OTOMASI PEMBENTUKAN ABSTRAK BAHASA INDONESIA BERDASARKAN KESELURUHAN DOKUMEN MENGGUNAKAN TERM FREQUENCY – INVERSE DOCUMENT FREQUENCY (TF-IDF)

Arie Prima Anggara[1]

[1]Teknik Informatika – Universitas Komputer Indonesia
Jl. Dipatiukur 112-114 Bandung
Email : arieprimaa@gmail.com

ABSTRAK

Abstrak adalah sebuah ringkasan dari keseluruhan dokumen penelitian agar pembaca tahu maksud dari penelitian tersebut. Dari abstrak, pembaca dapat mengetahui jalan pikiran penulis laporan/artikel ilmiah tersebut dan mengetahui gambaran umum tulisan secara lengkap. Dalam membentuk paragraf abstrak pada sebuah dokumen jurnal dari setiap sub bab tentunya membutuhkan pencarian kalimat yang singkat dan jelas. Pencarian kalimat pada setiap sub bab dalam dokumen jurnal yang banyak akan membantu penulis laporan/artikel ilmiah dalam membentuk abstrak dengan keterbatasan pembentukan abstrak harus singkat, bagian harus seimbang dan menghindari kalimat yang panjang.

Dalam penelitian ini dibangun sistem pembentukan abstrak secara otomatis dengan menerapkan metode *Term Frequency – Inverse Document Frequency*. Metode *Term Frequency – Inverse Document Frequency (TF-IDF)* akan mencari kalimat pada sub judul pendahuluan, analisis, dan kesimpulan sehingga akan terbentuklah kalimat abstrak secara otomatis dengan mengutamakan isi kalimat yang berada pada sub judul tersebut. Analisis sistem pembentukan abstrak otomatis yang dibangun memiliki 3 tahapan yaitu input jurnal, praproses, pembentukan abstrak. Input jurnal yang dipilih berupa jurnal berformat doc. Praproses yang dilakukan dalam tahap ini adalah pemecahan kalimat, case folding, filtering, tokenizing, synonim checking, dan stopword. Tahap selajutnya melakukan perhitungan menggunakan tf-idf dan tahap terakhir menentukan hasil abstrak dengan metode cosine similarity.

Berdasarkan hasil pengujian, maka dilakukan tahap pengujian terhadap sistem dan secara manual. Maka diperoleh jarak antara hasil manual dengan sistem rata-rata hanya berbeda 2 sampai 4 kalimat saja. Akurasi dengan persentase 100% didapat oleh jurnal ke 7 dan 9 sedangkan akurasi dengan pesentase terendah 60% didapat oleh jurnal ke 1.

Kata Kunci: Term Frequency – Inverse Document Frequency (TF-IDF), Kecerdasan Buatan, Pembentukan Abstrak Otomatis.

1. PENDAHULUAN

Abstrak adalah sebuah ringkasan dari keseluruhan dokumen penelitian agar pembaca tahu maksud dari penelitian tersebut. Abstrak digunakan sebagai jembatan untuk memahami uraian yang akan disajikan dalam suatu karangan biasanya laporan atau artikel ilmiah terutama untuk memahami ide-ide permasalahannya. Dari abstrak, pembaca dapat mengetahui jalan pikiran penulis laporan/artikel ilmiah tersebut dan mengetahui gambaran umum tulisan secara lengkap. Untuk itulah pembuat abstrak harus dapat mewakili isi karangan ilmiah secara keseluruhan, mulai dari latar belakang, analisis/analisa, dan hasil penelitian/kesimpulan. [1]

Dalam membentuk paragraf abstrak pada sebuah dokumen jurnal dari setiap sub bab tentunya membutuhkan pencarian kalimat yang singkat dan jelas. Di dalam dunia akademik, tulisan pendek ini digunakan oleh institusi/lembaga/organisasi pendidikan sebagai informasi awal atas sebuah penelitian ketika dimasukkan dalam jurnal, konferensi, lokakarya, atau yang sejenisnya. Pencarian kalimat pada setiap sub bab dalam dokumen jurnal yang banyak akan menyulitkan penulis laporan/artikel ilmiah dalam membentuk abstrak dengan keterbatasan pembentukan abstrak harus singkat, bagian harus seimbang dan menghindari kalimat yang panjang. Pembuatan abstrak dapat dilakukan secara otomatis dengan memanfaatkan metode *Term Frequency – Inverse Document Frequency (TF-IDF)*.

Untuk dapat membentuk kalimat abstrak secara otomatis maka akan diterapkan metode *Term Frequency – Inverse Document Frequency (TF-IDF)*, karena TF-IDF merupakan pembobotan yang mengukur seberapa penting sebuah kata dalam dokumen bila dilihat secara global pada seluruh dokumen.[5] Pada pembuatan abstrak penelitian akan mencari sub judul yang dibutuhkan yaitu pada bagian pendahuluan atau latar belakang, analisis atau analisa data dan kesimpulan. Dimana TF-IDF akan

mengukur kemiripan antara dua kalimat atau lebih pada bagian kalimat yang berada dalam sub judul yang dibutuhkan. Dalam mengukur sebuah kalimat dari kemiripan setiap sub bab pada sebuah dokumen akan menggunakan teknik *text mining*.

Dari permasalahan diatas, maka penelitian ini akan menerapkan metode *Term Frequency – Inverse Document Frequency (TF-IDF)* pada pembentukan sebuah paragraf abstrak secara otomatis.

1.1 Maksud dan Tujuan

Maksud dari pada penelitian tugas akhir ini adalah membuat sistem pembentukan abstrak secara otomatis pada keseluruhan dokumen jurnal penelitian dengan menggunakan metode *Term Frequency – Inverse Document Frequency (TF-IDF)*. Adapun tujuan dari penelitian tugas akhir ini adalah melakukan pembuatan kalimat abstrak yang baik secara otomatis untuk melengkapi suatu tulisan ilmiah dengan singkat dan jelas, sehingga membantu melengkapi laporan/artikel ilmiah dalam pembentukan kalimat abstrak secara cepat dan akurat.

1.2 Ringkasan

Ringkasan adalah suatu cara yang efektif untuk menyajikan suatu karangan yang panjang dalam bentuk singkat. Oleh karena itu membuat ringkasan atas sebuah karangan yang panjang dapat diumpamakan sebagai memangkas sebatang pohon sehingga tinggal batang, cabang-cabang dan ranting-ranting yang terpenting beserta daun-daun yang diperlukan, sehingga tampak bahwa esensi pohon masing dipertahankan. Dalam ringkasan keindahan gaya bahasa, ilustrasi, serta penjelasan-penjelasan yang terperinci dihilangkan, sedangkan sari kerangaknya dibiarkan tanap hiasan. Walaupun bentuknya ringkas, namun tetap mempertahankan pikiran pengarang dan pendekatan yang asli [11].

Dalam meringkas sebaiknya menggunakan kalimat tunggal daripada kalimat majemuk. Kalimat majemuk menunjukan bahwa ada dua gagasan atau lebih yang bersifat paralel. Bila memungkinkan ringkaslah kalimat menjadi frasa, selanjutnya frasa menjadi kata. Begitu pula rangkaian gagasan yang penjang hendaknya diganti dengan suatu gagasan sentral saja. Pada umumnya suatu ringkasan ditentukan panjang ringkasan akhirnya, ada yang meringkas seperdua total kata, sepertiga total kata, hingga sepersepuluh total kata. Untuk artikel yang memiliki jumlah kata yang sangat banyak tentu saja hasil ringkasan sepersepuluh menjadi pilihan yang tepat. Sedangkan untuk artikel yang jumlah katanya tidak terlalu banyak maka hasil ringkasan sebanyak seperdua total kata menjadi pilihan yang tepat sehingga hasil ringkasan nantinya masih dapat dibaca dengan struktur hasil ringkasan yang baik.

1.3 Abstrak

Abstrak merupakan sebuah ringkasan isi dari sebuah karya tulis ilmiah yang ditujukan untuk membantu seorang pembaca agar dapat dengan mudah dan cepat untuk melihat tujuan dari penulisannya. Di dalam dunia akademik, tulisan pendek ini digunakan oleh institusi/lembaga/organisasi pendidikan sebagai informasi awal atas sebuah penelitian ketika dimasukkan dalam jurnal, konferensi, lokakarya, atau yang sejenisnya. Dalam dunia maya (internet), sebuah abstrak digunakan sebagai gambaran singkat atas sebuah karya tulis ilmiah/penelitian untuk dibaca, sebagaimana halnya sebuah “display” model pakaian dipajang untuk dilihat atau diuji pakai sebelum dibeli. Selanjutnya, bagian lengkap sebuah penelitian dijual kepada mereka yang berminat untuk mendapatkannya.

Untuk membuat abstrak, hal-hal yang perlu diperhatikan adalah sebagai berikut:

1. Struktur paragraf

   Sebuah abstrak ditulis dalam satu paragraf yang menerangkan keseluruhan isi tulisan secara singkat dan jelas. Penulisannya tidak melakukan indentasi pada kalimat pertama paragraf. *Single space* adalah pilihan yang dimiliki oleh penulis untuk menyusun kalimat dalam paragrafnya. Lebih dalam, kadang seorang pembimbing Skripsi/Tesis/Disertasi mengatur hingga pada penggunaan jenis huruf dan ukuran tertentu.
2. Jumlah kata

   Idealnya sebuah paragraf terdiri dari 150 sampai dengan 200 kata. Namun, pertimbangan jumlah kata yang paling tepat dalam penulisan Skripsi, Tesis, ataupun disertasi biasanya bergantung pada pertimbangan pandangan pembimbing (supervisor) yang mendampingi seorang mahasiswa dalam penulisannya. Seorang supervisor harusnya tidak mempertimbangkan jumlah kata sebagai acuan utama penulisan paragraf, karena bagian utama justru isi (content) paragraf.
3. Isi paragraf

Pertama, identifikasi fokus penelitian dijelaskan secara singkat agar pembaca memahami apa yang diamati oleh seorang peneliti di dalam penelitiannya. Kedua, penulis perlu menggambarkan secara jelas desain penelitian yang dilakukan dalam proses pencarian jawaban atau solusi atas persoalan yang diangkat di dalam penelitiannya. Desain langkah penyelesaian masalah ini oleh mahasiswa lazim dikenal dengan istilah Metode Penelitian. Ketiga,

selanjutnya penulis akan menjelaskan hasil temuannya kepada pembaca. Beberapa peneliti menganggap hasil temuan yang diungkap tidak perlu mengungkap pembahasan yang dilakukan karena hal itu justru akan membuat pengulangan isi tulisan. Jelas maksudnya karena bagian pembahasan temuan penelitian juga diurai di dalam bagian kesimpulan. Keempat, perlunya bagian kesimpulan di dalam sebuah tulisan juga terlihat di dalam sebuah abstrak yang tetap mendapatkan perhatian penting sebagai bagian akhir dari paragraf. Pada bagian ini kadangkala sejumlah peneliti menyisipkan rekomendasi penelitian namun tanpa pembahasan atau uraian yang panjang. Lebih lanjut, tidaklah lazim sebuah abstrak diisi oleh nama si penulis serta para pembimbing tulisannya, apalagi hal itu ditulis dalam huruf cetak tebal.

1.4 Tahapan *Text Mining*

Pada tahap *text mining* ini terdapat lima langkah yaitu:

a. *Case Folding,* yaitu mengubah semua teks ke dalam huruf kecil.
b. *Filtering,* menghapus karakter yang tidak dipakai. Adapun karakter yang akan dihilangkan yaitu : '0', '1', '2', '3', '4', '5', '6', '7', '8', '9', ',', "", '-', '/', '{', '}', '+', '_', '!', '@', '#', '$', '%', '^', '&', '*', '(', ')', '?', '<', '>', '[', ']', '|', '~', '`', ';', ':', '=' dan "'s" untuk dokumen berbahasa Inggris akan dihapus kecuali tanda titik.
c. *Tokenizing,* yaitu memecah kalimat-kalimat yang ada kedalam kata.
d. *Synonim Checking,* yaitu mengganti kata yang ada yang memiliki arti yang sama. Penggantian kata yang dilakukan berdasarkan *database* yang dibuat oleh penulis.
e. *Stopwords,* yaitu Pada proses *stowords removal* ini proses yang dikerjakan yaitu menghapus kata yang kurang relevan atau kata yang tidak memiliki arti yang begitu penting dan berkaitan yang ada pada kalimat di dokumen dengan mencocokan *list* kata stopword yang pada *database* yang akan diuji. *Database* yang digunakan baik untuk Bahasa Indonesia maupun maupun Bahasa Inggris diambil dari sebuah situs yg menyediakan berbagain jenis *list stopword* untuk berbagai bahasa.

Pada umumnya tahap *preprocessing* terdapat 5 langkah yaitu *Case Folding, Filtering, Tokenizing, Stemming,* dan *Stopwords.* Ada satu langkah yang tidak penulis gunakan dalam pembuatan aplikasi peringkasan dokumen esktraktif ini yaitu proses *stemming*. *Stemming* yaitu proses menghilangkan imbuhan dari sebuah kata dan mengubahnya menjadi kata dasar. Tujuan dari penghilangan proses *Stemming* ini yaitu karena penulis memiliki pendapat bahwa dalam sebuah peringkasan dokumen ekstraktif imbuhan yang terdapat pada sebuah kata memiliki arti yang berbeda. Adapun contoh kalimat yaitu : "Bapak **membuatkan** sebuah rumah untuk istrinya", "Bapak **dibuatkan** secangkir kopi oleh ibu", "Bapak selalu **berbuat** baik kepada tetangga". Kata "buat" pada kalimat pertama dan kalimat ketiga memiliki arti bapak melakukan sebuah kegiatan tapi pada kalimat dua kata"buat" memiliki arti bapak mendapatkan sebuah perlakuan. Apabila proses *stemming* tetap dilakukan maka tidak akan didapatkan perbedaan antara satu kata dengan yang lainnya.

Pada tahap filtering penulis menambahkan karakter " 's " yang akan digunakan untuk menghapus tanda baca bagi dokumen yang menggunakan bahasa Inggris. Adapun contoh kalimatnya yaitu : "that is President's car" berubah menjadi "that is President car".

1.5 *Term Frequency – Inverse Document Frequency (TF-IDF)*

Metode *Term Frequency-Inverse Document Frequency* (TF-IDF) adalah cara pemberian bobot hubungan suatu kata (term) terhadap dokumen. Untuk dokumen tunggal tiap kalimat dianggap sebagai dokumen. Metode ini menggabungkan dua konsep untuk perhitungan bobot, yaitu *Term Frequency* (TF) merupakan frekuensi kemunculan kata (t) pada kalimat (s). *Document frequency* (DF) adalah banyaknya kalimat dimana suatu kata (t) muncul. Frekuensi kemunculan kata di dalam dokumen yang diberikan menunjukkan seberapa penting kata itu di dalam dokumen tersebut. [5]

Frekuensi kalimat yang mengandung kata tersebut menunjukkan seberapa umum kata yang ada pada dokumen uji tersebut. Bobot kata semakin besar jika sering muncul dalam suatu kalimat dan semakin kecil jika muncul dalam banyak kalimat. Pada Metode ini pembobotan kata dalam sebuah dokumen dilakukan dengan mengalikan nilai TF dan IDF. Pembobotan diperoleh berdasarkan jumlah kemunculan term dalam kalimat (TF) dan jumlah kemunculan term pada seluruh kalimat dalam dokumen (IDF). Bobot suatu istilah semakin besar jika istilah tersebut sering muncul dalam suatu dokumen dan semakin kecil jika istilah tersebut muncul dalam banyak dokumen. Nilai IDF sebuah term dihitung menggunakan persamaan (1) di bawah ini :

$$IDF = log(\frac{N}{Df}) \qquad (1)$$

keterangan :
N = Jumlah keseluruhan kalimat pada dokumen
Df = Jumlah kalimat yang mengandung keyword t

Menghitung bobot (W) masing-masing kalimat dengan persamaan (2) di bawah ini:

$$W_{d.t} = TF_{d.t} * IDF_t \qquad (2)$$

dengan :

d = kalimat ke-d
t = kata (*term*) ke-t
TF = *Term Frequency*
W = bobot kalimat ke-d terhadap kata (*term*) ke-t
IDF= *Inverse Document Frequency*

Kemudian baru melakukan proses pengurutan (sorting) nilai kumulatif dari W untuk setiap kalimat. Tiga kalimat dengan nilai W terbesar dijadikan sebagai hasil dari ringkasan atau sebagai output dari peringkasan teks otomatis. [5]

1.6 Pengujian Akurasi

Dalam penelitian ini, metode evaluasi yang digunakan adalah metode intrinsik penghitungan *F-Measure* berdasarkan perhitungan *Precision* dan *Recall* yang menurut Zaman B. dan E Winarko [12] merupakan standar evaluasi dalam penghitungan *information retrieval*. evaluasi perhitungan *information retrieval* dengan menggunakan *Precision* dan *Recall* juga dapat digunakan dalam evaluasi perhitungan peringkas teks otomatis.

$$recall = \frac{\#kalimat\ ringkasan\ sistem \cap ringkasan\ manual}{kalimat\ ringkasan\ manual} \qquad (4)$$

dimana, *recall* = tingkat keberhasilan

Precision adalah tingkat ketepatan hasil ringkasan. Perhitungan *precision* dapat dilihat pada persamaan berikut:

$$Precision = \frac{\#kalimat\ ringkasan\ sistem \cap ringkasan\ manual}{kalimat\ ringkasan\ sistem} \qquad (5)$$

dimana, *precision*= tingkat ketepatan

F-measure adalah gabungan antara *recall* dan *precision*. Perhitungan *f-measure* dapat dilihat pada persamaan berikut:

$$F - measure = \frac{2 \times recall \times precision}{recall + precision} \qquad (6)$$

2. ISI PENELITIAN

2.1 Analisis Masalah

Dalam membentuk paragraf abstrak pada sebuah dokumen jurnal dari setiap sub bab tentunya membutuhkan pencarian kalimat yang singkat dan jelas, pencarian kalimat pada setiap sub bab dalam dokumen jurnal yang banyak akan menyulitkan penulis dalam membentuk abstrak dengan keterbatasan pembentukan abstrak harus singkat, bagian harus seimbang dan menghindari kalimat yang panjang.

Dari permasalahan diatas, dibutuhkan metode *Term Frequency – Inverse Document Frequency (TF-IDF)* yang akan mengukur kemiripan antara dua kalimat atau lebih pada bagian kalimat yang berada dalam sub judul yang dibutuhkan.

2.2 Analisis Masukan

Data masukan yang digunakan berupa *file* jurnal dengan format *.doc. Dokumen tersebut dimasukan secara utuh kedalam aplikasi, yang kemudian oleh aplikasi akan dilakukan penguraian isi dari dokumen tersebut berupa beberapa kalimat yang mengacu pada aturan tertentu, seperti pengambilan kalimat hanya pada bagian pendahuluan, analisis, dan kesimpulan.

2.3 Analisis Preprocessing Data

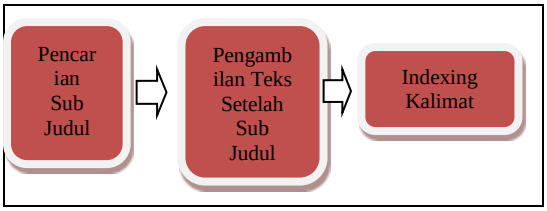
Tahap analisis *prepocessing* ini digunakan untuk menganalisa data apa saja yang sebenarnya dibutuhkan dalam proses *Term Frequency – Inverse Document Frequency (TF-IDF).* Adapun tahapan dari analisis *prepocessing* data dalam penelitian ini adalah sebagai berikut :

1. *Add File* Data

   Pertama-tama pilih dokumen terlebih dahulu. *File* dokumen yang di pilih harus berupa *file* berekstensi *.doc.
2. Ekstraksi Data

   Setelah file jurnal yang dimaksudkan didapat, kemudian lakukan proses ekstraksi text. Pada proses ini dilakukan pengambilan text dari jurnal dengan bantuan library di .net yaitu *Microsoft.Office.Interop.Word*.
3. Pengambilan Teks

Tahap pengambilan kalimat merupakan tahapan penggambilan teks yang dibutuhkan dari keseluruhan text yang ada pada jurnal tersebut. Teks yang dibutuhkan adalah teks pada bagian Pendahuluan atau Latar Belakang, Analisis atau Analisa Data dan Kesimpulan. Adapun alur proses yang dilakukan untuk mendapatkan teks yang dibutuhkan dapat dilihat pada gambar di bawah ini :



Gambar 1. Alur Seleksi Kalimat

a. Pencarian Sub judul :
   Mencari judul yang dibutuhkan yaitu : pendahuluan, analisis dan kesimpulan.
b. Pengambilan teks :
   Pengambilan teks yang indeksnya berada setelah judul yaitu pendahuluan atau latar

belakang, analisis atau analisa dan kesimpulan akan di simpan.

c. Indexing kalimat :

Setelah didapatkan teks diatas, dilakukan pemisahan tiap-tiap kalimat yang dibatasi oleh tanda titik pada tiap-tiap teks.

4. *Case Folding*

Pada tahapan ini, akan dilakukan pengecekan terhadap huruf-huruf kapital yang berada di tiap-tiap kalimat. Jika ditemukan huruf kapital tersebut, maka akan dilakukan *lower*, yaitu mengubah menjadi huruf kecil.

Tabel 1. Casefolding

	deskripsi kalimat
kalimat 1	di era otonomi daerah saat ini, sistem pemerintahan daerah sudah berbeda dibandingkan dengan sistem pemerintah diera orde baru.
Kalimat 2	kalau diera orde baru, organisasi pemerintah dan sistem informasinya ditentukan oleh pemerintah pusat, di era otonomi daerah ini pembentukan instansi pemerintah daerah termasuk sistem informasinya ditentukan oleh pemerintah daerah setempat
Kalimat 3	oleh karena itu sistem informasi pada setiap daerah bisa berbeda sesuai dengan perkembangan yang terjadi / kebutuhan di daerah masing-masing
Kalimat 4	pada awal otonomi daerah, pemerintah di daerah bisa membentuk dinas, badan dan lembaga tehnis sesuai dengan kebutuhan daerah setempat
Kalimat 5	adanya ketentuan ini membuat berbagai daerah membentuk dinas secara berlebihan untuk menampung sebanyak mungkin pejabat struktural
Kalimat 6	ketentuan mengenai pembentukan dinas dan lembaga tehnis tersebut kemudian disusul peraturan baru yang memberikan batasan jumlah dinas yang boleh dibentuk di pemerintah daerah.
Kalimat 7	bahkan penanganan informasi di suatu daerah cukup hanya dimasukkan dalam suatu seksi/bagian dari dinas dan setiap daeah menggunakan istilah yang berbeda seperti : hubungan masyarakat (humas) informasi komunikasi (infokom), badaninformasi komunikasi telematika (bikt).
…	…
Kalimat 21	3.kegiatan humas yang banyak berhubungan dengan unit kerja di pemda yang memiliki tingkat eselon yang lebih tinggi bisa menjadi hambatan bagi humas meski gengsi posisi humas cukup tinggi

5. *Filtering*

Menghapus karakter khusus yang tidak dipakai. Dalam pengujian ini karakter seluruh tanda baca seperti tanda seru, tanda tanya, tanda kutip, dan lain sebagainya akan dihapus kecuali kecuali tanda titik. Langkah yang dilakukan yaitu menghilangkan beberapa karakter yang tidak diperlukan selama proses perangkuman. Adapun karakter yang akan dihilangkan yaitu : '0', '1', '2', '3', '4', '5', '6', '7', '8', '9', ',', "", '-', '/', '{', '}', '+', '_', '!', '@', '#', '$', '%', '^', '&', '*', '(', ')', '?', '<', '>', '[', ']', '|', '~', '`', ';', ':', '=' dan "'s" untuk dokumen berbahasa Inggris.

Penghilangan karakter seperti angka akan dilakukan oleh sistem, karena sistem tidak dapat mengenali keterangan dari angka yang ada pada dokumen

Tabel 2. Filtering

	deskripsi kalimat
kalimat 1	di era otonomi daerah saat ini, sistem pemerintahan daerah sudah berbeda dibandingkan dengan sistem pemerintah diera orde baru.
Kalimat 2	kalau diera orde baru, organisasi pemerintah dan sistem informasinya ditentukan oleh pemerintah pusat, di era otonomi daerah ini pembentukan instansi pemerintah daerah termasuk sistem informasinya ditentukan oleh pemerintah daerah setempat
Kalimat 3	oleh karena itu sistem informasi pada setiap daerah bisa berbeda sesuai dengan perkembangan yang terjadi / kebutuhan di daerah masing-masing
Kalimat 4	pada awal otonomi daerah, pemerintah di daerah bisa membentuk dinas, badan dan lembaga tehnis sesuai dengan kebutuhan daerah setempat
Kalimat 5	adanya ketentuan ini membuat berbagai daerah membentuk dinas secara berlebihan untuk menampung sebanyak mungkin pejabat struktural
Kalimat 6	ketentuan mengenai pembentukan dinas dan lembaga tehnis tersebut kemudian disusul peraturan baru yang memberikan batasan jumlah dinas yang boleh dibentuk di pemerintah daerah.
Kalimat 7	bahkan penanganan informasi di suatu daerah cukup hanya dimasukkan dalam suatu seksi/bagian dari dinas dan setiap daeah menggunakan istilah yang berbeda seperti : hubungan masyarakat (humas) informasi komunikasi (infokom), badaninformasi komunikasi telematika (bikt).
…	…
Kalimat 21	3.kegiatan humas yang banyak berhubungan dengan unit kerja di pemda yang memiliki tingkat eselon yang lebih tinggi bisa menjadi hambatan bagi humas meski gengsi posisi humas cukup tinggi

6. *Tokenizing*

Pada proses *tokenizing* ini hal yang dilakukan adalah memecah kalimat-kalimat yang ada kedalam kata.

Tabel 3. Tokenezing

Kal1	Kal2	Kal3	…	Kal22
Era	Diera	Karena	…	Kegiatan
Ekonomi	Orde	Itu	…	Humas
Daerah	Baru	Sistem	…	yang
Saat	Organisasi	Informasi	…	Banyak

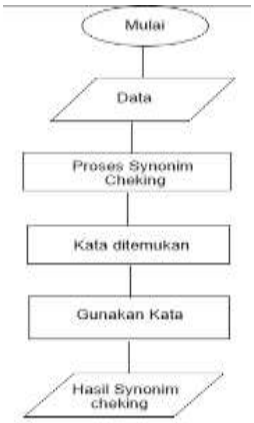
Ini	Pemerintah	Pada	…	Berhubungan
Sistem	Dan	Setiap	…	Dengan
Pemerintahan	Sistem	Daerah	…	Uniot
Daerah	Informasinya	Bisa	…	Kerja
Sudah	Ditentukan	Berbeda	…	Di
Berbeda	Oleh	Sesuai	…	Pemda
Dibandingkan	Pemerintah	Dengan	…	Yang
Dengan	Pusat	Perkembangan	…	Memiliki
Sistem	Di	Yang	…	Tingkat
Pemerintah	Era	Terjadi	…	Eselon
Diera	Otonomi	Kebutuhan	…	Yang
Orde	Daerah	Di	…	Lebih
baru	Ini	Daerah	..	Tinggi
	pembentukan	masingmasing	…	bisa

7. *Synonim Checking,*

Pada proses syinonim checking ini proses yang dikerjakan yaitu mengganti kata yang memiliki arti yang sama dengan sebuah kata induk yang telah dibuat dalam *database.* Pada dokumen uji kali ini kata yang masuk dalam proses sinonim yaitu misalkan pada kata "struktural, universal, skematis" yang nantinya akan digantikan oleh kata "sistematis". [9]

Tabel 4. Synonim Checking

Kal1	Kal2	Kal3	…	Kal22
Era	Diera	Karena	…	Kegiatan
Ekonomi	Orde	Itu	…	Humas
Daerah	Baru	Sistem	…	yang
Saat	Organisasi	Informasi	…	Banyak
Ini	Pemerintah	Pada	…	Berhubungan
Sistem	Dan	Setiap	…	Dengan
Pemerintahan	Sistem	Daerah	…	Uniot
Daerah	Informasinya	Bisa	…	Kerja
Sudah	Ditentukan	Berbeda	…	Di
Berbeda	Oleh	Sesuai	…	Pemda
Dibandingkan	Pemerintah	Dengan	…	Yang
Dengan	Pusat	Perkembangan	…	Memiliki
Sistem	Di	Yang	…	Tingkat
Pemerintah	Era	Terjadi	…	Eselon
Diera	Otonomi	Kebutuhan	…	Yang
Orde	Daerah	Di	…	Lebih
baru	Ini	Daerah	..	Tinggi
	pembentukan	masingmasing	…	bisa



Gambar 2. Proses Synonim checking

8. *Stopwords*

Pada proses *stowords removal* ini proses yang dikerjakan yaitu menghapus kata yang kurang relevan atau kata yang tidak memiliki arti yang begitu penting dan berkaitan yang ada pada kalimat di dokumen dengan mencocokan *list* kata stopword yang pada *database* yang akan diuji. *Database* yang digunakan baik untuk Bahasa Indonesia maupun Bahasa Inggris diambil dari sebuah situs yang menyediakan berbagain jenis *list stopword* untuk berbagai Bahasa.

Tabel 5. Stopword

Kal1	Kal2	Kal3	…	Kal22
Era	Diera	Karena	…	Kegiatan
Ekonomi	Orde	Itu	…	Humas
Daerah	Baru	Sistem	…	yang
Saat	Organisasi	Informasi	…	Banyak
Ini	Pemerintah	Pada	…	Berhubungan
Sistem	Dan	Setiap	…	Dengan
Pemerintahan	Sistem	Daerah	…	Uniot
Daerah	Informasinya	Bisa	…	Kerja
Sudah	Ditentukan	Berbeda	…	Di
Berbeda	Oleh	Sesuai	…	Pemda
Dibandingkan	Pemerintah	Dengan	…	Yang
Dengan	Pusat	Perkembangan	…	Memiliki
Sistem	Di	Yang	…	Tingkat

Pemerintah	Era	Terjadi	…	Eselon
Diera	Otonomi	Kebutuhan	…	Yang
Orde	Daerah	Di	…	Lebih
baru	Ini	Daerah	..	Tinggi
	pembentukan	masingmasing	…	bisa

4.3 Analisis Penerapan TF-IDF

Metode *Term Frequency-Inverse Document Frequency* (TF-IDF) adalah cara pemberian bobot hubungan suatu kata (term) terhadap dokumen. Untuk dokumen tunggal tiap kalimat dianggap sebagai dokumen. Metode ini menggabungkan dua konsep untuk perhitungan bobot, yaitu *Term Frequency* (TF) merupakan frekuensi kemunculan kata (t) pada kalimat (s). *Document frequency* (DF) adalah banyaknya kalimat dimana suatu kata (t) muncul.[3] Frekuensi kemunculan kata di dalam dokumen yang diberikan menunjukkan seberapa penting kata itu di dalam dokumen tersebut.
Langkah yang dilakukan adalah :

1. Hitung jumlah kemunculan dari masing-masing kata yang ada pada setiap kalimat yang ada di dokumen yang akan diuji. Pada pengujian kali ini, dokumen yang memiliki 22 kalimat (N).
2. Setelah menghitung jumlah kemunculan dari masing-masing kata di setiap kalimat yang ada pada dokumen uji tersebut, langkah selanjutnya yaitu menghitung total kemunculan dari masing-masing kata terhadap seluruh kalimat yang ada. Hasilnya disimpan pada kolom DF atau *Document Frequention.*
3. Langkah selanjutnya yaitu membagi nilai total keseluruhan kalimat yang ada pada dokumen uji dengan nilai total jumlah kemunculan dari masing-masing kata pada seluruh kalimat yang ada pada dokumen uji. Hasil dari pembagian tersebut disimpan pada kolom “N/DF”.
   N = Banyak kalimat = 22
   DF = Total kemunculan kata
   Maka
   N/DF [permasalahan] = 22/1 = 22
   N/DF [terjadi] = 22/1 = 22
   Hingga N/DF[ke n]..
4. Langkah selanjutnya yaitu melakukan perhitungan matematika dengan menggunakan rumus IDF yang berada pada persamaan 1.

$$IDF = log(\frac{N}{df}) \qquad (1)$$

   Perhitungan :
   IDF(1) = log (15/4) = 0.57403
   IDF(2) = log (15/1) = 1.17609
   IDF(3) = log (15/1) = 1.17609
   …… hingga IDF(idx ke n)

5. Langkah terakhir dari metode TF-IDF adalah menghitung bobot pada masing-masing kalimat, menggunakan persamaan 2.

$$W_{d,t} = TF_{d,t} * IDF_t \qquad (2)$$

dengan :

d = kalimat ke-d

t = kata (*term*) ke-t

TF = *Term Frequency*

W = bobot kalimat ke-d terhadap kata (*term*) ke-t

IDF= *Inverse Document Frequency*

Maka perhitungannya adalah :
Kalimat ke 1, kata ke 1 => W(1.1) = 2 * 0.176091259 = 0.3521825
Kalimat ke 2, kata ke 1 => W(2.1) = 0 * 0.176091259 = 0
Kalimat ke 3, kata ke 1 => W(2.1) = 0 * 0.176091259 = 0
Dst.. hingga
Kalimat ke (1,2,3), kata ke n => W(n.n) = x

Setelah selesai melakukan pembobotan dengan menggunakan metode *Term Frequention – Inverse Document Frequention* maka akan terbentuk matriks *term-by-sentence* dimana nilai-nilai yang sudah didapat pada proses ini akan digunakan untuk perhitungan nilai *cosine*

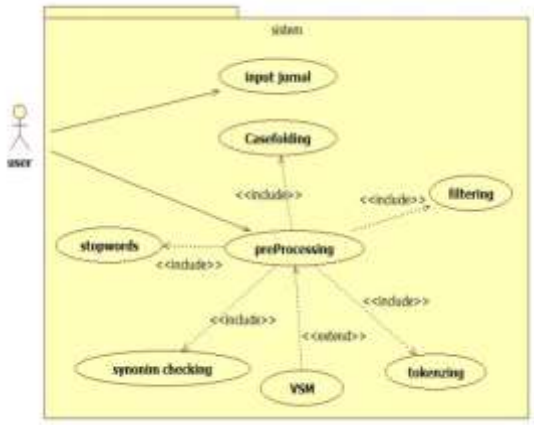
4.4 Pemodelan Sistem

Pemodelan sistem adalah bagian yang menjelaskan fungsi – fungsi yang dapat dilakukan oleh aplikasi berdasarkan analisis metode yang telah dilakukan sebelumnya dan mendeksripsikannya kedalam bentuk diagram diantaranya *usecase diagram, diagram scenario, activity diagram, sequence diagram,* dan *class diagram*

4.4.1 *Use Case Diagram*.

Use case Diagram berisi mengenai interaksi antara sekelompok proses dengan sekelompok aktor, menggambarkan fungsionalitas dari sebuah sistem yang dibangun dan bagaimana sistem berinteraksi dengan dunia luar. *Use case diagram* dapat digunakan selama proses analisis untuk menangkap kebutuhan sistem dan untuk memahami bagaimana sistem seharusnya bekerja. Pada gambar 3.4. menunjukan Diagram *Use Case* pada aplikasi

pencarian informasi berupa kata atau kalimat pada kumpulan dokumen menggunakan TF-IDF



Gambar 3. ***Use Case Diagram***

4.5 Skenario Pengujian

Skenario pengujian menggambarkan suatu tindakan yang akan dilakukan ketika pengujian terhadap sistem belangsung. Pada penelitian sistem pembentukan abstrak otomatis ini, akan dilakukan pengujian akurasi terhadap hasil abstrak pada sistem dan hasil abstrak manual. Hasil abstrak manual diperoleh dari 3 orang yang sedang menyusun tugas akhir yang membutuhkan kalimat abstrak, mengingat studi kasus yang diambil di dalam penelitian ini yaitu pembentukan abstrak pada dokumen jurnal. 3 orang tersebut akan menguji perbandingan hasil abstrak manual dengan hasil abstrak oleh sistem.

Data yang digunakan dalam penelitian skripsi ini berjumlah 15 dokumen jurnal berformat umum. Pembentukan abstrak dilakukan dengan mengambil kalimat pada setiap bagian yaitu latar belakang, analisis, dan kesimpulan dalam dokumen jurnal, baik untuk hasil yang dilakukan secara manual maupun hasil dari sistem. Jika > 2 orang memilih kalimat yang sama maka kesimpulannya kalimat tersebut terpilih, dan jika < 1 maka kalimat tersebut tidak terpilih. .

4.5.1 Kesimpulan Hasil Pengujian

Kesimpulan yang diperoleh dari hasil uji menyatakan bahwa hasil pembentukan abstrak secara manual dengan sistem menghasilkan persentase kalimat yang sama tidak terlalu jauh. Dimana jarak antara hasil manual dengan sistem rata-rata hanya berbeda 2 sampai 4 kalimat saja. Akurasi dengan persentase 100% didapat oleh jurnal ke 7 dan 9 sedangkan akurasi dengan pesentase terendah 60% didapat oleh jurnal ke 1.

Tabel 1. Kesimpulan Pengujian Akurasi

Jurnal	Akurasi kalimat yang sama	Persentase
1	1,14,16,19,20,22	60%
2	2,4,7,12,16,17,19	80%
3	1,3,5,7,12,14,16,17,19	90%
4	2,5,11,16,21,23,24	70%
5	4,6,14,18,20,22,25,26	70%
6	1,2,5,11,14,20,21	70%
7	3,5,8,9,11,12,14,16,18	100%
8	1,3,8,12,15,19,20,22	90%
9	2,4,6,10,11,14,15,18,20,22,25,27	100%
10	1,7,11,12,16,21,23,24	70%
11	1,7,11,12,16,18,19,21,22	90%
12	3,4,7,9,11,13,16,17,18	80%
13	1,4,7,11,14,16,17	70%
14	1,3,7,8,14,16,17,20,22	90%
15	3,8,9,12,13,15,16,18,19	90%

3. PENUTUP

Berdasarkan hasil yang didapatkan dalam penelitian dan penyusunan skripsi ini maka dapat disimpulkan bahwa pembentukan kalimat abstrak secara otomatis menggunakan metode *Term Frequency – Inverse Document Frequency (TF-IDF)* mampu membentuk paragraf abstrak pada sebuah dokumen jurnal secara otomatis. Kesimpulan yang diperoleh dari hasil uji menyatakan bahwa hasil pembentukan abstrak secara manual dengan sistem menghasilkan persentase kalimat yang sama tidak terlalu jauh. Dimana jarak antara hasil manual dengan sistem rata-rata hanya berbeda 2 sampai 4 kalimat saja. Akurasi dengan persentase 100% didapat oleh jurnal ke 7 dan 9 sedangkan akurasi dengan pesentase terendah 60% didapat oleh jurnal ke 1. Adapun saran untuk pengembangan sistem pembentukan abstrak ini yaitu penyempurnaan aplikasi masih harus dilakukan, saat ini dokumen yang digunakan hanya dalam format DOC.

DAFTAR PUSTAKA

[1] Gary Blake and Robert W. Bly, 1993 . *The Elements of Technical Writing*, pg. 117. New York:Macmillan Publishers, ISBN 0020130856

[2] Taufiq M. Isa 1) dan Taufik Fuadi Abidin 2)., "Mengukur tingkat kesamaan paragraf menggunakan vector space model untuk mendeteksi plagiarisme" Seminar Nasiaonal dan ExpoTeknik Elektro, 1) jurusan matematika. FMIPA, Universitas Syiah Kuala, 2) Jurusan informatika, FMIPA, Universitas Syiah Kuala, 2013.

[3] R. S. Pressman, Software Engineering: A Practitioner's Approach.

[4] Raymond J. Mooney. CS, 391L: Machine Learning Text Categorization. University of Texas at Austin, 2006.

[5] P. P. Widodo dan Heriawati. "Menggunakan

Unified Modeling Language (UML)". Bandung Informatika, 2011.

[6] Wahana Komputer, Shortcourse Series: Microsoft Visual C# 2010. Andi, 2011

[7] Deny Ocr.(2014, Aug.10) *Pengenalan DevExpress* [online]. Available : http://www.jagocoding.com/tutorial/600/Pengenalan_DevExpress_Windows_Form

[8] Sholechul Aziz, Jurus Andalan Menguasai EYD (Ejaan yang Disempurnakan). IV. Jakarta : Kunci Komunikasi, 2015.

[9] S Mohd. Ehmer Khan. "Different Approaches to White Box Testing Technique for Finding error", 2011

[10] Many, I. and Maybury. 1999. *Advance in Automatic Text Summarization*. The MIT Press: Cambrige.

[11] Zaman B. dan E Winarko. 2011. Analisis Fitur Kalimat untuk Peringkas Teks otomatis pada Bahasa Indonesia. *Indonesian Journal of Computing and*.