Application of The Nearest Neighbor Algorithm for Classification of Online Taxibike Sentiments In Indonesia In The Google Playstore Application

Sindhy Genjang Setyorini1*, Mustakim2

1,2 Department of Information System, Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, 28293, Indonesia
1,2 Puzzle Research Data Technology Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, 28293, Indonesia
sindhygenjang99@gmail.com1* mustakim@uin-suska.ac.id2

Abstract Google playstore is an official platform that provides applications to services, and not only that, user can perform assessments and reviews after using the Google playstore. One of the applications and services provided by Google playstore is online transportation. There’s a couple of online transports available but at this time's research just scanned the official transport account of gojek and grab online. The data used in this study were application review and service review data from the two online transportations which are processed using the K-Nearest Neighbor and Improved K-Nearest Neighbor algorithms by comparing the accuracy values. An experiment was conducted to divide the data with 5 K on the K-Fold Cross-Validation. The results of the comparison of the KNN application data were that Grab had a higher accuracy of 77.77% compared to Gojek which had an accuracy value of 66.24%. In terms of the Data Services using the IKNN algorithm, Grab had a higher accuracy value of 78.23% while Gojek had an accuracy value of 48.19%. Meanwhile, from the time complexity of the process, the IK-NN algorithm is better than the K-NN.

1. Introduction
Google Play and the App Store were first released in 2008 with more than 1 million applications that can be downloaded and can be reviewed or commented on [1]. Google Play is a digital content service owned by Google which consists of online product stores such as music/songs, books, applications, games, or cloud-based media players. Google Play can be accessed via the web, android application (Play Store), and Google TV [2]. Google informed that there were 1.4 billion Android devices activated in September 2015[3].

The reviews contained in the Google Playstore can be used as a reference and reference in various needs, especially in the field of Sentiment Analysis. Reviews from users are often used as an effective and efficient tool in finding information [4] on a product or service, therefore text mining techniques are needed in classifying reviews on Google Play[5].

Text mining is a process of mining large amounts of data, text mining has both structured and unstructured data types and can be analyzed using various tools[6]. Currently, text mining is a trend that is found in several research fields of computing such as linguistics, Information Retrieval (Ir)[7].
One of the goals of Text mining is to be able to examine new information to be able to find interesting patterns found in the review. With so much information that can be extracted from Google play and can be processed using text mining technology. The data processing stage in text mining can also adopt the data mining process[8].

Sentiment Analysis can be defined as a way to process public opinion data given to consumers [9] by using a combination of text mining and data mining that discusses the products and services of an agency [10]. This sentiment analysis was divided into 3 types of opinions, namely negative opinions, positive opinions, and neutral opinions to get information on how a product or service is marketed in the community.

This research was conducted using opinion data or reviews on the official Gojek and Grab online transportation accounts. Gojek's profit in 2019 reached $10 billion[11] while Grab had an even bigger profit of $10 trillion [12]. However, at the end of 2019, the world was shocked by the emergence of a deadly disease outbreak in Indonesia, in March 2020 Coronavirus Disease 2019 (Covid-19) was also growing rapidly throughout Indonesia. The spread of Covid-19 has caused a decline in people's economic conditions, including online transportation services with restrictions and policies for working from home or Work From Home (WFH) [13].

In analyzing or processing data, you can use the concept of text mining with classification methods, namely Improved k-Nearest Neighbor and k-Nearest Neighbor. KNN is an object classification algorithm based on the k value of the closest training data (nearest)[14] but with the condition that the k value cannot be greater than the amount of training data and the k value is an odd number that is more than one[15]. The KNN includes on classification algorithms that can process large data with high performance computing, and it also has a good value of accuracy and performance, and it is also a simple and easy algorithm to be learned with very fast training[16], but KNN also has a disadvantage is necessary to determine the optimal value of k to represent the number of nearest neighbors. Computational problems because in the classification process, all data are used to determine the class of the test pattern[17]. While Improved KNN is a development algorithm from KNN by using different k values for each category by adjusting the amount of training data it has[18].

A previous study conducted by Muslimah, et al (2019) regarding Film Classification Based on Synopsis Using Improved KNN involved 250 documents as training data and 50 documents as test data, the best results were obtained with precision = 1, recall = 0.88, f-measure = 0.93, and an accuracy rate of 88%, better than KNN[19].

Another study conducted by Onantya (2019) analyzed the sentiment on the BCA Mobile Application Review Using BM25 and Improved K-Nearest Neighbor. The test of this study resulted in the best k-values of 10, with a precision value of 0.946, recall of 0.934, f-measurement of 0.939, and accuracy of 0.942. The results of the tests carried out obtained fluctuating accuracy results because they were influenced by the magnitude of the k-values[20]. Therefore, in this study, a comparison of the Nearest Neighbor model with the KNN and IKNN algorithms was carried out to analyze sentiment in the Online Transportation application on Gojek and Grab. This study was also conducted to test the accuracy and advantages of the IKNN algorithm compared to KNN for the case of text mining. The main parameters in determining the algorithm were based on the accuracy and computation of each algorithm.

2. Material and Method

Text mining is a technology for extracting information from text by users by adopting processes[7]. Text mining focuses more on techniques and methodologies within the scope of information system retrieval and knowledge-based processes where users interact and work with a set of documents using several analytical tools. The process of processing and processing of data collection, preprocessing data, K-Fold Cross Validation, and implemented using KNN and IK-NN algorithms to get the accuracy and complexity value (Figure 1).
2.1 Text Preprocessing

Steps are needed in text mining to process text data to be more structured. One of the stages in text mining is preprocessing. This stage is the stage where the data is prepared to become data that can be analyzed[8]. The preprocessing stages in this study include Tokenizing, Filtering, and Stemming.

2.2 K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) is an algorithm used to classify objects based on the learning data that is closest to the object. KNN is a method that uses a supervised algorithm with the results of the newly classified query instance based on the majority of the categories in the KNN. KNN algorithm includes into the top ten data mining algorithms because it’s very simple and simple in the process but is quite effective in categorizing text. In achieving the goal of the KNN algorithm used the Euclidean distance formula in defining the distance between testing and training of objects[21]. The KNN algorithm is quite straightforward; it determines its KNN based on the distance between the query instance and the training sample. Near or far distance is usually calculated based on Euclidean Distance which is represented in equation (1).

\[
d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2}
\]  

(1)

Where \((x_i, x_j)\) were obtained from the subtraction of records \((x_i)\) from each attribute then squared, then the results of the squares are added up and sorted by the smallest value to see the closest distance to the test data.

---

**Figure 1.** Research methodology
2.3 Improved k-Nearest Neighbor (IKNN)
The distinction between Improved KNN and the KNN method is that the value of k is determined differently. In order to achieve a high level of accuracy in the document classification process, the value of k in KNN must be determined appropriately and precisely.[22] Determination of the value of k, was then be calculated to get a new value of k (n), determine the value of new k (n) can be calculated using equation (2).

\[
n = \frac{k \times N(cm)}{\text{Max}[N(cm)]_{j=1..Nc}}
\]  

Description:
- \( n \) = New k value
- \( k \) = Set value of k
- \( \text{Max}[N(cm)]_{j=1..Nc} \) = Most training documents in all categories

Furthermore, the calculation of the probability of the X test document was conducted including the \( dj \) training document as many as n neighbor values for each category in the X document in the \( dj \) training document as many as n neighbor values for the training set. Equation 3 can be used to calculate the probability of the X test document in the m category.

\[
\frac{\sum_{dj \in \text{Top}_n\text{kNN(cm)}} \text{Sim(x,dj)} \text{y(dj,cm)}}{\sum_{dj \in \text{Top}_n\text{kNN(cm)}} \text{Sim(x,dj)}}
\]

Description:
- \( P(x, cm) \) = Probability of document X
- \( \text{Sim}(x,dj) \) = Similarity between document X and training document
- \( \text{Top}_n\text{kNN} \) = n best neighbor value
- \( Y(dj,cm) \) = Attribute function that meets one of the categories

3. Result and Analysis

3.1 Data Collection
Data collection in this study was carried out by scraping on Google Play using the python programming language, using search keywords related to official online transportation accounts used to serve consumers or responses on Google Play. The data used in this study were a time span from March 2020 to December 2020 on Gojek and Grab accounts. The reviews that were taken and analyzed were only in Indonesian which were divided into two parts, namely: reviews related to applications and reviews related to services. Gojek's review data were 3443 with 2225 (64.62%) service data and 1149 (33.37%) application data recorded. Meanwhile, Grab's data were 3443 with 1518 (44.08%) application data recorded and 1925 (55.91%) service data recorded.

3.2 K-Nearest Neighbor (KNN)
The application of the KNN algorithm in this study was used to obtain accurate results from text mining classification using review data from Google Play. KNN in this study employed the Euclidean Distance formula contained in equation 2.2 and uses five K parameters [23], namely \( k = 5, k = 10, k = 15, k = 20, \) and \( k = 25 \) by determining the distance based on equation 2.

The comparison of KNN Accuracy using Application Data and Service Data of Gojek can be seen in the graph below showing that Application Data has higher accuracy results which lie in parameter K-5 of 66.27% compared to Service data of 63.63% lies in parameter K -5 (Figure 2). Meanwhile, the results of the comparison of accuracy using the KNN algorithm on Application Data and Grab Online Transportation Service Data show that Data Services have a higher accuracy value of 77.77% which is located in parameter K-10 while Service Data has a lower accuracy value of 62.13%, which lies in the K-15 parameter (Figure 3).
3.3 Improved K-Nearest Neighbor (IK-NN)

Improved K-Nearest Neighbor (IK-NN) is the development of the K-NN method, where there are differences in determining the value of k. This study using Online Transportation Data which consists of Application Data and Service Data obtained from Google Playstore by scrapping on the official Gojek and Grab accounts.

Tests using the Improved K-Nearest Neighbor algorithm on application data and Gojek Online Transportation Service Data using K Initial 5, 10, 15, 20, and 25 resulted that the highest accuracy value lies at K-25 of 48.92% with a new K value is Negative K-25, Neutral K-21, and Positive K-17. While the Gojek Service Data, the highest accuracy value lies in K-20 of 69.45%, and the new K is Negative K-20, Neutral K-13, Positive K-8 (Figure 4). Meanwhile, in the Application data and Grab Online Transportation Service data, the test results show the highest accuracy value, namely Service Data of 78.23%, which lies in the initial K parameter 20 and gets a new K value, Negative K-3, Neutral K-6, and Positive K-20, while the Data The application obtained an accuracy value of 77.39% at the initial K 20 also with the new K Negative K-17, Neutral K-20 and Positive K-17 (Figure 5).

Based on the time complexity of the algorithm processing on the combination performed using K-NN and IK-NN, the IK-NN algorithm is obtained which has a faster processing time. In the 4th experiment on Gojek Data, it took 180 seconds on K-NN, while on the IK-NN algorithm it took 124 seconds on the 4th experiment on Data Grab. Therefore, if it is observed from the resulting pattern, it does not necessarily mean that the higher the K value, the better the processing time. Table 1 provides additional information.
Table 1. The Complexity of Processing Time in the Algorithm (in seconds)

| No | Experiment (K) | K-NN Gojek Data | K-NN Grab Data | IK-NN Gojek Data | IK-NN Grab Data |
|----|----------------|----------------|----------------|----------------|----------------|
| 1  | 5              | 189            | 213            | 187            | 140            |
| 2  | 10             | 194            | 199            | 191            | 138            |
| 3  | 15             | 180            | 198            | 131            | 130            |
| 4  | 20             | 182            | 196            | 141            | 124            |

4. Conclusion

After experimenting with the classification model with the K-Nearest Neighbor algorithm and Improved K-Nearest Neighbor using the data sharing method with 5 K on the K-Fold Cross Validation, the results of the comparison of the accuracy values between Application Data and Service Data on Gojek and Grab Online Transportation were obtained. Experiments using the KNN algorithm on Application data showed that Grab has higher accuracy results of 77.77% compared to Gojek of 66.27%.

While the result of IKNN Grab algorithm has higher accuracy in the amount of 77.39% and Gojek has accuracy value in the amount of 48.92%. In the data of Online Transportation Service obtained accuracy result using KNN algorithm for Gojek with higher accuracy in the amount of 63.63% compared to Grab which has accuracy value 62.13%. For Service Data using IKNN algorithm, Grab has higher accuracy value in the amount of 78.23% compared to Gojek which has accuracy value in the amount of 69.45%. Likewise with the time complexity of IK-NN algorithm process which has better time process compared to K-NN namely 124 seconds.

References

[1] Aritonang J and Rokhim R 2020 Big Data Analysis of Paid and Free Applications in Google Playstore and Apple App Store to Know Application Characteristics and Monetization Opportunities for New Startup in Indonesia 160 205–210
[2] Frie K, Hartmann-Boyce J, Jebb S, Albury C, Nourse R and Aveyard P 2017 Insights from google play store user reviews for the development of weight loss apps: Mixed-method analysis JMIR mHealth uHealth 5(12) 1–14
[3] Effendi J and Ramadhan M J 2018 Analisa Cluster Aplikasi pada Google Play Store dengan Menggunakan Metode K-Means Pros. Annu. Res. Semin. 2018 4(1) 103–106
[4] Latif R M A, Abdullah M T, Aslam Shah S U, Farhan M, Ijaz F and Karim A 2019 Data scraping from google play store and visualization of its content for analytics 2019 2nd Int. Conf. Comput. Math. Eng. Technol. iCoMET 2019
[5] Deshpande G and Rokne J 2018 User Feedback from Tweets vs App Store Reviews: An Exploratory Study of Frequency, Timing and Content Proc. - 2018 5th Int. Work. Artif. Intell. Requir. Eng. AIRE 2018 15–21
[6] Bhardwaj P and Khosla P 2017 Review of Text Mining Techniques IITM J. Manag. IT 8(1) 27–31
[7] Salloum S A, Al-Emran M, Monem A A and Shaalan K 2017 A survey of text mining in social media: Facebook and Twitter perspectives Adv. Sci. Technol. Eng. Syst 2(1) 127–133

[8] Allahyari M et al 2017 A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques arXiv

[9] Xu G, Meng Y, Qiu X, Yu Z and Wu X 2019 Sentiment analysis of comment texts based on BiLSTM IEEE Access 7 51522–51532

[10] Ain Q T et al 2017 Radiotherapy is the gold standard in treating bone malignancy. Effective in 50-90 % expectancy months 8(6)

[11] Palevsky B 2019 Motorcycle Taxis in Jakarta, Indonesia The Case for Recognition and Integration

[12] Vera M, Dewi K and Darma G S 2019 The Role of Marketing & Competitive Intelligence In Industrial Revolution 4.0 J. Manaj. Bisnis 16(1) 1

[13] Susilawati S, Falefi R and Purwoko A 2020 Impact of COVID-19’s Pandemic on the Economy of Indonesia Budapest Int. Res. Critic Inst. Humanit. Soc. Sci. 3(2) 1147–1156

[14] Mustakim, Hidayat A, Efendi Z, Aszani, Novita R and Lestari E T 2018 Algorithm comparison of naive bayes classifier and probabilistic neural network for water area classification of fishing vessel in Indonesia J. Theor. Appl. Inf. Technol 96(13) 4114–4125

[15] Beckmann M, Ebecken N F F and Pires de Lima B S L 2015 A KNN Undersampling Approach for Data Balancing J. Intell. Learn. Syst. Appl 7(4) 104–116

[16] Mustakim, Hastarimasuci R, Papilo P, Zarkasih, Zaitun and Nazir A 2019 Variable Selection to Determine Majors of Student using K-Nearest Neighbor and Naïve Bayes Classifier Algorithm J. Phys. Conf. Ser 1363(1)

[17] Maillé J, Ramírez S, Triguero I and Herrera F 2017 kNN-IS: An Iterative Spark-based design of the k-Nearest Neighbors classifier for big data Knowledge-Based Syst 117 3–15

[18] Himeur Y, Alsalem A, Bensaali F and Amira A 2020 Smart non-intrusive appliance identification using a novel local power histogramming descriptor with an improved k-nearest neighbors classifier Sustain. Cities Soc 67 102764

[19] Muslimah N, Indriati and Wihandika R 2019 Klasifikasi Film Berdasarkan Sinopsis dengan Menggunakan Improved K-Nearest Neighbor (K-NN) J. Pengemb. Teknol. Inf. dan Ilmu Komput 3(1) 196–204

[20] Onantya I D, Indriati and Adikara P P 2019 Analisis Sentimen Pada Ulasan Aplikasi BCA Mobile Menggunakan BM25 Dan Improved K-Nearest Neighbor J. Pengemb. Teknol. Inf. dan Ilmu Komput 3(3) 2575–2580

[21] Okfalisa, Gazaliba I, Mustakim and Reza N G I 2018 Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification Proc. - 2017 2nd Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. ICITISEE 2017 294–298

[22] Rahutomo F, Rozi I F, Mirzha P N, Informatika T, Malang P N and Neighbor I K 2007 Evaluasi Implementasi Algoritma Improved K-Nearest Neighbor Pada Kategorisasi Lirik Lagu Bahasa 19–26

[23] Hidayat A, Mustakim, Fauzi M Z and Syukra I 2019 Implementasi Algoritma K-Nearest Neighbor dan Probabilistic Neural Network untuk Analisis Opini Masyarakat Terhadap Toko Online di Indonesia Semin. Nas. Teknol. Informasi, Komun. dan Ind 11 254–261