

Tourism Destinations Popularity Rating In Malang Raya using Naive Bayes Classifier and Selection Sort Based on Twitter Word Polarity

Yunifa Miftachul Arif¹, Mochammad Wahyu Firmansyah², Roro Inda Melani³,
Supriyono⁴

Department of Informatics, Faculty of Science and Technology, Universitas Islam
Negeri (UIN) Maulana Malik Ibrahim, Malang, Indonesia

¹yunif4@gmail.com, ²15650111@mhs.uin-malang.ac.id, ³roro.uin@gmail.com,
⁴priyono@ti.uin-malang.ac.id

Abstract

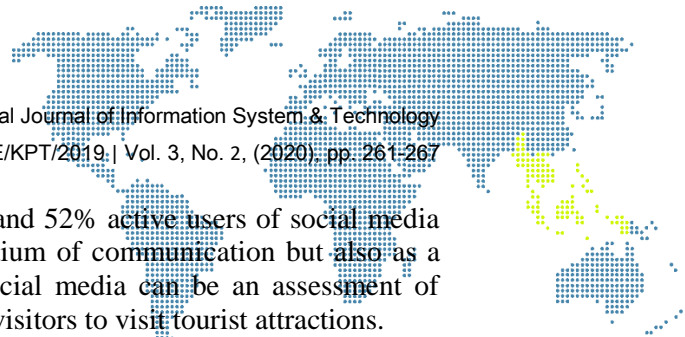
The development of tourism today has been supported by advances in information technology that can make it easier for everyone to get information about tourist attractions. Technology plays an essential role in improving the tourism industry sector. During the tour, tourists usually share moments by uploading photos or making a status on social media related to their experience visiting a tourist site. Malang, which has various types of tourism, makes it a tourist destination. However, the number of tours in Malang makes tourists confused to choose the trip to be visited. Because of this, we need a system that can provide information in the form of popular tourist rankings. In this research, a system that can determine the ranking of tourist attractions in Malang Raya was made. The data used comes from social media user tweets on Twitter using the keyword name of tourist attractions in Malang. The Naive Bayes Classifier method is used to help tweet classification, and the Selection Sort method is used to help the ranking process of tourist attractions. The final results obtained in the Batu City tourism ranking resulted in an accuracy of 86.3%, while in the tourism rating the artificial type of Batu City produced an accuracy of 100%. The difference in accuracy occurs because there are the same positive values at several tourist attractions, so the Selection Sort method cannot work. Because of this, further research is needed for ranking methods that can rank with the same positive value to produce a better ranking of tourist attractions.

Keywords: Naive Bayes, Ranking Tourism, Sentiment Analysis

1. Introduction

Malang Raya is an area in East Java province consisting of Malang City, Malang Regency, and Batu City [1]. Malang has various types of tourism, such as historical tourism, religious tourism, garden tourism, cultural tourism, and artificial tourism. Malang, which has multiple types of tourism, makes Malang have a nickname as a tourist city. So that not a few tourists make Malang as a tourist destination. According to LAW No. 10 of 2009 on tourism, tourist is an area that has much tourism and supported by a variety of services provided by entrepreneurs, people to the government. Nowadays tourism has developed with supported progress of information technology, making it easier for everyone to find information on tourist attractions that will be visited. Travellers are looking for information about photos, videos, and tourist reviews to see the latest conditions before travelling. Usually, tourists travel for picnics, fill leisure time, and enjoy the beauty of nature. The beauty of life is being served to make tourists interested in capturing and sharing to social media.

Social media users in Indonesia each year have increased. Based on the website social media users in Indonesia in 2019, 150 million users consist of 81% of Facebook active



users, 80% active users of Instagram social media, and 52% active users of social media Twitter. Nowadays, social media is not only a medium of communication but also as a media of information. Tourist information from social media can be an assessment of current conditions and reviews that can recommend visitors to visit tourist attractions.

Related research discusses Twitter's sentiments using the Twitter API and extracted raw data using natural language toolkit techniques. Views of online tweets evaluated based on the word score feature selection. To select the best features of the Chi-Square test are used, and Naïve Classifier Bayes is used for training and testing features and also evaluates sentimental polarity [2]. Related research is doing to classify tax complaints on Twitter automatically by using text mining [3].

Information and reviews from travellers are used in the sentiment analysis process that specifies positive and negative categories. Research on previous sentiment analysis has been conducted, among others, research [4] on tourism sentiment analysis in Malang city using the Naive Bayes method and Query Expansion Ranking using feature Selection. Related research [5] on the utilization of sentiment analysis for the popularity ranking of tourist destinations. Relevant research on the analysis of tourist attractions sentiment based on visitor comments using the Naive Bayes Classifier method. Proper research about the study of Twitter's opinion on the favourite tourist destination of Yogyakarta using the Naive Bayes Classifier method [6].

This study will use traveller reviews on social media to rank a tourist attraction. Surveys taken from social media Twitter contains tweets of travellers who have visited or learned about tourist attractions. Sentiment analysis is used to determine which reviews are positive, negative or neutral. The Naive Bayes Classifier method is used in this study for the tweet classification process obtained.

The purpose of this research is to determine the ranking of tourist spots using the Naive Bayes Classifier method and the Selection Sort method on tourist attractions in Malang Raya. While the limitations of this research problem is a tourist attraction in Malang Raya, the social media used is Twitter and tweets are taken only tweets Bahasa Indonesia.

2. Research Methodology

2.1. Place and Time research

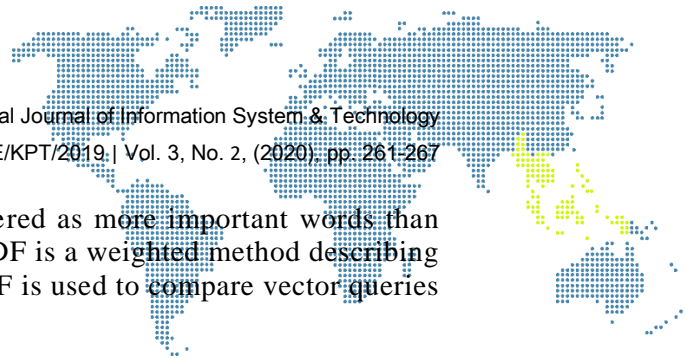
This research is done by crawling Twitter's social media user tweet data using the API (Application Programming Interface) that Twitter has provided. Retrieving this tweet data is done with a 1-hour interval in a day. As for the staging process, the tourist attraction is done every day during March 2020.

a) Sentiment analysis

Sentiment analysis is the process of grouping text and sentences within a document to determine the opinions in the document are either positive, neutral, or negative. Sentiment analysis can evaluate the sad, happy, or angry expression of a person [5]. In a broad field, the sentiment analysis consists of natural language processing, linguistic computing, and text analysis aimed at identifying the sentiment of the text, the document. Analysis of the task sentiment is used to group text in sentences. It aims to determine reports or penalties, including affirmative sentences, negative sentences, or neutral sentences [7].

b) TF-IDF

Term Frequency is a factor in determining the term weight in a document based on the number of occurrences in the document. The number of appearances in the word is calculated when the weight is given to a word. The higher the number of term occurrences of the document, the higher is also the weight in the document. Inverse Document Frequency is a process of breaking down a predominantly term that often appears in various documents. The process is required because the time often seems to be considered a general term, so it has an insignificant value. Words



or phrases that appear in a document are considered as more important words than those that often appear in many documents. TF-IDF is a weighted method describing vector space model documents. Generally, TF-IDF is used to compare vector queries and vector documents using similarities.

c) Preprocessing Text

Preprocessing text is a process for converting unstructured text into structured text. There are five stages of preprocessing text in this study, such as cleansing, case folding, recognized removal, stemming, and tokenizing.

1) Cleaning

In the cleaning stage, a process is carried out to clean the sound in the sentence from the crawl result. The omitted words that include noise are hyperlinks, emoji, usernames, hashtags, numbers, and special characters.

2) Case Folding

The case folding stage is the process of changing all the letters in the sentence to lowercase. For example, the word "tour of Malang" was changed to "Tourism of Malang City."

3) Stopword Removal

The stages of recognized removal are the process of removing words that are considered less critical. Words that include less important are not descriptive words, so they can be discarded using the bag-of-words approach. The examples of words that are less important are "or," "in," "are," and others. Stopword removal using the stoplist algorithm is throwing words based on a list of words that can be discarded. The list of words that can be discarded in this study comes from the literary library.

4) Stemming

The stemming stage is the process of changing the word to a basic word. For example, the word "travelling" is changed to the word "go." This stemming process is based on a pre-built list of English words. The basic English word list comes from the literary library.

5) Tokenizing

The tokenizing stage is the process of putting words in a sentence into a word unit. This process is done by cutting a sentence into words per word. Examples of tokenizing stages can be seen in Figure 1.

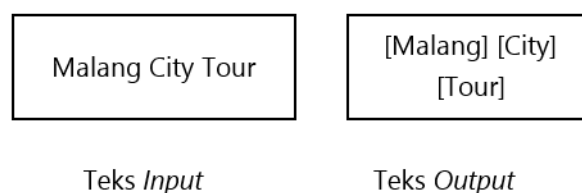
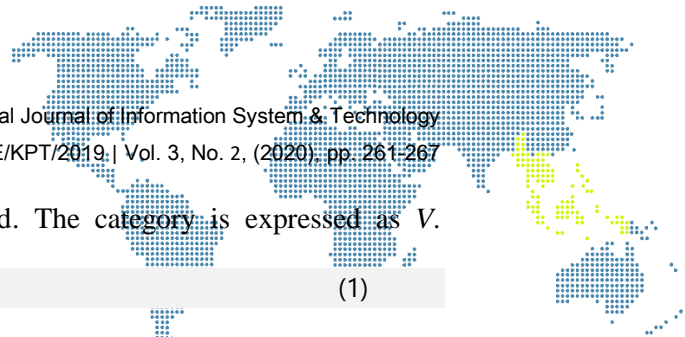


Figure 1. Example Tokenizing

2.2. Naive Bayes Classifier

Naive Bayes Classifier is a classification algorithm that is widely used for classifying documents. Naive Bayes Classifier Biasanaya used for text classification [8]. Naive Bayes Classifier is a classifying method based on Bayes' theorem that data is independent and not mutually related. Its independent nature makes the Naive Bayes Classifier method only require a small amount of training data for the classification process [5] Classification using the Naive Bayes Classifier has two stages. The first is the classification of data previously known by the category. The second is the classification of data, which has an unknown class. This review is considered a document with



attributes that contain the words of the first word. The category is expressed as V . Classification in Naive.

$$V_{MAP} = \text{arg}_{v_j \in V} \max P(a_1, a_2, \dots, a_n | v_j) \quad (1)$$

Bayes Classifier is done by finding the highest probability category against the P document ($V = v_j | D = di$). The so-called V_{MAP} ,

$$V_{MAP} = \text{arg}_{v_j \in V} \max P(a_1, a_2, \dots, a_n | v_j) P(v_j) \quad (2)$$

Following similarities:

$$V_{MAP} = \text{arg}_{v_j \in V} \max P(v_j) \prod_i P(a_i | v_j) \quad (3)$$

Description:

- v_j = Category
- a_i = Word in document
- $P(a_i/v_j)$ = Probability A_i against V_j categories $P(v_j) = \text{Probabilitas } v_j$

The values $P(v_j)$ and $P(a_i / v_j)$ are derived from the results of the calculation during training, the following similarities:

$$P(v_j) = \frac{|doc_j|}{|contoh|} \quad (4)$$

$$P(w_k | v_j) = \frac{n_k + 1}{n + |vocabulary|} \quad (5)$$

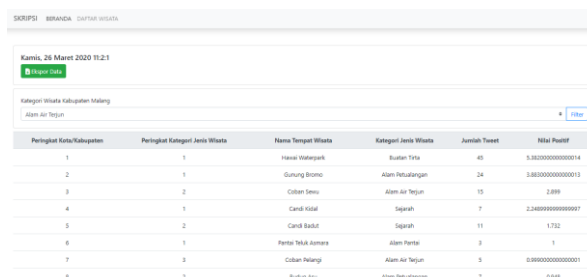
Description::

- $|doc_j|$ = number of documents in category j
- $|contoh|$ = Number of documents in the example
- n_k = frequency of each word occurrence
- n = total number of words in the v_j category
- $|vocabulary|$ = number of whole words from all categories

Selection sort is an algorithm for sorting data, starting with finding the smallest value, which is then placed in the initial position [9]. The Selection Sort method is called the maximum method or minimum method. The maximum method is based on the maximum data exchange process with elements that are at the very end to ascending, during the first element for the descending. The minimum method is based on the minimum data exchange process with elements that are at the very end to ascending, during the first element for the descending. The process of sorting data by retrieving the largest value of the data stack is then replaced with the rightmost data [10]. To process sort data by the Selection Sort method is depicted in the following image.

3. Result and Discussion

The result of tourist spot rating information is implemented web-based application, so the public can access it at the address <https://skripsi.wahyu.dev>. The ranking information on the webpage is updated daily. The Web page displays the rating of tourist spots in Malang, Batu, and Malang city, as shown in Figure 3, as well as tourist attractions rankings based on the category of tour type, as shown in Figure 4.



Peringkat Kota/Kabupaten	Peringkat Kategori Jenis Wisata	Nama Tempat Wisata	Kategori Jenis Wisata	Jumlah Tweet	Nilai Raster
1	1	Hawa Volcanik	Buana Toba	45	5.9200000000000014
2	1	Gurung Bromo	Alam Petalangan	24	5.8800000000000013
3	2	Coban Sempu	Alam Air Terjun	15	2.899
4	1	Candi Kidal	Sagarah	7	2.3400000000000007
5	2	Candi Bakid	Sagarah	11	1.732
6	1	Pantai Teluk Amara	Alam Pantai	3	1
7	3	Coban Palang	Alam Air Terjun	5	0.9900000000000001
8	2	Budug Aso	Alam Petalangan	7	0.946

Figure 2. City/District Tourist Spots Ranking



Beranda / Peringkat Wisata Alam Air Terjun Kabupaten Malang


Senin, 30 Maret 2020 21:7:51

Peringkat Kategori Jenis Wisata	Nama Tempat Wisata	Kategori Jenis Wisata	Jumlah Tweet	Nilai Positif
1	Coban Rondo	Alam Air Terjun	19	3.5800000000000005
2	Coban Sewu	Alam Air Terjun	12	1.433
3	Coban Pelangi	Alam Air Terjun	6	1.199
4	Coban Bidadari	Alam Air Terjun	2	0.619
5	Coban Jahe	Alam Air Terjun	8	0.45
6	Coban Cinde	Alam Air Terjun	0	0.2

Figure 3. Ranking Category Tour Type.

In Figure 3 and Figure 4 indicates when the name of the tourist spot in the Select, it will display the description information page, as shown in Figure 5.

Deskripsi Tempat Wisata



Hawaii Waterpark
 Peringkat 1 Wisata Kabupaten Malang Hari Ini
 Peringkat 1 Wisata Buatan Tirta Kabupaten Malang Hari Ini
 Jumlah Tweet : 45
 Nilai Positif : 5.3820000000000004

Deskripsi
 Hawaii Waterpark Malang dibangun di atas lahan dengan luas 28.000 meter persegi pada tahun 2014 dan mulai dioperasikan pada pertengahan tahun 2015. Wisata air ini terletak di Perumahan Graha Kencana Jl Raya Karanglo Malang dengan akses yang sangat strategis yang bisa diakses dari bandara, stasiun dan juga terminal. Dengan Rp. 50.000,00 selama soft opening anda bisa menikmati berbagai wahana-wahana yang berkualitas internasional. Ada 10 wahana yang terdapat di Hawaii Waterpark Malang, yaitu Hawaii Water House, Waikiki Beach, Waimea Stream River, Mavi Island, Rainbow Fall, Akaolu pool, Ekolu slide, Hula-hula slide, Wailele Slide. Dari sembilan wahana yang ada, ada dua wahana yang menjadi unggulan di Hawaii Waterpark Malang yaitu Hawaii Water House dan Tsunami pool (Waikiki Beach). Wisata air yang diperuntukkan untuk Keluarga beserta anak-anak, dewasa, remaja, kalangan bisnis, kelompok sekolah dan perusahaan ini akan buka soft opening pada Sabtu, 18 Juli 2015 jam 10.00 WIB dan selanjutnya akan buka mulai dari hari Senin sampai Minggu mulai dari jam 10.00 sampai jam 17.00 WIB. Dengan adanya Hawaii Waterpark Malang ini diharapkan bisa membuka wana baru bagi pariwisata yang ada di kota Malang terutama wisata air, serta dapat menarik wisatawan lokal maupun mancanegara untuk datang berkunjung ke kota Malang. Dan yang pasti Hawaii Waterpark Malang akan memberikan fasilitas dan pelayanan yang terbaik bagi pelanggannya.

Figure 4. Description of tourist attractions

The description page of the attractions shown is the name of the tourist attractions, ranking in the city/district, ranking in the category of travel type, number of crawling tweets, the positive value of the classification result and a brief description of tourist attractions. Data crawl results are done every day in March 2020, then the classification process to determine the positive value at the tourist attractions. The following data on the tweet classification result in Batu City on March 01, 2020.

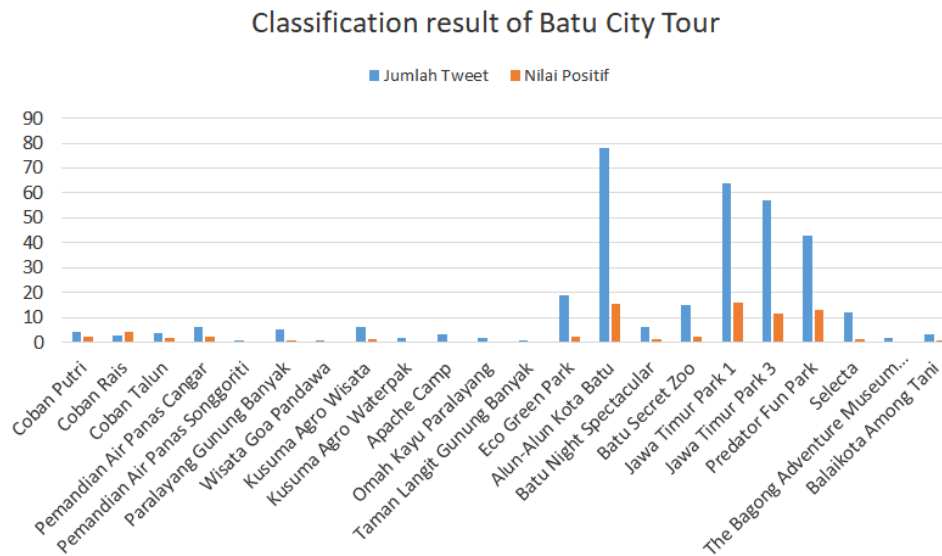
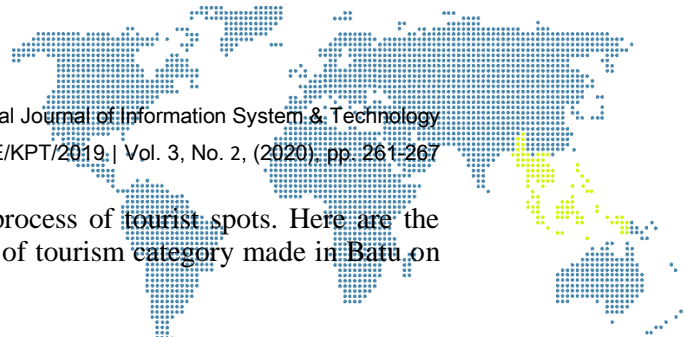


Figure 5. Classification result of Batu City Tour



After the classification process, then the rating process of tourist spots. Here are the rating data of Batu City tourist spot and the ranking of tourism category made in Batu on March 01 2020.

Table 2. Tourist spot rankings of Batu City March 01 2020

Ranking	Name of tourist attraction	Category Tour type	Number of tweets	Positive value	Time Rating
1	Jawa Timur Park 1	Buatan	64	15,938	01-03-2020
2	Alun-Alun Kota Batu	Buatan	78	15,226	01-03-2020
3	Predator Fun Park	Buatan	43	13,164	01-03-2020
4	Jawa Timur Park 3	Buatan	57	11,684	01-03-2020
5	Coban Talun	Alam Air Terjun	19	2,581	01-03-2020
6	Pemandian Air Panas Cangar	Alam Tirta	6	2,134	01-03-2020
7	Eco Green Park	Buatan Edukasi	19	2,065	01-03-2020
8	Batu Secret Zoo	Buatan	15	2,032	01-03-2020
9	Coban Rais	Alam Air Terjun	8	1,916	01-03-2020
10	Kusuma Agro Wisata	Alam Ekowisata	6	1,332	01-03-2020
11	Selecta	Buatan	12	1,249	01-03-2020
12	Batu Night Spectacular	Buatan	6	1,166	01-03-2020
13	Coban Putri	Alam Air Terjun	5	1,083	01-03-2020
14	Balaikota Among Tani	Event	3	0,666	01-03-2020
15	Paralayang Gunung Banyak	Alam Petualangan	5	0,583	01-03-2020
16	Omah Kayu Paralayang	Buatan Petualangan	2	0,4	01-03-2020
17	Kusuma Agro Waterpark	Buatan Tirta	2	0,333	01-03-2020
18	Pemandian Air Panas Songgoriti	Alam Tirta	1	0,286	01-03-2020
19	Apache Camp	Buatan Petualangan	3	0,25	01-03-2020

Table 3. Ranking of tourism category made in Batu 01 March 2020

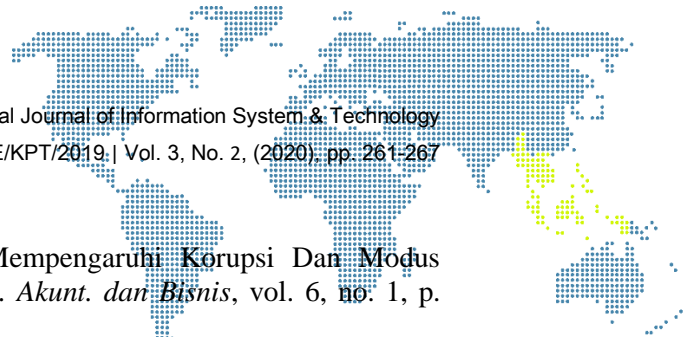
Ranking	Name of tourist	Number of tweets	Positive value	Time Rating
1	Jawa Timur Park 1	64	15,938	01-03-2020
2	Alun-Alun Kota Batu	78	15,226	01-03-2020
3	Predator Fun Park	43	13,164	01-03-2020
4	Jawa Timur Park 3	57	11,684	01-03-2020
5	Batu Secret Zoo	15	2,032	01-03-2020
6	Selecta	12	1,249	01-03-2020
7	Batu Night Spectacular	6	1,166	01-03-2020
8	The Bagong Adventure Museum Tubuh	2	0,333	01-03-2020

Based on the data collected from the results of classification and ranking of tourist attractions, can be seen in the rating of 22 tourist attractions. In comparison, the city level there are 19 tourist attractions. That happened because there was a positive value of 0.333 in the same place in the Bagong Adventure Body Museum, Kusuma Agro Waterpark, Gunung Sky Park many. When there is a similar value to a different tourist spot, the Selection Sort method cannot determine the largest and smallest value. So it is only used one of the tourist sites that have the same value, so as not to occur duplicate rankings.

Out of 22 tourist attractions, there are 19 tourist attractions including the Batu City tourist rating so that the accuracy of the obtained is 86.3%. For the category of artificial tourism, the classification and ranking of tourist objects are the same, resulting in 100% accuracy. The difference in accuracy occurs because there is a positive value of the same classification results in several tourist sites.

4. Conclusion

Conclusions that can be taken are the Naive Bayes Classifier, and Selection Sort methods can be used to rank tourist spots in Malang Raya. The ranking result of city/district tourism can be different from the category of tourist type because there is the same positive value in different tourism places. The suggestions can be done to make the results better, namely improved the method to rank in order to determine the rank when there is the same value resulting in a more accurate rating, and combining social media crawls of Instagram, Facebook and Twitter to get better results.



References

- [1] S. Bahri, “Analisis Faktor-Faktor Yang Mempengaruhi Korupsi Dan Modus Korupsi APBD Di Malang Raya,” *J. Manaj. Akunt. dan Bisnis*, vol. 6, no. 1, p. 110805, 2008.
- [2] M. Vadivukarassi, N. Puviarasan, and P. Aruna, “Sentimental Analysis of Tweets Using Naive Bayes Algorithm,” *World Appl. Sci. J.*, vol. 35, no. 1, pp. 54–59, 2017.
- [3] P. Dellia and A. Tjahyanto, “Tax Complaints Classification on Twitter Using Text Mining,” *IPTEK J. Sci.*, vol. 2, no. 1, p. 11, 2017.
- [4] S. Fanissa, M. A. Fauzi, and S. Adinugroho, “Analisis Sentimen Pariwisata di Kota Malang Menggunakan Metode Naive Bayes dan Seleksi Fitur Query Expansion Ranking | Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 8, pp. 2766–2770, 2018.
- [5] M. Murnawan, “Pemanfaatan Analisis Sentimen Untuk Peningkatan Popularitas Tujuan Wisata,” *J. Penelit. Pos dan Inform.*, vol. 7, no. 2, p. 109, 2017.
- [6] D. T. Hermanto, M. Ziaurrahman, M. A. Bianto, and A. Setyanto, “Twitter Social Media Sentiment Analysis in Tourist Destinations Using Algorithms Naive Bayes Classifier,” *J. Phys. Conf. Ser.*, vol. 1140, no. 1, 2018.
- [7] N. Yanti, R. Kurniawan, S. N. H. S. Abdullah, M. Z. A. Nazri, W. Hunafa, and M. Kharismayanda, “Tropical diseases web-based expert system using certainty factor,” *Proc. - 2018 2nd Int. Conf. Electr. Eng. Informatics Towar. Most Effic. W. Mak. Deal. with Futur. Electr. Power Syst. Big Data Anal. ICon EEI 2018*, no. October, pp. 62–66, 2018.
- [8] S. Mukherjee and P. K. Bala, “Sarcasm detection in microblogs using Naïve Bayes and fuzzy clustering,” *Technol. Soc.*, vol. 48, pp. 19–27, 2017.
- [9] M. E. Al Rivian, “Perbandingan Kecepatan Gabungan Algoritma Quick Sort dan Merge Sort dengan Insertion Sort, Bubble Sort dan Selection Sort,” *J. Tek. Inform. dan Sist. Inf.*, vol. 3, no. 2, 2017.
- [10] E. Retnoningsih, “Algoritma Pengurutan Data (Sorting) Dengan Metode Insertion Sort dan Selection Sort,” vol. 3, no. 1, pp. 95–106, 2018.

Authors



1st Author

Yunifa Miftachul Arif

Department of Islamic Education, Faculty of Tarbiyah and Teaching Training, Universitas Islam Negeri (UIN) Maulana Malik Ibrahim, Malang, Indonesia