Jurnal Teknik Informatika (JUTIF)
Vol. 4, No. 5, October 2023, pp. 1257-1265

DOI: https://doi.org/10.52436/1.jutif.2023.4.5.1462
p-ISSN: 2723-3863
e-ISSN: 2723-3871

TEXT CLUSTERING IN KARO LANGUAGE USING TF-IDF WEIGHTING AND K-MEANS CLUSTERING

Trisna Amanda Br Sembiring[*1], Muhammad Siddik Hasibuan[*2]

[1,2]Program Studi Ilmu Komputer, Universitas Islam Negeri Sumatera Utara, Indonesia
Email: [1]trisnaamnda366@gmail.com, [2]muhammadsiddik@uinsu.ac.id

(Article received: October 10, 2023; Revision: November 01, 2023; published: November 13, 2023)

Abstract

The aim of this research is to see how many presentations there are between dialects and look for clusters. There is also a method used for weighting, namely tf-idf, there are several steps used in this method, namely starting from the tokenizing process, transform cases, stopwords filter and token filter. to search for clusters using the k-means clustering method on rapidminer. The results of this research obtained a tf-idf weighting value, namely ginger dialect 37.5% for the number of word occurrences and 62.5% for the total of all words documented. Furthermore, for the Julu dialect, it was 37.5% for the number of word occurrences and 62.5% for the total of all words documented. The Singaporean Lau dialect accounts for 38% of the number of word occurrences and 62% of the total number of words documented. The singteruh deleng lau dialect accounts for 38% of the number of word occurrences and 62% of the total number of words documented. The Liang Melas dialect accounts for 38% of the number of word occurrences and 62% of the total number of words documented. Based on k-means clustering, it produces cluster 0: 68 items, cluster 1: 3 items, cluster 2: 15 items, cluster 3: 10 items, cluster 4: 4 items with a total sample of 100 items. The conclusion obtained is that the Ginger dialect and the Julu dialect are identical, while the Singaporean Lau dialect, the Teruh Deleng and Liang Melas dialects are also identical.

Keywords: *Clustering, Karo Language, K-means, RapidMiner, Tf-Idf.*

TEXT CLUSTERING DALAM BAHASA KARO MENGGUNAKAN PEMBOBOTAN TF-IDF DAN K-MEANS CLUSTERING

Abstrak

Tujuan dari penelitian ini untuk melihat berapa presentasi antara dialek dan mencari cluster. Ada pun metode yang digunakan untuk pembobotan yaitu *tf-idf*, ada beberapa langkah yang di gunakan dalam metode ini yaitu mulai dari proses *tokenizing, transform cases, filter stopwords* dan *filter token*. untuk mencari *cluster* menggunakan metode *k-means clustering* pada rapidminer. Hasil dari penelitian ini diperolah nilai pembobotan *tf-idf* yaitu dialek jahe 37,5% untuk jumlah kemunculan kata dan 62,5% total seluruh kata didokumen. Selanjutnya untuk dialek julu 37,5% untuk jumlah kemunculan kata dan 62,5% total seluruh kata didokumen. Dialek singalur lau 38% jumlah kemunculan kata dan 62% total seluruh kata didokumen. Dialek singteruh deleng lau 38% jumlah kemunculan kata dan 62% total seluruh kata didokumen. Dialek liang melas 38% jumlah kemunculan kata dan 62% total seluruh kata didokumen. Berdasarkan *k-means clustering* menghasilkan *cluster* 0: 68 items, cluster 1: 3 items, cluster 2: 15 items, cluster 3:10 items, cluster 4:4 items dengan total sempel 100 *items*. kesimpulan yang diperoleh yaitu antara dialek jahe dan dialek julu identik, sedangkan antara dialek singalur lau, dialek teruh deleng dan liang melas juga indentik.

Kata kunci: *Bahasa Karo, Clustering, K-means, RapidMiner, Tf-Idf.*

1. PENDAHULUAN

Clustering atau pengelompokan adalah teknik analisis data yang bertujuan untuk mengelompokkan objek-objek data yang serupa ke dalam kelompok atau cluster berdasarkan kemiripan atau kesamaan karakteristik tertentu. Clustering adalah teknik yang umum digunakan dalam berbagai bidang, seperti ilmu data, ilmu sosial, pengenalan pola, bioinformatika, ilmu kedokteran, dan sebagainya. Contoh aplikasi clustering adalah dalam pemasaran untuk memahami perilaku konsumen, dalam ilmu kedokteran untuk mengelompokkan pasien dengan penyakit yang sama, dan dalam analisis citra untuk mengelompokkan gambar-gambar berdasarkan karakteristik tertentu [1].

Clustering adalah teknik analisis data yang digunakan untuk mengelompokkan data yang serupa ke dalam kelompok-kelompok yang lebih kecil. Langkah-langkah dalam menentukan clustering antara lain memilih metode clustering yang tepat, memilih jumlah cluster yang diinginkan, menentukan atribut atau variabel yang akan digunakan, memilih nilai awal untuk setiap cluster, menghitung jarak antara data dan centroid, menetapkan data ke dalam cluster yang sesuai, melakukan evaluasi dan validasi, dan menerapkan hasil clustering [2].

Kajian terdahulu yang relevan dengan penelitian ini adalah penelitian oleh Aji Prasetya Wibawa, Hidayah Karima Fithri,Ilham Ari Elbaith Zaeni, Andrew Nafalski (2020) dengan judul "Generating Javanese Stopwords List using K-Means Clustering Algorithm" yang mana menyimpulkan bahwa K-Means berlaku untuk pembuatan daftar stopword bahasa Jawa. Algoritme menunjukkan lokasi stopword berada di cluster pertama dari daftar kata. Namun hasil yang menjanjikan saat ini masih memungkinkan untuk ditingkatkan [3].

Selanjutnya adalah penelitian yang dilakukan oleh Made Arta Purniawan, Gusti Made Arya Sasmita, Putu Agus Eka Pratama (2022) dengan judul "Clustering Berita Menggunakan Algoritma Tf-Idf Dan K-Means Dengan Memanfaatkan Sumber Data Crawling Pada Situs Detik.Com" Penelitian ini menggunakan bahasa pemrograman Python dan library BeautifulSoup4 untuk melakukan crawling dataset pada portal berita online detik.com. Hasil penelitian menunjukkan bahwa metode ini mampu dan handal dalam mengambil data teks dalam jumlah besar dengan total berita yang diperoleh sebanyak 124.509 [4].

Penelitian selanjutnya oleh Fitri Nuraeni, Dewi Tresnawati, Yoga Handoko Agustin, Gisna Fauzian Dermawan (2022) dengan judul "Optimization Of Market Basket Analysis Using Centroid-Based Clustering Algorithm And Fp-Growth Algorithm" hasil dari penelitian ini yaitu aturan asosiasi yang dibentuk dari model k-means dan FP-Growth menunjukkan bahwa cireng chili oil, kopi susu regal, pisang keju, dan vietnam drip sering dibeli bersamaan. Hal ini memungkinkan pemilik toko untuk mempersiapkan persediaan yang mencukupi untuk keempat item tersebut, sehingga dapat memastikan bahwa permintaan pelanggan terpenuhi secara konsisten [5].

Penelitian relevan lainnya adalah yang diteliti oleh Khairunnisa1, Juna Eska, Mustika Fitri Larasati dengan judul "Penggunaan K-Means Clustering Untuk Mengelompokkan Kemampuan Bahasa Inggris Siswa Lembaga Kursus Jason English Course" Kesimpulan yang dapat diambil adalah bahwa penggunaan k-means clustering mampu membagi Student English Skill atau kemampuan bahasa Inggris siswa Jason English Course dengan lebih cepat, jelas, dan akurat. Pengelompokan kemampuan bahasa Inggris siswa menggunakan metode algoritma K-means dilakukan dengan perhitungan hingga iterasi ketiga. Tingkat akurasi yang tercapai sebesar 80% mampu mengelompokkan data pada sampel yang telah di-cluster setelah dilakukan perhitungan secara manual menggunakan Microsoft Excel dan pengujian menggunakan RapidMiner[6].

Penelitian terakhir yaitu penelitian oleh Dewi Sinta Saputri, Guntur Maha Putra, Mustika Fitri Larasati dengen judul "Implementasi Algoritma K-Means Clustering Untuk Desa Tervaksinasi Covid-19 Pada Kecamatan Ujung Padang" Kesimpulan dari penelitian ini adalah bahwa penerapan algoritma k-means clustering dapat mempermudah pengelompokan desa yang telah divaksinasi Covid-19 di wilayah Kecamatan Ujung Padang[7].

Hasil clustering dapat digunakan untuk memahami karakteristik data, mengidentifikasi pola, dan menemukan wawasan baru, tetapi dalam melakukan hal tersebut sulit untuk menentukan jumlah cluster yang tepat dan selalu membutuhkan kompleksitas waktu yang tinggi [8]. Algoritma K-Means dengan jumlah cluster yang disesuaikan secara otomatis dapat menentukan jumlah cluster yang optimal memiliki efek yang dan akurasi yang lebih baik[9]. Selain itu dalam implementasinya algoritma ini memiliki beberapa kelemahan seperti : seringnya terjadi kekosongan cluster adanya outlier dan tingkat akurasi yang rendah [10][11].

Dalam kluster text terdapat beberapa metode yang dapat digunakan salah satunya adalah metode K-Means Clustering. K-Means Clustering adalah suatu metode yang digunakan untuk mengelompokkan data teks ke dalam beberapa kelompok berdasarkan kesamaan fitur yaitu suatu karakteristik yang membedakan suatu data teks dengan data teks yang lain. Dalam K-Means Clustering fitur yang digunakan adalah fitur yang terkandung dalam suatu teks [12].

Metode yang digunakan untuk mencari relevansi antar beberapa dokumen adalah metode TF-IDF [4]. Metode ini cukup mudah dipelajari dan mudah diterapkan untuk permasalahan keakuratan sebuah dokumen. TF-IDF melibatkan teknik tokenisasi, stopwords, dan stemming, serta frekuensi kemunculan kata dalam dokumen [13]. Hal ini menunjukkan betapa pentingnya kata tersebut di dalam sebuah dokumen. Dalam penelitian ini digunakan kamus bahasa Karo sebagai studi kasus.

Dalam kamus karo ini ada beberapa dialek yaitu : Julu, Teruh Deleng, Singalur Lau, Jahe dan Liang Melas. Dialek Julu terletak di Kecamatan Kabanjahe, Simpang Empat, Berastagi dan sekitarnya. Dialek Teruh Deleng di kecamatan Kuta Buluh dan sebagian dari Kecamatan Payung. Singalur Lau berada di Kecamatan Juhar dan Lau Baleng. Dialek Jahe berlaku di Deliserdang-Medan dan sebagain Langkat. Dan yang terakhir yaitu dialek Liang Melas berada di kecamatan Kuta Buluh, Kecamatan Bahorok dan sebagainya [14].

Pada dialek Singalur Lua, Liang Melas dan Teruh Deleng huruf “e” dibaca dengan bunyi “é”. Pada dialek Julu dibaca “ei”. Dialek Jahe dibaca “ai”. Untuk penggunaan kata ulang pada Dialek Liang Melas, Julu, Jahe dan Dialek Teruh Deleng tetap dibaca sebagai kata ulang [15]. Sedangkan pada Dialek Singalur Lau sering sekali bunyinya di singkat “beré-beré” dibaca “beberé”, “udan-udan” dibaca “uudan” dan sebagainya. Pada Dialek Jahe dan Julu huruf “o” yang terletak pada huruf kedua dari akhir dibaca “u”, contohnya “buloh” menjadi “buluh” [14].

Untuk mencari dokumen yang relevan, dilakukan klastering menggunakan metode *K-Means* sedangkan *tf-idf* dapat digunakan untuk mencari dokumen dengan tingkat relevansi yang tinggi. Oleh karena itu, metode ini dapat menjadi solusi untuk yang memerlukan pencarian dokumen yang akurat dan efisien [16].

Tujuan dari penelitian ini untuk melihat berapa presentasi antara dialek menggunakan metode *tf idf* dan mencari cluster menggunakan metode *k-menas clustering*. Hasil analisis ini diharapkan dapat menjelaskan bagaimana informasi dalam data teks dalam bahasa Karo dapat dikelompokkan berdasarkan kemiripan atau kecocokan antar data teks yang ada. Selain itu hasil analisis ini juga akan menunjukkan bagaimana informasi yang terkandung dalam data teks dalam bahasa Karo dapat digolongkan dan diklasifikasikan dengan kategori dan kelompok yang sesuai.

2. METODE PENELITIAN

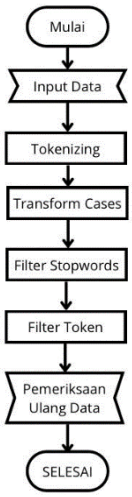
Penelitian ini menggunakan metode kuantitatif, yaitu "Tf-Idf K-Means Clustering" Metode ini melibatkan penghitungan nilai bobot (term frequency-inverse document frequency) untuk masing-masing kata dalam dokumen, dan kemudian mengelompokkan dokumen berdasarkan kesamaan bobot menggunakan algoritma K-Means [17]. Selain itu, penggunaan istilah "klaster" juga menunjukkan pengelompokkan yang dilakukan memiliki sifat kuantitatif.

2.1. Pengumpulan Data

Data yang digunakan pada penelitian ini merupakan hasil dari kata perkata yang disusun menjadi sebuah kalimat berdasarkan kata yang ada pada Kamus Karo Indonesia yang ditulis oleh Darwin Prinst dengan total 702 halaman dan melakukan wawancara sebanyak ± 45 masyarakat yang berada di bebrapa desa di Kabupaten karo, Sumatera Urata. Data yang digunakan pada penelitian ini ±5000 kata.

2.2. Analisi Data Menggunakan *TF-Idf*

Ada beberapa langkah yang digunakan dalam metode ini, untuk lebih jelas bisa dilihat pada *flowchart* dibawah ini:



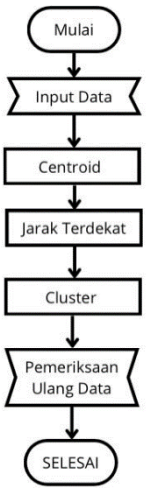
Gambar 1. *Flowchart Tf-Idf*

Penjelasan dari *flowchart* diatas yaitu mulai dari proses *tokenizing* yaitu proses memecah teks atau rangkaian karakter menjadi unit-unit yang lebih kecil disebut “token”, selanjutnya *Transform Cases* yaitu proses mengubah semua karakter huruf dalam teks menjadi huruf kecil atau huruf besar, tanpa mengubah struktur atau makna teks tersebut [18].

Selanjutkan dengan *Filter Stopwords* metode ini bertujuan untuk mencari kata dasar, yaitu penghilangan imbuhan. Imbuhan merupakan bunyi yang ditambahkan pada sebuah kata dasar, baik di awal, pertengahan maupun di akhir kata dengan tujuan untuk membentuk kata baru yang berhubungan dengan kata awal tersebut, contohnya seperti “ter, ber, ke, nya, kan dll”, yang terakhir yaitu *Filter Token* yang bertujuan untuk menghilangkan kata-kata yang terlalu banyak atau terlalu sedikit untuk menyaring token-token tertentu dari teks atau korpus teks [17].

2.3. Analisi Data Menggunakan K-means Clustering

Adapun tahadapan dalam metode ini bisa dilihat dari *flowchat* dibawah :



Gambar 2. *Flowchart k-means Clustering*

Penjelasan dari gambar diatas yaitu yang pertama menentukan Centroid yang bertujuan untuk mewakili pusat-pusat kelompok dalam data dan memudahkan pengelompokan data ke dalam cluster-cluster berdasarkan kedekatan dengan centroid terdekat. Dengan rumus :

$$D(x,y) = \sqrt{\sum_{i=1}^{n}(xi - yi)^2} \quad (1)$$

Keterangan :
D : Data
x : Centroid
y : Atribut
D (xi,yi) : Jarak antara data x1 dan y1

Selanjutnya menentukan jarak terdekat untuk mendapatkan cluster, disini saya memilih untuk membagi menjadi 5 cluster sesuai dengan data yang saya miliki.

3. HASIL DAN PEMBAHASAN

3.1. Implementasi Metode *Tf-Idf*

Yang pertama saya menampilkan proses dari *tokenizing* dapat dilihat dapat gambar dibawah ini:

Tabel 1. *Tokenising*

Data	Hasil Tokenizing
seh kel jilena kutaku. keliling kuta teridah denga seh buena batang kayu. sora bengkala erpagi-pagi datas uruk nari jelas denga terbegi seh kuta. sora sei-sei (sejenis sue-sue galang) rikut pe sora-perik-perik sierbage-bage nambahi riahna erpagi-pagi.	seh-kel-jilena-kutaku-keliling-kuta teridah-denga-seh-buena-batang-kayu-sora-bengkala-erpagi-pagi-datas-uruk-nari-jelas-denga-terbegi-seh-kuta-sora-sei-sei-(sejenis-sue-sue-galang)-rikut-pe-sora-perik-periksierbage-bage-nambahi-riahna-erpagi-pagi.

Sumber : Sapo Lau Limpek, Julianus Limbeng

Gambar 1. Terlihat pemecahan kata dari ±2000 kata yang digunakan dalam penelitian ini. Selanjutnya kita dapat melihat proses *tokenizing* dengan menggunakan aplikasi *rapidminer:*

Gambar 3. Proses Tokenize

Dari gambar diatas dapat dilihat pemanggilan *tokenizing* untuk mulai proses pengolahan data yang pertama, dan hasil output dari pemanggilan tersebut dapat kita lihat pada gambar dibawah ini:

Setelah dijalankan muncul Total *Occurences* (kemunculan kata dalam Paragraf) dan *Document Occurences* (kemunculan kata pada keseluruhan dokumen). Selanjutnya kita lanjutkan untuk proses yang kedua seperti dibawah ini:

Word	Attribute Name	Total Occurences	Document Occurences
Adi	Adi	11	11
Agama	Agama	1	1
Agressi	Agressi	1	1
Al	Al	1	1
Alkitab	Alkitab	1	1
Amburken	Amburken	1	1
Amerika	Amerika	1	1
Amin	Amin	4	4
Aminna	Aminna	2	2
Amsterda...	Amsterdam	3	3
Angin	Angin	1	1
Antariksa	Antariksa	1	1

Gambar 4. Output Proses Tokenize

Tabel 2. Transform Cases

Hasil Tokenizing	Hasil Transform Cases
Seh kel jilena kutaku. Keliling kuta teridah denga seh buena batang kayu. Sora bengkala erpagi-pagi datas uruk nari jelas denga terbegi seh kuta. Sora sei-sei (sejenis sue-sue galang) rikut pe sora-perik-perik sierbage-bage nambahi riahna erpagi-pagi.	seh kel jilena kutaku. keliling kuta teridah denga seh buena batang kayu. sora bengkala erpagi-pagi datas uruk nari jelas denga terbegi seh kuta. sora sei-sei (sejenis sue-sue galang) rikut pe sora-perik-perik sierbage-bage nambahi riahna erpagi-pagi.

Sumber : Sapo Lau Limpek, Julianus Limbeng

Dalam *Transform Cases*, setiap huruf besar diubah menjadi huruf kecil, sehingga tidak ada perbedaan antara huruf besar dan kecil dalam teks. Selanjutnya kita lakukan di *rapidminer:*

Gambar 5. Proses *Transform Cases*

Dari gambar diatas terlihat sudah ada dua proses, selanjutnya kita lihat proses outputmya:

Word	Attribute Name	Total Occurences	Document Occurences
a	a	8	8
abit	abit	1	1
ade	ade	1	1
adi	adi	27	18
adon	adon	1	1
agama	agama	1	1
agressi	agressi	1	1
ah	ah	2	2
ahli	ahli	1	1
ajarina	ajarina	1	1
akan	akan	1	1
akao	akao	6	6

Gambar 6. *Output* Proses *Cases*

Setelah dijalankan muncul Total *Occurences* (kemunculan kata dalam Paragraf) dan *Document Occurences* (kemunculan kata pada keseluruhan dokumen). Selanjutnya mulai proses yang ketiga yaitu :

Tabel 3. *Filter Stopwords*

Hasil Transform Cases	Hasil Stopwords
seh kel jilena kutaku. keliling kuta teridah denga seh buena batang kayu. sora bengkala erpagi-pagi datas uruk nari jelas denga terbegi seh kuta.	seh kel jile kuta keliling kuta idah denga seh bue batang kayu sora bengkala pagi datas nari jelas denga begi seh kuta sora sei-sei (jenis sue-sue

sora sei-sei (sejenis sue-sue galang) rikut pe sora-perik-perik sierbage-bage nambahi riahna erpagi-pagi.	galang) ikut pe sora perik erbage nambah riah pagi.

Sumber : Sapo Lau Limpek, Julianus Limbeng

Pada table 3 diatas bertujuan untuk mencari kata dasar, yaitu penghilangan imbuhan. Imbuhan merupakan bunyi yang ditambahkan pada sebuah kata dasar, baik di awal, pertengahan maupun di akhir kata. Selanjutnya proses pada *rapidminer :*

Gambar 7. Proses *Filter Stopword*

Pada gambar 7 terlihat sudah 3 proses yang dilakukan, untuk selanjutnya kita akan melihat hasil output dari gambar diatas, sebagai berikut:

Word	Attribute Name	Total Occurences	Document Occurences
a	a	8	8
abit	abit	1	1
ade	ade	1	1
adi	adi	27	18
adon	adon	1	1
agama	agama	1	1
agressi	agressi	1	1
ah	ah	2	2
ahli	ahli	1	1
ajarina	ajarina	1	1
akan	akan	1	1
akap	akap	6	6

Gambar 8. *Output Filter Stopword*

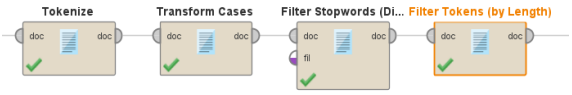
Setelah dijalankan muncul Total *Occurences* (kemunculan kata dalam Paragraf) dan *Document Occurences* (kemunculan kata pada keseluruhan dokumen) dan dilanjutkan untuk proses terakhir seperti gambari dibawah ini:

Tabel 4 Penerjemahan

Hasil Stopword	Hasil Penerjemahan
seh kel jile kuta keliling kuta idah denga seh bue batang kayu sora bengkala pagi datas nari jelas denga begi seh kuta sora sei-sei (jenis sue-sue galang) ikut pe sora perik erbage nambah riah pagi.	Cantik sekali kampung keliling kampung lihat masih sampai banyak batang kayu suara kera pagi atas dari jelas masih dengar sampai kampung suara gareng ikut juga suara burung macam tambah seru pagi

Sumber : Sapo Lau Limpek, Julianus Limbeng

Tabel 4 merupakan proses mengubah teks atau ucapan dari bahasa karo ke bahasa indonesia dengan menjaga arti, makna, dan pesan yang sama sebisa mungkin. Selanjutnya kita akan melakukan proses *filter token* menggunakan aplikasi *rapidminer* sebagai berikut :



Gambar 9. Proses Filter Token

Dari gambar 9 diatas terlihat bahwa sudah ada empat proses dari *tf-idf* dimana proses ini merupakan tahap terakhir pada metode ini dan dapat dilihat pada gambar dibawah ini:

Word	Attribute Name	Total Occuren...	Document Occurences
agama	agama	1	1
agressi	agressi	1	1
ajarina	ajarina	1	1
akhirnya	akhirnya	1	1
akrab	akrab	1	1
akitab	akitab	1	1
amburken	amburken	1	1
amerika	amerika	1	1
anajar	anajar	1	1
anakndu	anakndu	1	1
anaknya	anaknya	1	1
analisa	analisa	1	1

Gambar 10. Output Filter Token

Setelah dijalankan muncul Total *Occurences* (kemunculan kata dalam Paragraf) dan *Document Occurences* (kemunculan kata pada keseluruhan dokumen). Pada table dibawah kita akan melihat 100 kata yang sudah kita proses menggunakan metode *tf-idf*, tapi dibawah ini saya hanya akan menampilkan 1 – 15 kata. Kata yang dimaksud sebagai berikut:

Gambar 11. Hasil Wordcloud Tf Idf

Dari hasil gambar 11 terlihat jelas bahwa kata yang sering muncul merupakan gambar yang katanya paling besar seperti kata “emaka,enggo,dll“.

3.2. Contoh Sub-Bab Kedua

Menggunakan metode k-means clustering, dari data berikut:

NO	DIALEK									
	JAHE		JULU		SINGALUR LAU		TERUH DELENG		LIANG MELAS	
	WORD	IN DOCUMENTS	WORD	IN DOCUMENTS	WORD	IN DOCUMENTS	WORD	IN DOCUMENTS	WORD	IN DOCUMENTS
TOTAL	489	814	489	814	468	763	468	763	468	763

Gambar 12. *K-means Clustering*

Tabel 5. Hasil Tf Idf Menggunakan *RapitMiner*

No	Word	In Documents	Total
1.	emaka	27	36
2.	enggo	23	36
3.	kalak	21	28
4.	doktor	21	25
5.	kenca	17	23
6.	lalang	20	21
7.	rhoda	14	20
8.	belanda	13	19
9.	denga	13	17
10.	janah	17	17
11.	bagepe	13	16
12.	tahun	11	16
13.	kerina	11	15
14.	mesin	10	15
15.	banci	11	14

Dari tabel diatas dapat dilihat bahwa seperti kata "emaka" in documents sebagai 27 dari 36 total dokumen yang ada dan seperti itu seterusnya. Pada tahap selanjutnya kita akan melihat gambar dari hasil *Wordcloud* pada aplikasi *rapidminer*, sebagai berikut:

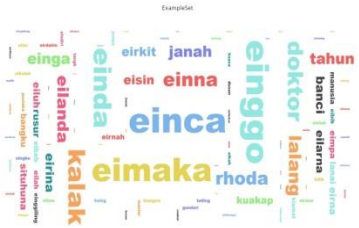
Tabel diatas memperlihatkan total dari perdialek menggunakan Rapidminer.

Persentase dialek jahe
Di dokumen = 489 : 1303 x 100 = 37, 5 %
Total = 814 : 1303 x 100 = 62,5%
Perentase dialek julu
Di dokumen = 489 : 1303 x 100 = 37,5 %
Total = 814 : 1303 x 100 = 62,5 %
Pesentase dialek singalur lau
Di dokumen = 468 : 1231 x 100 = 38 %
Total = 763 : 1231 x 100 = 62 %
Persentase dialek teruh deleng
Di dokumen = 468 : 1231 x 100 = 38 %
Total = 763 : 1231 x 100 = 62 %
Prsentase dialek liang melas
Di dokumen = 468 : 1231 x 100 = 38 %
Total = 763 : 1231 x 100 = 62 %

Setelah menemukan hasil persen perdialek, selanjutnya kita akan melihat *wordcloud* menggunkaan *rapidminer* agar terlihat secara jelas perbedaannya, sebagai berikut :

Gambar 13. Hasil Wordcloud Dialek Jahe

Dari gambar 13 diatas terlihat bahwa pada dialek jahe kata yang paling tinggi nilainya adalah kata "ainca, ainggo, dll" , selanjutnya kita akan melihat hasil untuk dialek julu sebagai berikut :



Gambar 14. Hasil Wordcloud Dialek Julu

Pada gambar diatas kata paling sering muncul atau yang memiliki nilai paling banyak adalah kata "einca, eimaka,dll" dilanjutkan dengan dialeh singalur lau sebagai berikut :

Gambar 15. Hasil Wordcloud Dialek Singalur Lau

Untuk gambir 15 kata yang muncul paling banyak adalah kata "enggo, emaka, dll" pada *wordcloud* selanjutnya ada dialek teruh deleng sebagai berikut :

Gambar 16. Hasil Wordcloud Dialek Teruh Deleng

Pada dialek teruh deleng diatas kata yang paling sering mencul adalah kata "emaka,enggo,dll" *wordcloud* yang ke lima atau terakhir ada dialek liang melas, dapat dilihat pada gambar dibawah ini :

Gambar 17. Hasil Wordcloud Dialek Liang Melas

Dari gambar 17 terlihat tidak terlihat perbedaan yang signifikan antara dialek singalur lau karena kata yang paling besar adalah kata "enggo, emaka, dll". Untuk lebih lanjut kita akan membuat titik pusat centroid untuk mencari cluster dari dialek dialek diatas, untuk lebih jelasnya dapat dilihat pda table dibawah ini :

Tabel 6. Titik pusat Centroid

Centroid	Jahe	Julu	Singalur Lau	Teruh Deleng	Liang Melas
C1	29	29	25	25	25
C2	15	15	14	14	14
C3	7	7	7	7	7
C4	5	6	5	5	5
C5	20	21	19	19	19

Dari tabel 6 selanjutnya kita akan mengolah data yang diatas menggunakan rumus seperti dibawah ini:

$$D(x,y) = \sqrt{\sum_{i=1}^{n}(xi - yi)^2} \qquad (2)$$

Jarak antara No 1 dengan C1:
D(C1, No 1) = √((29 - 38)^2 + (29 - 38)^2 + (25 - 36)^2 + (25 - 36)^2 + (25 - 36)^2)
D(C1, No 1) = √((-9)^2 + (-9)^2 + (-11)^2 + (-11)^2 + (-11)^2)
D(C1, No 1) = √(81 + 81 + 121 + 121 + 121)
D(C1, No 1) = √(525)
D(C1, No 1) = **22.92**

Jarak antara No 1 dengan C2:
D(C2, No 1) = √((15 - 38)^2 + (15 - 38)^2 + (14 - 36)^2 + (14 - 36)^2 + (14 - 36)^2)
D(C2, No 1) = √((-23)^2 + (-23)^2 + (-22)^2 + (-22)^2 + (-22)^2)
D(C2, No 1) = √(529 + 529 + 484 + 484 + 484)
D(C2, No 1) = √(2506)
D(C2, No 1) = **50.06**

Jarak antara No 1 dengan C3:
D(C3, No 1) = √((7 - 38)^2 + (7 - 38)^2 + (7 - 36)^2 + (7 - 36)^2 + (7 - 36)^2)
D(C3, No 1) = √((-31)^2 + (-31)^2 + (-29)^2 + (-29)^2 + (-29)^2)
D(C3, No 1) = √(961 + 961 + 841 + 841 + 841)
D(C3, No 1) = √(4445)
D(C3, No 1) = **66.69**

Jarak antara No 1 dengan C4:
D(C4, No 1) = √((5 - 38)^2 + (6 - 38)^2 + (5 - 36)^2 + (5 - 36)^2 + (5 - 36)^2)
D(C4, No 1) = √((-33)^2 + (-32)^2 + (-31)^2 + (-31)^2 + (-31)^2)
D(C4, No 1) = √(1089 + 1024 + 961 + 961 + 961)
D(C4, No 1) = √(5096)
D(C4, No 1) = **71.38**

Jarak antara No 1 dengan C5:
D(C5, No 1) = √((20 - 38)^2 + (21 - 38)^2 + (19 - 36)^2 + (19 - 36)^2 + (19 - 36)^2)
D(C5, No 1) = √((-18)^2 + (-17)^2 + (-17)^2 + (-17)^2 + (-17)^2)
D(C5, No 1) = √(324 + 289 + 289 + 289 + 289)
D(C5, No 1) = √(1480)
D(C5, No 1) = **38.48**

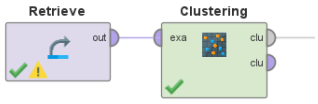
Setelah menemukan nilai dari *centroid* dari masing masing dialek kita dapat menguji kembali menggunakan aplikasi *rapidminer* apakah hasilnya sudah sesuai apa belum, dapat dilihat pada gambar dibawah ini :

Row No.	Jahe	Julu	Singalur Lau	Teruh Deleng	Liang Melas
1	22.913	50.100	66.671	70.682	38.471
2	22.159	49.194	65.750	69.771	37.577
3	11.180	38.341	54.818	58.720	26.907
4	0	27.477	44.045	48.010	15.906
5	5.385	22.627	39.154	43.081	11.045
6	10.630	17.205	33.719	37.656	5.657
7	13.711	13.892	30.463	34.453	2.236
8	15.906	11.662	28.231	32.218	0
9	19.313	8.246	24.759	28.705	3.742

ExampleSet (100 examples, 0 special attributes, 5 regular attributes)

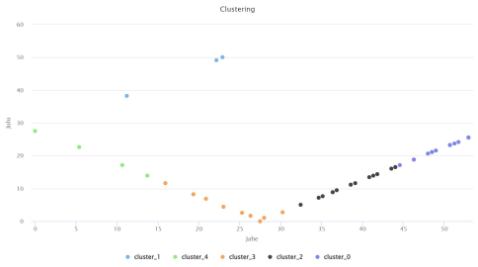
Gambar 18. Hasil *Centroid*

Setelah menemukan *centroid* tahap selanjutnya kita akan menetukan cluster dari data diatas, adapun cluster yang dilakukan menggunakan aplikasi *rapidminer* seperti gambar dibawah ini:



Gambar 19. Proses Clustering

Setelah proses pada gambar 19 selesai maka muncullah cluster 0: 68 items, cluster 1: 3 items, cluster 2: 15 items, cluster 3:10 items, cluster 4:4 items dengan total sempel 100 items. Untuk lebih jelas dapat dilihat pada gambar dibawah ini :

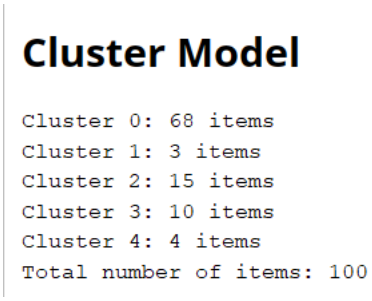


Gambar 20. grafik cluster pada rapidminer

Hasil dari gambar 20 diatas menampilkan hasil akhir dari proses metode k-means yang telah kita lakukan.

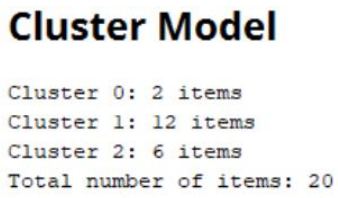
4. DISKUSI

Maka dari penelitian ini di dapat rekap cluster dari 100 kata yang sudah melewati tahap tahap seperti, *tokenizing, transform cases, filter stopword, filter token, menentukan centroid*,jarak terdekat yang menghasilkan *cluster*. Untuk lebih jelas dapat dilihat pada gambar dibawah ini:



Gambar 21. Hasil *Cluster* Dialek Pada *Rapidminer*

Dari dataset yang awalnya ±5000 kata dapat kita perkecil sedemikian rupa agar mudah dipahami. Disini saya menggunakan aplikasi *rapidminer* untuk mendapatkan hasil cluster, sama hal nya dengan penelitian dengan judul "implementation of the k-means clustering algorithm for the covid-19 vaccinated village in the ujung padang sub-district"[7] sama sama menggunakan metode cluster dan menggunakan aplikasi *rapidminer*. Dapat dilihat pada gambar di bawah ini.



Gambar 22. Model cluster tervaksinasi covid-19 pada rapidminer

Pada gambar diatas dapat dilihat bahwa ada 20 total dari *cluster* yang dibagi dari cluster 0 – *cluster* 2. Ini menandakan bahwa melakukan penelitian dengan data yang tidak sedikit dapat di lakukan dengan metode *k-means clustering* agar hasil dari data yang begitu banyak dapat kita lihat berdasarkan *cluster* atau kemiripan antar kata.

Menurut saya hasil yang ada diatas sudah cukup jelas bahwa metode *tf-idf* dan *k-means clustering* dapat membantu kita untuk mengelolah data yang besar menggunakan aplikasi *rapidminer*.

5. KESIMPULAN

Dari hasil penelitian yang dilakukan terdapat 5 cluster yaitu dialek jahe, julu, singalur lau, teruh deleng dan liang melas. Dari hasil perhitungan/pembobotan tf-idf terdapat persentase untuk dialek jahe 37,5% untuk jumlah kemunculan kata dan 62,5% total seluruh kata didokumen. Selanjutnya untuk dialek julu 37,5% untuk jumlah kemunculan kata dan 62,5% total seluruh kata didokumen. Dialek singalur lau 38% jumlah kemunculan kata dan 62% total seluruh kata didokumen. Dialek singteruh deleng lau 38% jumlah kemunculan kata dan 62% total seluruh kata didokumen. Dialek liang melas 38% jumlah kemunculan kata dan 62% total seluruh kata didokumen. Maka dapat disimpulkan bahwa antara dialek jahe dan dialek julu identik, sedangkan antara dialek singalur lau, dialek teruh deleng dan liang melas juga indentik.

Pada cluster model menggunakan RapidMiner terdapat cluster 0: 68 items, cluster 1: 3 items, cluster 2: 15 items, cluster 3:10 items, cluster 4:4 items dengan total sempel 100 items.

DAFTAR PUSTAKA

[1] R. A. Kurniawan, M. S. Hasibuan, P. Piramida, And R. S. Ramadhan, "Penerapan Algoritma K-Means Untuk Clustering Tempat Makan Di Batubara," *J. Comput. Sci. Informatics Eng.*, Vol. 01, No. 1, Pp. 10–18, 2022, Doi: 10.55537/Cosie.V1i1.27.

[2] A. I. Abdullah, E. Winarko, And A. Musdholifah, "Metode Boost-K-Means Untuk Clustering Puskesmas Berdasarkan Persentase Bayi Yang Diimunisasi," *Jrst (Jurnal Ris. Sains Dan Teknol.*, Vol. 4, No. 2, P. 89, 2020, Doi: 10.30595/Jrst.V4i2.7546.

[3] A. P. Wibawa, H. K. Fithri, I. A. E. Zaeni, And A. Nafalski, "Generating Javanese Stopwords List Using K-Means Clustering Algorithm," *Knowl. Eng. Data Sci.*, Vol. 3, No. 2, P. 106, 2020, Doi: 10.17977/Um018v3i22020p106-111.

[4] M. Purniawan Arta, G. Sasnita Arya, And P. Pratama Eka Agus, "Clustering Berita Menggunakan Algoritma Tf-Idf Dan K-Means Dengan Memanfaatkan Sumber Data Crawling Pada Situs Detik.Com.*Jurnal Ilmiah Teknologi Dan Komputer*. 2022 "

[5] F. Nuraeni, D. Tresnawati, Y. Handoko Agustin, And G. Fauzi, "Optimization Of Market Basket Analysis Using Centroid-Based Clustering Algorithm And Fp-Growth Algorithm," *J. Tek. Inform.*, Vol. 3, No. 6, Pp. 1581–1590, 2022, Doi: 10.20884/1.Jutif.2022.3.6.399.

[6] J. Eska, M. Fitri Larasati, P. Studi Sistem Informasi, And S. Tinggi Manajemen Informatika Dan Komputer Royal Kisaran, "Application Of K-Means Clustering Method To Cluster Students' English Skill Jason English Course," *J. Tek. Inform.*, Vol. 3, No. 3, Pp. 479–485, 2022, [Online]. Available: Https://Doi.Org/10.20884/1.Jutif.2022.3.3.167.

[7] D. S. Saputri, G. M. Putra, And M. F. Larasati, "Implementation Of The K-Means Clustering Algorithm For The Covid-19 Vaccinated Village In The Ujung Padang Sub-District Implementasi Algoritma K-Means Clustering Untuk Desa Tervaksinasi Covid-19 Pada Kecamatan Ujung Padang," Vol. 3, No. 2, Pp. 261–267, 2022 *Jurnal Teknik Informatika*.

[8] R. Rosmini, A. Fadlil, And S. Sunardi,

"Implementasi Metode K-Means Dalam Pemetaan Kelompok Mahasiswa Melalui Data Aktivitas Kuliah," *It J. Res. Dev.*, Vol. 3, No. 1, Pp. 22–31, 2018, Doi: 10.25299/Itjrd.2019.Vol3(1).1773.

[9] S. Fitriani, "Implementasi Data Mining Dalam Pengelompokkan Minat Baca Pengunjung Pada Perpustakaan Stmik Triguna Dharma Medanmenggunakan Metode K-Means," *J. Cybertech*, 2020.

[10] A. Iriansyah And M. F. Y. Gafallo, "Budaya Partisipasi Dan Resistensi Komunitas Keagamaan Di Media Sosial Participatory Culture And Resistance Of Religious Communities Kementerian Agama Republik Indonesia Merilis Tayangan Live Telekonferensi Sidang Isbat Pada," Pp. 17–30, 2022, Doi: 10.17933/J. Kominfo 2022.4780.

[11] P. A. E. P. Made Arta Purniawan, Gusti Made Arya Sasmita, "Clustering Berita Menggunakan Algoritma Tf-Idf Dan K-Means Dengan Memanfaatkan Sumber Data Crawling Pada Situs Detik.Com," *J. Ilm. Teknol. Dan Komput. Vol.*, Vol. 3, No. 1, 2022.

[12] E. Ikhsan, "Penerapan K-Means Clustering Dari Log Data Moodle Untuk Menentukan Perilaku Peserta Pada Pembelajaran Daring," *J.sistemasi*, Vol. 10, No. 2, P. 414, 2021, Doi: 10.32520/Stmsi.V10i2.1285.

[13] M. Darwis, G. T. Pranoto, And Y. E. Wicaksana, "Implementation Of Tf-Idf Algorithm And K-Mean Clustering Method To Predict Words Or Topics On Twitter," Vol. 03, No. 02, Pp. 49–55, *Jurnal Informatika dan Sains.* 2020.

[14] D. Prinst, *Kamus Karo Indonesia*, 4th Ed. Medan: Bina Media Printis, 2014.

[15] A. Siregar Samin, P. Sukapiring, S. Tarigan, M. Sembiring Cikappen, And Zulkifly, Eds., *Kamus Bahasa Karo-Indonesia*. Jakarta: Balai Pustaka, 2001.

[16] M. A. Rofiqi, A. C. Fauzan, A. P. Agustin, And A. A. Saputra, "Implementasi Term-Frequency Inverse Document Frequency (Tf-Idf) Untuk Mencari Relevansi Dokumen Berdasarkan Query," *Ilk. J. Comput. Sci. Appl. Informatics*, Vol. 1, No. 2, Pp. 58–64, 2019, Doi: 10.28926/Ilkomnika.V1i2.18.

[17] Y. Muhammad Darwis , Gatot Tri Pranoto , Yusuf Eka Wicaksana, "Implementation Of Tf-Idf Algorithm And K-Mean Clustering Method To Predict Words Or Topics On Twitter," *Jisa (Jurnal Inform. Dan Sains*, Vol. 03, Pp. 49–55, 2020.

[18] R. T. Wahyuni, D. Prastiyanto, And E. Supraptono, "Penerapan Algoritma Cosine Similarity Dan Pembobotan Tf-Idf Pada Sistem Klasifikasi Dokumen Skripsi," *J. Tek. Elektro Univ. Negeri Semarang*, Vol. 9, No. 1, Pp. 18–23, 2019, [Online]. Available: Https://Journal.Unnes.Ac.Id/Nju/Index.Php/Jte/Article/Download/10955/6659.