years. The emergence of the TikTokShop feature has also become a special attraction for the community (Yuniarti et al., 2020). The features offered by TikTokShop have their advantages over other social media. In addition to being a social media, Tiktok also provides shopping features so that marketing or promotion of interest in product purchases is getting closer without the need to go to other market places (Sulistianti & Sugiarta, 2022). But TikTokShop also has drawbacks in terms of applications and terms of sellers. Many negative responses have been given, such as incorrect delivery of goods, poor product packaging, fraud during transactions and applications that suddenly malfunction when used (Artanti et al., 2018).

Public opinion on TikTokShop is widely expressed on other social media such as Twitter. This is one of the things that became the basis for the authors to conduct an analysis of public sentiment towards the TikTokShop they wrote on the twitter page. Sentiment analysis is a technique to find information in a data in the form of text or public opinion. Generally, this opinion refers to certain products, figures or public figures, services, institutions, politics or something that is trending at a certain time. Sentiment is usually classified into three parts, namely positive, negative and neutral sentiment (Tri Romadloni et al., 2019).

Classification is one of the techniques in machine learning. We can use several methods to classify a dataset, one of which is K Nearest Neighbor. This method can also be combined with other methods to obtain better results. Several additional techniques are used in data preprocessing such as feature selection and data balancing. In this study, two approaches will be used, namely feature selection by implementing n-grams and undersampling techniques to obtain balanced data.

2. RESEARCH METHOD

2.1 Crawling Data

Crawling data in this study uses python and the Twitter API. The crawling of data was carried out on April 1 – May 14, 2022, with "TikTokShop" as the search keyword. The total data obtained were 875 data.

2.2 Labeling

Labeling of crawled data is carried out for validation and to determine the original value of the data class. The labeling is assisted by experts who work as Indonesian language lecturers. The results of labeling can be seen in Figure 1 below:

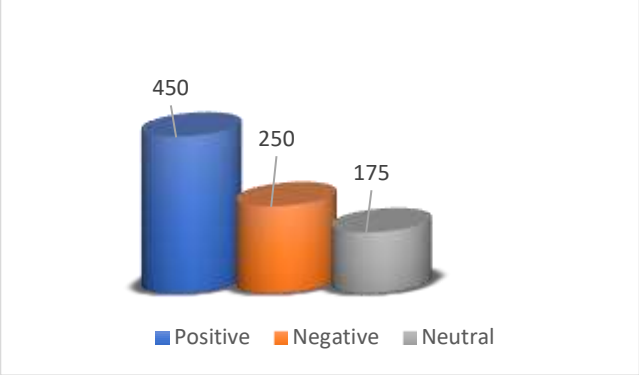

Figure 1. Label validation results

2.3 Text Preprocessing

The techniques used in the text preprocessing stage are as follows:

1. Cleansing or cleaning text data from symbols, numbers, links, and other things that interfere.
2. Case folding or converting text data into lowercase letters.
3. Stopword removal or deleting words that are considered to not affect sentiment.
4. Tokenizing or dividing sentences into words by word.
5. Stemming or changing words into basic tenses.

2.4 Data Partition

The clean data will be partitioned into two parts, namely training data and validation test data. The total data is 875 tweets, but in this study only tweets with positive and negative sentiments were used. The total data that will be used as research material is 700 tweets. The data will be divided into 20% test data and 80% training data.

2.5 Implementation method

The partitioned data will be selected using the N-gram feature. The N-gram values used are unigram, bigram, and trigram. Each of the three N-grams will implement k-fold cross-validation as a modeling scenario with a value of k = 10 and an undersampling technique using the near-miss algorithm will be implemented to overcome data imbalances. The classification method used is K nearest neighbor with neighboring values k = 3, 5, 7 and 9 and the distance metric used is cosine similarity.

The method used in feature selection is N-Gram. N-gram is a method to predict the occurrence of words. This method works by calculating the frequency of occurrence in order and will be worth 0 (zero) if the frequency of occurrence of the word does not exist (Prayogo, 2018). The equation for solving the number of N-gram features is stated as follows:

$$Ngrams_k = X - (N - 1) \tag{1}$$

The method used in word weighting is TF-IDF. TF-IDF is a technique in feature extraction to know the weight of each word. The technique consists of two parts, namely the TF value or term frequency and the IDF value or inverse document frequency (Candra & Nanda Rozana, 2020). TF or term frequency is the word weight obtained by counting the occurrence of certain words in a document. The more the word appears, the higher the value (Fitri, 2013). IDF or inverse document frequency is the opposite of document frequency, the more a word appears, the less unique the word will be (Saadah et al., 2013). The following equations are used in TF and IDF:

$$Wtf_{t,d} = \begin{cases} 1 + log_{10}\, tf_{t,d} & if\ tf_{t,d} > 0 \\ 0 & then \end{cases} \tag{2}$$

$$idf_t = log_{10}\, \frac{N}{Df_t} \tag{3}$$

$$W_{t,d} = Wtf_{(t,d)}.idf_{(t)} \tag{4}$$

The classification method used is K Nearest Neighbor. In measuring distance using metric cosine similarity. K Nearest Neighbor is one of the classification methods and includes supervised learning. This method works by labeling based on the nearest neighbor (Tan & Zhang, 2008). Cosine similarity is a method to find similarities or similarities between documents. This method works by comparing test data to training data. The closer the resulting angle is to zero degrees, the higher the similarity level, and vice versa if the resulting angle widens, the lower the similarity level (Yasni et al., 2018). The results of the similarity measurement obtained by the test data against the training data will be sorted to determine the nearest neighbor (Herwijayanti et al., 2018). The following is the Cosine similarity equation:

$$CosSim\ (q, d_j) = \frac{d_j.\, q}{|d_j|\ |q|} = \frac{\sum_{i=1}^{n} W_{iq} \times W_{ij}}{\sqrt{\sum_{i=1}^{t} (W_{ij})^2} \times \sqrt{\sum_{i=1}^{n} (W_{iq})^2}} \tag{5}$$

2.6 Evaluation

The prediction results of the K nearest neighbor method to the actual data will be compared using a confusion matrix and for performance, measurement using measurements of accuracy, precision, and recall. The confusion matrix is a technique to analyze or evaluate the performance of an algorithm. The prediction results of an algorithm will be compared with the actual data to find the level of accuracy and others (Han et al., 2012). The comparison of the prediction results to the actual data will be divided into four categories, namely true positive, true negative, false positive, and false negative. The following equations are used to measure the level of accuracy, precision, and recall (Bowes et al., 2012):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

3. RESULTS AND DISCUSSIONS

3.1 Implementation of the K Nearest Neighbor method

1. Modeling

a. Unigram

Based on the modeling results in the k-fold cross-validation scenario on the training data with the unigram selection feature, it is shown in Figure 2 below.

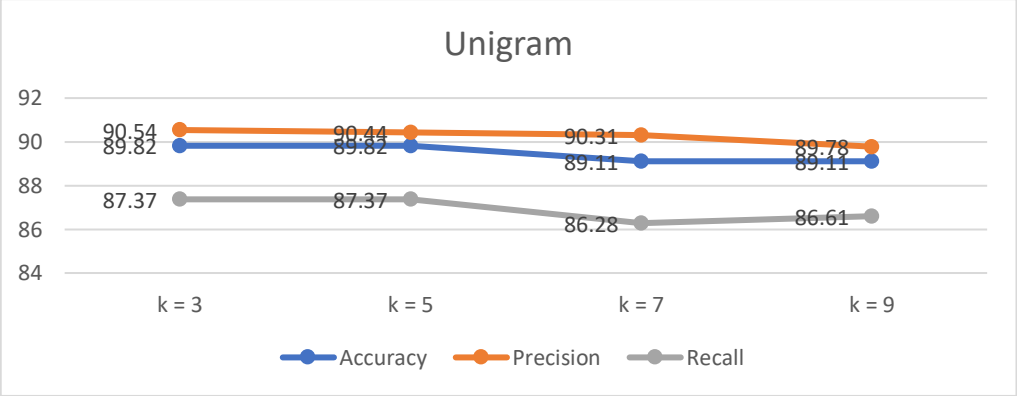

Figure 2. The average value of accuracy, precision, and recall on nearest neighbor k = 3, 5, 7, and 9 with unigram feature selection

Based on the results of measuring the level of accuracy, precision, and recall from the modeling carried out on the training data using the unigram selection feature described in Figure 2 above, it can be explained that k = 3 is the best nearest neighbor value in the compiled model. This is evidenced by the values of accuracy, precision, and recall which have higher values than the other k values of neighbors.

b. Bigram

Based on the modeling results in the k-fold cross-validation scenario on the training data with the bigram selection feature, it is shown in Figure 3 below.

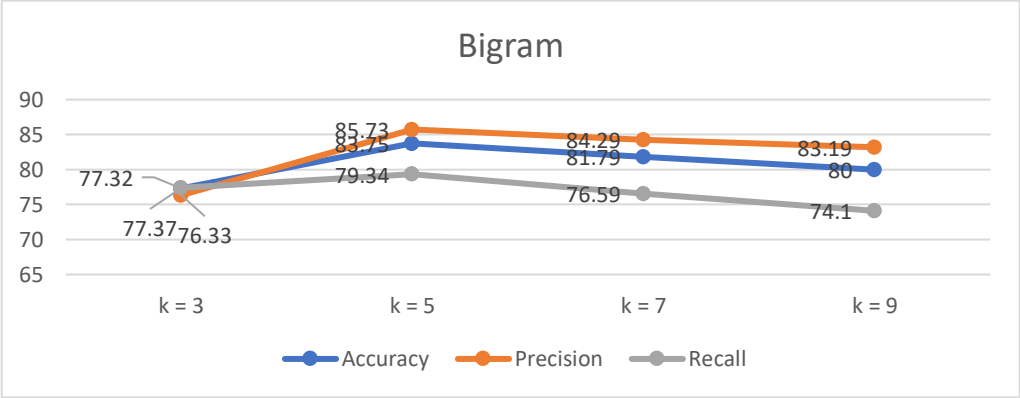

Figure 3. The average value of accuracy, precision, and recall on nearest neighbor k = 3, 5, 7, and 9 with bigram feature selection

Based on the results of measuring the level of accuracy, precision, and recall from the modeling carried out on the training data using the bigram selection feature described in Figure 3 above, it can be explained that k = 5 is the best nearest neighbor value in the compiled model. This is evidenced by the values of accuracy, precision, and recall which has higher values than the other k values of neighbors.

c. Trigram

Based on the modeling results in the k-fold cross-validation scenario on the training data with the trigram selection feature, it is shown in Figure 4 below.

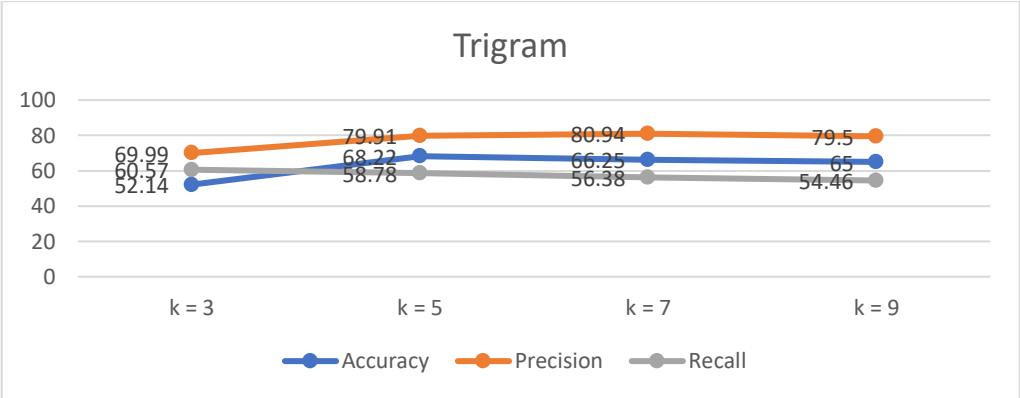

Figure 4. The average value of accuracy, precision, and recall on nearest neighbor k = 3, 5, 7, and 9 with trigram feature selection

Based on the results of measuring the level of accuracy, precision, and recall from the modeling carried out on the training data using the bigram selection feature described in Figure 4 above, it can be explained that k = 5 is the best nearest neighbor value in the compiled modeling. This is evidenced by the values of accuracy, precision, and recall which has higher values than the other k values of neighbors.

2. Testing

The test is carried out on the test data with a portion of 20% of the initial data partition. The following are the results of measuring the level of accuracy, precision, and recall of test data from the k nearest neighbor method which are presented in table 1 below.

Table 1. Measurement of accuracy, precision, and recall on the prediction results of K-NN

	Unigram			Bigram			Trigram		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Positive	0.91	0.9	0.97	0.85	0.81	1	0.69	0.68	1
Negative		0.93	0.8		1	0.57		1	0.12
		0.92	0.89		0.91	0.79		0.84	0.56

Based on the results of measuring the level of accuracy, precision, and recall from the testers who were carried out on test data using the unigram, bigram, and trigram selection features presented in table 1 above, it can be explained that the unigram selection feature is the best. This is evidenced by the values of accuracy, precision, and recall which has higher values than other selection features.

3.2 Implementation Of K Nearest Neighbor And Undersampling Method

The condition of the data that is considered unbalanced is balanced using the undersampling technique and the algorithm used is a near-miss. The undersampling technique trims the majority data to have the same amount as the minority data. The results of this application obtained a total of 402 data with 201 positive data and 201 negative data conditions.

1. Modeling

The modeling was carried out using a k-fold cross-validation scenario with a value of k = 10. This modeling scenario was also carried out on three selection features, namely unigram, bigram, and trigram.

a. Unigram

Based on the modeling results in the k-fold cross-validation scenario on training data with the unigram selection feature, it is shown in Figure 5 below.

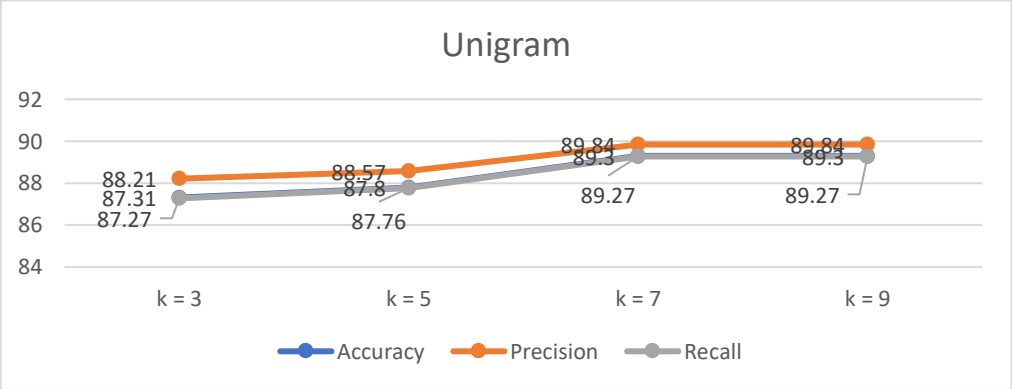

Figure 5. The average value of accuracy, precision, and recall on nearest neighbor k = 3, 5, 7, and 9 with unigram feature selection on undersampling data

Based on the results of measuring the level of accuracy, precision, and recall from the modeling carried out on the training data after the undersampling technique was

implemented using the unigram selection feature described in Figure 5 above, it can be explained that k = 7 and 9 are the best closest neighbor values in the modeling that is compiled. This is evidenced by the values of accuracy, precision, and recall which has higher values than the other k values of neighbors.

b. Bigram

Based on the modeling results in the k-fold cross-validation scenario on the training data with the unigram selection feature, it is shown in Figure 6 below:

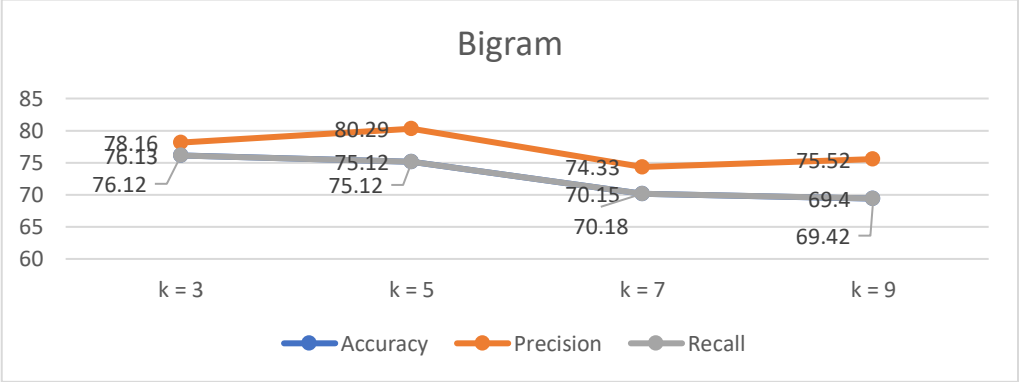

Figure 6. Average acquisition of accuracy, precision, and recall values at nearest neighbor k = 3, 5, 7 and 9 with bigram feature selection on undersampling data

Based on the results of measuring the level of accuracy, precision and recall from the modeling carried out on the training data after the undersampling technique was implemented using the bigram selection feature described in Figure 6 above, it can be explained that k = 5 is the best nearest neighbor value in the modeling that is compiled. This is evidenced by the values of accuracy, precision, and recall which has higher values than the other k values of neighbors.

c. Trigram

Based on the modeling results in the k-fold cross-validation scenario on training data with the unigram selection feature, it is shown in Figure 7 below.

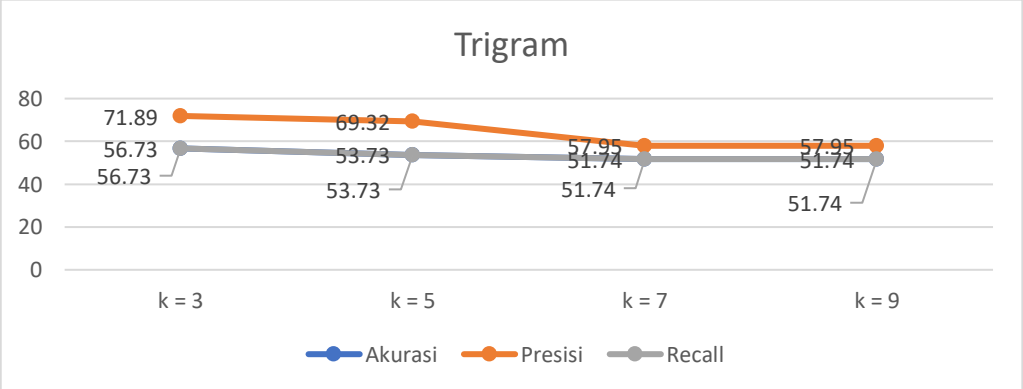

Figure 7. Average acquisition of accuracy, precision and recall values at nearest neighbor k = 3, 5, 7, and 9 with trigram feature selection on undersampling data

Based on the results of measuring the level of accuracy, precision, and recall from the modeling carried out on the training data after the undersampling technique was implemented using the trigram selection feature described in Figure 7 above, it can be explained that k = 3 is the best nearest neighbor value in the compiled model. This is evidenced by the values of accuracy, precision, and recall which has higher values than the other k values of neighbors.

2. Testing

The test is carried out on the test data with a portion of 20% of the initial data partition. The following are the results of measuring the level of accuracy, precision, and recall of test data from the k nearest neighbor method which are presented in table 2 below.

Table 2. Measurement results on the k nearest neighbor test using the undersampling data model

	Unigram			Bigram			Trigram		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Positive	0.87	0.86	0.96	0.75	0.92	0.9	0.51	1	1
Negative		0.9	0.71		0.59	0.67		0.42	0.25
		0.88	0.84		0.76	0.79		0.71	0.63

Based on the results of measurements of the level of accuracy, precision, and recall from the examiner carried out on the test data using the unigram, bigram, and trigram selection features presented in table 2 above, it can be explained that the unigram selection feature is the best. This is evidenced by the values of accuracy, precision, and recall which has higher values than other selection features.

4. CONCLUSION

Based on the research results obtained, here are some things that can be concluded from a series of processes carried out: from the application of the N-gram selection feature, the unigram is the selection feature that has the best performance compared to the bigram and trigram selection features. This is based on the results of modeling that the unigram selection feature has the highest accuracy, precision, and recall with the highest average accuracy value of 89.82%, the highest average precision value of 90.54, and the highest recall average value of 87.37%, based on the results of modeling performed on each nearest neighbor k = 3, 5, 7, and 9, k = 3 is the best model for the nearest neighbor value. This is based on the results obtained by an average accuracy of 89.82%, an average precision of 90.54%, and an average recall of 87.37%, based on the modeling performed on the data before the undersampling technique was implemented, the highest average value was 89.82%, the highest average precision value was 90.54% and the highest recall average value was 87.37%. While the test results obtained the highest accuracy measurement of 91%, the highest precision of 0.92%, and the highest recall of 0.89%. Based on the modeling carried out on the data after the undersampling technique was implemented, the highest average value was 89.3%, the highest average precision value was 89.84% and the highest recall average value was 89.27%. While the test results obtained the highest accuracy measurement of 0.87%, the highest precision of 0.88%, and the highest recall of 0.84%. So it can be concluded that the initial data before the undersampling technique was implemented had better modeling and testing results.

REFERENCES

Artanti, D. P., Syukur, A., Prihandono, A., & Setiadi, D. R. I. M. (2018). *Analisa Sentimen Untuk Penilaian Pelayanan Situs Belanja Online Menggunakan Algoritma Naïve Bayes*. 8–9.

Bowes, D., Hall, T., & Gray, D. (2012). Comparing the performance of fault prediction models which report multiple performance measures: Recomputing the confusion matrix. *ACM International*

Conference Proceeding Series, 109–118. https://doi.org/10.1145/2365324.2365338

Candra, R. M., & Nanda Rozana, A. (2020). Klasifikasi Komentar Bullying pada Instagram Menggunakan Metode K-Nearest Neighbor. *IT Journal Research and Development*, *5*(1), 45–52. https://doi.org/10.25299/itjrd.2020.vol5(1).4962

Fitri, M. (2013). Perancangan Sistem Temu Balik Informasi Dengan Metode Pembobotan Kombinasi Tf-Idf Untuk Pencarian Dokumen Berbahasa Indonesia. *Jurnal Sistem Dan Teknologi Informasi*, *Vol. 1*(1), 1–6.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. In *Data Mining: Concepts and Techniques*. Elsevier Inc. https://doi.org/10.1016/C2009-0-61819-5

Herwijayanti, B., Ratnawati, D. E., & Muflikhah, L. (2018). Klasifikasi Berita Online dengan menggunakan Pembobotan TF-IDF dan Cosine Similarity. *Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *2*(1), 306–312. https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/796

Prayogo, R. D. (2018). Implementasi Bigram Model dan Markov Chain dalam Add-in Microsoft Word untuk Membentuk Kalimat Bahasa Inggris dari Bag of Words. *Tangerang: Universitas Mutltimedia Nusantara*. https://kc.umn.ac.id/5069/

Saadah, M. N., Atmagi, R. W., Rahayu, D. S., & Arifin, A. Z. (2013). SISTEM TEMU KEMBALI DOKUMEN TEKS DENGAN PEMBOBOTAN TF-IDF DAN LCS. *JUTI: Jurnal Ilmiah Teknologi Informasi*, *11*(1), 19. https://doi.org/10.12962/j24068535.v11i1.a16

Sulistianti, R. A., & Sugiarta, N. (2022). *Konstruksi Sosial Konsumen Online Shop Di Media Sosial Tiktok (Studi Fenomenologi Tentang Konstruksi Sosial Konsumen Generasi Z Pada Online Shop Smilegoddess Di Media Sosial Tiktok)*. *6*(1), 3456–3466.

Tan, S., & Zhang, J. (2008). An empirical study of sentiment analysis for chinese documents. *Expert Systems with Applications*, *34*(4), 2622–2629. https://doi.org/10.1016/j.eswa.2007.05.028

Tri Romadloni, N., Santoso, I., Budilaksono, S., & Ilmu Komputer STMIK Nusa Mandiri Jakarta, M. (2019). Perbandingan Metode Naive Bayes, Knn Dan Decision Tree Terhadap Analisis Sentimen Transportasi Krl Commuter Line. *Jurnal IKRA-ITH Informatika*, *3*(2), 1–9.

Yasni, L., Subroto, I. M. I., & Haviana, S. F. C. (2018). IMPLEMENTASI COSINE SIMILARITY MATCHING DALAM PENENTUAN DOSEN PEMBIMBING TUGAS AKHIR. *Transmisi*, *20*(1), 22. https://doi.org/10.14710/transmisi.20.1.22-28

Yuniarti, N., Ismawati, A., & Aini, A. N. (2020). Pengaruh Promosi Online Melalui Tiktok Terhadap Peningkatan Penjualan Produk Usaha di Masa Pandemi Covid-19. *Proceedings The 1st UMYGrace 2020 (Universitas Muhammadiyah Yogyakarta Undergraduate Conference)*.