

Rancang Bangun Aplikasi Korpus Bahasa Dayak Ma'anyan

Camila Aprillinsa¹⁾, Ariesta Lestari²⁾, Nahumi Nugrahaningsih³⁾, Septian Geges⁴⁾, Efrans Christian⁵⁾

¹⁾²⁾³⁾⁴⁾⁵⁾ Teknik Informatika, Fakultas Teknik, Universitas Palangka Raya
Jl. Yos Sudarso, Kota Palangka Raya, Kalimantan Tengah

¹⁾camilaaprilinsa@gmail.com

²⁾ariesta@it.upr.ac.id

³⁾nahumi@it.upr.ac.id

⁴⁾septian.geges@it.upr.ac.id

⁵⁾efrans@it.upr.ac.id

Abstrak

Seiring perkembangan zaman, banyak bahasa daerah terancam punah, hal ini disebabkan karena semakin berkurangnya penutur bahasa daerah asli. Oleh karena itu diperlukan langkah-langkah konkret untuk menjaga dan melestarikan kekayaan bahasa daerah, salah satunya adalah pembuatan Korpus Bahasa Daerah. Penelitian ini bertujuan merancang dan membangun Aplikasi Korpus Bahasa Dayak Ma'anyan untuk mendokumentasikan kosakata secara digital dan memfasilitasi pembelajaran bahasa tersebut.

Dalam merancang dan membangun aplikasi ini menggunakan 2 metode, yaitu metode *Waterfall* sebagai metode pengembangan perangkat lunak dan metode *Knowledge Discovery in Database Process* (KDD) untuk proses pembuatan korpus. Sumber data utama yang digunakan adalah teks Alkitab dalam Bahasa Dayak Ma'anyan.

Hasil penelitian ini berupa korpus digital yang dapat diakses oleh publik, yang dapat digunakan untuk mendukung pembelajaran dan pelestarian bahasa Dayak Ma'anyan. Metode *Text Mining* yang diterapkan pada korpus memungkinkan analisis yang lebih mendalam terkait penggunaan kata, frekuensi, dan pola kalimat. Fitur frekuensi, konkordansi, kalimat dan N-gram dalam korpus ini juga membantu dalam menilai kata-kata penting serta memahami konteks penggunaannya. Dengan demikian, aplikasi ini berkontribusi terhadap pengembangan linguistik dan penerjemahan bahasa alami, serta diharapkan dapat menarik minat masyarakat dan generasi muda untuk mempelajari dan melestarikan bahasa Dayak Ma'anyan.

Kata Kunci: Korpus, *Knowledge Discovery in Databases* (KDD), *Waterfall*

Abstract

Along with the development of the times, many regional languages are threatened with extinction, this is due to the decreasing number of native speakers of the regional language. Therefore, concrete steps are needed to maintain and preserve the richness of regional languages, one of which is the creation of a Regional Language Corpus. This study aims to design and build a Dayak Ma'anyan Language Corpus Application to digitally document vocabulary and facilitate language learning.

In designing and building this application, 2 methods are used, namely the *Waterfall* method as a software development method and the *Knowledge Discovery in Database Process* (KDD) method for the corpus creation process. The main data source used is the Bible text in the Dayak Ma'anyan language.

The results of this study are in the form of a digital corpus that can be accessed by the public, which can be used to support learning and preservation of the Dayak Ma'anyan language. The *Text Mining* method applied to the corpus allows for a more in-depth analysis of word usage, frequency, and sentence patterns. The frequency, concordance, sentence and N-gram features in this corpus also help in assessing important words and understanding the context of their use. Thus, this application contributes to the development of linguistics and natural language

translation, and is expected to attract the interest of the community and the younger generation to learn and preserve the Dayak Ma'anyan language.

Keywords: Corpus, Knowledge Discovery in Databases (KDD), Waterfall

1. PENDAHULUAN

Bahasa daerah atau bahasa regional adalah bahasa yang dituturkan secara turun-temurun di suatu wilayah dalam sebuah negara berdaulat, yaitu di suatu daerah asli, negara bagian federal, provinsi, atau teritori yang lebih luas. Keberadaan sebuah bahasa lokal atau bahasa daerah sangat erat dengan eksistensi suku bangsa yang melahirkan dan menggunakan bahasa tersebut. Bahasa menjadi unsur pendukung utama tradisi dan adat istiadat. Dewasa ini, sebagai dampak dari pengaruh perubahan dan perkembangan zaman yang terjadi pada saat ini keberadaan bahasa daerah mulai terancam pudar/punah. Badan Pengembangan dan Pembinaan Bahasa (Badan Bahasa) melalui Kelompok Kepakaran dan Layanan Profesional (KKLP) Pelindungan Bahasa dan Sastra melakukan kajian vitalitas bahasa dari tahun 2011 sampai 2019. Dari 718 bahasa yang dipetakan, baru 94 bahasa daerah atau sekitar 10 persen yang dikaji vitalitas bahasanya. Hasil kajian itu menunjukkan bahwa 8 bahasa dikategorikan punah, 5 bahasa kritis, 24 bahasa terancam punah, 12 bahasa mengalami kemunduran, 24 bahasa dalam kondisi rentan (stabil, tetapi terancam punah), dan 21 bahasa berstatus aman. Oleh karena itu diperlukan langkah-langkah konkret untuk menjaga dan melestarikan kekayaan bahasa daerah. Salah satu upaya penting dalam melestarikan dan memahami bahasa daerah adalah melalui pembuatan korpus bahasa daerah. Di sinilah *text mining* mengambil bagian, yaitu mengolah data teks yang telah di kumpulkan. *Text mining* menganalisis *text* yang tidak terstruktur yang terkait satu sama lain dan yang berhubungan dengan prinsip dan aturan lain. Hasil yang diharapkan adalah pemahaman baru yang yang tidak diketahui dan belum jelas sebelumnya.

Berdasarkan latar belakang di atas penulis ingin membuat Rancang Bangun Aplikasi Korpus Bahasa Dayak Ma'anyan dengan harapan dapat menjadi salah satu alternatif atau solusi bagi pemerintah untuk mengembangkan Bahasa Dayak Ma'anyan dan dapat menarik minat masyarakat untuk mengajarkan bahasa Bahasa Dayak Ma'anyan untuk generasi-generasi muda di Kalimantan, khususnya KalimantanTengah.

2. TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Dalam melakukan penelitian ini, terdapat beberapa penelitian terkait yang telah dilakukan sebagai referensi, pada penelitian, yaitu :

Table 1. Tinjauan Pustaka

Judul	Isi Jurnal
"Pengaruh Penambahan Korpus Paralel pada Mesin Penerjemah Statistik Bahasa Indonesia ke Bahasa Lampung Dialek Nyo". [1]	Penelitian ini bertujuan untuk membangun model penerjemahan bahasa Indonesia ke dalam dialek bahasa Lampung Nyo dengan pendekatan Mesin Statistik Terjemahan. Penambahan jumlah korpus paralel berdampak pada peningkatan akurasi hasil terjemahan Bahasa Indonesia ke Bahasa Lampung Dialek Nyo berdasarkan nilai BLEU. Hasil penelitian menunjukkan bahwa nilai akurasi saat menggunakan 1000 kalimat, 2000 kalimat dan 3000 kalimat merupakan nilai akurasi evaluasi bilingual yang diteliti berturut-turut yaitu 25.91%, 43.18% dan 45.26%.
"Penyusunan Model Korpus Al-Qur'an Digital".[2]	Penelitian ini bertujuan ntuk membangun sebuah model file korpus Al-Qur'an digital. Hasil penelitian ini adalah format isi teks digital berbahasa Arab yang telah diolah berisi karakter huruf sepenuhnya, tanpa menyertakan tanda-tanda diakritik (harakat) yang dapat mengganggu teknis

	pengolahan file dan Secara khusus, model file digital semacam ini dapat digunakan oleh para peneliti pemula yang ingin menggunakan Al-Qur'an sebagai korpus data penelitiannya.
“Analisa Topik Pendidikan dalam Al-Quran dengan Pendekatan <i>Text Mining</i> ”. [3]	Penelitian ini berfokus pada perancangan untuk mengelompokan data berita ke suatu kategori dengan menggunakan <i>Clustering</i> . Hasil yang diperoleh dari penelitian ini adalah pertama kata kunci “cipta” 3,05 %, kemudian “Ingat” 2,25 %, selanjutnya “Ajar” 1,96 %, lalu “Lihat” 0,82 %, terakhir “Dengar” 0,62 % serta “Bicara” 0,34% dengan total bobot kata sebanyak 3.516 kata yang di filter. Hasil persentase keseluruhan adalah 9,04 % dari total 38.761 kata bersih dalam Al-Quran.

Dengan merujuk pada penelitian-penelitian diatas, penelitian ini bertujuan untuk merancang dan membangun sebuah Aplikasi Korpus Bahasa Dayak Ma'anyan.

2.2 Korpus

Menurut Tony McEnery dan Andrew Hardie (2011). Dalam bukunya "Corpus Linguistics: Method, Theory, and Practice," mereka mendefinisikan korpus sebagai "koleksi besar teks elektronik yang dapat digunakan oleh peneliti linguistik untuk menyelidiki aspek-aspek bahasa.". [4]

2.3 Metode Pengembangan Perangkat Lunak Waterfall

Model SDLC air terjun (waterfall) sering disebut juga model sekuensial linear (sequential linear) atau alur hidup klasik (classic life cycle). Model air terjun menyediakan pendekatan alur hidup perangkat lunak secara sekuensial atau terurut dimulai dari analisis, desain, pengkodean, pengujian, dan tahap pendukung (support) (Rosa A. S. dan M. Shalahuddin, 2013). [5]

2.4 Text Mining

Text mining dapat didefinisikan sebagai penemuan informasi baru dan tidak diketahui sebelumnya oleh komputer, yang secara otomatis mengekstrak informasi dari sumber-sumber teks tak terstruktur yang berbeda. Kunci dari proses ini adalah menggabungkan informasi yang berhasil diekstraksi dari berbagai sumber. [6]

2.5 Frekuensi

Frekuensi atau kekerapan pada linguistik korpus mengacu pada jumlah kemunculan suatu kata dalam sebuah korpus atau teks.

Metode *Term Frequency-Inverse Document Frequency* (TF-IDF) merupakan suatu cara untuk memberikan bobot hubungan suatu kata (term) terhadap dokumen. Metode ini menggabungkan dua konsep untuk perhitungan bobot, yaitu frekuensi kemunculan sebuah kata di dalam sebuah dokumen tertentu dan inverse frekuensi dokumen yang mengandung kata tersebut. Frekuensi kemunculan kata di dalam dokumen yang diberikan menunjukkan seberapa penting kata itu di dalam dokumen tersebut. Frekuensi dokumen yang mengandung kata tersebut menunjukkan seberapa umum kata tersebut. Sehingga bobot hubungan antara sebuah kata dan sebuah dokumen akan tinggi apabila frekuensi kata tersebut tinggi di dalam dokumen dan frekuensi keseluruhan dokumen yang mengandung kata tersebut yang rendah pada kumpulan dokumen.[7]

2.6 Konkordansi

Konkordansi merupakan daftar atau urutan contoh-contoh dari kata, bagian dari kata atau kombinasi kata yang berada dalam konteksnya yang diambil dari korpus teks (Baker, 2012).

2.7 N-Gram

Sebuah N-Gram mewakili kelompok kata atau karakter n (ditunjukkan sebagai gram menunjukkan tata bahasa) yang mengikuti satu sama lain. N-Gram dapat digunakan untuk memprediksi kata berikutnya berdasarkan N -1 kata sebelumnya. N-gram banyak

digunakan dalam penambangan data dan tugas pengolah kata. Misalnya Jika "Saya bekerja di Kantor" berarti $n = 4$ (jumlah kata dari kalimat). Kata tersebut digunakan untuk membuat indeks tentang seberapa sering kata-kata mengikuti satu sama lain (Mulyani, Muhamad, and Cahyanto 2021). [8]

2.8 Knowledge Discovery in Databases (KDD)

Knowledge Discovery in Databases (KDD) adalah proses yang melibatkan ekstraksi informasi berguna, yang sebelumnya tidak diketahui, dan berpotensi berharga dari kumpulan data besar. Tahapan KDD ini terdiri dari *Selection* (Seleksi Data), *Pre-processing* (Praproses Data), *Transformation* (Transformasi Data), dan Data Mining (Penemuan Pola).

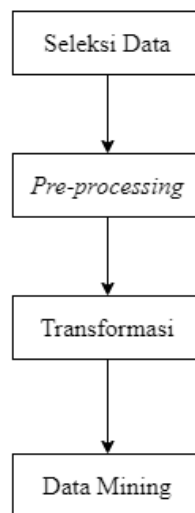
3. METODE PENELITIAN

3.1 Pengumpulan Data

Data Alkitab Bahasa Dayak Ma'anyan yang digunakan dalam penelitian ini berasal dari website GospelGo yang dapat diakses di https://gospelgo.com/p/maanyan_nt.htm. Daftar kata dasar yang digunakan dalam penelitian ini berasal dari Kamus Dwibahasa Dayak Ma'anyan-Indonesia yang didapatkan dari Balai Bahasa Kalimantan Tengah.

3.2 Knowledge Discovery in Databases (KDD)

Adapun urutan dari pembuatan korpus Bahasa Dayak Ma'anyan adalah sebagai berikut.



Gambar 1. Flowchart Proses Pembuatan Korpus

1. Seleksi Data (*Selection*)

Pada tahap ini peneliti akan menggunakan data Alkitab Bahasa Dayak Ma'anyan, untuk keperluan penelitian, peneliti memilih hanya akan menggunakan 5 data kitab dari kitab perjanjian baru sebagai data yang akan digunakan untuk proses KDD. Kitab yang digunakan adalah kitab Matius, Markus, Lukas, Yohanes dan Roma.

2. Pemilihan Data (*Pre-processing/Cleaning*)

Pada tahap preprocessing peneliti hanya melakukan 2 proses yaitu:

- Cleaning*, merupakan proses untuk menghilangkan atribut yang tidak berpengaruh terhadap klasifikasi yaitu menghapus angka, karakter khusus, tanda baca kecuali titik (yang memisahkan kalimat) dan menghapus *spasi* yang tidak perlu.
- Case Folding*, merupakan proses untuk mengubah semua huruf yang ada pada dokumen menjadi huruf kecil.

3. Transformation

Transformasi Data adalah proses yang terdiri dari serangkaian instruksi untuk mengubah data *input* menjadi *output* yang diinginkan (*input-proses-output*). Transformasi data dapat mencakup pemilihan atribut yang relevan, agregasi data, penggabungan data dari berbagai sumber, pengelompokan data ke dalam kategori tertentu, atau pembuatan fitur baru berdasarkan informasi yang tersedia. Tujuan dari tahap transformasi adalah untuk mempersiapkan data yang akan digunakan dalam proses data mining, sehingga data tersebut siap untuk dianalisis lebih lanjut.

4. Data Mining

Data Mining pada tahap ini, peneliti menerapkan teknik *Text Mining* untuk melakukan berbagai analisis pada korpus teks Bahasa Dayak Ma'anyan. Teknik-teknik ini meliputi: mencari jumlah kemunculan suatu kata dalam kategori tertentu (frekuensi), menemukan daftar kemunculan suatu kata dalam kalimat (konkordansi), menentukan penggunaan dan posisi kata yang dicari dalam kalimat (kalimat), dan mengidentifikasi daftar kluster atau kelompok kata yang ditemukan dalam teks (N-Gram).

a. Frekuensi

Langkah-langkah pencarian frekuensi dalam aplikasi korpus Bahasa Dayak Ma'anyan adalah sebagai berikut:

1. Menggunakan data kitab yang telah melalui proses pra-pemrosesan.
2. Menghitung nilai *Inverse Document Frequency* (IDF).
3. Menghitung bobot (W) masing-masing dokumen terhadap kata kunci (*query*).
4. Menampilkan hasil bobot (W) masing-masing dokumen terhadap kata kunci (*query*).

b. Konkordansi

Langkah-langkah pencarian konkordansi dalam aplikasi korpus Bahasa Dayak Ma'anyan adalah sebagai berikut:

1. Kata yang dimasukkan dalam variabel *input* (kata dalam Bahasa Dayak Ma'anyan) digunakan untuk mencari kata yang sama di kitab.
2. Mencari semua ayat yang mengandung kata yang sama dengan kata pada variabel *input*.
3. Jika ditemukan ayat yang sesuai pada langkah (2), maka akan dicek indeksnya untuk menggambarkan posisi ayat tersebut pada grafik.
4. Proses pada langkah (3) diulang terus hingga tidak ada lagi ayat yang mengandung kata yang sama dengan kata pada variabel *input*.
5. Menampilkan hasil akhir dari pencarian konkordansi.

c. Kalimat

Langkah-langkah pencarian kalimat dalam aplikasi korpus Bahasa Dayak Ma'anyan adalah sebagai berikut:

1. Kata yang ada di variabel *input* (kata dalam Bahasa Dayak Ma'anyan) digunakan untuk mencari kata yang sama di kitab.
2. Mencari semua ayat yang mengandung kata yang sama dengan kata pada variabel *input*.
3. Jika ditemukan ayat yang sesuai pada langkah (2), maka akan dicek indeksnya untuk menggambarkan posisi ayat tersebut pada grafik.
4. Proses pada langkah (3) diulang hingga tidak ada lagi ayat yang mengandung kata yang sama dengan kata pada variabel *input*.
5. Menampilkan hasil akhir dari pencarian kalimat.

d. N-Gram

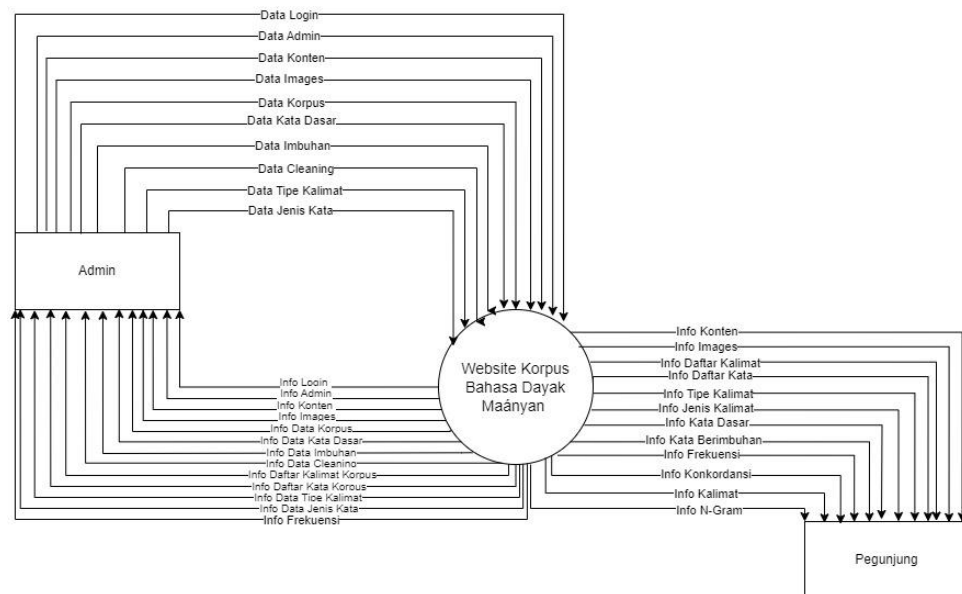
Langkah-langkah pencarian N-Gram dalam aplikasi korpus Bahasa Dayak Ma'anyan adalah sebagai berikut:

1. Dilakukan pengecekan terhadap ketentuan tambahan yang dituliskan oleh pengguna pada formulir lanjutan, kemudian dilakukan pembaruan daftar ayat.
2. Menginisialisasi variabel '*range*' dengan nilai rentang minimal.
3. Mengecek apakah rentang dari suatu kombinasi kata tertentu masih kurang dari '*rangeMax*'. Selama kondisi ini terpenuhi, kombinasi kata tersebut akan ditampilkan dalam daftar.
4. Ketika kondisi tidak lagi terpenuhi, maka akan ditampilkan nilai kumulatif frekuensi kemunculan, peringkat (*rank*), rentang (*range*), dan klaster.

3.3 Desain Sistem

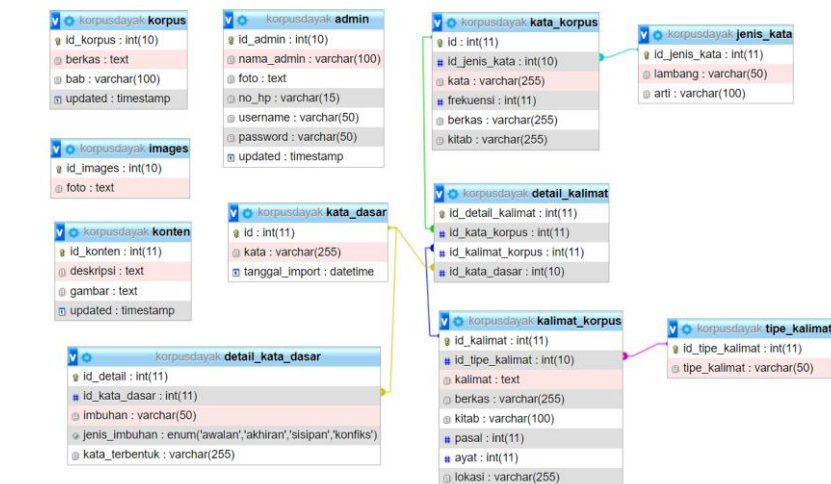
1. Diagram Konteks

Diagram konteks memberikan gambaran tentang bagaimana informasi dan data mengalir melalui sistem, serta interaksi antara pengguna dan sistem. Diagram konteks yang digunakan dalam pembuatan Aplikasi Korpus Bahasa Dayak Ma'anyan memiliki 2 aktor, yaitu admin dan pengunjung.



Gambar 2. Diagram Konteks

2. Entity Relationship Diagram (ERD)

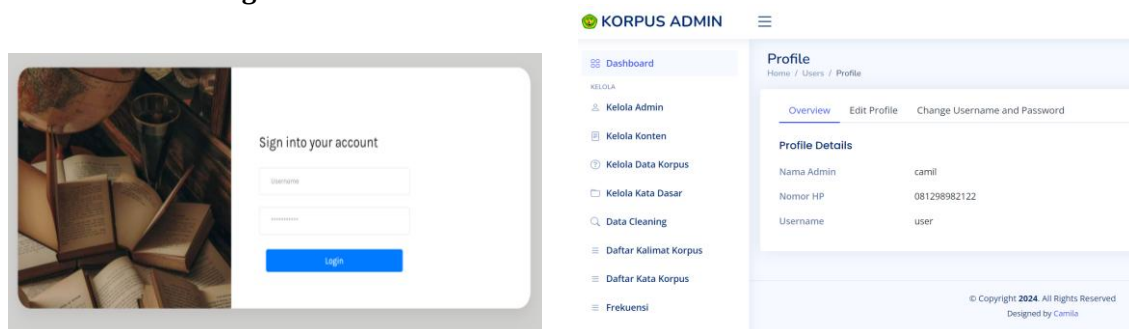


Gambar 3. Entity Relationship Diagram (ERD)

4. HASIL DAN PEMBAHASAN

4.1 Implementasi Aplikasi

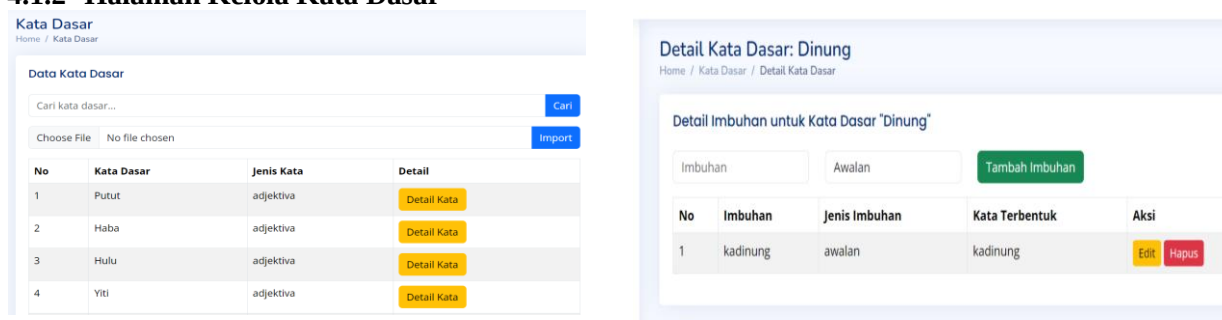
4.1.1 Halaman Login Dan Halaman User Admin



Gambar 4. Halaman Login dan Halaman Dashboard Admin

Halaman *login* merupakan halaman awal saat seseorang admin ingin mengakses website. Halaman ini akan ditampilkan ketika *admin* ingin masuk dan mengelola halaman admin. Pada halaman *login* ini terdapat *form login* yang diisi dengan *username* dan *password*. Halaman *dashboard* admin berisikan data mengenai pengelolaan data admin, data konten, data korpus, data kata dasar, data *cleaning*, data kalimat korpus dan data kata korpus.

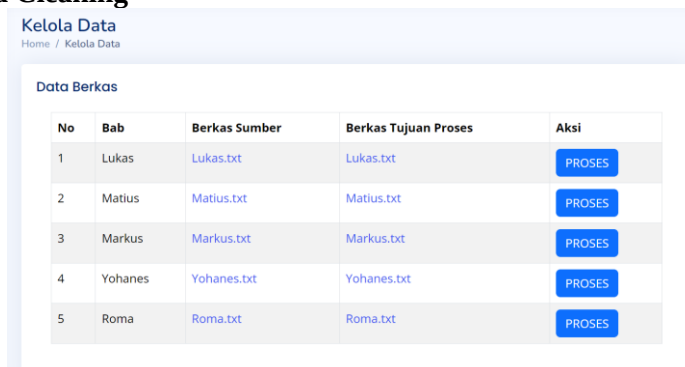
4.1.2 Halaman Kelola Kata Dasar



Gambar 5. Halaman Kata Dasar dan Halaman Detail Kata Dasar

Halaman kelola kata dasar adalah halaman yang dapat digunakan oleh admin untuk menambahkan kata dasar dalam korpus dan menambahkan imbuhan pada kata dasar.

4.1.3 Halaman Data Cleaning



Gambar 6. Halaman Data Cleaning

Halaman *data cleaning* adalah halaman yang digunakan untuk melakukan proses *pre-processing* dalam Aplikasi Korpus Bahasa Dayak Ma'anyan. Pada proses ini data akan

dibersihkan dan dipecah menjadi kalimat dan kata yang akan di simpan dalam table `kalimat_korpus` dan `kata_korpus`.

[illegible][illegible]

Gambar 7. Data yang belum dibersihkan dan data yang sudah dibersihkan

4.1.4 Halaman Daftar Kata Korpus dan Daftar Kalimat Korpus

Isi Tabel Kata Korpus Home / Kata Korpus					Daftar Kalimat Korpus Home / Kalimat Korpus																																				
Data Kata Korpus Cari kata...					Isi Tabel Kalimat Korpus <table> <tr> <th>No</th><th>Kalimat</th><th>Kitab</th><th>Pasal</th><th>Ayat</th><th>Lokasi</th><th>Berkas</th><th>Tipe Kalimat</th></tr> <tr> <td>1</td><td>teofilus sa maeh budi rama ulun haut ngakali nyurat anri taatur neu kawan kajadian sa haut jari hang penah takam.</td><td>Lukas</td><td>1</td><td>1</td><td>Lukas.txt</td><td>Lukas.txt</td><td>pa </td></tr> <tr> <td>2</td><td>here nyurat sameh kala sa nasarita ma takam daya kawan ulun sa kadinung raeral kawan kajadian yiru teka pamulaanni balalu iwara kabar yiru.</td><td>Lukas</td><td>1</td><td>2</td><td>Lukas.txt</td><td>Lukas.txt</td><td>pa </td></tr> <tr> <td>3</td><td>udi aku anri bahimat nyalidik katuluhni yiru teka pamulaanni aku ngangap patut nampin nyurat isa laporan sa sasusun maeh ma hamyu.</td><td>Lukas</td><td>1</td><td>3</td><td>Lukas.txt</td><td>Lukas.txt</td><td>pa </td></tr> </table>					No	Kalimat	Kitab	Pasal	Ayat	Lokasi	Berkas	Tipe Kalimat	1	teofilus sa maeh budi rama ulun haut ngakali nyurat anri taatur neu kawan kajadian sa haut jari hang penah takam.	Lukas	1	1	Lukas.txt	Lukas.txt	pa	2	here nyurat sameh kala sa nasarita ma takam daya kawan ulun sa kadinung raeral kawan kajadian yiru teka pamulaanni balalu iwara kabar yiru.	Lukas	1	2	Lukas.txt	Lukas.txt	pa	3	udi aku anri bahimat nyalidik katuluhni yiru teka pamulaanni aku ngangap patut nampin nyurat isa laporan sa sasusun maeh ma hamyu.	Lukas	1	3	Lukas.txt	Lukas.txt	pa
No	Kalimat	Kitab	Pasal	Ayat	Lokasi	Berkas	Tipe Kalimat																																		
1	teofilus sa maeh budi rama ulun haut ngakali nyurat anri taatur neu kawan kajadian sa haut jari hang penah takam.	Lukas	1	1	Lukas.txt	Lukas.txt	pa																																		
2	here nyurat sameh kala sa nasarita ma takam daya kawan ulun sa kadinung raeral kawan kajadian yiru teka pamulaanni balalu iwara kabar yiru.	Lukas	1	2	Lukas.txt	Lukas.txt	pa																																		
3	udi aku anri bahimat nyalidik katuluhni yiru teka pamulaanni aku ngangap patut nampin nyurat isa laporan sa sasusun maeh ma hamyu.	Lukas	1	3	Lukas.txt	Lukas.txt	pa																																		

Gambar 8. Halaman Daftar Kata Korpus dan Halaman Daftar Kalimat Korpus

Halaman daftar kata korpus dan daftar kalimat korpus adalah halaman yang menampilkan hasil data yang sudah diekstrak dalam proses *pre-processing*. Di halaman daftar kata korpus admin dapat menentukan jenis kata yang terdapat di dalam kitab dan di halaman daftar kalimat korpus, admin dapat menentukan tipe kalimat berdasarkan tag yang menunjukan jenis kata dalam kalimat tersebut.

4.1.5 Halaman User Pengunjung



Gambar 9. Dashboard Pengunjung

Halaman *dashboard* pengunjung adalah halaman yang menampilkan halaman awal saat pengunjung mengakses *website*. Halaman ini berisi fitur-fitur yang dapat diakses oleh pengunjung. Di dalam halaman ini juga terdapat informasi singkat tentang korpus dan grafik distribusi 10 kata dengan frekuensi tertinggi di dalam korpus.

4.1.6 Tampilan Halaman Frekuensi



No	Kata	Bab	Frekuensi
1	dinung	Lukas	21
2	dinung	Matius	23
3	dinung	Markus	10
4	dinung	Yohanes	16
5	dinung	Roma	2

No	Kata	Kitab	TF	IDF	TF-IDF
1	dinung	Lukas	0.000911	0.817678	0.000745
2	dinung	Markus	0.000705	0.817678	0.000577
3	dinung	Matius	0.001042	0.817678	0.000852
4	dinung	Roma	0.000187	0.817678	0.000153
5	dinung	Yohanes	0.000941	0.817678	0.000770

Gambar 10. Halaman Frekuensi dan Detail Frekuensi

Halaman frekuensi adalah halaman yang digunakan untuk menampilkan frekuensi dari kata yang dicari, dalam halaman frekuensi juga terdapat halaman detail frekuensi yang berisi hasil perhitungan dari metode *Term Frequency-Inverse Document Frequency* (TF-IDF).

4.1.7 Tampilan Halaman Konkordansi

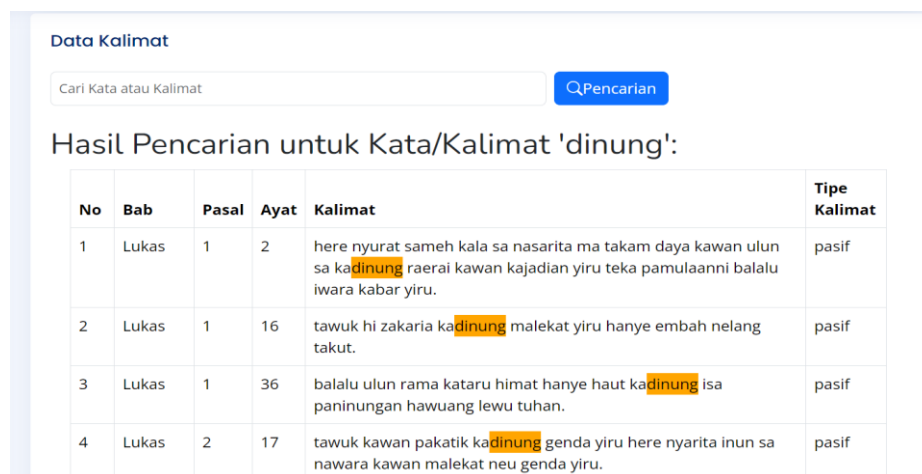


No	Kitab	Pasal	Ayat	Kalimat	Jenis Kalimat	Jenis Kata
1	Lukas	2	20	kawan pakatik yiru mulek hampi padang rikut nelang nawat nante alatalla daya katuluh sa haut rengei nelang dinung here pas kala sa nawara daya malekat.	pasif	adjektiva
2	Lukas	2	20	kawan pakatik yiru mulek hampi padang rikut nelang nawat nante alatalla daya katuluh sa haut rengei nelang dinung here pas kala sa nawara daya malekat.	pasif	adjektiva

Gambar 11. Halaman Konkordansi

Halaman konkordansi adalah halaman yang digunakan untuk menampilkan konkordansi dari kata yang dicari.

4.1.8 Tampilan Halaman Kalimat

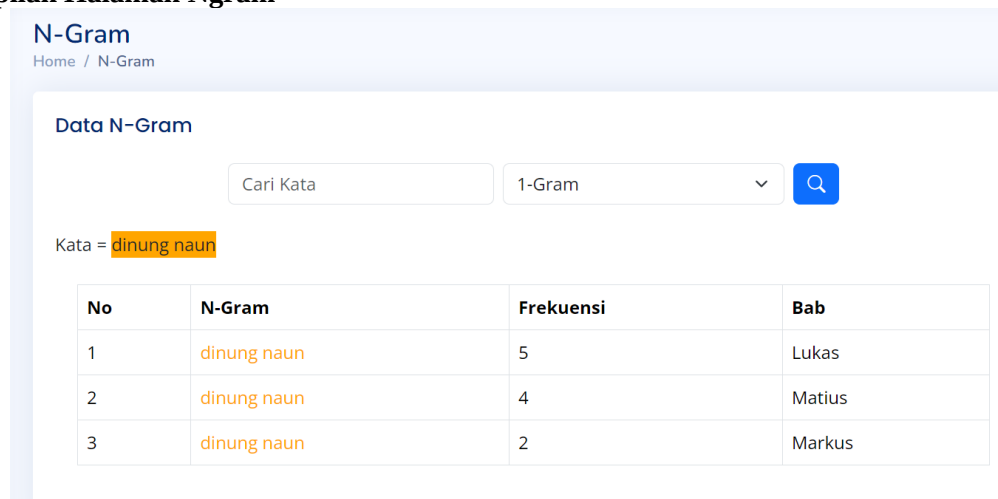


No	Bab	Pasal	Ayat	Kalimat	Tipe Kalimat
1	Lukas	1	2	here nyurat sameh kala sa nasarita ma takam daya kawan ulun sa ka dinung raerai kawan kajadian yiru teka pamulaanni balalu iwara kabar yiru.	pasif
2	Lukas	1	16	tawuk hi zakaria ka dinung malekat yiru hanye embah nelang takut.	pasif
3	Lukas	1	36	balalu ulun rama kataru himat hanye haut ka dinung isa paninungan hawuang lewu tuhan.	pasif
4	Lukas	2	17	tawuk kawan pakatik ka dinung genda yiru here nyarita inun sa nawara kawan malekat neu genda yiru.	pasif

Gambar 12. Halaman Kalimat

Halaman kalimat adalah halaman yang digunakan untuk menampilkan imbuhan dari kata yang dicari.

4.1.8 Tampilan Halaman Ngram



No	N-Gram	Frekuensi	Bab
1	dinung naun	5	Lukas
2	dinung naun	4	Matius
3	dinung naun	2	Markus

Gambar 13. Halaman N-Gram

Halaman N-Gram adalah halaman yang digunakan untuk menampilkan N-Gram dari kata yang dicari.

5. KESIMPULAN

5.1 Kesimpulan

Kesimpulan yang dapat ditarik dari penelitian ini adalah bahwa pengembangan aplikasi Korpus Bahasa Dayak Ma'anyan telah berhasil dilakukan dengan menggunakan dua metode utama, yaitu metode Waterfall dan Knowledge Discovery in Database (KDD). Metode Waterfall diterapkan dalam tahapan pengembangan perangkat lunak yang meliputi analisis, desain, pengkodean, dan pengujian. Sementara itu, metode KDD digunakan untuk proses pembuatan korpus, melibatkan beberapa tahapan seperti Selection (Seleksi Data), Pre-processing (Praproses Data), Transformation (Transformasi Data), dan Data Mining (Penemuan Pola).

Hasil penelitian ini berupa korpus digital yang dapat diakses oleh publik, yang dapat digunakan untuk mendukung pembelajaran dan pelestarian bahasa Dayak Ma'anyan. Metode Text Mining yang diterapkan pada korpus memungkinkan analisis yang lebih mendalam terkait penggunaan kata, frekuensi, dan pola kalimat. Fitur Term Frekuensi-Invers Dokumen

Frekuensi (TF-IDF), konkordansi, kalimat dan N-gram dalam korpus ini juga membantu dalam menilai kata-kata penting serta memahami konteks penggunaannya. Dengan demikian, aplikasi ini berkontribusi terhadap pengembangan linguistik dan penerjemahan bahasa alami, serta diharapkan dapat menarik minat masyarakat dan generasi muda untuk mempelajari dan melestarikan bahasa Dayak Ma'anyan

Penelitian ini diharapkan dapat menjadi langkah awal yang signifikan dalam pelestarian bahasa daerah, khususnya Bahasa Dayak Ma'anyan, yang terancam punah seiring dengan berkurangnya penutur asli. Melalui pengembangan korpus digital ini, diharapkan ada solusi konkret untuk menjaga keberlangsungan bahasa dan budaya Dayak Ma'anyan di masa mendatang.

5.2 Saran

Berdasarkan hasil penelitian yang dilakukan pada skripsi ini, ada beberapa saran yang dapat diberikan untuk menyempurnakan pengembangan sistem ini agar lebih baik lagi, yaitu:

1. Peningkatan Proses Pra-pemrosesan Data: disarankan untuk memperbaiki proses pra-pemrosesan data, dengan fokus pada perbaikan kesalahan ejaan dan penggunaan kata yang ada, sehingga kualitas data yang digunakan dalam sistem dapat lebih akurat.
2. Otomatisasi Input Data Imbuhan: saat ini, fitur detail kata dasar pada website masih memerlukan input manual untuk data imbuhan. Untuk meningkatkan efisiensi dan akurasi, sebaiknya sistem dikembangkan agar dapat melakukan proses ini secara otomatis di masa mendatang.

DAFTAR PUSTAKA

- [1] Abidin, Zaenal, and Permata Permata. 2021. "PENGARUH PENAMBAHAN KORPUS PARALEL PADA MESIN PENERJEMAH STATISTIK BAHASA INDONESIA KE BAHASA LAMPUNG DIALEK NYO." *Jurnal Teknoinfo* 15(1):13. doi: 10.33365/jti.v15i1.889.
- [2] Hizbullah, Nur, Fuzi Fauziah, Program Studi Sastra Arab, Fakultas Sastra, Universitas Al Azhar Indonesia Jalan Sisingamangaraja, Kebayoran Baru, and Jakarta Selatan. 2016. *Penyusunan Model Korpus Al-Qur'an Digital Penulis Untuk Korespondensi*. Vol. 3.
- [3] Yusuf, Bustami, Muhammad Zaeki, Hendri Ahmadian, Khairan Ar, Sri Wahyuni, Prodi Teknologi Informasi, Fakultas Sains dan Teknologi, Uin Ar-Raniry Banda Aceh, Prodi Pendidikan Teknologi Informasi, and Fakultas Tarbiyah dan Keguruan. 2021. "Analisa Topik Pendidikan Dalam Al-Quran Dengan Pendekatan Text Mining." *Serambi Engineering* VI(1).
- [4] Agungnesia. 2018. "Pengertian Korpus Dalam Penelitian Bahasa, Berikut Sumber Korpus Yang Perlu Diketahui." <https://Agungnesia.Com/Pengertian-Korpus-Dalam-Penelitian-Bahasa-Berikut-Sumber-Korpus-Yang-Perlu-Diketahui/>.
- [5] Rosa, A.S., dan M. Shalahudin. 2014. *Rekayasa Perangkat Lunak*. Bandung: Informatika.
- [6] Zaki Izzani Akbar. 2021. "Apa Itu Text Mining?" <https://Sis.Binus.Ac.Id/2021/04/23/Apa-Itu-Text-Mining/>.
- [7] Musfiroh, Hamdani, Inda Fitri Astuti. 2013. PENERAPAN ALGORITMA TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF) UNTUK TEXT MINING. <https://e-journals.unmul.ac.id/index.php/JIM/article/view/113>.
- [8] Mulyani, Esti, Fachrul Pralienka Bani Muhamad, and Kurnia Adi Cahyanto. 2021. "Pengaruh N-Gram Terhadap Klasifikasi Buku Menggunakan Ekstraksi Dan Seleksi Fitur Pada Multinomial Naïve Bayes." *JURNAL MEDIA INFORMATIKA BUDIDARMA* 5(1):264. doi: 10.30865/mib.v5i1.2672.