Penerapan Algoritma *Stemming* Nazief & Adriani Pada Proses Klasterisasi Berita Berdasarkan Tematik Pada Laman (Web) Direktorat Jenderal HAM Menggunakan Rapidminer

Septian Firman S[1*], Wahyu Desena[2], Arief Wibowo[3]
[1,2,3]Magister Ilmu Komputer, Universitas Budi Luhur
Fakultas Teknologi Informasi, Universitas Budi Luhur
Jl. Raya Ciledug, Petukangan Utara, Pesanggrahan, Jakarta Selatan 12260
Telp. (021) 5853753, Fax. (021) 5853752
Email: [1*]2111600058@student.budiluhur.ac.id, [2*]2111600207@student.budiluhur.ac.id,
[3*]Arief.Wibowo@budiluhur.ac.id

Abstrak. Situs web merupkan media yang digunakan untuk menyampaikan informasi. saat ini berita pada situs web Direktorat Jenderal HAM belum terkategorisasi dengan baik. hanya ada tiga kategori berita yaitu berita highlight, berita, kegiatan, dan info kanwil namun belum ada informasi terkait kategori berita berdasarkan tematiknya. penelitian ini bertujuan untuk melakukan klasterisasi berita pada situs web ham.go.id berdasarkan tematiknya menggunakan rapidminer, pada rapidminer tersedia fitur stemporter namun belum tersedia dalam bahasa indonesia oleh karena itu penulis melakukan proses *stemming* dengan memanfaatkan algoritma *stemming* Nazief & Adriani untuk meningkatkan performa klasterisasi. untuk menentukan jumlah klaster terbaik penulis menggunakan nilai DBI terendah dan melakukan pengujian eksternal dengan menggunakan Confusion Matrix dari penelitian ini didapati nilai DBI tanpa melalui proses *stemming* sebesar 4.351 dengan akurasi sebesar 81,58%, recall 83,15%, precision 80,59%. setelah melakukan *stemming* dengan menggunakan algoritma Nazief & Adriani didapati nilai DBI 3.935 dengan nilai akurasi sebesar 86,84%, recall 85,71%, precision 82,50%.

Kata kunci: *Clustering, K-Means, DBI, Confusion Matrix.*

1 Pendahuluan

Direktorat Jenderal HAM yang memiliki sebuah situs web yang digunakan untuk menyampaiakan informasi capaian kegiatan dan materi substansi terkait hak asasi manusia. Saat ini berita pada situs web Direktorat Jenderal HAM hanya ada empat kategori, yaitu berita highlight, berita, kegiatan, dan info kanwil namun belum ada informasi terkait kategori berita berdasarkan tematiknya. belum

Diterima 14 Oktober 2022 Direvisi 25 Oktober 2022 Diterima untuk publikasi 27 Oktober 2022

terkategorisasinya berita menyulitkan pengguna maupun pengelola informasi untuk menemukan atau mencari kembali informasi.

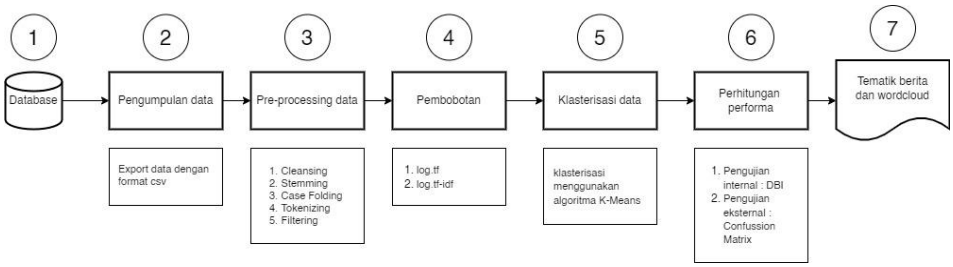
Dalam melakukan proses klasterisasi, penulis menggunakan perangkat lunak Rapidminer. rapidminer merupakan perangkat lunak yang digunakan untuk melakukan penambangan data[1], tersedia banyak fitur yang dapat digunakan dalam proses pengolahan data sebelum melakukan penambangan data, salah satunya adalah fitur stemporter yang digunakan pada proses *stemming*. *stemming* adalah proses untuk mencari kata dasar dari sebuah kata derivatif yang memiliki imbuhan[2]. kekurangan dari versi Rapidminer saat ini adalah belum tersedianya stemporter dengan bahasa indonesia, untuk menggunakan fitur stemporter dalam pengolahan teks berbahasa indonesia harus terlebih dahulu mendaftarkan seluruh kata-kata berbahasa indonesia kedalam bentuk kata dasarnya secara manual, hal tersebut sangat sulit untuk dilakukan dikarenakan jumlah kata pada bahasa indonesia yang jumlahnya terdiri dari 52,923 kata dasar, 27,670 kata turunan, dan 32,607 gabungan kata [3].

Penelitian ini bertujuan untuk melakulan klasterisasi berita pada situs web ham.go.id di tahun 2022 periode (januari – juni) dengan jumlah data sebanyak 266 berita menggunakan metode textmining dengan menerapkan algoritma K-Means menggunakan rapidminer. dalam melakukan klasterisasi dokumen teks penulis melakukan dua proses pengujian yaitu melakukan proses klasterisasi tanpa melakukan proses *stemming* dokumen dan yang kedua melakukan proses klasterisasi dengan terlebih dahulu melakukan proses *stemming* pada data, selain itu penulis juga membandingkan hasil pembobotan data dengan menggunakan metode TF dan TF-IDF dan membandingkan hasil dari kedua proses ujicoba tersebut.

Telah banyak penelitian sejenis terkait pemanfaatan algoritma K-Means dalam melakukan klasterisasi data berbasis teks dengan menggunakan rapidminer, diantaranya adalah penelitian yang dilakukan oleh Hafiz Irsyad dan M Rizky Pribadi [4] dengan judul "Implementasi *Text mining* Dalam Pengelompokan Data Tweet Pertanian Indonesia Dengan K-Means" dalam penelitian tersebut dihasilkan 5 (lima) klaster data namun pada penelitian tersebut tidak melakukan proses *stemming* pada tahapan preprocessing data, pada penelitian yang dilakukan oleh Muhammad A Ayub [5] dengan judul "Analisis Topik Ekonomi Dengan Algoritma K-Means Pada Media Online Era Pandemi Covid-19 Di Sulawesi Tenggara" tidak melakukan proses *stemming* dikarenakan dapat mengurangi akurasi.

2 Metode Penelitian

Penelitian ini adalah hasil eksperimen penulis dengan menggunakan pendekatan kuantitatif, gambaran umum alur proses penelitian dijabarkan pada Gambar 1.



Gambar 1. Alur Proses Penelitian

2.1 Website ham.go.id

Website ham.go.id terdapat beragam kategori informasi yaitu berita, data statistik terkait substansi hak asasi manusia, dan informasi terkait instansi Direktorat Jenderal HAM. pada kategori informasi berita terdiri dari empat kategori yaitu : (1) *Berita*, menampilkan informasi kegiatan tokoh pejabat pada Direktorat jenderal HAM, serta informasi seputar kegiatan instansi baik yang terkait substansi hak asasi manusia atau umum. (2) *Kegiatan*, yaitu informasi berupa kegiatan yang dilaksanakan oleh seluruh unit kerja dibawah Direktorat Jenderal HAM. (3) *Info Kanwil*, merupakan informasi yang berkaitan dengan kegiatan pemajuan hak asasi manusia di wilayah yang dilakukan oleh Kantor Wilayah Kementerian Hukum dan HAM.

2.2 Pengumpulan Data

Dalam penelitian ini penulis menggunakan data pada website ham.go.id dengan kategori berita, kegiatan dan info kanwil di tahun 2022 periode data januari sampai dengan juni. proses pengumpulan data dapat dilihat pada Gambar 2.

Gambar 2. Proses Pengumpulan Data

2.3 Pre-processing Data

menurut wishnu hardi dalam [6] *Pre-Processing Data* merupakan tahapan penting dan mendasar dalam *text mining* untuk melakukan transformasi bentuk teks menjadi lebih bermakna dan mudah untuk dipahami, hal tersebut dikarenakan teks merupakan data yang tidak terstruktur sehingga memerlukan pemprosesan data yang umumnya terdiri dari tahapan *case folding, filtering, stopwords removal*, dan *stemming.*

2.3.1 *Cleansing* / Pembersihan data

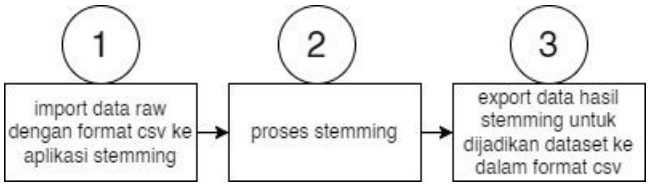
Pembersihan data merupakan hal penting dalam pre-processing data, pembersihan data dilakukan untuk menghilangkan noise dari data yang berupa simbol, angka, atau pola-pola karakter yang selalu berulang disetiap data. jumlah noise yang banyak akan menyebabkan proses klasterisasi menjadi tidak optimal.

2.3.2 *Stemming*

Stemming merupakan proses mengubah kata berimbuhan menjadi bentuk kata dasarnya. Algoritma *stemming* tidak dapat digunakan untuk berbagai macam bahasa dikarenakan memiliki morfologi yang berbeda.

Pada penelitian yang dilakukan Dwi dalam [7] telah membandingkan algoritma *stemming* Porter dan algoritma *stemming* Nazief & Adriani didapati hasil bahwa algoritma Nazief & Adriani memberikan hasil yang lebih baik untuk proses *stemming* pada dokumen berbahasa Indonesia. berdasarkan hasil tersebut penulis menggunakan algoritma Nazief & Adriani dalam proses *stemming* yang dilakukan pada penelitian ini.

jika pada umumnya proses *stemming* dilakukan di tahap ahir pada proses pre-processing data, namun dalam penelitian ini dilakukan pendekatan yang berbeda yaitu menempatkan proses *stemming* di awal setelah proses pembersihan data dikarenakan pada proses *stemming* dilakukan diluar aplikasi Rapidminer yaitu dengan membuat program sederhana berbasis PHP dengan mengimplementasikan algoritma *Stemming* Nazief & Adriani. Tahapan proses *Stemming* penelitian ini dilihat pada Gambar 3.



Gambar 3. Alur Proses *Stemming*

2.3.3 *Case Folding*

Case Folding merupakan proses transformasi data yang bertujuan untuk menyamakan karakter yang ada pada data yaitu dengan merubah huruf kapital menjadi huruf kecil atau sebaliknya.

2.3.4 Tokenizing

Tokenizing adalah proses pemisahan kalimat dalam dokumen teks menjadi token yang berisi kata-kata [8]. spasi digunakan sebagai delimeter pemisah antar token.

2.3.5 Filtering

Filtering merupakan proses menseleksi kata berdasarkan kata-kata yang memiliki makna, untuk kata-kata yang tidak memiliki makna akan di hapus. pada *filtering* terdiri dari proses *Filter Tokens by length*, dan *Filter Stopwords. Filter Tokens by Length* adalah proses membuang kata yang memiliki jumlah karakter kurang atau lebih dari jumlah karakter yang telah ditentukan, dalam penelitian ini, kata-kata yang memiliki karakter kurang dari 3 dan lebih dari 20 maka akan dihapus. *Filter Stopwords* adalah proses membuang kata-kata yang tidak penting dan tidak memiliki makna, kata-kata yang tidak penting didaftarkan terlebih dahulu dengan cara mengumpulkan kata-kata dalam file berekstensi txt yang kemudian dimasukan kedalam parameter *Filter Stopwords* pada Rapidminer.

2.4 Pembobotan

Terdapat dua jenis pembobotan kata yang diujicobakan pada penelitian ini yaitu TF (*term frequency*) dan TF-IDF(*term frequency-inverse document frequency*). TF digunakan untuk menghitung bobot kata berdasarkan jumlah kemunculan kata pada suatu dokumen, sedangkan menurut ramos dalam[9] TF-IDF merupakan merupakan statistik numerik yang sering digunakan sebagai faktor pembobotan dalam text mining, information retrieval, dan user modelling untuk menggambarkan seberapa penting satu kata bagi sebuah dokumen.

2.5 Algoritma Klasterisasi K-Means

Penelitian ini menggunakan metode K-Means yang merupakan algoritma klastering yang banyak digunakan pada berbagai bidang [10]. Menurut *Macqueen* dalam [11] dalam pengimplementasiannya K-Means memiliki 7 tahapan yaitu : (1) menyiapkan dataset, (2) menentukan jumlah klaster (k), (3) menentukan titk centroid awal secara acak, (3) Menghitung jarak tiap dataset dengan pusat klaster menggunakan euclidian distance, (4) mengelompokkan data dengan klaster terdekat berdasarkan nilai jarak terpendek, menghitung nilai Sum of Square Error, (5) menghitung ulang nilai centroid dengan anggota klaster, kembali ke langkah 3-6, (6) Proses selesai jika nilai centroid tidak berubah.

2.6 Evaluasi

Penelitian ini menggunakan dua tahapan evaluasi yaitu evaluasi internal dan evaluasi eksternal. Evaluasi internal digunakan untuk menentukan jumlah kluster yang ideal berdasarkan nilai *Davies-Bouldin index* (DBI) terendah, dalam tahapan ini penulis melakukan pengujian bertahap dimulai dari klaster 1 sampai 20 dan mencari DBI yang terendah untuk menentukan jumlah klaster terbaik. Pada tahapan evaluasi eksternal menggunakan *Confusion Matrix* dengan memberikan label secara manual pada data berdasarkan kesesuaian tematiknya.

3 Hasil dan Pembahasan

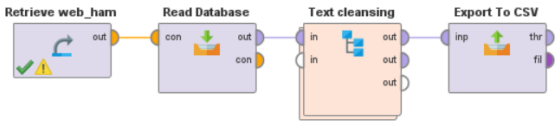
Pada penelitian ini akan membandingkan nilai akurasi hasil klastering dengan melakukan eksperimen sebagai berikut :

- Data tanpa melalui proses *stemming* dengan pembobotan TF
- Data tanpa melalui proses *stemming* dengan pembobotan TF-IDF
- Data melalui proses *stemming* dengan pembobotan TF
- Data melalui proses *stemming* dengan pembobotan TF-IDF

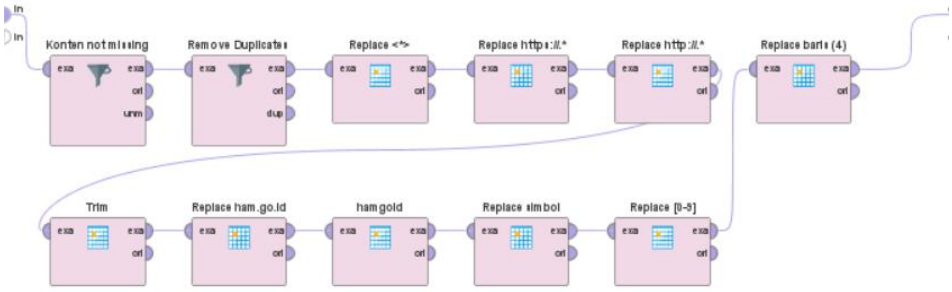
untuk melakukan eksperimen disiapkan dua dataset yaitu dataset melalui proses *stemming* dan dataset tanpa melalui proses *stemming*.

3.1 Penyiapan Data

Pengumpulan data dilakukan dengan cara melakukan ekspor data pada database ham.go.id pada periode januari sampai juni tahun 2022 dengan jumlah 266 data. Pada tabel konten terdapat 23 field namun penulis hanya mengunakan field judul dan isi berita saja yang kemudian dilanjutkan pada tahapan *cleansing* data yang terdiri dari beberapa proses yaitu : menghilangkan simbol, angka, tanda baca dan pola karakter yang tidak memiliki makna. pada tahapan ini penulis menggunakan aplikasi Rapidminer. yang dapat dilihat pada Gambar 4 dan Subproses cleansing data pada Gambar 5.



Gambar 4 Proses Pengumpulan Data dan Data Cleansing

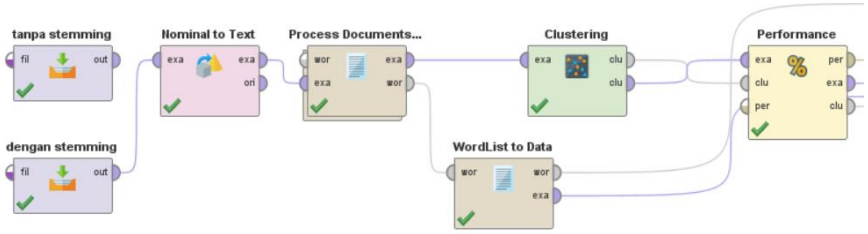


Gambar 5 Subproses *Data Cleansing*

Setelah data melewati tahap *Cleansing* selanjutnya dilakukan proses stemming menggunakan program berbasis PHP yang telah memanfaatkan algoritma Nazief & Adriani yang kemudian diekspor kedalam format .csv untuk dijadikan dataset.

3.2 Eksperimen

Pada tahapan ini dilakukan eksperimen terhadap dua dataset yang dapat dilihat pada gambar dengan menggunakan nilai Davies Bouldin untuk pengukuran performance klastering.



Gambar 6 Eksperimen Klastering

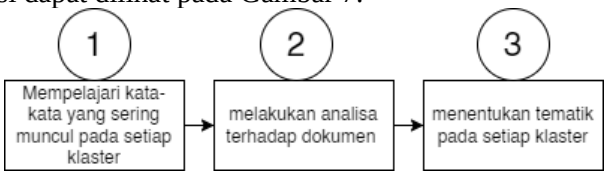
Penentuan awal jumlah klaster sebanyak 12. Pada tabel memperlihatkan perbandingan nilai Davies Bouldin dan rata-rata jarak centroid antara masing-masing eksperimen yang dilakukan.

Tabel 1 Pengujian Internal

Eksperimen	DBI	Rata-rata jarak centroid
Data tanpa melalui proses *stemming* dengan pembobotan TF	2.799	0.551
Data tanpa melalui proses *stemming* dengan pembobotan TF-IDF	4.351	0.838

Data melalui proses *stemming* dengan pembobotan TF	2.992	0.540
Data melalui proses *stemming* dengan pembobotan TF-IDF	3.935	0.826

Hasil pengujian internal pada Tabel 3 memperlihatkan nilai DBI pada eksperimen pertama memiliki nilai DBI terkecil yaitu dengan nilai 2.799 dan nilai DBI tertinggi adalah pada eksperimen ke-2 dengan nilai DBI sebesar 4.351. Tahapan ahir pada penelitian ini adalah interpretasi data hasil klasteriasi untuk mendapatkan tematik pada setiap klaster yang terbentuk. tahapan interpretasi data hasil kasterisasi dapat dilihat pada Gambar 7.



Gambar 7. Interpretasi data hasil klasterisasi

Selanjutnya dilakukan perhitungan akurasi untuk mengetahui kesesuaian data pada setiap klaster, hasil perhitungan akurasi dapat dilihat di tabel 2.

Tabel 2 Hasil Pengujian Eksternal

Eksperimen	**Accuracy**	**Precision**	**Recall**
Data tanpa melalui proses *stemming* dengan pembobotan TF	80,45%	84,45%	78,66%
Data tanpa melalui proses *stemming* dengan pembobotan TF-IDF	81,58%	80,59%	83,15%
Data melalui proses *stemming* dengan pembobotan TF	77,82%	72,15%	85,42%
Data melalui proses *stemming* dengan pembobotan TF-IDF	86,84%	82,50%	85,71%

Dari hasil pengujian eksternal didapati hasil bahwa eksperimen ke-4 yaitu Data melalui proses *stemming* dengan pembobotan TF-IDF mendapatkan hasil akurasi yang terbaik dengan nilai 86,84% dan diluar dugaan bahwa pada eksperimen ke-3 yaitu Data melalui proses *stemming* dengan pembobotan TF mendapatkan nilai akurasi hanya 77,82%.

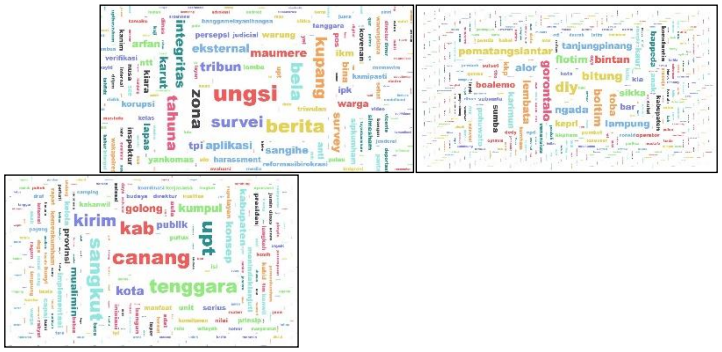
3.3 Hasil

Dari hasil pengujian eksternal didapati hasil bahwa eksperimen ke-4 yaitu Data melalui proses *stemming* dengan pembobotan TF-IDF mendapatkan hasil akurasi yang terbaik dengan nilai 86,84% dan diluar dugaan bahwa pada eksperimen ke-3 yaitu Data melalui proses *stemming* dengan pembobotan TF mendapatkan nilai akurasi hanya 77,82%. berdasarkan hasil pengamatan pada eksperimen ke-4 tematik yang terbentuk dapat dilihat pada tabel 4.

Tabel 3. Klaster yang terbentuk pada eksperimen ke-4

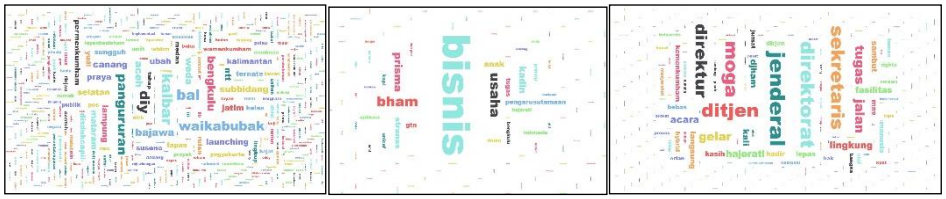
Klaster	Tematik
1	kinerja organisasi
2	kabupaten kota peduli HAM
3	Pencanangan pelanyanan publik berbasis HAM
4	perancangan produk hukum
5	indeks HAM
6	tindak lanjut penanganan dugaan pelanggran HAM
7	pelayanan publik berbasis HAM
8	bisnis dan HAM
9	berita kegiatan diluar substansi HAM
10	yankomas
11	ranham
12	instrumen dan pemenuhan hak penyandang disabilitas

terdapat subtematik yang terbentuk dari hasil klastering yaitu terkait pencanangan pelayanan publik berbasis HAM yang merupakan subtematik dari pelayanan publik berbasis HAM dan tematik terkait tindak lanjut penanganan dugaan pelanggaran HAM yang merupakan subtematik dari yankomas.

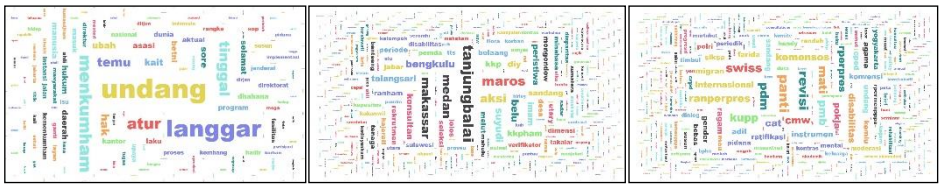


Gambar 8. Klaster 1, 2 dan 3

Gambar 9. Klaster 4, 5 dan 6



Gambar 10. Klaster 7, 8 dan 9



Gambar 11. Klaster 10, 11 dan 12

4 Kesimpulan

Berdasarkan serangkaian eksperimen yang telah dilakukan didapati bahwa proses *stemming* justru dapat menurutkan performa nilai DBI baik menggunakan metode pembobotan TF maupun TF-IDF, serta metode pembobotan TF menghasilkan DBI yang lebih baik dibanding dengan metode pembobotan TF-IDF namun setelah dilakukan pengujian eksternal dengan menggunakan *Confussion Matrix* didapati hasil yang berbeda yaitu data yang telah dilakukan proses *stemming* dengan metode pembobotan TF-IDF memiliki performa akurasi yang paling baik. kualitas klaster dalam penelitian ini masih dinilai kurang baik disarankan pada penelitian selanjutnya untuk melakukan peringkasan terhadap data untuk mendapatkan hasil klaster yang optimal.

5 Referensi

[1] Aprilla Dennis, “Belajar Data Mining dengan RapidMiner,” *Innov. Knowl. Manag. Bus. Glob. Theory Pract. Vols 1 2*, vol. 5, no. 4, pp. 1–5, 2013, [Online].

Available: http://esjournals.org/journaloftechnology/archive/vol1no6/vol1no6_6.pdf%5Cn http://www.airccse.org/journal/nsa/5413nsa02.pdf.

[2] A. Guterres, Gunawan, and J. Santoso, "Stemming Bahasa Tetun Menggunakan Pendekatan Rule Based," *Teknika*, vol. 8, no. 2, pp. 142–147, 2019, doi: 10.34148/teknika.v8i2.224.

[3] Kementerian Pendidikan Kebudayaan Riset dan Teknologi Republik Indonesia, "Halaman Statistik - KBBI Daring," 2022. https://kbbi.kemdikbud.go.id/Beranda/Statistik (accessed Jul. 22, 2022).

[4] H. Irsyad and M. R. Pribadi, "Implementasi Text Mining Dalam Pengelompokan Data Tweet Pertanian Indonesia Dengan K-Means," *KURAWAL J. Teknol. Inf. dan Ind.*, vol. 3, no. 2, pp. 164–172, 2020, [Online]. Available: https://t.co/FXtzMcbdHp.

[5] M. Arifiansyah Ayub, "Analisis Topik Ekonomi Dengan Algoritma K-Means Pada Media Online Era Pandemi Covid-19 Di Sulawesi Tenggara," *JIKO (Jurnal Inform. dan Komputer)*, vol. 4, no. 2, pp. 133–138, 2021, doi: 10.33387/jiko.v4i2.3235.

[6] W. Hardi, W. A. Kusuma, and S. Basuki, "Pengelompokan Topik Dokumen Berbasis Text Mining Dengan Algoritme K-Means : Studi Kasus Pada Dokumen Kedutaan Besar Australia Jakarta," *Visi Pustaka*, vol. 21, no. 1, pp. 67–76, 2019.

[7] D. Wahyudi, T. Susyanto, and D. Nugroho, "Implementasi Dan Analisis Algoritma Stemming Nazief & Adriani Dan Porter Pada Dokumen Berbahasa Indonesia," *J. Ilm. SINUS*, vol. 15, no. 2, pp. 49–56, 2017, doi: 10.30646/sinus.v15i2.305.

[8] F. A. Muttaqin and A. M. Bachtiar, "Implementasi Teks Mining Pada Aplikasi Pengawasan Penggunaan Internet Anak 'Dodo Kids Browser,'" *J. Ilm. Komput. dan Inform.*, pp. 1–8, 2016.

[9] E. Ogi, I. Pratiwi1, and W. Yustanti2, "Analisis Sentimen Kualitas Layanan Teknologi Pembayaran Elektronik pada Twitter (Studi Kasus Ovo dan Dana)," *Jeisbi*, vol. 02, no. 03, pp. 47–54, 2021.

[10] H. S. Behera, J. Nayak, B. Naik, and A. Abraham, "Computational intelligence in data mining," in *Informatica (Ljubljana)*, 2019, vol. 711, no. 1, doi: https://doi.org/10.1007/978-981-10-8055-5.

[11] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.

[12] S. Retno, “Peningkatan Akurasi Algoritma K-Means Dengan Clustering Purity Sebagai Titik Pusat Cluster Awal (Centroid),” *Tesis*, pp. 1–86, 2019, [Online]. Available: https://repositori.usu.ac.id/bitstream/handle/123456789/16782/177038001.pdf?sequence=1&isAllowed=y.