SITEKIN: Jurnal Sains, Teknologi dan Industri, Vol. 20, No. 1, Desember 2022, pp.116 - 121
ISSN 2407-0939 print/ISSN 2721-2041 online

Optimalisasi *K-Means* Dalam Pengelompokan Ancaman Insiden Aplikasi Yang Dilaporkan Melalui *Service Desk* TIK

Rimba Prasasti[1], Rifki Sadikin[2], Eni Heni Hermallani[3]
[1-3]Program Studi Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Nusa Mandiri
Jalan Kramat Raya No. 18, Senen, Jakarta Pusat
Email: masr1mba@gmail.com[1], rifki.rdq@ nusamandiri.ac.id[2], enie_h@nusamandiri.ac.id[3]

ABSTRAK

Layanan click, call, counter (3C) merupakan bentuk transformasi layanan digital Perpajakan. Insiden layanan 3C yang terjadi ini dilaporkan melalui Service Desk TIK. Banyaknya laporan insiden membuat kendala dalam penanganan penyelesaian permasalahan. Dengan menggunakan K-Means secara unsupervised learning untuk pengelompokan ancaman insiden diharapkan dapat membantu penyelesaian lebih efektif. Optimalisasi untuk meningkatkan nilai akurasi yang lebih baik dicari menggunakan word embedded dengan algortima Elkan dan algortima Lloyd pada K-Means. Hasil optimal didapatkan pada jumlah kluster 4 yang dievaluasi menggunakan Silhouette Score, Calinski Harabasz dan Davies-Bouldin Index. Hasil optimal dari penerapan model pada algoritma K-Means dan parameter algoritma Elkan dengan word embedding CountVectorizer didapatkan sebesar 71,94% pengelompokan yang sesuai.

Kata Kunci: *unsupervised learning, k-means, word embedded,* insiden, *service desk*

ABSTRACT

Click, call, counter (3C) service is a form of digital tax service transformation. This 3C service incident is reported through the ICT Service Desk. The number of incident reports creates obstacles in handling problem solving. By using K-Means in unsupervised learning for incident threat clustering, it is hoped that it can help solve more effectively. Optimization to increase the value of better accuracy is sought using word embedded with Elkan's algorithm and Lloyd's algorithm on K-Means. Optimal results were obtained in the number of clusters 4 which were evaluated using Silhouette Score, Calinski Harabasz and Davies-Bouldin Index methods. The optimal results from the application of the model to the K-Means algorithm and the Elkan algorithm parameters with word embedding CountVectorizer obtained 71.94% the appropriate classification.

Keywords: *unsupervised learning, k-means, word embedded, incident, service desk*

Pendahuluan

Layanan *click*, *call*, *counter* (3C) merupakan bentuk transformasi layanan digital oleh Direktorat Jenderal Pajak (DJP). Layanan *click* adalah kegiatan pelayanan perpajakan yang diberikan melalui situs *web*, aplikasi *desktop*, atau layanan lainnya tanpa melalui bantuan petugas pajak. Sedangkan layanan *call* adalah layanan perpajakan yang diberikan oleh pusat kontak (*contact center*), dan layanan *counter* adalah layanan perpajakan yang dilakukan secara manual melalui Kantor Pelayanan Pajak [1]. Salah satu jenis layanan 3C yang diterapkan DJP yaitu pendaftaran wajib pajak yang dapat dilayani melalui aplikasi *e-Registration*. Aplikasi ini, selain digunakan oleh wajib pajak atau calon wajib pajak untuk melakukan pendaftaran atau perubahan data wajib pajak, juga digunakan oleh petugas pajak untuk memproses pendaftaran atau perubahan data wajib pajak [2].

Pengaruh yang positif memberikan peningkatan kualitas layanan yang lebih baik kepada kepatuhan wajib pajak dengan indikator pelayanan seperti *Reliability*, *Responsiveness*, *Competence*, *Access*, *Communication*, *Credibility*, *Security*, dan *Understanding* [3]. Dalam rangka menjaga kelangsungan layanan 3C sesuai indikator pelayanan tersebut, DJP membentuk *service desk* TIK dalam penanganan insiden aplikasi melalui aplikasi LasisOnline [4]. Harapan penanganan insiden dengan baik yakni memastikan penyelesaian sesuai dengan waktu yang diharapkan, untuk meminimalisir insiden yang sama berulang terjadi dan diharapkan layanan elektronik dapat berfungsi dengan normal kembali [5].

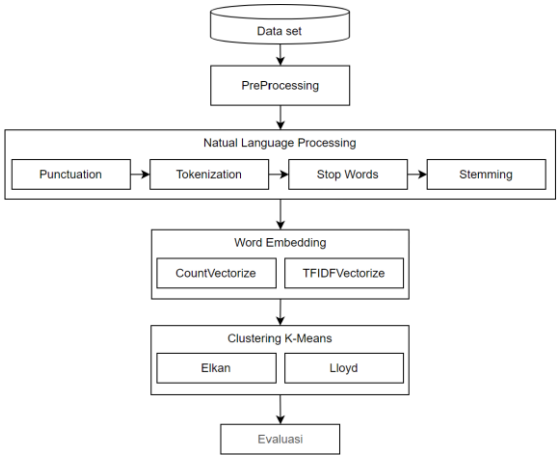
Beragam dan banyaknya laporan insiden yang disampaikan melalui *service desk* TIK menjadi kendala dalam penanganan insiden dan perbaikan kesalahan aplikasi di DJP. Kekurangan jumlah personil yang terlibat juga menjadi salah satu permasalahan penanganan insiden. Pengelompokan insiden diharapkan dapat membantu penyelesaian penanganan insiden.

Pada penelitian sebelumnya, perbandingan penanganan insiden berdasarkan kerangka kerja ITIL dengan penanganan insiden yang terjadi di *Service Desk* dan menghasilkan rekomendasi perbaikan *Service Desk*. Penelitian lainnya adalah penelitian model klasifikasi kategori pada insiden dengan menggunakan *machine learning* seperti pengukuran performa algoritma *Random Forest*, *Support Vector Machine*, *Multilayer Perceptron*, *Recurrent Neural Network*, *Long Short-Term Memory* (LSTM), *Gated Recurrent Unit* (GRU) [6], dan penelitian prediksi resolusi menggunakan algoritma *Multinomial Naive Bayes, SVM, KNN, Decision Tree and Logistic Regression* [7] yang menghasilkan akurasi tiap algoritma berbeda berdasar pada dataset dan pengaturan parameter. Sedangkan penelitian sebelumnya pada aplikasi *e-Registration* yakni mengenai pengaruh aplikasi kepada kepatuhan wajib pajak [8].

Penelitian ini mengelompokan permasalahan pada penggunaan aplikasi *e-Registration* yang dilaporkan melalui *Service Desk* TIK dengan menggunakan algoritma *K-Means* dengan menggunakan *word embedding* untuk mencari ancaman insiden dan mengevaluasi menggunakan beberapa metode seperti *Silhouette Score*, *Calinski Harabasz* dan *Davies-Bouldin Index.*

Metode Penelitian

Metodologi pada penelitian ini menggunakan model *Machine Learning* algoritma *Natural Language Processing* (NLP) untuk memproses data insiden dan algoritma *clustering* untuk mengidentifikasi dan *clustering* data insiden. Algoritma *Natural Language Processing* yang digunakan adalah *Punctuations*, *Tokenize*, *Stopword*, *Stemming* dan *Word Embedding*. Sedangkan algoritma *clustering* yang digunakan adalah *K-Means* dengan algoritma *Lloyd* dan *Elkan*. Melalui metodologi tersebut penulis menggunakan Google Colab dengan Python versi 3.7.12 untuk memperoleh hasilnya. Sebagai *backend* menggunakan *library* numpy, re, pandas, nltk, sastrawi dan scikit-learn. Pada penelitian ini akan dilakukan beberapa tahapan untuk mendapatkan model paling baik sebagaimana Gambar 1.



Gambar 1. Alur proses penelitian

Dataset yang digunakan untuk penelitian ini adalah data insiden aplikasi aplikasi *e-Registration* yang berasal dari penanganan insiden aplikasi yang dilaporkan melalui aplikasi LasisOnline. Insiden yang dilaporkan pada tahun 2021 sampai dengan Maret 2022. Aplikasi *e-Registration* memiliki 15.872 insiden.Sebaran dataset insiden untuk feature judul terlihat pada gambar 4.2 dengan menggunakan *word cloud*. Kalimat yang mengandung "NPWP", "NIK" dan "EFIN" menjadi kata yang sering muncul pada setiap insiden.

Gambar 2. Sebaran dataset insiden yang dilaporkan

Preprocessing ini untuk mengetahui adanya *missing value*, data yang tidak standar, dan *imbalance* data [9]. Semua data insiden aplikasi yang dilaporkan kemudian dipilih dataset sesuai dengan insiden aplikasi yang akan diteliti. Pemilihan dataset insiden aplikasi *e-Registration* didasarkan pada beberapa kriteria yakni data insiden sesuai kategori dan data insiden yang telah selesai ditindaklajuti. Data insiden sesuai kategori ini dipilih karena masih terdapat kesalahan pelapor dalam memilih katagori insiden, sedangkan data insiden telah ditindaklajuti dipilih karena untuk memastikan insiden telah terjadi.

Selanjutnya proses *Natural Language Processing* untuk melakukan pekerjaan ini secara efektif dan dengan akurasi, seperti yang dilakukan manusia [10]. Proses yang dilakukan meliputi *punctuation*, *tokenization*, *stop words*, *stemming*. *Punctuation* pada penelitian ini untuk menghilangkan karakter html [11], karakter special dan tanda baca,. *Tokenization* untuk memisahkan teks sebuah kalimat menjadi beberapa kata atau token [12]. *Stop words* untuk mengeluarkan kata yang sering muncul dan umum digunakan pada bahasa sehari-hari [13]. *Stemming* yang dipakai untuk mengubah bentuk kara yang berbeda menjadi bentuk dasar [14]. Proses ini menggunakan *library* sastrawi untuk mengatasi masalah perubahan kata menjadi kata dasar [15].

Proses *Word Embedding* digunakan untuk merepresentasi vektor kata yang bernilai dengan menanamkan makna semantik dan sintaksis yang diperoleh dari korpus besar yang tidak berlabel [16]. Proses yang dijalankan untuk memperoleh hasil terbaik adalah dengan menggunakan model *Bag of Word* (BOW) dan model *Term frequency-inverse document frequency* (TFIDF). Model BOW terdiri dari kata unik yang telah digunakan pada sebuah kalimat [17]. Sedangkan TFIDF digunakan untuk menilai secara numerik relevansi kata dalam kalimat dan frekuensi skor yang diberikan ini menentukan pentingnya kata dalam kalimat berdasarkan frekuensi kata [18].

Kemudian dilakukan pemodelan *clustering* K-Means dengan algoritma Elkan dan algoritma Lloyd yang dipadukan dengan dengan hasil proses *Word Embedded* sebelumnya. *Clustering* K-Means merupakan salah satu analisis data untuk mendapatkan dan memahami struktur data dengan cara mengidentifikasi kelompok data yang homogen hingga mendapatkan titik data di setiap kelompok berdasarkan ukuran kesamaan seperti jarak berbasis Euclidean atau jarak berbasis korelasi [19].

Algoritma Lloyd dimulai dengan k *centroid* yang diberikan sebagai titik untuk kluster k, dan mengulanginya hingga konvergensi [20]. Sedangkan algoritma Elkan ini dapat mempercepat kinerja K-Means dengan menghindari perhitungan jarak yang berlebihan [21].

Percobaan kombinasi tersebut menghasilkan nilai akurasi yang dievaluasi pada masing-masing pemodelan menggunakan *silhouette_score*, *calinski_harabasz_score* dan *davies_bouldin_score.* Metode *silhouette_score* merupakan ukuran kualitas kluster, nilai tertinggi menunjukkan jumlah cluster yang optimal [22]. Nilai indeks *davis-bouldin* bertujuan untuk mengukur kualitas pengelompokan dengan mengukur kesamaan intracluster dan perbedaan antar-kluster [23]. Sedangkan nilai *calinski_harabasz* untuk mengevaluasi dampak pengelompokan kerapatan kluster dan kerapatan antara kluster [24].

Hasil dan Pembahasan

Dataset yang digunakan untuk penelitian setelah melalui proses Preprocessing didapatkan informasi berjumlah 15.872 data. Pada tabel 1 adalah contoh dataset yang akan digunakan, namun atas dasar kerahasian dan perlindungan data wajib pajak maka judul dan isi yang mengandung Nama, NIK dan NPWP akan ditutup atau disamarkan.

Tabel 1. Hasil dataset setelah dilakukan *prepocessing*

Data set	Judul	Isi
1	update NIK	mohon bantuan teman2 lasis agar mengupdate NIK pada NPWP : 07.XXX.XXX.0-XXX.000 An. AXXXD RXXXXXI dengan NIK : 35XXXX10XXXX0018 / WP tsb ganda dengan NPWP 31.XXX.XXX.1-XXX.000 status NE (WP PEN PENERBITAN JABATAN) sehingga NPWP : 07.XXX.XXX.0-XXX.000 tidak bisa kami update NIK karena nabrak dgn NPWP satunya, terima kasih.
2	Tidak bisa tindak lanjut Pengaktifan WP NE	selamat siang, WP mengajukan permohonan Pengaktifan WP NE di KPP tetapi pada saat tindak lanjut muncul notifikasi seperti ini dan tidak ada BPS yang belum diselesaikan selain permohonan tersebut. mohon bantuanny, terima kasih.
3	Ada permohonan yang nyantol padahal sudah selesai	Selamat Pagi, terdapat permohonan yang nyantol di menu ereg aktiasi akun pkp , padahal sudah selesai hingga aktivasi akun pkp di menu e-nofa , untuk permohonan S-3XXXX/WPJ.XX/KP.XXXX/2020 HXXXX PXXXX SEXXXXS , bukti terlampir

Untuk menghasilkan dataset yang akan diproses *clustering* dilakukan tahapan *Natural Language Processing.* Tahapan yang dijalankan yakni tahapan *punctuation*, tahapan *tokenization*, tahapan *stop words*, dan tahapan *stemming*. Pada tabel 2 merupakan hasil dari tahapan *Natural Language Processing.*

Tabel 2. Hasil *natual language prepocessing*

Data set	*Natural language processing*
1	'nik', 'nik', 'axxxd', 'rxxxxxi', 'nik', 'tsb', 'ganda', 'ne', 'jabat', 'nik', 'nabrak', 'satu'
2	'aktif', 'ne', 'aktif', 'ne', 'bps', 'bantuanny'
3	'aktiasi', 'akun', 'pkp', 'aktivasi', 'akun', 'pkp', 'nofa', 'permohoanan', 'hxxxx', 'pxxxx', 'sexxxxs', 'bukti'

Dengan menggunakan *library* dari scikit-learn dan modul *CountVectorizer*, dijalankan pada hasil *Natural Language Processing* sesuai tabel 2 yang digabung menjadi kesatuan kalimat dan menghasilkan data hasil model BOW sebagaimana pada tabel 3 di bawah ini.

Tabel 3. Model bow

Data set	*Natural language processing*	Kata yang muncul			
		nik	ne	bps	pkp
1	nik nik axxxxd rxxxxxi nik tsb ganda ne jabat nik nabrak satu	4	1	0	0
2	aktif ne aktif ne bps bantuanny	0	2	1	0
3	aktiasi akun pkp aktivasi akun pkp nofa permohonan hxxxxa pxxxx sexxxxs bukti	0	0	0	2

Sedangkan menggunakan library dari scikit-learn dan modul *TfidfVectorizer*, dijalankan pada hasil stemming sesuai tabel 2 yang digabung menjadi kesatuan kalimat dan menghasilkan data model TFIDF sebagaimana pada tabel 4.

Tabel 4. Model tfidf

Data set	*Natural language processing*	Kata yang muncul			
		nik	ne	bps	pkp
1	nik nik axxxxd rxxxxxi nik tsb ganda ne jabat nik nabrak satu	0.354	0.237	0	0
2	aktif ne aktif ne bps bantuanny	0	0,579	0,177	0
3	aktiasi akun pkp aktivasi akun pkp nofa permohonan hxxxxa pxxxx sexxxxs bukti	0	0	0	0.289

Hasil *word embedded* yang dihasilkan seperti pada tabel 3 dan tabel 4 dilakukan masing-masing *clustering* K-Means dengan algoritma Elkan dan algortima Lloyd. Untuk mencari jumlah kluster yang terbaik, dilakukan evaluasi menggunakan beberapa metode seperti *Silhouette Score*, *Calinski Harabasz* dan *Davies-Bouldin Index.*

Pada tabel 5 merupakan hasil dari evaluasi yang telah dilakukan terhadap jumlah kluster antara tiga sampai dengan delapan kluster. Jumlah kluster optimal pada *clustering* K-Means tersebut adalah empat kluster dengan menggunakan algoritma Elkan dan model BOW untuk *word embedded.* Nilai yang dihasilkan untuk *Silhouette Score* adalah 0,14953 dan nilai *Calinski Harabasz Score* adalah 1390,09225 serta nilai *Davies Bouldin Score* yaitu 2,12115.

Tabel 5. Hasil evaluasi kluster

Algoritma	Word embedding	Jumlah kluster	Silhouette score	Calinski harabasz score	Davies bouldin score
Elkan	CV	3	0.13934	1673.24613	2.44271
Elkan	CV	4	0.14953	1390.09225	2.12115
Elkan	CV	5	0.12331	1145.17331	2.53231
Elkan	CV	6	0.11994	1001.91673	2.50384
Elkan	CV	7	0.10465	920.79755	2.70819
Elkan	CV	8	0.04207	781.82811	2.74592
Elkan	Tfidf	3	0.03257	411.54424	5.17636
Elkan	Tfidf	4	0.03587	346.36241	4.82911
Elkan	Tfidf	5	0.03867	299.75545	4.86976
Elkan	Tfidf	6	0.02592	252.30801	4.91933
Elkan	Tfidf	7	0.02839	236.43307	4.97737
Elkan	Tfidf	8	0.03561	208.88629	5.37977
Lloyd	CV	3	0.13933	1673.2461	2.44270
Lloyd	CV	4	0.14901	1388.7819	2.16460
Lloyd	CV	5	0.09497	1164.7541	2.45293
Lloyd	CV	6	0.04961	1029.8699	2.87535
Lloyd	CV	7	0.03877	892.60812	2.66921
Lloyd	CV	8	0.05302	806.92330	3.10394
Lloyd	Tfidf	3	0.03255	411.54550	5.17549
Lloyd	Tfidf	4	0.03594	346.36383	4.83054
Lloyd	Tfidf	5	0.03877	287.98634	4.51729
Lloyd	Tfidf	6	0.03354	263.68657	4.87269
Lloyd	Tfidf	7	0.03499	228.66161	4.73176
Lloyd	Tfidf	8	0.03253	219.61552	4.73272

Dengan model optimal tersebut dilakukan pencarian titik pusat *cluster* dan jumlah sebaran data pada tiap *cluster* seperti pada tabel 6. Titik pusat nik mempunyai jumlah data paling banyak sejumlah 7371 data dibandingkan dengan titik pusat pkp dengan jumlah 572 data.

Tabel 6. Titik pusat dan sebaran data

Titik pusat	Jumlah Data
pkp	572
efin	1168
nik	7371
bps	6761

Setelah mendapatkan model optimal berdasarkan hasil evaluasi performa beberapa model penelitian dan diketahui masing-masing titik pusat *cluster*. Kemudian dilakukan penerapan model untuk dataset guna mengetahui kelompok ancaman insiden. Sebagaimana hasil pengelompokan ancaman insiden aplikasi disajikan pada tabel 7. Hasil pengelompokan dilakukan konfirmasi kepada petugas *service desk* TIK yang menangani dan menyelesaikan laporan insiden. Hasil tersebut seperti pada tabel 8.

Tabel 7. Hasil evaluasi kluster

Judul	Isi	Ancaman Insiden
update NIK	mohon bantuan teman2 lasis agar mengupdate NIK pada NPWP : 07.XXX.XXX.0-XXX.000 An. AXXXD RXXXXXI dengan NIK : 35XXXX10XXXX0018 / WP tsb ganda dengan NPWP 31.XXX.XXX.1-XXX.000 status NE (WP PEN PENERBITAN JABATAN) sehingga NPWP : 07.XXX.XXX.0-XXX.000 tidak bisa kami update NIK karena nabrak dgn NPWP satunya, terima kasih.	nik
Tidak bisa tindak lanjut Pengaktifan WP NE	selamat siang, WP mengajukan permohonan Pengaktifan WP NE di KPP tetapi pada saat tindak lanjut muncul notifikasi seperti ini dan tidak ada BPS yang belum diselesaikan selain permohonan tersebut. mohon bantuanny, terima kasih.	bps

Tabel 8. Hasil konfirmasi pengelompokan

Ancaman Insiden	Jumlah	Sesuai	Tidak Sesuai	Persentase
pkp	572	572	0	100%
efin	1.168	1.168	0	100%
nik	7.371	7.371	0	100%
bps	6.761	2.308	4.453	34,14%
Jumlah	15.872	11.419	4.453	71,94%

Kesimpulan

Penggunaan algoritma *clustering* seperti K-Means dapat membantu untuk pengelompokan data *unsupervised* dalam mengidentifakasi ancaman insiden aplikasi yang dilaporkan melalui *service desk* TIK.

Prepocessing data membantu untuk memilih dataset yang sesuai dengan kategori permasalahan. Karena pemilihan kategori oleh pengguna masih ada yang belum sesuai dengan kategori permasalahan. Penggunaan *natural language processing* perlu dilakukan untuk mengoptimalkan klasifikasi unsupervised learning.

Penerapan model pengelompokan menggunakan K-Means menghasilkan nilai sebesar 71,94% sesuai dengan keadaan yang sebenarnya.

Daftar Pustaka

[1] Y. I. Santoso, "Ditjen Pajak optimalisasi layanan digital, demi kejar pendapatan di tengah pandemi," 2020. https://newssetup.kontan.co.id/news/ditjen-pajak-optimalisasi-layanan-digital-demi-kejar-pendapatan-di-tengah-pandemi (accessed Nov. 02, 2021).

[2] K. P. Kinanti and D. Pratomo, "Pengaruh Penerapan Pendaftaran Npwp Secara Online (E- Registration), E-Billing Dan E-Filing Terhadap Kepatuhan Wajib (Survei pada Wajib Pajak Orang Pribadi Non Karyawan di KPP Pratama Depok Cimanggis Tahun 2019)," *e-Proceeding Manag.*, vol. 8, no. 6, pp. 1–8, 2021.

[3] M. Zuraeva and N. Rulandari, "Analisis Kualitas Pelayanan Perpajakan dalam Rangka Meningkatkan Kepatuhan Wajib Pajak," *J. Pajak Vokasi*, vol. 2, no. 1, pp. 37–44, 2020.

[4] Direktorat Jenderal Pajak, "SURAT EDARAN DIREKTUR JENDERAL PAJAK NOMOR SE - 37/PJ/2013," no. Agustus, 2013.

[5] D. Safitri and S. P. Silalahi, "Pengaruh Kualitas Pelayanan Fiskus, Pemahaman Peraturan Perpajakan Dan Penerapan Sistem E-Filling Terhadap Kepatuhan Wajib Pajak: Sosialisasi Perpajakan Sebagai Pemoderasi," *J. Akunt. dan Pajak*, vol. 20, no. 2, 2020, doi: 10.29040/jap.v20i2.688.

[6] M. A. Prihandono, R. Harwahyu, and R. F. Sari, "Performance of machine learning algorithms for IT incident management," *2020 11th Int. Conf. Aware. Sci. Technol. iCAST 2020*, pp. 2–7, 2020, doi: 10.1109/iCAST51195.2020.9319487.

[7] R. R. COSTA, Jorge, Rubén PEREIRA, "ITSM

Automation - Using Machine Learning to Predict Incident Resolution Category," no. 351, 2021.

[8] F. Alamri and A. Widyatama, "TAM Sebagai Solusi Atas Minat Penggunaan Layanan E-Registration Wajib Pajak," vol. 10, no. 2, pp. 89–99, 2019.

[9] J. Luengo, D. García-Gil, S. Ramírez-Gallego, S. García, and F. Herrera, *Big Data Preprocessing*. 2020. doi: 10.1007/978-3-030-39105-8.

[10] K. R. Chowdhary, *Fundamentals of Artificial Intelligence*. 2020. [Online]. Available: https://doi.org/10.1007/978-81-322-3972-7_19

[11] Gong, Nan, Chunxiao Fan, Yuexin Wu, Yue Ming, "A Web Content Extraction Method Base on Punctuation Distribution and HTML Tag Similarity," *Proc. 3rd Int. Conf. Logist. Informatics Serv. Sci.*, 2013.

[12] G. N. R Prasad Sr Asst professor, "Identification of Bloom's Taxonomy level for the given Question paper using NLP Tokenization technique," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 13, pp. 1872–1875, 2021.

[13] D. Na and C. Xu, "Automatically generation and evaluation of stop words list for Chinese patents," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 13, no. 4, pp. 1414–1421, 2015, doi: 10.12928/TELKOMNIKA.v13i4.2389.

[14] A. Jabbar, S. Iqbal, M. I. Tamimy, S. Hussain, and A. Akhunzada, "Empirical evaluation and study of text stemming algorithms," *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5559–5588, 2020, doi: 10.1007/s10462-020-09828-3.

[15] K. K. Purnamasari and I. S. Suwardi, "Rule-based Part of Speech Tagger for Indonesian Language," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 407, no. 1, 2018, doi: 10.1088/1757-899X/407/1/012151.

[16] W. bin, angela wang, fenxiao chen, yuncheng wang and c.-c. jay kuo, "Evaluating word embedding models : methods and experimental results," vol. 8, 2019, doi: 10.1017/ATSIP.2019.12.

[17] M. Hamisu and A. Mansour, "Detecting Advance Fee Fraud Using NLP Bag of Word Model," pp. 94–97, 2020.

[18] A. Addiga and S. Bagui, "Sentiment Analysis on Twitter Data Using Term Frequency-Inverse Document Frequency," pp. 117–128, 2022, doi: 10.4236/jcc.2022.108008.

[19] D. K. Hashim and L. A. N. Muhammed, "Performance of K-means algorithm based an ensemble learning," *Bull. Electr. Eng. Informatics*, vol. 11, no. 1, pp. 575–580, 2022, doi: 10.11591/eei.v11i1.3550.

[20] K. Aoyama, K. Saito, and T. Ikeda, "Accelerating a Lloyd-Type k-Means Clustering Algorithm with Summable Lower Bounds in a Lower-Dimensional Space," no. 11, pp. 2773–2783, 2018.

[21] C. Elkan, "Using the Triangle Inequality to Accelerate k-Means," *Proceedings, Twent. Int. Conf. Mach. Learn.*, vol. 1, pp. 147–153, 2003.

[22] K. R. Shahapure and C. Nicholas, "Cluster Quality Analysis Using Silhouette Score," pp. 2020–2021, 2020, doi: 10.1109/DSAA49011.2020.00096.

[23] A. K. Singh, S. Mittal, P. Malhotra, and Y. V. Srivastava, "Clustering Evaluation by Davies-Bouldin Index(DBI) in Cereal data using K-Means," *Proc. 4th Int. Conf. Comput. Methodol. Commun. ICCMC 2020*, no. Iccmc, pp. 306–310, 2020, doi: 10.1109/ICCMC48092.2020.ICCMC-00057.

[24] X. Wang and Y. Xu, "An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 569, no. 5, 2019, doi: 10.1088/1757-899X/569/5/052024.