PENERAPAN METODE *COSINE SIMILARITY* DAN PEMBOBOTAN *TF-IDF* PADA SISTEM KLASIFIKASI SINOPSIS BUKU DI PERPUSTAKAAN KEJAKSAAN NEGERI JEMBER

[1]Moh. Mahdi Sya'bani (1510651015), [2]Reni Umilasari, S.Pd, M.Si
Jurusan Teknik Informatika
Fakultas Teknik
Universitas Muhammadiyah Jember
E-mail : [1]mahdiumj@gmail.com

Abstrak – Selama ini perpustakaan Kejaksaan Negeri Jember belum dikelola dengan baik. Pada saat pegawai perpustakaan ingin mengetahui macam-macam judul buku sesuai kategori yang mereka inginkan, pegawai perpustakaan mencari satu persatu di katalog bukunya. Sehingga kondisi demikian akan menyulitkan pegawai perpustakaan dalam mencari judul buku sesuai kategori yang diinginkan. Hal ini dapat mengakibatkan pegawai perpustakaan kewalahan. Pelayanan yang sangat baik jika pengguna perpustakaan merasa puas dengan pelayananya. Semakin banyaknya dokumen buku yang ada di perpustakaan semakin banyak tenaga dan waktu yang diperlukan. Oleh karena itu penelitian ini membangun sebuah sistem aplikasi yang mampu mengklasifikasikan dokumen buku berdasarkan kategori buku secara otomatis. Untuk mendapatkan hasil yang optimal dalam mengklasifikasikan sebuah dokumen maka diperlukan sebuah metode untuk mengklasifikasikan dokumen. Metode yang digunakan adalah pembobotan *TF/IDF* dan *cosine similarity* pada *model vector space model*. Untuk mengukur tingkat kemiripan suatu dokumen dengan menggunakan sinopsis buku. Sinopsis merupakan sebuah ikhtisar karangan atau abtraksi yang biasanya diterbitkan bersama-sama dengan karangan asli yang menjadi dasar.

Data latih yang digunakan pada pengujian aplikasi ini berjumlah 120 dokumen sinopsis dengan 10 kategori yang berbeda dan menghasilkan nilai *precision* sebesar 90,91% pada *threshold 0,1* dan nilai *recall* sebesar 100% pada *threshold 0,1* dan *0,2*. Tingkat ketepatan klasifikasi sistem sebesar 80,83%.

Kata kunci – Perpustakaan Kejaksaan Negeri Jember, Klasifikasi Dokumen, Sinopsis Buku, Pembobotan *TF/IDF* dan *Cosine Similarity.*

I. PENDAHULUAN

Kejaksaan Negeri terdapat sebuah perpustakaan yang dapat menjadi suatu sarana meningkatkan kinerja terutama dalam menangani suatu perkara. Tujuannya adalah mengembangkan sarana yang ada dan memberikan pencerahan terhadap pegawai yang ada di Kejaksaan Negeri, termasuk di Kejaksaan Negeri Jember. Perpustakaan yang berisi tentang hukum dan dokumentasi perundangan-udangan dalam bentuk unit satuan kerja untuk menangani khusus dalam ilmu hukum. Terdapat 1.966 judul buku yang di perpustakaan tersebut dan di dalamnya terdapat banyak kategori buku yang di antaranya HAM Nasional, Hukum Asing, Hukum Pidana, Hukum Perdata dan lain-lain.

Dokumen buku yang ada di perpustakaan Kejaksaan Negeri Jember belum dikelola dengan baik. Pada saat pegawai perpustakaan ingin mengetahui macam-macam judul buku sesuai kategori yang mereka inginkan, pegawai perpustakaan mencari satu persatu di katalog bukunya. Sehingga kondisi demikian akan menyulitkan pegawai perpustakaan dalam pencarian judul buku sesuai kategori yang diinginkan. Hal ini dapat mengakibatkan pegawai perpustakaan kewalahan. Pelayanan suatu perpustakaan dikatakan prima (sangat baik) jika para pengguna perpustakaan merasa puas atas pelayanan yang diberikan [1]. Semakin banyak jumlah buku atau dokumen yang tersedia maka semakin banyak pula waktu dan tenaga yang diperlukan. Dari sinilah faktor penyebabnya menjadi topik menarik untuk sebuah penelitian. Dengan demikian diperlukan sebuah sistem yang dapat mengklasifikasikan dokumen secara otomatis. Untuk mendapatkan hasil yang optimal dalam mengklasifikasikan dokumen ini mengambil dataset dari sebuah sinopsis buku yang ada di perpustakaan.

Penerapan klasifikasi dokumen membutuhkan metode yang dapat digunakan untuk mengklasifikasikan dokumen secara otomatis sesuai kategorinya adalah dengan menggunakan metode text mining. Metode ini dapat mengklasifikasikan dokumen. Terdapat banyak algoritma salah satunya adalah cosine similiraty pada *model vector space model*. Metode tersebut dapat mengklasifikasikan yaitu memberi pembobotan pada suatu dokumen dengan menghitung *TF* (*Term Frekuensi*) atau *IDF* (*Inverse Dokumen Frekuensi*). Tiap *term* diasumsikan memiliki nilai kepentingan yang sebanding dengan jumlah kemunculan term tersebut pada teks [2].

Tujuan dari penelitian ini adalah harapannya dengan dibuatnya sistem ini bertujuan dapat membantu pegawai perpustakaan dalam menimalisir pekerjaannya dalam mencari kategori buku.

II. TINJAUAN PUSTAKA

A. Klasifikasi Dokumen

Pada tahapan proses klasifikasi dokumen secara keselurahan, terdapat beberapa tahapan yang akan diawali dengan identifikasi dokumen yakni dimana masing-masing

dokumen akan diidentifikasi secara kata atau term yang terdapat didalamnya. Sehingga tahap pertama adalah tokenizing pada kata yang terdapat di dalam dokumen untuk mendapatkan kata yang mampu berdiri sendiri, dan terbebas dari tanda-tanda baca, spasi dan sebagainya [3]. Selanjutnya tahap filtering (*wordlist/stoplist*) untuk menghilangkan kata yang tidak berpotensi sebagai indikasi topic dalam dokumen [4]. Setelah itu dilakukan stemming pada kata yang tersisa untuk mendapatkan kata dasar.

B. Sinopsis

Sinopsis merupakan ikhtisar karangan yang biasanya diterbitkan bersama-sama dengan karangan asli yang menjadi dasar sinopsis itu, ringkasan atau abstraksi [5]. Sinopsis mengandung tiga pengertian yaitu ikhtisar karangan, ringkasan, atau abstraksi, [6] menyatakan bahwa ringkasan sumarry précis adalah suatu cara yang efektif Funtuk menyajikan suatu karangan yang panjang dalam bentuk pendek. Kata précis berarti memotong atau meringkas. Dengan demikian meringkas ibarat memangkas sebatang pohon yang akhirnya tinggal batang dan cabang-cabang yang terpenting.

C. *Text Mining*

Text mining adalah sebuah proses pengetahuan intensif dimana pengguna berinteraksi dan bekerja dengan sekumpulan dokumen dengan menggunakan beberapa alat analisis. Text mining mencoba untuk mengekstrak informasi yang berguna dari sumber data melalui identifikasi dan eksplorasi pola yang menarik.

Tahap proses awal terhadap teks untuk mempersiapkan teks menjadi data yang diolah lebih lanjut. Sekumpulan karakter yang bersambungan (teks) harus dipecah-pecah menjadi unsur yang lebih berarti. Suatu dokumen dapat dipecah menjadi bab, sub-bab, paragraf, kalimat, kata dan bahkan suku kata. Tokenizing adalah proses memecah teks menjadi kalimat dan kata/token [7]. Fitur ini terdiri dari tipe kapitalisasi, keberadaan digit, tanda baca, karakter spesial dan lain-lain. Tahap pre *processing* yang dilakukan secara umum dalam text mining pada dokumen yaitu *casefolding*, *tokenizing*, *filtering* dan *stemming*.

D. Pembobotan *TF/IDF* dan *Cosine Similarity*

Metode *TF/IDF* (*Term Frequency* / *Invers Document Frequency*) merupakan suatu cara untuk memberikan bobot hubungan suatu kata (term) terhadap dokumen [8]. 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 tersebut. Frekuensi dokumen yang mengandung kata menunjukkan seberapa umum kata itu muncul. Sehingga bobot hubungan antara sebuah kata dan sebuah dokumen akan tinggi apabila frekuensi kata itu tinggi didalam dokumen dan frekuensi keseluruhan dokumen yang mengandung unsur kata yang rendah pada kumpulan dokumen (*database*).

Rumus umum untuk *TF-IDF* dan *Cosine Similarity*:

$$IDF = tf_{ti}\ x\ IDF \quad (1)$$

$$IDF = tf_{ti}\ x\ log\frac{N}{df_{ti}} + 1 \quad (2)$$

$$\omega_D\ (ti) = \frac{tf_{ti} x \log\left(\frac{N}{df}\right)+1}{\sqrt{\sum(tf_{ti} x \log\left(\frac{N}{df_{ti}}\right)+1)}} \quad (3)$$

$$cos(Q,D) = \sum_{r=1}^{M} \omega_Q(t_i)\ x\ \omega_D(t_i) \quad (4)$$

III. METODE PENELITIAN

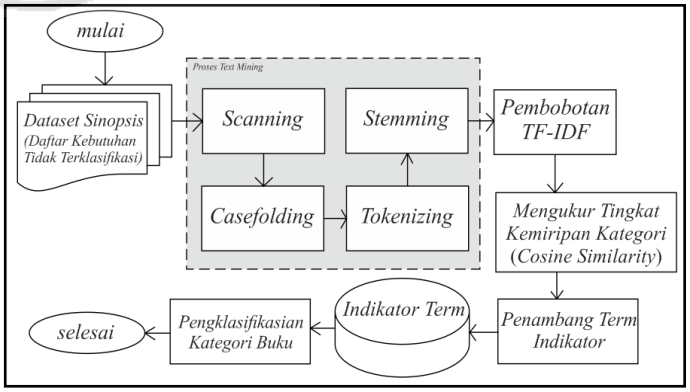
A. Tahap Penelitian

Pada tahap ini, penelitian dilakukan dengan mengumpulkan data-data yang berisi dari sinopsis suatu judul buku yang ada di perpustakaan Kejaksaan Negeri Jember. Dataset ini bertipe excel tetapi belum dikelola dengan baik oleh perpustakaan Kejaksaan Negeri Jember. Dataset yang digunakan 120 data buku yang berisi atribut yaitu sinopsis, judul buku, penerbit, jumlah buku dan tahun terbit. Untuk kategori atau jenis bukunya terdiri 10 kategori judul buku yang dapat dilihat pada Tabel 1.

TABEL I. JENIS KATEGORI BUKU

No	KATEGORI	No	KATEGORI
1	HAM NASIONAL	6	HUKUM TATA NEGARA
2	HAM ASING	7	HUKUM INTERNASIONAL
3	HUKUM UMUMNYA	8	SOSIAL DAN POLITIK
4	HUKUM PIDANA	9	PENELITIAN PENGKAJIANNYA
5	HUKUM PERDATA	10	PERUNDANGAN-UNDANGAN

Pada Gambar 1 dijelaskan bahwa *flowchart* metode *cosine similarity* ini menjelaskan kerangka kerja sistem aplikasi :



Gambar 1. Rancangan Sistem Aplikasi

B. Metode

Tahap data *sample* digunakan untuk melihat ketepatan pada klasikasi pada dokumen. Untuk mendapatkan klasifikasi yang efektif, membutuhkan *keyword* (kata kunci) setiap

kategorinya. Tahap ini akan menggunakan 2 dokumen setiap kategori buku yang ditunjukan Tabel I.

Tahap *training* dalam penelitian ini berjumlah 120 dokumen sinopsis buku dengan kategori yang berbeda. Sistem ini bertujuan untuk mengklasifikasikan dokumen seacra otomatis berdasarkan kategori yang ada.

1. Sebelumnya, setiap baris kebutuhan yang akan diklasifikasikan dilakukan proses *scanning* sampai proses *stemming* dan pembuangan *stopword* agar memperoleh *term-term* dasar.
2. Proses pembobotan pada setiap daftar kata sebagai berikut.

a. Metode *TF*

Menentukan nilai frekuensi kemunculan kata.

TABEL II. DAFTAR FREKUENSI KATA (*TF*)

Kategori HAM Nasional	*Term*	*TF*
Dokumen 1	angkat	*1*
	kitab	*2*
	regional	*2*
Dokumen 2	kitab	*2*
	tambah	*1*
	regional	*2*

b. Metode *IDF*

Berdasarkan daftar kata dalam Tabel II, maka dihitung *IDF* untuk setiap kata menggunakan persamaan 2.

Sebagai contoh hubungan kata *angkat* pada Dokumen 1 :

$$IDF = tf_{ti}\ x\ log\ \frac{N}{df_{ti}} + 1$$

$$= 1 \text{ x } log\ \frac{2}{1} + 1 = 1{,}301$$

Dengan cara yang sama untuk semua kata yang lain, maka akan diperoleh hasil perhitungan IDF sebagaimana Tabel III.

TABEL III. HASIL PERHITUNGAN *IDF*

Kategori HAM Nasional	*Term*	*IDF*
Dokumen 1	angkat	*1, 301*
	kitab	*1*
	regional	*1*
Dokumen 2	kitab	*1*
	tambah	*1, 301*
	regional	*1*

c. Metode *TF-IDF*

Berdasarkan Tabel II dan III, dilakukan pembobotan *TF-IDF* melalui persamaan 3 sehingga dihasilkan bobot .

Sebagai contoh kata angkat pada Dokumen 1:

$$\omega_D\ (ti) = \frac{tf_{ti} x \log\left(\frac{N}{df}\right)+1}{\sqrt{\sum(tf_{ti} x \log\left(\frac{N}{df_{ti}}\right)+1)}}$$

$$= \frac{1\ x \log\left(\frac{2}{1}\right)+1}{\sqrt{\sum(1\ x \log\left(\frac{2}{1}\right)+1)}} = \frac{1{,}301}{\sqrt{1\ x\ 3.301}} = 0{,}7161$$

Dengan cara yang sama untuk semua kata yang lain, maka akan diperoleh hasil perhitungan IDF sebagaimana Tabel IV.

TABEL IV. HASIL PERHITUNGAN TF-*IDF*

Kategori HAM Nasional	*Term*	*TF-IDF* $\omega_D\ (ti)$
Dokumen 1	angkat	*0, 7161*
	kitab	*0, 5503*
	regional	*0, 5503*
Dokumen 2	kitab	*0, 5503*
	tambah	*0, 7161*
	regional	*0, 5503*

d. Metode *Cosine Similarity*

Pengukuran tingkat kemiripan kebutuhan dalam pengklasifian kategori berdasarkan persamaan 4.

$$cos(Q, D) = \sum_{r=1}^{M} \omega_Q(t_i)\ x\ \omega_D(t_i)$$

$$= (\omega_{Q1T1} * \omega_{D1T1}) + (\omega_{Q1T2} * \omega_{D1T2}) + (\omega_{Q1T3} * \omega_{D1T3}) + (\omega_{Q1T4} * \omega_{D1T4})$$

$$= (0{,}7161 * 0) + (0{,}5503 * 0{,}5503) + (0{,}5503 * 0{,}5503) + (0 * 0{,}7161)$$

$$= \mathbf{0{,}6056}$$

Pada penambang *term* indikator *Q* ditetapkan sebagai tipe kategori dokumen tertentu, dan *D* ditetapkan sebagai kebutuhan yang terklasifikasi dalam tipe kategori *Q*. *Term-term* indikator dari tipe kategori *Q* ditemukan dengan mempertimbangkan sekumpulan kebutuhan *D* dari semua tipe kategori *Q* pada suatu data *training*.

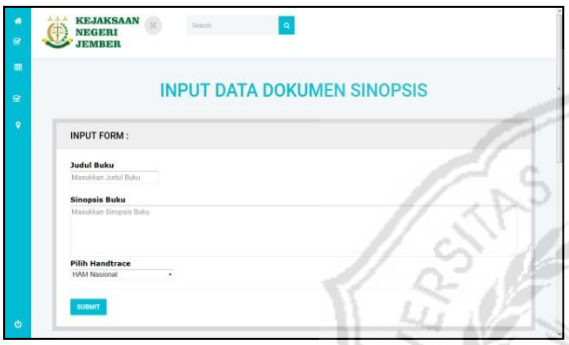
Masing-masing *term* diberikan suatu nilai pembobotan *TF-IDF*, dan *term-term* ini diurutkan secara menurun. Term akan di rangking dan teratas diidentifikasi sebagai *term-term* indikator pada tipe kategori *Q* tertentu.

Perhitungan manual berdasarkan kesusaian makna dan term potensial dari kebutuhan, sehingga total jumlah kebutuhan yang terambil (retrived) dan jumlah relevan berdasarkan handtrace (relevan) akan membentuk kinerja rata-rata precision dan recall dari semua dataset dengan menggunakan nilai ambang batas atau threshold. Disini menguji 2 dokumen sinopsis setiap kategorinya, hasil dari perhitungan manual akan dibandingkan dengan handtrace yang ada.

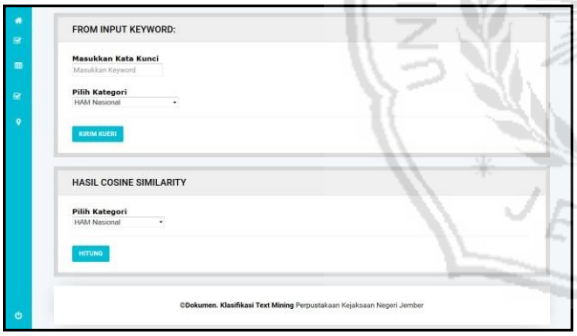
IV. HASIL DAN PEMBAHASAN

A. Implementasi Halaman Aplikasi

Pada tahap implementasi ini dilakukan terhadap dataset sebesar 120 data yang berisikan sinopsis buku. Dataset tersebut akan diolah oleh aplikasi dengan fungsi-fungsi yang ada yaitu *create* data, *delete* data dan *read* data. Tahap penelitian ini akan menggunakan perhitungan pembobotan *TF/IDF* dan *cosine similarity* yang akan berfungsi untuk menghasilkan klasifikasi sinopsis buku berdasarkan kategori buku. Pengujian aplikasi ini menekankan seberapa kuat akan menghasilkan kebutuhan-kebutuhan yang terklasfikasi dalam setiap kategorinya memiliki nilai ambang batas (*threshold*). Tampilan halaman aplikasi menginputkan data pada Gambar 2 dan Gambar 3.

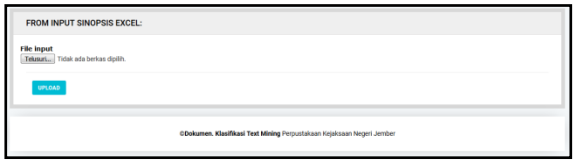


Gambar 2. Tampilan Input 1



Gambar 3. Tampilan Input 2

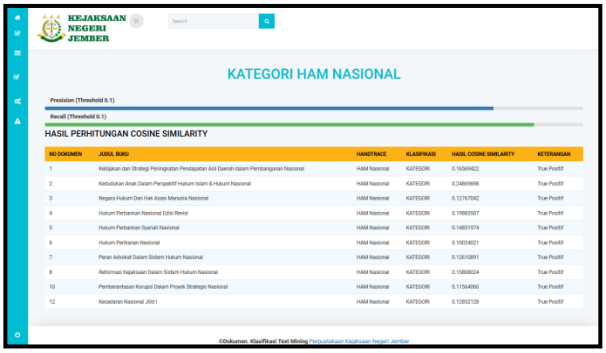
Berikutnya Gambar 4 terdapat input-an untuk memasukan kata kunci baru (keyword) dan sesuai dengan kategori yang akan dipilih dari 10 kategori yang ada. Untuk implementasi aplikasi kali ini menggunakan data training sebanyak 120 data.



Gambar 4. Tampilan *file import*

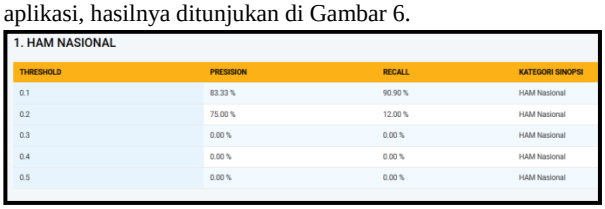
B. Implementasi Halaman Pengujian Klasifikasi

Pengujian aplikasi mengambil contoh kategori yaitu HAM Nasional dan beserta Perhitungan *Threshold*. Data yang digunakan 120 sinopsis dan dapat dilihat pada Gambar 5 dan Gambar 6.



Gambar 5. Hasil Klasifikasi HAM Nasional

Selanjutnya hasil perhitungan *threshold* menggunakan aplikasi, hasilnya ditunjukan di Gambar 6.



Gambar 6. Hasil *threshold* HAM Nasional

Dari 120 sinopsis menghasilkan nilai *threshold* dari kategori HAM Nasional dan rata-rata nilai *precision* tertinggi terletak pada *threshold* 0,1 sebesar **83,33%**. Kinerja *recall* tertinggi terletak pada *threshold* 0,1 sebesar **90,90%**.

C. Tabulasi Perbandingan Kinerja Rata-rata *Threshold*

Hasil dari keseluruhan sinopsis, akan membuat suatu tabulasi perbandingan dari 10 kategori yang diuji ke aplikasi.

TABEL V. TABULASI *THRESHOLD*

Threshold	HAM Nasional		HAM Asing		Hukum pada Umunya		Hukum Pidana		Hukum Perdata	
	presisi	*recall*	*presisi*	*recall*	*presisi*	*recall*	*presisi*	*recall*	*presisi*	*recall*
0,1	83,33 %	90,90 %	80,65 %	100,00 %	79,49%	80,00%	90,91 %	60,00 %	87,32 %	73.81 %
0,2	75,00 %	12,00 %	85,71 %	100,00 %	82,61%	66,66%	72,73 %	33,33 %	85,71 %	57.69 %
0,3	0 %	0%	0 %	0%	66,67%	44,44%	0 %	0%	75,00 %	12,00 %
0,4	0 %	0%	0 %	0%	0 %	0%	0 %	0%	0 %	0%
0,5	0 %	0%	0 %	0%	0 %	0%	0 %	0%	0 %	0%
Threshold	Hukum Tata Negara		Hukum Internasional		Sosial dan Politk		Penelitian & Pengkajiannya		Perundang – undangan	
	presisi	*recall*	*presisi*	*recall*	*presisi*	*recall*	*presisi*	*recall*	*presisi*	*recall*
0,1	73,91 %	100,00 %	80,85 %	71,42 %	83,33 %	70,31 %	83,93 %	68,12 %	74,60 %	68,12 %
0,2	85,00 %	100,00 %	82,61 %	62,50 %	83,33 %	51,28 %	85,29 %	56,86 %	82,86 %	56,39 %
0,3	0 %	0%	75,00 %	38,46 %	0 %	0%	75,00 %	12,00 %	0 %	0%
0,4	0 %	0%	0 %	0%	0 %	0%	0 %	0%	0 %	0%
0,5	0 %	0%	0 %	0%	0 %	0%	0 %	0%	0 %	0%

Evaluasi dari perbandingan 10 kategori sinopsis buku hasil pengujiannya yang ditunjukan dalam tabel 4.1 dapat ditemukan bahwa nilai precision tertinggi yang telah diarsir sebesar 90,91% pada threshold 0,1 dan nilai recall yang diarsir tertinggi sebesar 100% pada threshold 0,1 dan 0,2. Kinerja rata-rata threshold yang ditunjukan Tabel 4.1 menunjukan bahwa aplikasi telah berhasil mengindentifikasi kebutuhan-kebutuhan klasifikasi synopsis dengan 10 kategori.

D. Akurasi

Hasil pada penelitian ini dari 120 data sinopsis buku yang telah berhasil di uji dengan 10 kategori dan setiap

kategori memiliki kata kunci (keyword) masing-masing menggunakan aplikasi tersebut sehingga dapat menghasilkan tingkat akurasi algoritma Cosine Similarity dan Pembobotan TF-IDF.

$$\text{Tingkat akurasi} = \frac{TP+TN}{TP+TN+FP+FN} x\ 100\%$$

Sehingga dari 120 data yang di uji dengan 10 kategori yang dinyatakan benar dalam mengklasifikasikan sinopsis buku ada 97 record.

$$\text{Tingkat Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} x\ 100\%$$

$$= \frac{82+15}{82+15+19+4} x\ 100\% = \mathbf{80{,}83\%}$$

Berikut tampilan perhitungan akurasi keseluruhan :

Gambar 7. Tampilan Akurasi Sistem Aplikasi

V. KESIMPULAN

Berdasarkan pengujian yang telah dilakukan, terdapat beberapa kesimpulan yang dapat diambil sebagai berikut :

1. Perbandingan keseluruhan dari pengujian aplikasi menunjukan bahwa nilai *precision* tertinggi sebesar 90,91% pada *threshold* 0,1 dengan kategori Hukum Pidana dan nilai *recall* tertinggi sebesar 100% pada threshold 0,1 dan *threshold* 0,2 dengan kategori HAM Asing dan Hukum Tata Negara.
2. Dari pengujian aplikasi yang telah dilakukan dengan 10 tipe kategori berbeda maka dapat ditemukan threshold yang terbaik adalah 0,1.
3. Dari sejumlah 120 data yang terdiri dari 10 kategori, aplikasi mampu melakukan proses analisa dengan sejumlah 97 data yang berhasil diklasifikasikan dengan benar dan 23 data yang tidak valid, maka akurasi ketepatan klasifikasi aplikasi sebesar 80,83%.

Penelitian ini tentunya masih perlu banyak pengembangan adalah penelitian bisa menggunakan metode klasifikasi lainnya dan pada aplikasi, untuk penelitian selanjutnya bisa mengembangkan aplikasi tersebut dengan lebih efisien dalam mengeksekusi waktu untuk menghitung sebuah proses sistem klasifikasi

VI. REFERENSI

[1] Prastowo, A. (2013). Manajemen perpustakaan sekolah profesional.Yogyakarta: DIVA Press.

[2] Mark A. Hall., & Lloyd A. Smith. (1999). *Feature Selection for Learning: Comparing a Correlation a Corelation-based Filter Approach to the Wrapper* . In FLAIRS Conference.

[3] Kaplan, R.M. (1995). *A Methode for Tokenizing Text*. Palo Alto Research Center (Festscrift in The Honor of Prof. Kimmo Koskenniemi's 60 th Anniversary).

[4] Porter, M.F.2001. *Snowball: A language for Stemming Algorithms Computer Laboraty*, Cambridge (England).

[5] Kamus Besar Bahasa Indonesi (KBBI) [online]. Tersedia di https://kbbi.web.id/sinopsis. Diakses 24 Juni 2019

[6] Keraf, Gorys. 1977. Tata Bahasa Indonesia, Ende: Nusa Indah. Hal. 84.

[7] Feldman, R dan Sanger, J., (2007). *Text Mining Handbook: Advanced Approaches in Analyzing Unstructered Data*. New York: Cambridge University Press.

[8] Robertson, S.. (2005). *Understanding Inverse Document Frequency: On Theoretical Arguments for IDF*. England: Journal of Documentation, Vol. 60, 502-520.