IMPLEMENTASI EKSTRAKSI INFORMASI PADA SURAT MASUK MENGGUNAKAN METODE NAIVE BAYES

Chandra Ratiwi[1], Nelly Indriani[2]

Program Studi Teknik Informatika
Fakultas Teknik dan Ilmu Komputer Universitas Komputer Indonesia
Jl. Dipati Ukur 114 Bandung
E-mail : chandrartw@gmail.com[1], indiwidi@gmail.com[2]

ABSTRAK

Ekstraksi informasi merupakan suatu metode dari text mining yang bertujuan untuk mencari, mengambil ataupun mengolah informasi - informasi yang terkandung dalam suatu objek. Pada hasil penelitian ekstraksi informasi pada dokumen sebelumnya memiliki kelemahan yang karena metode *rule-based* harus memiliki dokumen yang terstruktur dan terformat, maka dalam penelitian ini akan digunakan metode *Naive Bayes* sebagai algoritma machine learning untuk mengekstraksi informasi dengan pengklasifikasian setiap kalimat yang terdapat pada surat masuk,karena surat masuk tidak selalu identik dengan format yang sudah ada. Sebelum surat di klasifikasikan oleh *Naive Bayes* surat akan melalui proses preprocessing terlebih dahulu yang terdiri dari *filtering*, *tagging*, dan *tokenization*. Setelah preprocessing dilakukan data diberi nilai bobot pada setiap fitur yang terdapat pada teks kalimat untuk perhitungan probabilitas *Naive Bayes*. Berdasarkan hasil pengujian confusion matrix didapatkan hasil yang cukup baik dan dapat disimpulkan bahwa algoritma *Naive Bayes* dapat diterapkan untuk sistem ekstraksi informasi pada surat masuk dengan hasil akurasi yang sangat tinggi yakni 96,96%.

Kata Kunci : Ekstraksi Informasi, Supervised Learning, Klasifikasi, *Naive Bayes*, Surat Masuk

1. PENDAHULUAN

Surat merupakan salah satu alat yang digunakan untuk berkomunikasi antara satu pihak dan pihak lainnya untuk menyampaikan suatu informasi yang tertulis. Setiap surat memiliki nilai informasi yang sudah pasti seperti nama instansi pengirim, alamat, tanggal, perihal dan sebagainya. Untuk mendapatkan informasi tersebut maka timbul kebutuhan untuk mengekstraksi informasi secara otomatis dari data yang berupa teks pada surat tersebut agar dapat disimpan dalam bentuk data[1].

Ekstraksi informasi adalah pencarian otomatis pada informasi yang terstruktur seperti entitas, hubungan antar entitas, dan atribut yang menggambarkan entitas dari sumber yang tidak terstruktur [2]. Pencarian dari suatu ekstraksi informasi merupakan sebuah sistem yang berfungsi untuk mencari data spesifik dalam *Natural Language Text.* Dalam sebuah *template* formulir atau tabel serta komponen – komponen teks yang ada diambil datanya untuk diekstraksi menjadi data yang memiliki kegunaan [3]. Ekstraksi informasi pada teks berpola sebelumnya pernah diteliti oleh Dimas Mustaqwa dengan judul "Implementasi Ekstrasi Informasi Pada Dokumen Skripsi Menggunakan Metode Ruled Based" akan tetapi walaupun tingkat akurasi yang dihasilkan oleh metode tersebut sudah 100% namun metode tersebut harus memiliki pola yang sama persis dengan dataset sehingga tidak cocok dengan teks yang tidak selalu memiliki pola yang sama [4]. Berdasarkan masalah pada rule based tersebut maka menurut Marina Dorash bahwa solusi yang dapat dilakukan adalah dengan menggunakan *machine* learning[5]. Salah satu metode dari *machine learning* adalah *Naïve Bayes. Naïve Bayes* memiliki beberapa keunggulan dan kelebihan dari algoritma lainnya yaitu sederhana, cepat dan memiliki tingkat akurasi yang tinggi [6]. Ekstraksi informasi pernah dilakukan oleh Oman Sumantri dalam penelitiannya bahwa *Naïve Bayes* memiliki tingkat akurasi 14% lebih baik dibandingkan dengan SVM [7]. Hal tersebut diperkuat bahwa berdasarkan penelitian yang dibuktikan oleh Xhemali, Hinde dan Stone dalam penelitiannya mengatakan bahwa *Naïve Bayes Classifier* memiliki tingkat akurasi yang lebih baik dibandingkan *Decision Tree* dan *Neural Network* [8].

2. LANDASAN TEORI

Ekstraksi Informasi adalah pencarian otomatis pada informasi yang terstruktur seperti entitas, hubungan antar entitas, dan atribut yang menggambarkan entitas dari sumber yang tidak terstruktur [2]. Pencarian dari suatu ekstraksi informasi merupakan sebuah sistem yang berfungsi untuk mencari data spesifik dalam *Natural Language Text*. Dalam sebuah template formulir atau tabel serta komponen – komponen teks yang ada diambil datanya untuk diekstraksi menjadi data yang memiliki kegunaan [3]. Ekstraksi informasi berperan sebagai sistem yang akan mengenali data yang tidak terstruktur yang memiliki informasi yang belum dikategorikan dan belum memiliki arti yang spesifik. Tidak terstruktur bukan berarti secara data yang membingungkan atau tidak jelas dengan artian bahwa informasi yang terkandung didalamnya tidak dapat langsung diterjemahkan oleh komputer sehingga membutuhkan sistem yang dinamakan ekstraksi informasi.

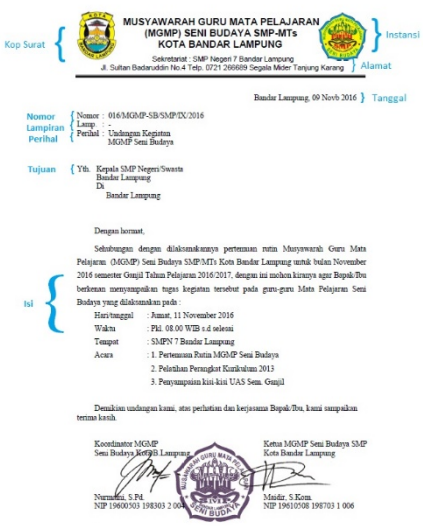
3. METODE PENELITIAN

3.1 Analisis Surat

Sumber data surat masuk bersifat resmi didapat secara fisik maupun internet dengan aturan surat resmi dari dinas. Dalam data surat masukan memiliki batasan yakni :

1. Format dokumen surat berekstensi .pdf
2. Format surat harus sesuai aturan surat resmi.
3. Logo instansi, cap, dan tanda tangan tidak termasuk dalam ekstraksi informasi pada penelitian.

Berikut adalah hal yang terdapat dalam surat:



Gambar 1. ***Surat***

3.2 Analisis Sistem

Analisis sistem dapat didefinisikan sebagai penguraian dari suatu sistem yang utuh kedalam bagian – bagian komponennya dengan maksud untuk mengidentifikasikan dan mengevaluasi permasalahan dan hambatan yang ada sehingga dapat di usulkan perbaikan – perbaikannya dalam membantu proses perancangan sistem tersebut. Analisis sistem Pengimplementasi Ekstraksi Informasi pada Surat Masuk dengan Metode Naïve Bayes meliputi beberapa bagian yaitu :

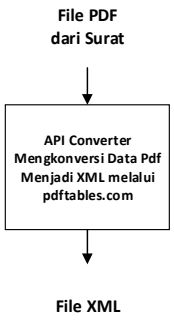
1. Pengubahan format data (*Converting*).
2. Analisis Text *Preprocessing* (*Filtering*, *Tagging*, dan *Tokenization*)
3. Analisis Pembobotan
4. Analisis Algoritma *Naïve Bayes*
5. Hasil Ekstraksi Informasi



Gambar 2. Deskripsi Sistem Ekstraksi Surat Masuk

3.3 Proses Pengubahan Data (*Converting*)

Proses *converting* (konversi) adalah proses pengubahan format data menjadi format data yang diperlukan untuk menuju tahap proses *preprocessing*. Konversi yang dilakukan dalam penelitian ini berupa format file *.pdf menjadi *.xml. Alasan dipilihnya *converting* karena berdasarkan percobaan konversi yang dilakukan ke berbagai format seperti docx , csv atau pun txt memiliki kendala pada hasil yang didapatkan dan xml dipilih karena teks yang dihasilkan memiliki tag html, dari awal hasil konversi dan memiliki tag tiap barisnya sehingga menjaga bentuk surat dan sesuai untuk surat yang memiliki data kalimat per baris untuk tanggal,nomor,lampiran dan tujuan sehingga lebih mudah untuk pengolahan text preprocessing. Proses converting dapat dilihat pada gambar 3, dan hasil pada gambar 4 dengan contoh sebagian hasil dari proses konversinya.



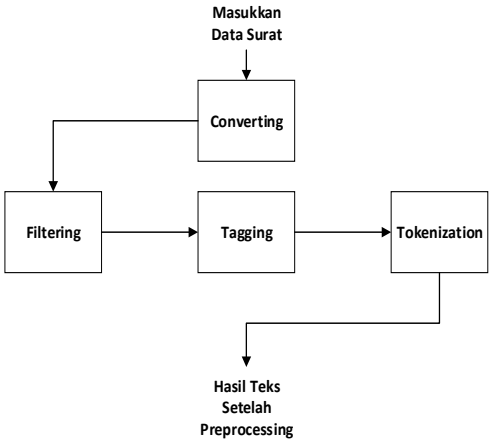
Gambar 3. Proses *Converting*



Gambar 4. Hasil Converting

3.4 Analisis *Preprocessing*

Preprocessing merupakan suatu proses untuk menyiapkan data untuk diolah pada proses utama dengan menggunakan *Naive Bayes*. Terdapat 3 (tiga) proses yang ada dalam preprocessing pada penelitian ini yaitu *filtering*, *tagging*, *tokenization*. Penggambaran preprocessing dapat dilihat pada gambar 5.



Gambar 5. *Preprocessing*

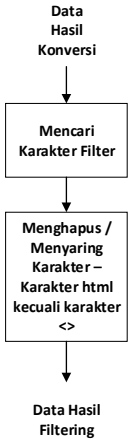
3.4.1 *Filtering*

Proses *filtering* bisa disebut juga *stopwords*. Motif bahwa stopwords harus dihapus dari sebuah teks adalah bahwa mereka membuat teks terlihat lebih berat

dan kurang penting bagi penelitian. Menghapus stopword mengurangi dimensi ruang term. Kata-kata ini diperlakukan sebagai kata-kata henti. *Stopwords* dihapus dari dokumen karena kata-kata tersebut tidak diukur sebagai kata kunci dalam aplikasi penambangan teks [9]. Karakter yang akan difilter dapat dilihat pada tabel 2 dan proses dapat dilihat pada gambar 6 dan hasil pada gambar 7.

Tabel 1. Filtering

Keadaan	Aksi
<td>,<tr>,</tr>,</td>	Diganti <>
Colspan, Style text	Colspan, Style text
Kata/tag dalam HTML	Dihapus dengan fungsi strip_tags
Selain Kata Diatas	Tidak ada aksi



Gambar 6. Proses *Filtering*

<document page-count="1">
<page number="1">
<table data-filename="php2E3E.tmp" data-page="1" data-table="1"><
colspan="5">MUSYAWARAH GURU MATA PELAJARAN < colspan="4">(MGMP)
SENI BUDAYA SMP-MTs KOTA BANDAR LAMPUNG <>Sekretariat : SMP Negeri 7
Bandar Lampung <><>< colspan="7"><>Jl. Sultan Badaruddin No.4 Telp. 0721 266689
Segala Mider Tanjung Karang < colspan="5">Bandar Lampung, 09 Novb 2016<>Nomor
:016/MGMP-SB/SMP/IX/2016 Lamp. :< style="text-align: right">- <>Perihal :Undangan
Kegiatan <><>MGMP Seni Budaya <>Yth.Kepala SMP Negeri/Swasta <>Bandar
Lampung Di <><>Bandar Lampung <><>Dengan hormat, <><>Sehubungan
dengan<>dilaksanakannya<>pertemuan<>rutin<>Musyawarah<>Guru<>Mata<>Pelajaran<
colspan="7">(MGMP) Seni Budaya SMP/MTs Kota Bandar Lampung untuk bulan
November<>< colspan="9">2016 semester Ganjil Tahun Pelajaran 2016/2017, dengan ini
mohon kiranya agar Bapak/Ibu<>berkenanm enyampaikan tugas kegiatan
tersebut<>pada<>guru-guru<>Mata<>Pelajaran<>Seni<>< colspan="4">Budaya yang
dilaksanakan pada : Hari/tanggal : Jumat, 11 November 2016 <>Waktu<>: Pkl. 08.00
WIB s.d selesai <>Tempat<>: SMPN 7 Bandar Lampung <>Acara: 1. Pertemuan Rutin
Acara: 1. Pertemuan Rutin MGMP Seni Budaya <>2. Pelatihan Perangkat Kurikulum 2013
<>3. Penyampaian kisi-kisi UAS Sem. Ganjil <><>< colspan="8"><>Demikian undangan
kami, atas perhatian dan kerjasama Bapak/Ibu, kami sampaikan<>terima kasih.
<><>Koordinator MGMP<><>< colspan="4">Ketua MGMP Seni Budaya SMP<><>Seni
Budaya Kota B.Lampung<>Kota Bandar Lampung GURU MATA<><
colspan="4"><>Nurmaini, S.Pd.<>S<>< colspan="4">Maidir, S.Kom.<>NIP 19600503
198303 2 004<>E N<>A< colspan="4">NIP 19610508 198703 1 006<><>I<>AY<
colspan="4"><>< colspan="9">B U D</table></page>
</document>

Gambar 7. Hasil *Filtering*

3.4.2 *Tagging*

Tagging adalah aplikasi kondisional yang digunakan untuk memberikan *tag* kondisional barang atau kata atau frasa tertentu dalam sebuah kalimat tergantung dari kata kunci yang terkandung dalam kalimat yang diidentifikasi oleh *sed command* [10].

Tagging merupakan tahapan untuk memberikan tanda pada kata – kata kunci yang akan menjadi awal kalimat dan sebagai pembatas/pembendung (*stemming*) yang akan di ekstraksi. Namun dalam surat sering kali terjadi perbedaan penulisan ataupun karakter tambahan sehingga dibutuhkan kata penyesuaian karakter untuk menambahkan tag. Kata kunci dalam tagging pada proses ini dapat dilihat pada tabel 3 dan kata penyesuaian pada tabel 4.

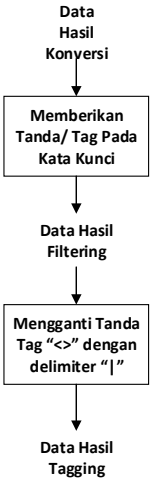
Tabel 2. Kata Kunci *Tagging*

Kata Kunci	
Nomor	Perihal
No	H a l
Hal	Lampiran
Lamp	Dengan Hormat
Lamp.	Lamp:

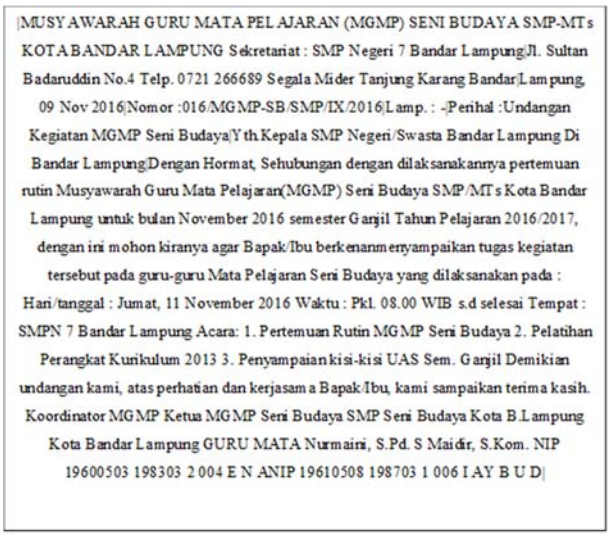
Kepada	Assalam
Yth.	Dalam rangka
Jalan	Tempat
Tag	
<>	

Tabel 3. Kata Penyesuaian

Kata Kunci	Kata Tambahan Penyesuaian
Lamp:	Lamp : atau _Lamp:
Dengan Hormat	Dengan hormat
No	No. / _No / No_
Perihal	_Perihal
Jalan	Jln / Jln. / Jl / Jl. / jl / jl.
Yth.	_Yth. / yth. / YTH. / Yth
Nomor	Nomor: / _Nomor:



Gambar 8. Proses *Tagging*



|MUSYAWARAH GURU MATA PELAJARAN (MGMP) SENI BUDAYA SMP-MTs KOTA BANDAR LAMPUNG Sekretariat : SMP Negeri 7 Bandar Lampung|Jl. Sultan Badaruddin No.4 Telp. 0721 266689 Segala Mider Tanjung Karang Bandar|Lampung, 09 Nov 2016|Nomor :016/MGMP-SB/SMP/IX/2016|Lamp. : -|Perihal :Undangan Kegiatan MGMP Seni Budaya|Yth.Kepala SMP Negeri/Swasta Bandar Lampung Di Bandar Lampung|Dengan Hormat, Sehubungan dengan dilaksanakannya pertemuan rutin Musyawarah Guru Mata Pelajaran(MGMP) Seni Budaya SMP/MTs Kota Bandar Lampung untuk bulan November 2016 semester Ganjil Tahun Pelajaran 2016/2017, dengan ini mohon kiranya agar Bapak/Ibu berkenanmenyampaikan tugas kegiatan tersebut pada guru-guru Mata Pelajaran Seni Budaya yang dilaksanakan pada : Hari/tanggal : Jumat, 11 November 2016 Waktu : Pkl. 08.00 WIB s.d selesai Tempat : SMPN 7 Bandar Lampung Acara: 1. Pertemuan Rutin MGMP Seni Budaya 2. Pelatihan Perangkat Kurikulum 2013 3. Penyampaian kisi-kisi UAS Sem. Ganjil Demikian undangan kami, atas perhatian dan kerjasama Bapak/Ibu, kami sampaikan terima kasih. Koordinator MGMP Ketua MGMP Seni Budaya SMP Seni Budaya Kota B.Lampung Kota Bandar Lampung GURU MATA Nurmaini, S.Pd. S Maidir, S.Kom. NIP 19600503 198303 2 004 E N ANIP 19610508 198703 1 006 I AY B U D|

Gambar 9. Hasil *Tagging*

3.4.3 *Tokenization*

Tokenization adalah proses pemecahan aliran konten tekstual menjadi kata-kata, istilah, simbol, atau beberapa elemen bermakna lainnya yang disebut token [9]. Proses *Tokenization* dimulai dari membaca tag delimiter kemudian sistem akan memisahkan kalimat per kalimat berdasarkan delimiter "|" dan kalimat akan diinputkan kedalam database. Proses tokenization dapat dilihat pada gambar 10 dan hasil pada gambar 11.



Gambar 10. Proses *Tokenization*

No	Kalimat
1	MUSYAWARAH GURU MATA PELAJARAN (MGMP) SENI BUDAYA SMP-N
2	Jl. Sultan Badaruddin No.4 Telp. 0721 266689 Segala Mider Tanjung Karang Bandar
3	Lampung, 09 Novb 2016
4	Nomor :016/MGMP-SB/SMP/IX/2016
5	Lamp. : -
6	Perihal :Undangan Kegiatan MGMP Seni Budaya
7	Yth.Kepala SMP Negeri/Swasta Bandar Lampung Di Bandar Lampung
8	Dengan Hormat, Sehubungan dengan dilaksanakannya pertemuan rutin Musyawara

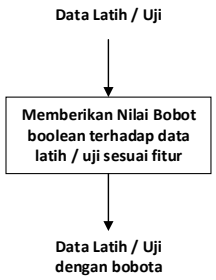
Gambar 11. Hasil *Tokenization*

3.5 Pembobotan

Pembobotan dilakukan setelah data surat masuk sudah melalui seluruh tahapan *preprocessing*. Pembobotan atau ekstraksi fitur dilakukan menggunakan algoritma *Rule – Based* dengan memberi nilai boolean pada fitur. Beberapa Fitur lokal disesuaikan berdasarkan penelitian Peng [3] dan beberapa fitur pengembangannya menjadi 11 parameter. Berikut adalah tabel fitur yang digunakan pada tabel 4 dan proses pembobotan pada gambar 12.

Tabel 4. Tabel Ekstraksi Fitur

Fitur Lokal	
EMAIL	Teks memiliki email
ALLCAPS	Seluruh karakter teks adalah huruf besar (Uppercase)
CONTAINDIGIT	Teks memiliki sedikitnya 1 karakter angka
CONTAINDASH	Teks memiliki sedikitnya 1 strip
CONTAINMORESLASH	Teks memiliki sedikitnya 1 garis miring
INITIALKEYWORD	Teks memiliki sedikitnya 1 kata kunci
WORD	Teks memiliki kata itu sendiri
CONTAINCOLON	Teks memiliki sedikitnya 1 karakter titik dua
STRINGLENGTH	Panjang dari teks lebih 270 karakter
CONTAINCOMMA	Teks memiliki sedikitnya 1 karakter koma
CONTAINDATE	Teks memiliki teks tanggal



Gambar 12. Proses Pembobotan

3.6 Klasifikasi *Naive Bayes*

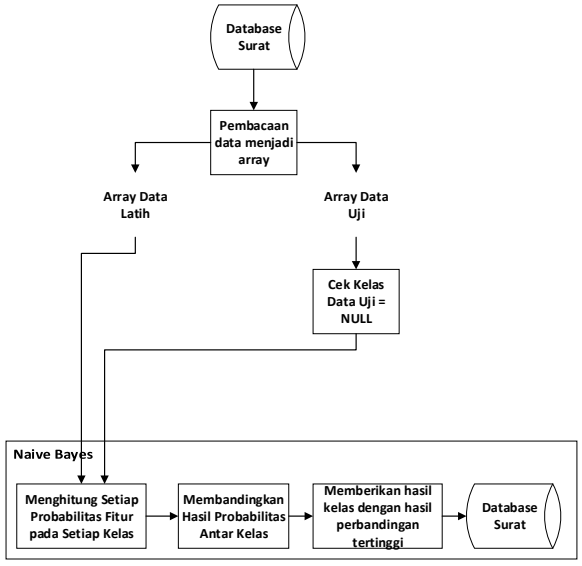
Pada Algoritma *Naive Bayes* dibutuhkan data latih untuk tahap pembelajaran dengan menghitung setiap probabilitas yang ada pada data latih dari setiap ekstraksi fitur yang terdapat pada data latih sebelum algoritma mengklasifikasikan setiap data uji. Misalkan ada data dengan hasil pembobotan seperti pada gambar 13. Kemudian data tersebut dihitung dengan mengalikan setiap probabilitas fitur yang ada pada setiap kelas, data akan memiliki hasil dan membandingkan peluang pada setiap kelas. Setiap hasil probabilitas tertinggi maka data akan diklasifikasi kan sebagai hasil probabilitas kelas tertinggi tersebut seperti pada gambar 14. Proses Klasifikasi dapat dilihat pada gambar 15.

	Kalimat	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
1	PT ROCKET MANA	0	1	0	0	0	0	0	0	0	0	0

Gambar 13. Data Uji

DATA UJI	Fitur	Nilai Fitur	DI KELAS							
			1	2	3	4	5	6	7	8
1	x1	0	1	2/5	1	1	1	1	1	1
	x2	1	4/5	0	0	0	0	0	0	0
	x3	0	2/5	0	0	0	3/5	4/5	3/5	0
	x4	0	1	1	1	1	3/5	1	1	2/5
	x5	0	1	1	1	0	1	4/5	4/5	4/5
	x6	0	1	0	1	0	0	0	0	2/5
	x7	0	1	1	1	1	1	0	1	2/5
	x8	0	1	4/5	1	0	0	1/5	1	0
	x9	0	1	1	1	1	1	1	1	0
	x10	0	1	1/5	0	1	1	1	4/5	0
	x11	0	1	1	0	1	1	1	4/5	1/5
Peluang Pada Setiap Kelas			8/25	0	0	0	0	0	0	0
Persentase Peluang			100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Gambar 14. Perhitungan Klasifikasi



Gambar 15. Proses Klasifikasi

4. PENGUJIAN DAN PEMBAHASAN

Pengujian penelitian ini dilakukan untuk mengetahui tingkat akurasi klasifikasi *naive bayes*

dalam pengekstraksian informasi. Pengujian dilakukan dengan *confusion matrix*. Terdapat pengujian utama dengan data yang balance dengan 27 data surat yang setiap suratnya memiliki 8 data dengan total 216 data latih. Pengujian ini pun dilakukan dengan skenario data latih yang tidak seimbang untuk membandingkan hasil yang didapat. Pengujian data yang balance bisa dilihat pada gambar 16

		PREDIKSI							
		1	2	3	4	5	6	7	8
AKTUAL	1	9	0	0	0	0	0	1	0
	2	0	9	0	0	0	0	0	0
	3	0	0	7	0	0	0	0	0
	4	0	0	0	9	1	0	0	0
	5	0	0	0	0	7	0	0	0
	6	0	0	0	0	0	10	0	0
	7	0	4	0	0	0	0	6	0
	8	3	0	0	0	0	0	0	7

Gambar 16. Confusion Matrix

$$\text{Akurasi} = \frac{\frac{9+60}{9+60+3+1} + \cdots + \frac{6+62}{6+62+1+4} + \frac{7+63}{7+63+0+3}}{8} * 100\%$$

$$= \frac{0.94+0.94+1+0.98+0.98+1+0.93+0.96}{8} * 100\%$$

$$= 96{,}96\%$$

$$\textit{Precision} = \frac{9+9+7+9+7+10+6+7}{(3+9)+(4+9)+(0+7)+(0+9)+(1+7)+(0+10)+(1+6)+(0+7)}$$

$$= \frac{64}{73} * 100\%$$

$$= 87{,}67\%$$

$$\textit{Recall} = \frac{9+9+7+9+7+10+6+7}{(1+9)+(0+9)+(0+7)+(1+9)+(0+7)+(0+10)+(4+6)+(3+7)}$$

$$= \frac{64}{73} * 100\%$$

$$= 87{,}67\%$$

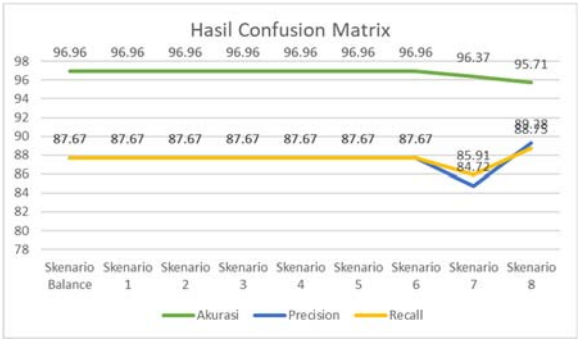
Didapatkan akurasi 96,96% dengan Precision dan Recall yang sama yakni 87,67%. Kemudian membandingkan nilai akurasi yang didapat dengan skenario imbalance data dengan data latih yang memiliki komposisi seperti pada tabel 5 dan hasil *confusion matrix* tiap skenario terdapat pada tabel 6 serta grafik pada gambar 17:

Tabel 5. Komposisi Skenario

Skenario	Kelas							
	1	2	3	4	5	6	7	8
1	40	25	30	15	30	25	19	22
2	20	40	30	20	30	25	19	22
3	20	30	40	20	30	25	20	21
4	23	23	24	40	30	25	24	27
5	30	23	20	25	40	27	25	26
6	25	27	27	29	20	40	23	25
7	27	14	32	27	22	25	40	25
8	25	27	20	24	30	25	25	40

Tabel 6. Hasil Akurasi

No	Skenario	Precision	Recall	Akurasi
1	Data *Balance*	87,67	87,67	96,96%
Skenario *Imbalance* Data				
2	Skenario 1	87,67%	87,67%	96,96%
3	Skenario 2	87,67%	87,67%	96,96%
4	Skenario 3	87,67%	87,67%	96,96%
5	Skenario 4	87,67%	87,67%	96,96%
6	Skenario 5	87,67%	87,67%	96,96%
7	Skenario 6	87,67%	87,67%	96,96%
8	Skenario 7	84,72%	85,91%	96,37%
9	Skenario 8	89,28%	88,75%	95,71%



Gambar 17. Grafik Hasil Confusion Matrix

4.1 Kesimpulan Pengujian

Berdasarkan hasil pengujian dengan perhitungan yang dilakukan masing – masing skenario didapatkan hasil dengan data latih yang *balance* yaitu akurasi 96,96 % dengan *precision* 87,67% dan *recall* 87,67% serta hasil yang sama diperoleh dengan skenario 1 sampai 6, namun pada skenario 7 dan 8 memiliki perbedaan yakni pada skenario 7 mengalami

penurunan akurasi 96,37% dan bahkan pada skenario 8 menurun ke 95,71%. Precision dan Recall pada skenario 7 pun menurun ke 84,72 dan 85,71. Pada skenario 8 *Precision* dan *Recall* naik ke 89,28% dan 88,75%. Hal tersebut disebabkan ketidakseimbangan data pada kelas 7 dan 8 yang sangat mempengaruhi hasil klasifikasi sehingga mengalami kesalahan dalam klasifikasi. Pada skenario 7 didapatkan perubahan data sebanyak 4 data yang memiliki kelas aktualnya adalah kelas 7 tetapi diklasifikasikan kelas 5. Penurunan *precision* terjadi untuk mengklasifikasikan kelas 2, sehingga data *true positive* nya berkurang. Kemudian pada skenario 8 untuk kelas 8 bertambah 1 yang menjadi *true positive* dan pengklasifikasian kelas 7 yang menjadi data *false negative* atau data yang terklasifikasikan kelas 2 oleh sistem bertambah. Alasan lainnya adalah bahwa dalam data uji didapatkan kelas 7 dengan fitur x3, yaitu data kelas 7 yang terkandung angka serta pada kelas 8 yang mengandung *e-mail* didalamnya, menyebabkan data terklasifikasikan sebagai kelas 1. Sedangkan pada data latih, memiliki data yang kurang cukup untuk tujuan atau kelas 7 yang memiliki fitur x3 atau angka, dan isi atau kelas 8 yang memiliki fitur x1 atau yang mengandung e-mail. Jadi dapat disimpulkan bahwa data latih sangat berpengaruh dalam pengklasifikasian naive bayes serta keseimbangan datanya.

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Pada penelitian yang telah dilakukan ini dapat ditarik kesimpulan bahwa :

1. *Naive bayes* yang menjadi algoritma untuk ekstraksi informasi memiliki tingkat akurasi yang sangat tinggi dan *naive bayes* layak digunakan untuk sistem ekstraksi informasi pada surat masuk.

5.2 Saran

Berdasarkan penelitian yang telah dilakukan ini, terdapat permasalahan pada proses konversi melalu API dimana ditemukan simbol – simbol absurd yang menyebabkan proses ekstraksi informasi tidak berhasil. Adapun saran untuk kajian lebih lanjut adalah sebagai berikut.

1. Dibutuhkan API konversi yang lebih baik lagi dalam merekognisi tulisan dalam surat sehingga dapat meminalisir kesalahan yang ada dalam konversi.
2. Dibutuhkan data yang memiliki kualitas gambar yang bagus agar proses konversi berjalan dengan lancar tanpa kendala.

DAFTAR PUSTAKA

[1] F. Peng, “Accurate Information Extraction from Research Papers Using Conditional Random Fields,” *Inf. Process. Manag. an Int. J.*, vol. 42, 2004.

[2] S. Sarawagi, “Information extraction,” *Commun. ACM*, vol. 39, 2008.

[3] A. Riaddy, “6 Ekstraksi Informasi pada Makalah Ilmiah dengan Pendekatan Supervised Learning,” *e-Proceeding Eng.*, vol. 3, no. 1, 2016.

[4] D. Mustaqwa, “Implementasi Ekstraksi Informasi Pada Dokumen Teks Skripsi Menggunakan Metode Rule Based,” Universitas Komputer Indonesia, 2017.

[5] M. Dorash, “Machine Learning vs. Rule Based Systems in NLP.” unpublished.

[6] A. Syarifah, “Pemanfaatan Naive Bayes Untuk Merespon,” vol. 4, 2015.

[7] O. Somantri, “Text Mining Untuk Klasifikasi Kategori Cerita Pendek Menggunakan Naïve Bayes (NB),” *J. Telemat.*, vol. 12, 2017.

[8] D. Xhemali, C. J. Hinde, and R. G. Stone, “Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages,” *Int. J. Comput. Sci.*, vol. 4, no. 1, 2009.

[9] S. Vijayarani, J. Ilamathi, and M. Nithya, “Preprocessing Techniques for Text Mining - An Overview,” *Int. J. Comput. Sci. Commun. Networks*, vol. 5, no. 1, 2015.

[10] A. Kao and S. R. Poteet, *Natural language processing and text mining*. New York: Springer, 2007.