SCIENTIST
Journal of Security, Computer, Information, Embedded, Network, and Intelligence System

P-ISSN : 3024-8701
E-ISSN : 3024-8329

(SCIENTIST) Journal of Security, Computer, Information, Embedded, Network, and Intelligence System
Vol. 1, No. 1, 2023

Clustering Data Cuitan pada Media Sosial Twitter Menggunakan Metode K-Means

Dewi Fatmarani Surianto[1*]
Program Studi Teknik Komputer
Fakultas Teknik Universitas Negeri Makassar
Kota Makassar, Indonesia
dewifatmaranis@unm.ac.id

ARTICLE INFO

Received : 30 October 2023
Accepted : 27 November 2023
Published : 09 December 2023

ABSTRACT

Twitter, as a popular social media with millions to billions of global users, stores a wide variety of information. This study focuses on the use of Text Mining to analyze tweet content through the application of clustering techniques, specifically using the K-Means algorithm. The implementation process involves several stages of text processing, including casefolding, tokenizing, stopword removal, and stemming. Feature extraction is performed to provide input for the K-Means algorithm. The clustering evaluation uses the Silhouette coefficient method. The test results show that different K values result in a variation of the silhouette value. In a particular test scenario, a value of K=2 resulted in a silhouette of 0.5000421, K=5 had a value of 0.0501051, and K=9 had a value of 0.501893. From these values, the data structure of the dataset taken can be categorized as medium structure, because the silhouette value is in the range of 0.5 to 0.7. These results show that cluster quality is influenced by the K value, with the silhouette value being the main determinant.
Keywords: clustering, evaluation, k-means, social media, twitter

ABSTRAK

Twitter, sebagai media sosial yang populer dengan jutaan hingga miliaran pengguna global, menyimpan beragam informasi. Studi ini fokus pada penggunaan Text Mining untuk menganalisis konten tweet melalui penerapan teknik clustering, khususnya menggunakan algoritma K-Means. Proses implementasi melibatkan beberapa tahap text processing, termasuk casefolding, tokenizing, stopword removal, dan stemming. Ekstraksi fitur dilakukan untuk menyediakan input untuk algoritma K-Means. Evaluasi clustering menggunakan metode Silhouette coefficient. Hasil pengujian menunjukkan bahwa nilai K yang berbeda menghasilkan variasi nilai silhouette. Dalam skenario uji tertentu, nilai K=2 menghasilkan silhouette 0,5000421, K=5 memiliki nilai 0,0501051, dan K=9 memiliki nilai 0.501893. Dari nilai-nilai ini, struktur data dari dataset yang diambil dapat dikategorikan sebagai medium structure, karena nilai silhouette berada di rentang 0,5 hingga 0,7. Hasil ini menunjukkan bahwa kualitas cluster dipengaruhi oleh nilai K, dengan nilai silhouette menjadi penentu utama.
Keywords : klaster, evaluasi, k-means, media sosial, twitter

I. PENDAHULUAN

Twitter merupakan salah satu media sosial yang popular dan memiliki jutaan hingga milyaran pengguna di seluruh dunia. Kementerian Komunikasi dan Informatika (Kemenkominfo) mengungkapkan Indonesia menempati peringkat 5 pengguna Twitter terbesar di dunia. Pengguna Twitter, berdasarkan data PT Bakrie Telecom, memiliki 19,5 juta pengguna di Indonesia. Twitter dapat dimanfaatkan oleh pengguna dengan berbagai tujuan, salah satunya media kampanye.

Pada tahun 2019 lalu, Indonesia memasuki tahun pemilihan umum, salah satu yang menjadi perhatian adalah fenomena pemilihan presiden. Pengguna Twitter dapat saling berhubungan, mengutarakan opini ataupun saling merespon ke berbagai pengguna lainnya di seluruh penjuru dunia. Pengguna Twitter cukup membuat pesan pendek yang disebut dengan tweet. Terdapat beberapa fitur umum yang disediakan oleh Twitter, seperti retweet, reply, likes, dan lainya. Dengan menggunakan perintah retweet pada Twitter, pengguna Twitter dapat menyebarkan tweet dari pengguna lain tanpa harus mengetiknya kembali.

Pengumpulan data tweet dari Twitter dapat dilakukan dengan mengintegrasikan bahasa pemrograman Python dan Twitter API. Untuk mempermudah mengetahui jenis konten dari sejumlah data tweet, maka dibutuhkan proses Text Mining terhadap data tweet salah satunya dengan menerapkan teknik clustering. Pada Text Mining, teknik clustering digunakan untuk mengelompokkan data tekstual berdasarkan kesamaan konten yang dimiliki kedalam beberapa kelompok/klaster, sehingga pada setiap klaster berisikan data-data tekstual yang memiliki karakteristik dan konten semirip mungkin [1].

Beberapa metode yang paling umum digunakan untuk melakukan clustering seperti DBScan, K-Means, dan lainnya. K-Means merupakan salah satu metode clustering yang dikenal dengan jenis unsupervised learning, yang dimana digunakan pada data yang belum memiliki label (data tanpa kategori atau grup). Tujuan algoritma ini adalah menemukan kelompok dalam data, dengan jumlah kelompok direpresentasikan melalui variabel K [2].

Twitter juga dapat digunakan untuk mengetahui peran seseorang, seberapa penting dan cepatnya informasi dapat tersebar melalui tweet yang dituliskan oleh seseorang. Terdapat beberapa tokoh penting yang menjadi perhatian oleh publik Indonesia seperti Joko Widodo dan Prabowo. Baik Joko Widodo ataupun Prabowo keduanya menggunakan media sosial Twitter sebagai sarana berkomunikasi dengan masyarakat Indonesia. Kedua tokoh tersebut dikaitkan sebagai calon Presiden dalam pemilihan Presiden nantinya. Tak sedikit orang pun yang ikut memberikan komentar, opini mengenai pandangan mereka melalui media sosial Twitter.

II. METODE

A. Autentikasi

Pada saat menggunakan Twitter API untuk dapat digunakan dalam pengambilan data, diperoleh beberapa kode berupa consumer key, customer secret, access token, dan access key. Kode tersebut digunakan untuk proses inegrasi antar Twitter API dengan software PyCharm. Pada proses autentikasi, peneliti menggunakan library Tweepy dengan fungsi tweepy.AppAuthHandler dan tweepy.API

B. Pengambilan Data dari Twitter

Pengambilan data dilakukan secara real time dari Twitter dengan query = “Pilpres 2019”, tweets yang mengandung query tersebut menghasilkan sebesar 5054 tweets. Tweets tersebut merupakan tweets yang mengandung kata “Pilpres 2019” dengan terdapat beberapa tweets yang memiliki konten yang sama. Berikut adalah contoh data tweet yang didapatkan dari Twitter dengan query = “Pilpres 2019”.

Berikut adalah Tabel 1 salah satu contoh data *tweet* dengan *query* = “Pilpres 2019”.

Tabel 1. Contoh data *tweet* dengan *query* = “Pilpres 2019”

Nomor *Tweet*	**Teks *Tweet***
1	20 tahun reformasi di Tahun Politik 2018 dan menjelang pilpres 2019. \n\nAda beberapa orang ya mungkin bisa dianggap? https://t.co/pAH5HsCreo'
2	Waketum Gerindra Fadli Zon membidik PKB untuk diajak berkoalisi di Pilpres 2019. PKB belum mau dirayu Gerindra. Ini? https://t.co/8FCo5EhVdj'
	@semiaji_w: *Padahal waktu PKS dikabinet SBY...sering bikin reseh SBY?* https://t.co/lgtXj08n4P\n\nPaai Demokrat intens menjalin komunik?'

C. *Preparation* Data

1) *Case Folding*

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use

abbreviations in the title or heads unless they are unavoidable.

Pada proses *case folding*, peneliti menggunakan *library preprocessor* untuk membersihkan tweets dari karakter-karakter yang dianggap tidak valid, khususnya pada data Twitter, karakter yang dihapus adalah angka, URL, emoji, smiley, hashtag, mention. Penghapus karakter tersebut menggunakan fungsi:

```
datas = data[i]
makeitastring = ' '.join(map(str,
datas))
p.set_options(p.OPT.URL, p.OPT.EMOJI,
p.OPT.HASHTAG, p.OPT.NUMBER,
p.OPT.MENTION, p.OPT.RESERVED,
p.OPT.SMILEY)
cleaning = p.clean(makeitastring)
```

Adapun Tabel 2. menunjukkan contoh hasil proses *cleaning* dengan menggunakan *library preprocessor.*

Tabel 2. Contoh hasil proses *cleaning* pada *tweets*

Nomor *Tweet*	**Teks *Tweet***
1	tahun reformasi di Tahun Politik dan menjelang pilpres 2019.Ada beberapa orang ya mungkin bisa dianggap?
2	Waketum Gerindra Fadli Zon membidik PKB untuk diajak berkoalisi di Pilpres. PKB belum mau dirayu Gerindra. Ini?
3	semiaji_w: Padahal waktu PKS dikabinet SBY...sering bikin reseh SBY?Paai Demokrat intens menjalin komunik?

2) *Tokenizing*

Pada proses ini, peneliti menggunakan library *nltk* untuk pemotongan kalimat pada tweet berdasarkan tiap kata yang menyusunnya dengan menggunakan fungsi berikut:

```
from nltk.tokenize import word_tokenize
all_word =
word_tokenize(dataset.lower())
tokens = [w for w in all_word if w not
in symbols]
x = ' '.join(tokens)
print(x)
```

Berikut adalah Tabel 3. merupakan contoh hasil proses *tokenizing* pada beberapa *tweets:*

Tabel 3. Contoh hasil proses *Tokenizing* pada *tweets*

Nomor *Tweet*	**Hasil *Tokenizing***
1	tahun reformasi di tahun politik dan menjelang pilpres ada beberapa orang ya mungkin bisa dianggap
2	waketum gerinda fadli zon membidik pkb untuk diajak berkoalisi di pilpres pkb belum mau dirayu gerindra ini

3) *Stopword Removal*

Pada proses stopword removal, penghilangan kata-kata yang dianggap tidak penting atau tidak menggambarkan isi dari sebuah tweet. Proses ini menggunakan library Sastrawi untuk memproses stopword removal dalam bahasa Indonesia. Penghilangan kata-kata tersebut menggunakan fungsi.

```
from
Sastrawi.StopWordRemover.StopWordRemove
rFactory import StopWordRemoverFactory
```

```
factoryStopword =
StopWordRemoverFactory()
stopword = factoryStopword.
          create_stop_word_remover()
outputstopword = stopword.remove(x)
```

Adapun contoh hasil proses *Stopword Removal* pada beberapa *tweets* terlihat pada Tabel 4 di bawah ini:

Tabel 4. Contoh hasil proses *Stopword Removal* pada *tweets*

Nomor *Tweet*	Hasil *Tokenizing*
1	tahun reformasi tahun politik menjelang pilpres beberapa orang mungkin dianggap
2	waketum gerinda fadli zon membidik pkb diajak berkoalisi pilpres pkb mau dirayu gerindra

4) *Stemming*

Pada proses Stemming, yaitu merubah berbagai kata yang berimbuhan menjadi kata dasarnya. Peneliti juga menggunakan library Sastrawi yang menjadi library untuk teks bahasa Indonesia. Proses ini dilakukan dengan menggunakan fungsi:

```
from    Sastrawi.Stemmer.StemmerFactory
import StemmerFactory
factory = StemmerFactory()
stemmer = factory.create_stemmer()
output = stemmer.stem(outputstopword)
```

Di bawah ini adalah contoh hasil proses *Stemming* pada beberapa *tweets*:

Tabel 5. Contoh hasil proses *Stemming* pada *tweets*

Nomor *Tweet*	Hasil *Tokenizing*
1	tahun reformasi tahun politik jelang pilpres beberapa orang mungkin anggap
2	waketum gerinda fadli zon bidik pkb ajak koalisi pilpres pkb mau rayu gerindra

5) Pemberian Bobot pada Kata

Pada proses ini, peneliti menggunakan metode TF-IDF dengan memanfaatkan library sklearn. Adapun dalam proses metode tersebut menggunakan fungsi:

```
from    sklearn.feature_extraction.text
import TfidfVectorizer
vectorizer = TfidfVectorizer()
strg = dataset
response = vectorizer.transform([strg])
```

Hasil yang diperoleh adalah terdapat 2372 *term* yang masing-masing memiliki bobot yang telah didapatkan dari hasil perhitungan. Berikut adalah Tabel 6. merupakan contoh tf-idf *term weighting:*

Tabel 6. Contoh hasil proses *Stopword Removal* pada *tweets*

Feature	***Weights***
airlangga	0.0013209259976689886
ahok	0.002496752381081031
agusyudhoyono	0.0013209259976689886
ahy	0.050530689393760216
pilpres	0.4048682709896633
politik	0.1185595839144633
popularitas	0.00047026236826035195

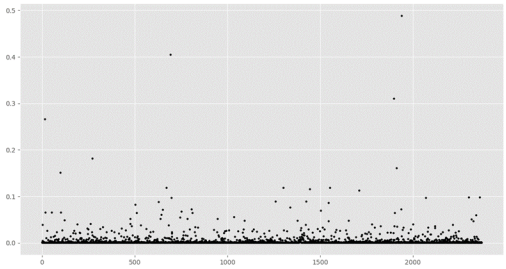
Adapun dari nilai tf-idf *term weighting,* diketahui *top-words* yang ada pada suatu dataset. *Top-words* merupakan kata-kata yang paling banyak muncul pada suatu dokumen *tweets*.Tabel 7. menunjukkan beberapa *top-words* beserta nilai masing-masing pada contoh 2 *tweets:*

Tabel 7. *Top Words*

Nomor	***Words***	***Value***
1	pilpres	2495.946928
2	Tahun	887.000000
3	Koalisi	542.000000
4	indonesia	517.000000
5	Politik	516.000000
6	Jokowi	497.984229
7	Gerindra	468.037620
…	…	…
…	…	…
…	…	…
18	Sandiaga	322.150010
…	…	..
…	…	..
…	…	..
23	Prabowo	274.959240

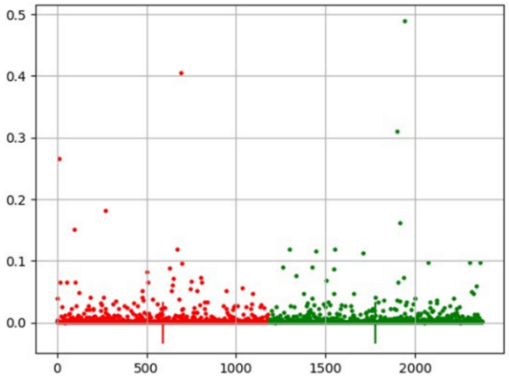
6) *Clustering* pada data

Pada proses *clustering*, peneliti menggunakan metode k-means. Adapun dataset yang digunakan pada proses ini adalah berupa x, dan y, dimana x merepresentasikan *term* dan y merepresentasikan nilai bobot tiap term menggunakan tf-idf. Berikut ini adalah hasil *plotting* nilai bobot tiap *term*:



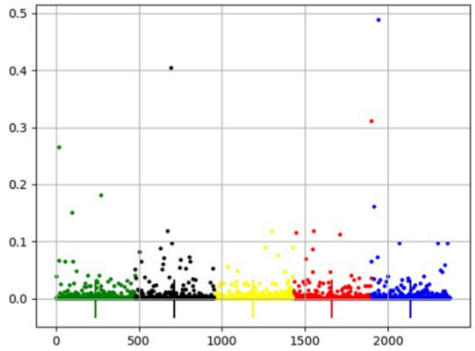
Gambar. 1. *Plotting* pada Penyebaran Data

Kemudian, selanjutnya memilih nilai K yaitu jumlah kluster yang dibutuhkan pada clustering. Pada awalnya nilai K yang dipilih adalah 2, dengan hasil *plotting* ditunjukkan pada Gambar 2.



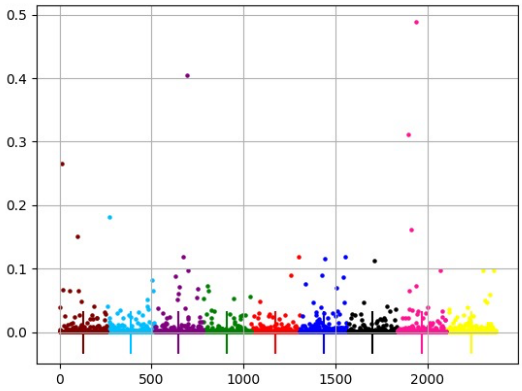
Gambar. 2. *Plotting clustering* dengan K = 2

Lalu, untuk nilai K = 5, juga dilakukan eksperimen dan menghasilkan hasil pada Gambar 3. di bawah ini:



Gambar.3. *Plotting clustering* dengan K = 5

Sedangkan, hasil *clustering* untuk nilai K = 9, seperti berikut ini terlihat pada Gambar 4.



Gambar 4. *Plotting clustering* dengan K=9

III. HASIL DAN PEMBAHASAN

Berdasarkan penelitian eksperimen yang telah dilakukan, seperti yang terlihat pada Gambar 2, nilai K dipilih yaitu 2, K = 2. Pada gambar tersebut merupakan hasil *plotting* dengan jumlah klaster adalah 2. Ketika nilai klaster tidak sama dengan 2, maka data-data tersebut juga akan membentuk kelompok sesuai dengan nilai K yang diatur. Salah satu metode yang digunakan untuk menguji kualitas klaster yang dihasilkan dari proses *clustering* adalah *Silhoutte Coefficients.*

Adapun hasil yang diperoleh terlihat pada Tabel 8. Hasil Perhitungan *Silhoutte Coefficients* Untuk K = 2.

Tabel 8. Hasil Perhitungan *Silhoutte Coefficients* Untuk K = 2.

No	Nomor Klaster	Nilai *Silhoutte*
1	1 dan 2	0.500421

Jika nilai K = 5, maka hasil menunjukkan perbedaan dapat dilihat pada Tabel 9. Hasil Perhitungan *Silhoutte Coefficients* Untuk K = 5.

Tabel 9. Hasil Perhitungan *Silhoutte Coefficients* Untuk K = 5.

No	Nomor Klaster	Nilai *Silhoutte*
1	1,2,3,4, dan 5	0.50105

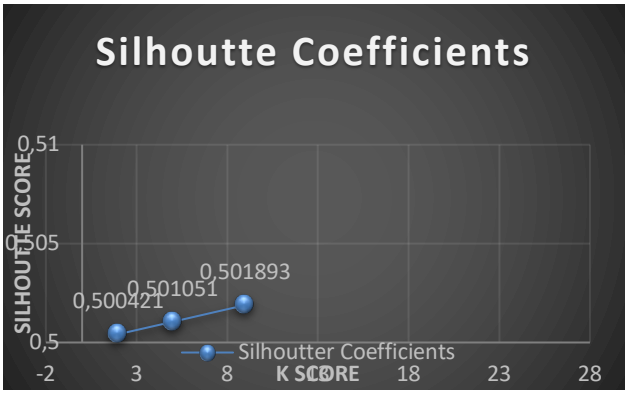
Sedangkan, untuk nilai K = 9, hasil yang diperoleh ditunjukkan pada Tabel 10. Hasil Perhitungan *Silhoutte Coefficients* Untuk K = 9. sebagai berikut:

Tabel 10. Hasil Perhitungan *Silhoutte Coefficients* Untuk K = 2.

No	Nomor Klaster	Nilai *Silhoutte*
1	1,2,3,4,5,6,7,8, dan 9	0.50189

Dari ketiga percobaan nilai K pada *clustering* k-means, diperoleh informasi bahwa nilai K pada algoritma k-means mempengaruhi kualitas klaster berdasarkan nilai *silhouette*. Pada nilai K = 2, nilai kualitas klaster sebesar 0.5004. Tetapi jika dibandingkan dengan nilai K=9, kualitas klaster menjadi lebih baik yaitu sebesar 0.50189. Berbeda dengan K sebelumnya, jika K bernilai 5 diperoleh nilai kualitas klaster sebesar 0.50105. Oleh sebab itu nilai K pada algoritma k-means klastering sebaiknya mengacu pada jumlah dataset yang dimiliki sehingga persebaran data dapat diperoleh lebih baik.

Untuk melihat perbandingan nilai *Silhoutte Coefficients,* berikut adalah Gambar 5. Grafik *Silhoutte Score* berdasarkan nilai K:



Gambar 5. Grafik Perbandingan Nilai *Silhoutte Coefficients*

Terdapat beberapa faktor yang membuat hasil *clustering* dengan metode k-means pada data *tweet* masih belum optimal, diantaranya adalah sebagai berikut:

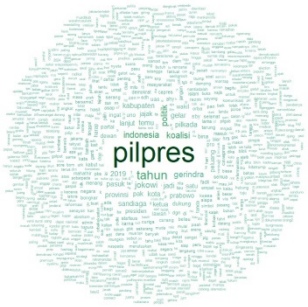
1) Pada hasil *pre-processing*, masih terdapat beberapa *tweet* yang masih belum bersih dan mengandung kata-kata yang berupa singkatan. Contohnya: kata “yang” menjadi “yg”, kata “tidak” menjadi “tdk, kata “datang” menjadi “dtg” dan lain sebagainya. Penggunaan kata singkatan memberikan pengaruh terhadap proses stopword removal, dimana kata yang seharunya dihilangkan menjadi tidak terdeteksi sehingga tetap ada pada kalimat *tweet*.

2) Masih terdapatnya angka-angka yang seharusnya hilang pada tweet tapi tidak mampu diproses oleh proses cleaning yaitu case folding.
3) Terdapat kata-kata yang harusnya terpisah menjadi beberapa kata tetapi masih menjadi suatu kata, seperti "nantiadalah" yang seharusnya menjadi "nanti" dan "adalah" dan lain sebagainya.

Faktor-faktor diatas sangat berpengaruh terhadap hasil dari proses *clustering*. Hal ini dikarenakan setiap kata dari sebuah tweet menjadi variabel-variabel/ *term* yang digunakan pada proses perhitungan jarak antar tiap *tweet*, sehingga hasil *clustering* dapat menempatkan *tweet* pada klaster yang tidak tepat.

D. Analisis Pengaruh Jenis Konten

Gambar 5. *Word Cloud,* merupakan daftar kata-kata yang muncul pada data *tweet*. Berdasarkan kata-kata tersebut, dapat diketahui top words, yaitu kata-kata yang paling banyak muncul pada tweet diantaranya pilpres, Indonesia, politik, tahun, gerindra, dan lainnya. Berikut adalah *word cloud* yang dihasilkan pada sekumpulan data tweet.



Gambar. 5. *Word Cloud* Pada Dataset

Pada Gambar 5, diketahui bahwa kata "pilpres" menjadi *top-words* pertama yang memiliki nilai paling besar, diikuti dengan kata "tahun" kemudian "indonesia" dan lain sebagainya. Kata "pilpres" menjadi top-words pertama dikarenakan pada saat pengambilan data tweet, salah satu yang harus dipenuhi adalah hanya tweet-tweet yang mengandung kata "pilpres" lah yang diambil sebagai dataset. Selain itu, berdasarkan Tabel 7, kata "Jokowi" menjadi kata yang lebih sering muncul pada *tweet* dibandingkan dengan kata "Prabowo". Bahkan nama "Prabowo" memiliki nilai lebih rendah jika dibandingkan dengan nama "Sandiaga". Oleh karena itu, dapat dilihat bahwa nama "Jokowi" merupakan nama tokoh yang lebih popular dan memiliki pengaruh yang lebih tinggi di kalangan publik daripada nama "Prabowo".

IV. KESIMPULAN

Berdasarkan hasil implementasi yang telah dilakukan, dapat disimpulkan bahwa:
1) Penerapan metode clustering pada data tweet dengan query = "Pilpres 2019" menghasilkan perbedaan klaster dengan diberikan nilai K = 2, K = 5, dan K = 9.
2) Berdasarkan nilai silhouette coefficients dengan masing masing nilai K, bahwa nilai K pada klastering memiliki pengaruh terhadap kualitas klaster yang dihasilkan.
3) Menentukan nilai K sebaiknya menyesuaikan dengan jenis data yang dimiliki, sehingga persebaran data dapat terlihat ketika proses klastering.
4) Nilai K terbaik pada dataset twitter ini, adalah 9 dengan nilai kualitas silhouette coefficients sebesar 0.50189.

REFERENSI

[1] D. Anggoro and A. W. Hidayat, "Rancang Bangun Sistem Informasi Perpustakaan Sekolah Berbasis Web Guna Meningkatkan Efektivitas Layanan Pustakawan," Edumatic J. Pendidik. Inform., 2020, doi: 10.29408/edumatic.v4i1.2130.

[2] J. Fernando, I. R. I. Astutik, and H. Setiawan, "RANCANG BANGUN APLIKASI SISTEM INFORMASI CARWASH &Amp; AUTOCARE BERBASIS ANDROID PADA MAHKOTA MOTOR JAYA MAKMUR," J. Teknoinfo, 2023, doi: 10.33365/jti.v17i1.2294.

[3] M. F. Syawalludin and M. E. A. Rivan, "Sistem Informasi Perpustakaan Berbasis Website Di Sekolah DasarNegeri 240 Palembang," MDP-Sc, 2023, doi: 10.35957/mdp-sc.v2i1.4476.

[4] S. DwiDara, A. Y. Rindarwati, R. N. Fadillah, and Y. Iskandar, "Evaluasi Sistem Penyimpanan Obat Berdasarkan Standar Pelayanan Kefarmasian Pada Salah Satu Apotek Di Kota Bandung," J. Pharm. Sci., 2023, doi: 10.36490/journal-jps.com.v6i1.67.

[5] E. S. Dasopang, A. Utami, F. Hasanah, D. N. Siahaan, and N. S. Harefa, "Profil Penyimpanan Obat LASA (Look Alike Sound Alike) Pada Beberapa Apotek Di Kota Medan," Jfionline | Print Issn 1412-1107 | E-Issn 2355-696x, 2022, doi: 10.35617/jfionline.v14i2.97.

[6] R. E. G. Rahayu and P. Marup, "Rancang Bangun Sistem Informasi Pelayanan Administrasi Publik Terpadu Berbasis Web," J. Algoritm., 2021, doi: 10.33364/algoritma/v.18-1.826.

[7] Y. Septiana, W. Baswardono, and R. E. N. Awaludin, "Rancang Bangun Sistem Informasi Administrasi Klinik Berbasis Website Menggunakan Metode Extreme Programming," J. Algoritm., 2022, doi: 10.33364/algoritma/v.19-2.1151.

[8] S. Kosasi, "Pembuatan Sistem Informasi Geografis Berbasis Web Untuk Persebaran Lokasi Apotek," Csrid (Computer Sci. Res. Its Dev. Journal), 2016, doi: 10.22303/csrid.8.2.2016.99-108.

[9] I. Soraya, W. R. Adawiyah, and E. Sutrisna, "Pengujian Model Hot Fit Pada Sistem Informasi Manajemen Obat Di Instalasi Farmasi RSGMP Unsoed Purwokerto," J. Ekon. Bisnis Dan Akunt., 2019, doi: 10.32424/jeba.v21i1.1261.

[10] A. S. Wulandari and N. Ahmad, "Hubungan Faktor Sosiodemografi Terhadap Tingkat Pengetahuan Swamedikasi Di Beberapa Apotek Wilayah Purworejo," Inpharnmed J. (Indonesian Pharm. Nat. Med. Journal), 2021, doi: 10.21927/inpharnmed.v4i1.1764.