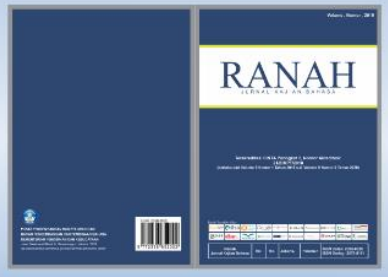


Ranah: Jurnal Kajian Bahasa
ojs.badanbahasa.kemdikbud.go.id/jurnal/index.php/jurnal_ranah

p-ISSN: 2338-8528
e-ISSN: 2579-8111

K-MEDOIDS CLUSTERING UNTUK PEMBENTUKAN DATABASE STOPWORD BAHASA JAWA

K-Medoids Clustering for the Establishment of Javanese Language Stopword Database

Aji Prasetya Wibawa, Farid Miftahuddin dan Suyono
Universitas Negeri Malang
Jalan Semarang No.5, Sumbersari, Kec. Lowokwaru, Kota Malang, Jawa Timur, Indonesia
aji.prasetya.ft@um.ac.id

Naskah Diterima Tanggal 23 September 2019—Direvisi Akhir Tanggal 26 April 2021—Disetujui Tanggal 1 Desember 2021
doi: https://doi.org/10.26499/rnh/v9i2.1490

Abstrak

Stopword merupakan kata yang bisa diabaikan dalam permrosesan bahasa alami. Proses penghapusan kata ini ini tidak mempengaruhi proses analisis teks. Teknik yang digunakan untuk menghapus *stopword* disebut *Stopword Removal*. Teknik ini mencocokkan kata dengan daftar *stopword* (*stoplis*t). Apabila kata tersebut terdapat pada daftar maka akan dihapus. Bahasa jawa sampai saat ini masih memiliki daftar *stopword* yang terbatas. Penelitian ini bertujuan membentuk daftar *stopword* menggunakan teknik *cluster* yakni K-*medoids clustering*. Teknik ini mengelompokkan kata berdasarkan kemunculan dalam teks bahasa Jawa. Dalam penerapannya, metode yang digunakan dalam penelitian ini terdiri dari lima tahap. Tahapan penelitian tersebut dimulai dari pengumpulan *dataset, preprocessing data*, *clustering,* dan terakhir adalah evaluasi. Setiap hasil *cluster* diuji dengan mencocokkannya dengan *stopword* hasil identifikasi ahli bahasa Jawa. Hasil penelitian ini menunujkkan bahwa *stopword* yang dihasilkan *k-medoids clustering* dengan nilai K=13 yang memiliki akurasi sebesar 70,5%.
Kata-kata kunci: *stopword*, bahasa Jawa, *K-medoids, clustering*

Abstract

Stopword is a word that can be ignored in the natural language process. This word removal process does not affect the text analysis process. The technique used to remove stopword is called Stopword Removal. This technique matches words to a stopword list. If the word is in the list it will be deleted. Javanese language to date still has a limited list of stopword. This study aims to form a list of stopword using cluster techniques namely K-medoids clustering. This technique groups words by occurrence in Javanese text. Each cluster result is tested by matching it with a stopword of javanese expert identification. The results of this study suggest that the stopword produced by k-medoids clustering with a value of K=13 has an accuracy of 70.5%.
Keywords: *stopword, Javanese Language, K-medoids, clustering*

How to Cite: Wibawa, Aji Prasetya, dkk. (2021). K-Medoids Clustering untuk Pembentukan Database Stopword Bahasa Jawa. *Ranah: Jurnal Kajian Bahasa. 10*(2). 261—269. Doi: https://doi.org/10.26499/rnh/v9i2.1490

PENDAHULUAN

Bahasa Jawa merupakan salah satu bahasa Austronesia, yang termasuk dalam subkelompok Melayu-Polinesia Barat dan rumpun Sunda. Sesuai dengan anggota subkelompok lainnya, kebanyakan akar kata Jawa terdiri dari dua suku kata, dan dari varian tata bahasa ini diucapkan melalui imbuhan (Oakes, 2016). Bahasa Jawa merupakan bahasa pergaulan sehari-hari di

daerah Jawa, yang digunakan untuk berinteraksi antar individu dan memungkinkan terjadinya komunikasi dan perpindahan informasi. Tidak hanya di wilayah pulau Jawa dan Indonesia, bahasa Jawa juga digunakan pada daerah lain, seperti Sumatera, Kalimantan, Sulawesi, dan pulau lainnya di Indonesia, serta di luar negeri seperti Suriname, Kaledonia Baru, dan kampung Jawa di Malaysia (Saddhono & Hartanto, 2021).

Dari latar belakang tersebut, bahasa Jawa merupakan salah satu warisan budaya Indonesia yang harus dilestarikan dan dijaga (Kridalaksana, 2001). Berdasarkan survey pada tahun 2021, Bahasa Jawa merupakan bahasa yang digunakan oleh sekitar 68 juta orang di seluruh dunia. Hal ini menjadikan bahasa Jawa sebagai bahasa yang paling sering digunakan di urutan ke-26 dunia (Ethnologue, 2021). Berdasarkan fakta tersebut, bahasa Jawa dipilih dalam penelitian ini karena memiliki daya tarik kepada orang lain untuk dipelajari. Selain itu, membuat dan membaca dokumen berbahasa Jawa merupakan salah satu cara melestarikan penggunaan bahasa Jawa.

Bahasa Jawa memiliki aksen, dialek, intonasi, dan diksi yang kuat dan unik. Secara umum, bahasa Jawa diucapkan dengan jelas, tegas, dengan intonasi pendek dan penekanan di beberapa bagian (Wardani & Suwartono, 2019). Dalam penerapannya, terdapat tiga dialek utama bahasa Jawa yang saling mempunyai keterkaitan (Oakes, 2016). Dialek daerah Solo dan Yogyakarta, pusat sejarah budaya Jawa, disebut Kejawen, dan dianggap sebagai bentuk standar bahasa Jawa. Bahasa Jawa Timur dituturkan di Surabaya, Malang dan Pasuran (Gordon Jr, 2005). Sedangkan bahasa Jawa Barat dituturkan di Banten, Cirebon dan Tegal yang banyak dipengaruhi oleh bahasa Sunda. Dalam memahami teks bahasa Jawa banyak kata kata yang tidak penting yang berpengaruh dalam kalimat/bacaan. Oleh karena itu, perlu mengurangi jumlah kata dalam sebuah dokumen yang nantinya akan berpengaruh dalam kecepatan dan performa atau biasa disebut dengan stopword removal. Maka diperlukan *database stopword* yang baik yang digunakan pada proses *stopword removal* ini.

Algoritma *stopword removal* dalam sistem *Information Retrival*, merupakan proses penghapusan semua kata yang tidak memiliki makna (Manning et al., 2009). Dengan kata lain sebuah proses untuk menghilangkan kata yang ‘tidak relevan’ pada hasil parsing sebuah dokumen teks dengan membandingkannya dengan *Stoplist* (*Stopword List*) yang ada (Budhi et al., 2006). Proses pengelompokkan dapat dilakukan dengan mengimplementasikan metode *clustering*. *K-means* telah digunakan untuk membentuk *stopword* bahasa Jawa. Akurasi yang dihasilkan dari metode ini adalah 78,28% (Wibawa et al., 2020). Penelitian ini bertujuan untuk menerapkan *K-medoids* sebagai *alternative* pembuatan *stopwordlis*t bahasa Jawa berbasis klastering. Metode ini dipilih karena performa yang lebih kuat terhadap *noise* dan outlier dibandingkan dengan k-means pada penelitian sebelumnya. Hal ini dikarenakan metode *K-medoids* meminimalkan jumlah ketidakmiripan berpasangan umum daripada jumlah jarak Euclidean kuadrat. Selain itu, metode ini dipilih karena berbagai kelebihannya antara lain: kecepatan komputasi (Velmurugan, 2010), bebas dari pengaruh *outlier* (Madhulatha, 2011).

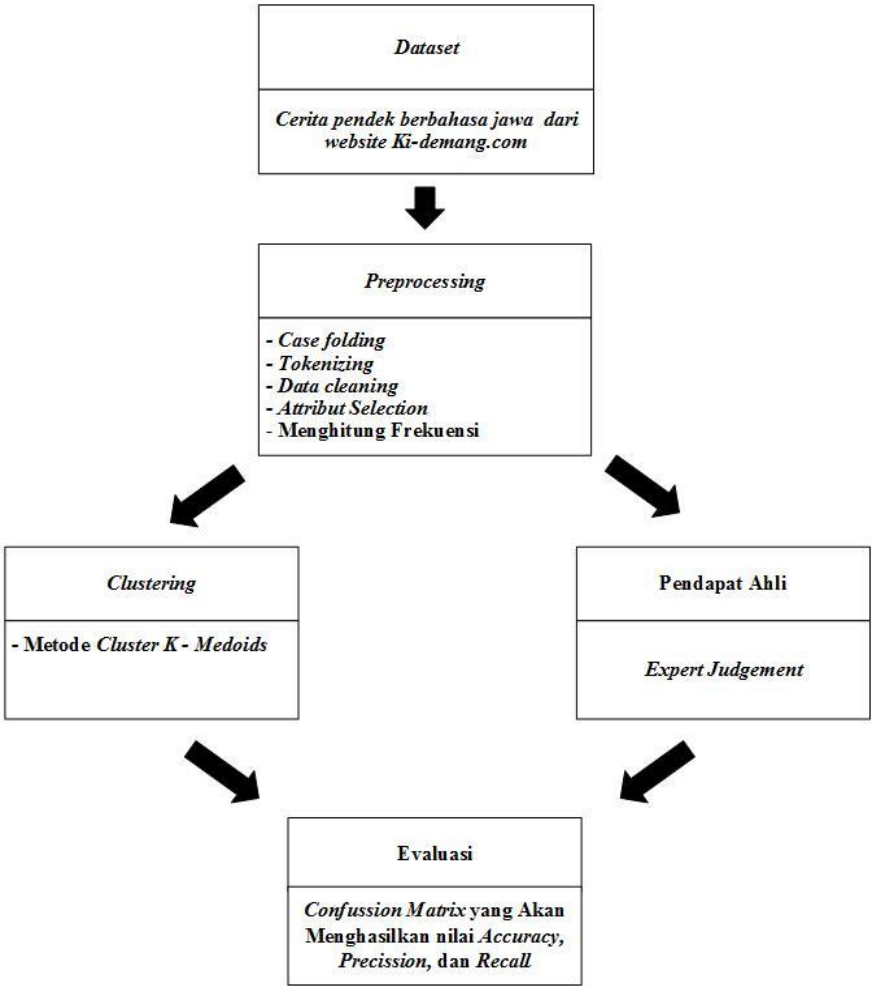
LANDASAN TEORI

Clustering atau klaterisasi adalah metode pengelompolan data. *Clustering* adalah sebuah proses untuk mengelompokan data ke dalam beberapa *cluster* atau kelompok, sehingga data dalam satu *cluster* memiliki tingkat kemiripan yang maksimum dan data antar *cluster* memiliki kemiripan yang minimum (Tan et al., 2006). Metode untuk melakukan *clustering* dapat dikategorikan menjadi empat metode, yaitu: *partitioning, hierarchical, grid-based and model-based. Clustering* berbasis *partitioning* menghasilkan partisi dari data sehingga objek dalam *cluster* lebih mirip satu sama lain daripada objek yang ada dalam *cluster* lain. *K-means* dan *K-medoids* adalah contoh dari metode *partitioning* (Triyanto, 2015).

Metode *K-medoids* merupakan metode clustering yang berfungsi untuk memecah dataset menjadi kelompok-kelompok. Kelebihan dari metode ini mampu mengatasi kelemahan dari metode K-Means yang sensitive terhadap outlier (Pramesti et al., 2017). *K-medoids* atau *Partitioning Around Medoids* (PAM) adalah algoritma clustering yang mirip dengan *K-means*. Perbedaan dari kedua algoritma ini yaitu algoritma *K-Medoids* atau PAM menggunakan objek sebagai perwakilan (*medoid*) sebagai pusat *cluster* untuk setiap *cluster*, sedangkan *K-means* menggunakan nilai rata-rata (*mean*) sebagai pusat *cluster* (Madhulatha, 2011). Algoritma *K-medoids* memiliki kelebihan untuk mengatasi kelemahan pada algoritma *K-means* yang *sensitive* terhadap *noise* dan *outlier*, dimana objek dengan nilai yang besar yang memungkinkan menyimpang dari distribusi data. Kelebihan lainnya yaitu hasil proses *clustering* ini tidak bergantung pada urutan masuk *dataset* (Pramesti et al., 2017).

METODE PENELITIAN

Studi ini menggunakan pendekatan penelitian terapan untuk mengetahui kinerja algoritma K-medoids untuk clustering. Pendekatan penelitian terapan ini dipilih karena bertujuan untuk mencari solusi, dalam kasus ini adalah pembentukan database stopword Bahasa Jawa. Penelitian ini terdiri dari lima tahapan penelitian. Tahapan penelitian ditunjukkan pada gambar 1.



Gambar 1. Tahap penelitian

Tahap pertama dimulai dengan pengambilan *dataset* yang digunakan sebagai input dalam penelitian, *dataset* diambil dari situs *www.ki-demang.com* dengan kategori Cerita Pendek

berbahasa Jawa berjumlah 106 cerita. Cerita yang diambil adalah isi cerita tanpa nomor dan judul. Kumpulan cerita tersebut digabung menjadi satu dokumen teks. Dokumen teks ini yang akan digunakan sebagai *dataset* penelitian.

Tahap kedua dilakukan proses *preprocessing*, tahap *preprocessing* pertama yang dilakukan adalah *case folding* yaitu tahapan yang mengubah semua huruf dalam dokumen menjadi huruf kecil (Amalia et al., 2018). Hanya huruf 'a' sampai dengan 'z' yang diterima. Kemudian menghilangkan karakter tanda baca dan angka pada *dataset*. Selain itu, data akan dibersihkan dari kata yan salah tulis, tidak memiliki makna, nama orang, dan kata yang menggunakan selain bahasa Jawa.

Tahap ketiga, menghitung kata dengan jumlah frekuensi yang terdapat pada *dataset* dan *attribut selection* untuk menentukan atribut yang paling berpengaruh dan akan digunakan untuk *cluster*. *Dataset* kemudian akan diserahkan ke ahli bahasa Jawa. Ahli mengelompokkan *dataset* tersebut menjadi dua kelas yaitu kata umum dan kata khusus. Kata umum yaitu kata yang dianggap tidak memiliki makna, seperti kata tambahan dan kata hubung. Sedangkan kata khusus adalah kata yang memiliki makna seperti kata kerja.

Tahap keempat adalah *Clustering* yang dilakukan untuk mengelompokkan berdasarkan frekuensi setiap jenis kata menggunakan metode *K-Medoids*. Tahapan dari proses clustering ini dapat dilihat pada Gambar 2.

Input:
D = {d1, d2,....., dn} // Data yang digunakan
k = {1, 2,.....n} // Jumlah *cluster* yang diinginkan
Output:
Satu set *k cluster*.
Tahap 1:
Memilih dan menentukan nilai *k* sebagai inisial centroid atau nilai tengah (medoids) sebanyak *k cluster*.
Tahap 2:
Mencari semua poin yang paling dekat dengan medoids, dengan cara menghitung jarak vektor antar dokumen dengan menggunakan Euclidian Distance.
Tahap 3:
Memilih poin atau nilai *k* secara acak yang mana nilai *k* bukanlah nilai medoids.
Tahap 4:
Menghitung total jarak antar medoids.
Tahap 5:
Jika Total Distance (TD) baru < TD awal, maka akan menukar posisi medoids awal dengan medoids baru.
Tahap 6:
Mengulangi langkah 2-5 hingga nilai medoids tidak berubah.

Gambar 2. Tahapan *K-Medoids Clustering*

Berdasarkan gambar 2, salah satu tahapan *K-Medoids Clustering* adalah menentukan nilai *k* atau jumlah *cluster*, pada penelitian ini nilai *k* yang digunakan adalah 3,5,7,9,11,13,15,17,19,21,23,25 dengan metode pemilihan *random* yaitu nilai *k* ditentukan tanpa menggunakan perhitungan khusus.

Tahap kelima adalah evaluasi. Evaluasi bertujuan untuk menguji ketepatan hasil *clustering* dengan metode yang digunakan sehingga dapat diketahui *cluster* mana yang memiliki *stopword* terbaik (Ramadhani & Januarita AK, 2019), mendekati identifikasi pakar. Dalam penelitian ini pendapat ahli bahasa Jawa dianggap mutlak kebenarannya dalam memberikan kelas pada *dataset*. Setiap hasil *clustering* akan diuji dengan data ahli untuk evaluasi. Evaluasi dilakukan dengan menggunakan *Confussion Matrix*. Pengukuran terhadap kinerja suatu sistem klasifikasi merupakan hal yang penting. Kinerja sistem klasifikasi menggambarkan seberapa baik sistem dalam mengklasifikasikan data. *Confussion matrix* merupakan salah satu metode yang dapat digunakan untuk mengukur kinerja suatu metode klasifikasi. Pada dasarnya *confussion matrix* mengandung informasi yang membandingkan

hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi yang seharusnya (Zayuka et al., 2017).

Pada tahapan ini akan dilakukan uji keakurasian dari algoritma yang digunakan untuk *clustering* dataset tersebut. Pengujian dilakukan dengan menggunakan parameter berikut.

Akurasi

Akurasi adalah perbandingan data yang diidentifikasi benar dengan jumlah semua data. Dalam penelitian ini nilai akurasi metode didapatkan dengan membagi jumlah dokumen benar bernilai *true* dengan jumlah semua dokumen yang diklasifikasikan. Nilai *true* berarti apabila hasil *clustering* menyatakan bahwa suatu kata adalah *stopword* dan pada hasil ahli kata tersebut memang *stopword*, begitu pula untuk kata *non stopword.*

Presisi

Presisi adalah tingkat ketepatan antara data yang diminta dengan jawaban yang diberikan oleh sistem. Perhitungan rasio jumlah data dalam *dataset* yang benar bernilai *true positive* terhadap jumlah data *true positive* dan jumlah data *false negative*. Nilai *true positive* berarti hasil *clustering* menyatakan bahwa suatu kata adalah *stopword* dan pada hasil ahli kata tersebut memang *stopword* sedangkan *false negative* berarti hasil *clustering* menyatakan bahwa suatu kata adalah *stopword* akan tetapi pada hasil ahli kata tersebut termasuk *nonstopword.*

PEMBAHASAN

Berdasarkan metode yang telah dijelaskan pada bagian sebelumnya, *dataset* yang digunakan dalam penelitian merupakan 106 cerita pendek dari situs www.ki-demang.com. Setelah semua *dataset* terkumpul, dilanjutkan pada tahap kedua yaitu *preprocessing*. Pada tahap kedua ini, *dataset* dipotong per kata (*token)* sehingga menghasilkan 17.763 jenis kata dan frekuensinya. Data hasil token tersebut dibersihkan dari kata yang salah tulis, kata tidak memiliki makna, nama orang dan kata yang menggunakan bahasa Jawa selain Ngoko seperti bahasa Indonesia, bahasa Inggris dan bahasa Jawa Krama dihapus, sehingga menjadi 14.384 jenis. Penghapusan ini berdasarkan kamus terjemahan Jawa-Indonesia dan Indonesia-Jawa. Contoh kata yang dihapus dapat dilihat pada tabel 1.

Tabel 1.
Contoh Kata Yang Dihapus Dalam Tahap Preprocessing

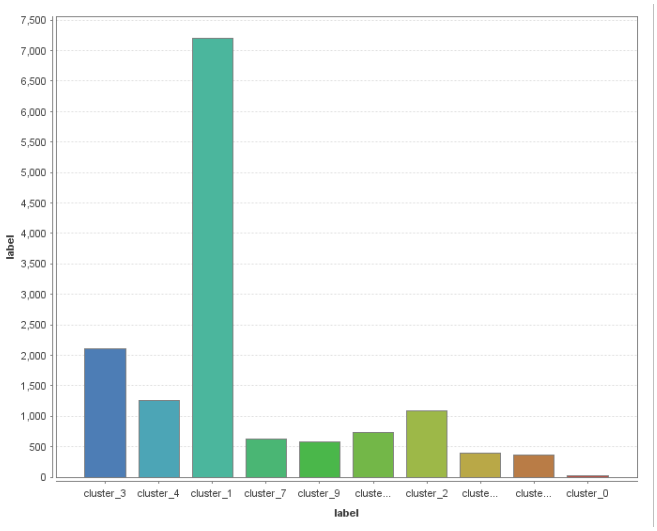
Kata salah tulis	Kata tidak bermakna	Nama Orang	Kata selain Ngoko
sepiraa	met	Ezza	Inbox
rilaaaaa	We	Siti	Out
ewoesemono	Rin	Laras	Awalnya
ngetikkkkkk	ooh	Irvan	Apabila
pendhafatran	Loh	Dini	Mouse
sept	Srat	Yono	In

Pada tahap ketiga, *dataset* yang berjumlah 14.384 jenis kata diserahkan ke ahli bahasa Jawa. Ahli mengelompokkan dataset tersebut menjadi dua kelas yaitu kata umum dan kata khusus. Kata umum yaitu kata yang dianggap tidak memiliki makna, seperti kata tambahan dan kata hubung. Sedangkan kata khusus adalah kata yang memiliki makna seperti kata kerja. Selanjutnya kata umum dianggap sebagai *stopword* berjumlah 3224 jenis kata dan kata khusus dianggap sebagai *non stopword* berjumlah 11160. Hasil klasifikasi dari ahli kata dapat dilihat pada tabel 2.

Tabel 2.
Contoh Kata Hasil Klasifikasi Dari Ahli Bahasa Jawa

Stopword	Non Stopword
Abdine	Abang
Abote	Alasane
Apa	Beritane
Iki	Karal
Jan	Dandang
Jane	Kawin

Pada tahap keempat, data yang di*cluster* berjumlah 14.384 jenis kata dan frekuensinya. pada penelitian ini nilai *k* yang digunakan adalah 3,5,7,9,11,13,15,17,19,21,23,25 dengan metode pemilihan *random* yaitu nilai *k* ditentukan tanpa menggunakan perhitungan khusus. Hasil setiap *cluster* akan dikecilkan menjadi 2 kelas, yaitu kelas *stopword* dan kelas *non stopword.* Kelas *non stopword* adalah seluruh kata pada kelas yang mempunyai nilai anggota tertinggi di setiap *cluster* nya. Sedangkan kelas *stopword* adalah seluruh kata pada kelas yang mempunyai nilai anggota lebih rendah dari kelas *non stopword.* Sebagai contoh pada *k*=19, yang dianggap *non stopword* adalah *k1* karena mempunyai nilai anggota yang paling tinggi, sedangkan *stopword* adalah *k 0,2,3,4,7,9,11,13,17* karena nilai *k* tersebut lebih rendah dari *k1*, seperti gambar 3.



Gambar 3. Diagram *Cluster 19*

Kemudian pada tahap terakhir yaitu evaluasi. Berdasarkan tahapan penelitian yang telah dilakukan, evaluasi dilaporkan menggunakan *confusion matrix* berdasarkan data uji yang telah di tentukan dengan mencocokkan *stopword* dan *non stopword* hasil *clustering* dengan *stopword* dan *non stopword* milik ahli didapatkan hasil perhitungan seperti pada tabel 3.

Tabel 3.
Hasil *Cluster Stopword* dengan Berbagai Nilai K

k	Akurasi	Presisi	Stopword	Non Stopword
3	70%	69,9%	3811	10572
5	68,9%	69%	3811	10572
7	70,4%	70,4%	3811	10572
9	68,8%	69,9%	7118	7205
11	69,1%	68,8%	7118	7205
13	70,5%	70,2%	7118	7205
15	69%	68,7%	7118	7205
17	68,8%	68,5%	7178	7205
19	69.1%	68,7%	7178	7205
21	69%	68,7%	7178	7205
23	68,8%	68,5%	7178	7205
25	51,3%	81,2%	7178	7205

Pada tabel 3 diketahui akurasi tertinggi adalah 70,5% dengan presisi 70,2% yang dimiliki *cluster* dengan nilai K=13, pada K tersebut *stopword* berjumlah 7118 kata. Evaluasi didapat dengan mencocokkan *stopword* milik ahli yang berjumlah 3224 kata terhadap data uji (*data testing*) berupa teks berita berbahasa jawa yang berjumlah 387 kata. Diinisialisasikan bahwa *true positive* merupakan kata yang menurut ahli benar dan hasil *cluster* benar (benar *stopword*), *false positive* merupakan kata yang menurut ahli salah tetapi hasil *cluster* benar, *false negative* merupakan kata yang menurut ahli benar tetapi hasil *cluster* salah, *true negative* merupakan kata yang menurut ahli salah dan hasil *cluster* juga salah (bukan *stopword*) pada data uji.

Berdasarkan tahapan penelitian yang telah dilakukan, hasil perhitungan dengan *Confussion Matrix* terdapat pada tabel 3. Pada tabel ini sudah diketahui nilai TP (*True Positive*), FP (*False Negative*), TN (*True Negative*) dan FN (*False Negative*).

Tabel 4.
Hasil Perhitungan Confussion Matrix

K	TP	FP	TN	FN	Akurasi	Presisi	Recall
3	212	91	17	7	70%	69,9%	96,8%
5	210	94	17	8	68,9%	69%	96,3%
7	215	90	16	7	70,4%	70,4%	96,8%
9	219	99	8	4	68,8%	69,9%	98,2%
11	218	99	8	2	69,1%	68,8%	99%
13	224	95	8	2	70,5%	70,2%	99,1%
15	217	99	8	2	69%	68,7%	99%
17	217	100	8	2	68,8%	68,5%	99,1%
19	217	99	9	2	69.1%	68,7%	99%
21	217	99	8	2	69%	68,7%	99%
23	217	100	8	2	68,8%	68,5%	99%
25	82	19	80	135	51,3%	81,2%	37,8%

Pada tabel 4, hasil akurasi tertinggi pada *cluster* dengan nilai K=13 dengan nilai akurasi sebesar 70,5% dan nilai presisi sebesar 70,2%. Nilai akurasi didapat dari 224 kata bernilai *true positive* ditambah 8 kata bernilai true negative dibagi dengan 329 kata yang didapatkan dari jumlah nilai pada setiap *cell* pada *confusion matrix* (TP+FP+FN+TN). Nilai presisi didapat dari 224 kata bernilai *true positive* yang dibagi dengan 224 kata bernilai *true positive* itu sendiri ditambah dengan 95 kata bernilai *false positive*.

Hasil akurasi terendah terdapat pada *cluster* dengan nilai K=25 dengan nilai akurasi sebesar 51,3% dan nilai presisi sebesar 81,2%. Nilai akurasi didapat dari 82 kata bernilai *true positive* ditambah 80 kata bernilai *true negative* dibagi dengan 316 kata yang didapatkan dari jumlah nilai pada setiap *cell* pada *confusion matrix* (TP+FP+FN+TN). Nilai presisi didapat dari 82 kata bernilai *true positive* yang dibagi dengan 82 kata bernilai *true positive* itu sendiri ditambah dengan 19 kata bernilai *false positive*.

Hasil peneltian menggunakan metode *K-medoids* lebih ideal dibandingkan dengan metode *K-means* yang digunakan dalam penelitian sebelumnya. Hal ini dikarenakan *K-Medoids* mempunyai kinerja lebih kuat dibandingkan dengan *K-Means*. Dalam *K-Medoids*, k digunakan sebagai objek representatif untuk meminimalkan jumlah ketidaksamaan objek data sedangkan, *K-Means* menggunakan jumlah jarak Euclidean kuadrat untuk objek data. metrik jarak ini mengurangi *noise* dan *outlier*. Namun, dalam penerapannya perlu adanya sumber teks yang lebih banyak. Hal ini digunakan untuk meningkatkan akurasi dalam proses *clustering*. Pembentukan *stopword* bahasa Jawa dapat digunakan untuk mengembangkan beberapa aplikasi bidang bahasa khususnya yang berbentuk teks antara lain: klasifikasi genre (Muslimah et al., 2018), analisis sentimen (Hayuningtyas & Sari, 2019), dan klasifikasi jenis bahasa(Nursirwan, 2012). Dampak pengembangan yang lain adalah mempekaya khasanah pelestarian bahasa dan satra daerah melalui teknologi. Harapannya, generasi muda yang akrab dengan teknologi akan

semakin mencintai budayanya karena budaya yang bersifat tradisional dapat berasimilasi dengan teknologi.

PENUTUP

Berdasarkan penelitian yang telah dilakukan mengenai penerapan metode *K-medoids Clustering* pada pembuatan *stoplist* bahasa Jawa berdasarkan kemunculan kata dalam teks berbahasa Jawa dilakukan dengan beberapa tahapan antara lain, *scraping, case folding, tokenizing, data cleaning*, perhitungan frekuensi dan atribut *selection* dapat disimpulkan bahwa metode *K-medoids* mampu digunakan untuk pembentukan *stopword list* (*stoplist*) berbahasa Jawa. Penelitian ini menghasilkan nilai akurasi sebesar 70,5% yang berada pada *cluster* dengan nilai K=13, sehingga *stopword* pada K=13 dipilih untuk membentuk *stoplist* dikarenakan kebenarannya mendekati *stopword* yang ditentuka oleh ahli bahasa Jawa. Kedepan, perlu dibuat *sopword* bahasa Jawa dengan jumlah sumber teks yang lebih banyak, sehingga hasilnya dapat digunakan untuk menunjang pengembangan aplikasi yang terkait pelestarian bahsa dan sastra daerah. Selain itu perlu juga diaplikasikan pada bahasa daerah yang lain untuk mengukur akurasi dan mengetahui apakah karakteristik dari bahasa tersebut bisa digunakan dengan metode yang sama.

DAFTAR PUSTAKA

Amalia, A., Lydia, M. S., Fadilla, S. D., & Huda, M. (2018). Perbandingan Metode Klaster dan Preprocessing Untuk Dokumen Berbahasa Indonesia. *Jurnal Rekayasa Elektrika*, *14*(1), 35–42. https://doi.org/10.17529/jre.v14i1.9027

Budhi, G. S., Gunawan, I., & Yuwono, F. (2006). Algoritma Porter Stemmer For Bahasa Indonesia Untuk Pre-Processing Text Mining Berbasis Metode Market Basket Analysis. *PAKAR Jurnal Teknologi Informasi dan Bisnis*, *7*(3).

Ethnologue. (2021). *What are the top 200 most spoken languages?* Ethnologue Languages of the World. https://www.ethnologue.com/guides/ethnologue200

Gordon Jr, R. G. (2005). *Ethnologue, languages of the world*. SIL International.

Hayuningtyas, R. Y., & Sari, R. (2019). Analisis Sentimen Opini Publik Bahasa Indonesia terhadap Wisata TMII Menggunakan Naïve Bayes dan PSO. *Jurnal Techno Nusa Mandiri*, *16*(1), 37–42. https://doi.org/10.33480/techno.v16i1.115

Kridalaksana, H. (2001). *Wiwara: pengantar bahasa dan kebudayaan Jawa*. Gramedia Pustaka Utama.

Madhulatha, T. S. (2011). Comparison between K-Means and K-Medoids Clustering Algorithms. In D. C. Wyld, M. Wozniak, N. Chaki, N. Meghanathan, & D. Nagamalai (Eds.), *Advances in Computing and Information Technology* (pp. 472–481). Springer. https://doi.org/10.1007/978-3-642-22555-0_48

Manning, C., Nayak, P., & Raghavan, P. (2009). *Introduction to Information Retrieval: Probabilistic information retrieval*. https://doi.org/10.1017/CBO9780511809071

Muslimah, N., Indriati, I., & Wihandika, R. C. (2018). Klasifikasi Film Berdasarkan Sinopsis dengan Menggunakan Improved K-Nearest Neighbor (K-NN). *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, *3*(1), 196–204.

Nursirwan, N. (2012). Klasifikasi Leksikostatistik Bahasa Melayu Langkat, Bahasa Melayu Deli, Dan Bahasa Dairi Pakpak. *Fakultas Ilmu Budaya Universitas*, 10.

Oakes, M. P. (2016). Javanese. In B. Comrie (Ed.), *The World's Major Languages* (p. 14). Routledge. https://doi.org/10.4324/9781315084862-76

Pramesti, D. F., Furqon, M. T., & Dewi, C. (2017). Implementasi Metode K-Medoids Clustering Untuk Pengelompokan Data Potensi Kebakaran Hutan/Lahan Berdasarkan Persebaran Titik Panas (Hotspot). *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer E-ISSN*, *1*(9), 723–732.

Ramadhani, R. D., & Januarita AK, D. (2019). Evaluasi K-Means dan K-Medoids pada Dataset Kecil. *SNIA (Seminar Nasional Informatika dan Aplikasinya)*.

Saddhono, K., & Hartanto, W. (2021). A dialect geography in Yogyakarta-Surakarta isolect in Wedi

District: An examination of permutation and phonological dialectometry as an endeavor to preserve Javanese language in Indonesia. *Heliyon*, *7*(7), e07660. https://doi.org/10.1016/j.heliyon.2021.e07660

Tan, P. N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining*. Pearson.

Triyanto, W. A. (2015). Algoritma K-Medoids Untuk Penentuan Strategi Pemasaran Produk. *Simetris: Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer*, *6*(1), 183–188. https://doi.org/10.24176/simet.v6i1.254

Velmurugan. (2010). Computational Complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points. *Journal of Computer Science*, *6*(3), 363–368. https://doi.org/10.3844/jcssp.2010.363.368

Wardani, N. A., & Suwartono, T. (2019). Javanese Language Interference in the Pronunciation of English Phonemes. *Celtic: A Journal of Culture, English Language Teaching, Literature and Linguistics*, *6*(2), 14–25. https://doi.org/10.22219/celtic.v6i2.8589

Wibawa, A. P., Fithri, H. K., Zaeni, I. A. E., & Nafalski, A. (2020). Generating Javanese Stopwords List using K-means Clustering Algorithm. *Knowledge Engineering and Data Science*, *3*(2), 106. https://doi.org/10.17977/um018v3i22020p106-111

Zayuka, H., Nasution, S. M., & Purwanto, Y. (2017). Perancangan Dan Analisis Clustering Data Menggunakan K-medoids Untuk Berita Berbahasa Inggris. *EProceedings of Engineering*, *4*(2).