

**IMPLEMENTASI ALGORITMA MULTINOMIAL NAIVE BAYES CLASSIFIER  
PADA SISTEM KLASIFIKASI SURAT KELUAR  
(Studi Kasus : DISKOMINFO Kabupaten Tangerang)**

**Dea Herwinda Kalokasari<sup>1</sup>, Dr. Imam Marzuki Shofi<sup>2</sup>, Anif Hanifa Setyaningrum<sup>3</sup>**

Program Studi S1 Teknik Informatika, Fakultas Sains dan Teknologi  
UIN Syarif Hidayatullah Jakarta

*E-mail:* dewinka42ti@mhs.uinjkt.ac.id<sup>1</sup>, imam@uinjkt.ac.id<sup>2</sup>, anifhanifa@yahoo.com<sup>3</sup>

**ABSTRAK**

Informasi diperkirakan lebih dari 80% tersimpan dalam bentuk teks tidak terstruktur. Oleh karena itu, dibutuhkan sistem pengelolaan teks yaitu dengan metode *text mining* yang diyakini memiliki potensial nilai komersial tinggi. Salah satu implementasi dari *text mining* yaitu klasifikasi teks. Tidak hanya dokumen, pemanfaatan klasifikasi juga digunakan pada surat. Peneliti mengkaji *Multinomial Naive Bayes Classifier* untuk mengklasifikasi surat keluar sehingga dapat menentukan nomor surat secara otomatis. Sistem klasifikasi didukung dengan *confix-stripping stemmer* untuk menemukan kata dasar dan TF-IDF untuk pembobotan kata. Pengujian diukur dengan menggunakan *confusion matrix*. Dari hasil pengujian menunjukkan bahwa implementasi *Multinomial Naive Bayes Classifier* pada sistem klasifikasi surat memiliki tingkat *accuracy*, *precision*, *recall*, dan *F-measure* berturut-turut sebesar 89,58%, 79,17%, 78,72%, dan 77,05%.

**Kata kunci:** Surat Keluar, Klasifikasi, Text Mining, Multinomial Naive Bayes Classifier, Confix-stripping Stemmer, TF-IDF

**ABSTRACT**

The information estimated that more than 80% is stored in the form of unstructured text. Therefore, it takes a text management system, namely text mining method is believed to have high potential commercial. One of text mining implementation is text classification. Not only documents, the use of classification is also used in official letter. Researcher examined Multinomial Naive Bayes Classifier to classify the letter so it can determine the letters classification code automatically. The classification system is supported by *confix-stripping stemmer* to find root and TF-IDF for term weighting. The test used by *confusion matrix* of a classified as a measure of its quality. The test results showed that the implementation of Multinomial Naive Bayes Classifier on letter classification system has a level of *accuracy*, *precision*, *recall*, and *F-measure* respectively for 89.58%, 79.17%, 78.72% and 77.05%.

**Keywords:** Letter; Classification; Text Mining; Multinomial Naive Bayes Classifier; Confix-stripping Stemmer; TF-IDF

DOI: 10.15408/jti.v10i2.6822

## I. PENDAHULUAN

Semakin banyaknya kumpulan data yang berisi informasi-informasi penunjang kegiatan mengakibatkan kesulitan dalam pengolahannya. Menurut Clara Bridge (2011), dikarenakan banyaknya informasi (perkiraan umum mengatakan lebih dari 80%) saat ini disimpan sebagai teks, *text mining* diyakini memiliki potensi nilai komersial tinggi [1]. Hal yang lebih menyulitkan dalam analisis adalah bahwa sekitar 80% sampai 85% bentuk informasi tersebut dalam format tidak terstruktur (*unstructured data*).

Salah satu implementasi dari text mining yaitu klasifikasi. Tanpa adanya klasifikasi dokumen, proses pencarian dokumen akan melakukan pencarian menyeluruh terhadap semua dokumen yang ada, sehingga memakan waktu yang lama dan memberikan peluang hasil temu data yang meluas dari hasil topik yang dibutuhkan.

Penentuan klasifikasi surat dalam lembaga nasional memiliki aturan yang diatur dalam PERKA ANRI nomor 19 tahun 2012 tentang pedoman penyusunan klasifikasi arsip yang telah ditetapkan pada Undang-Undang nomor 43 Tahun 2009 tentang Kearsipan. Klasifikasi surat digunakan sebagai dasar penomoran surat dengan kode klasifikasi berjumlah 681 buah.

Penelitian yang sama dilakukan oleh Akhmad Pandhu Wijaya dan Heru Agus Santoso melakukan penelitian dengan judul “*Naive Bayes Classification* pada Klasifikasi Dokumen Untuk Identifikasi Konten *E-Government*” menghasilkan akurasi keseluruhan klasifikasi sebesar 85% [2]. Oleh karena itu, penulis mengkaji algoritma *Multinomial Naive Bayes Classifier* (NBC) untuk kinerja sistem klasifikasi terhadap dokumen surat.

Rumusan masalah yang diangkat dalam sistem ini adalah “Bagaimana mengimplementasikan algoritma *Multinomial Naive Bayes Classifier* pada sistem klasifikasi surat keluar?”

Batasan masalah dari penelitian ini diantaranya adalah

- a) Sistem klasifikasi surat yang dibangun berbasis *desktop* dengan bahasa pemrograman *Java* dan basis data *MySQL*.
- b) Metode yang digunakan dalam klasifikasi dokumen adalah dengan algoritma *Multinomial Naive Bayes Classifier* (MNBC).

- c) Pengambilan sampel berupa surat keluar yang terdiri dari 4 kategori yaitu cuti, undangan, perintah, dan peminjaman ruang dengan format word berekstensi (.doc) atau (.docx) atau (.txt).
- d) Pengujian dilakukan dengan memperhitungkan *accuracy*, *recall*, *precision*, dan *F-measure* pada setiap klasifikasi.

## II. TINJAUAN PUSTAKA

### 2.1 Studi Literatur

Berikut studi literatur sejenis dengan sistem klasifikasi surat dari *proceeding* maupun jurnal.

- a.) Penelitian yang dilakukan oleh Bambang Kurniawan, Syahril Effendi, dan Opim Salim Sitompul mengenai “Klasifikasi Konten Berita dengan Metode *Text Mining*” (2012).
- b.) Penelitian tentang “Klasifikasi Dokument Teks Menggunakan Algoritma *Naive Bayes* dengan Bahasa Pemrograman *Java*” (2013) yang dilakukan oleh Silfia Andini.
- c.) Penelitian mengenai “*KNN based Machine Learning Approach for Text and Document Mining*” (2014) oleh Vishwanath Bijalwan, Vinay Kumar, Pinki Kumari, dan Jordan Pascual
- d.) Acmad Nurhadi meneliti tentang “Klasifikasi Konten Berita Digital Bahasa Indonesia Menggunakan *Support Vector Machines* ( *SVM* ) Berbasis *Particle Swarm Optimization* ( *PSO* )” (2015)
- e.) Akhmad Pandhu Wijaya melakukan penelitian dengan judul “*Naive Bayes Classification* Pada Klasifikasi Dokumen Untuk Identifikasi Konten *E-Government*” (2016)
- f.) Peneliti : “Implementasi Algoritma *Multinomial Naive Bayes Classification* (MNBC) pada Sistem Klasifikasi Surat Keluar” (2017)

Tabel 1. Kesimpulan hasil evaluasi studi literatur sejenis

No.	Text Mininng	Algoritma Stemming	Algoritma Klasifikasi	TF-IDF	Pengujian	Objek	Tool
a.)	<ul style="list-style-type: none"> <li>• Tokenizing,</li> <li>• Tolowercase</li> <li>• Stopword removal,</li> <li>• Stemming</li> </ul>	Confix-stripping stemmer	Naïve Bayes Classifier	x	x	dokumen berita 4 kategori	PHP dan MySQL
b.)	<ul style="list-style-type: none"> <li>• Tokenizing,</li> <li>• Folding (tolowercase),</li> <li>• Removal Stopword</li> </ul>	x	Naïve Bayes Classifier	x	x	Dokumen teks 2 kategori	Java
c.)	<ul style="list-style-type: none"> <li>• Bag of word,</li> <li>• Stopword removal,</li> <li>• Case folding</li> </ul>	x	K-Nearest Neighbors, Naïve Bayes Classifier, Term-Graph Support Vector Machine	✓	Accuracy	Dokumen teks 5 kategori	Java
d.)	<ul style="list-style-type: none"> <li>• Tokenizing,</li> <li>• Tolowercase</li> </ul>	x	Naive Bayes Classifier	✓	Recall, Precision, Accuracy	Konten berita 3 kategori	RapidMiner 5.3
e.)	<ul style="list-style-type: none"> <li>• Tokenizing,</li> <li>• Stopword Removal,</li> <li>• Filtering,</li> <li>• Stemming</li> </ul>	tidak dijelaskan penggunaan algoritma	Naive Bayes Classifier	✓	Accuracy	Dokumen 2 kategori	tidak dijelaskan
f.)	<ul style="list-style-type: none"> <li>• Tolowercase</li> <li>• Tokenizing,</li> <li>• Stopword removal,</li> <li>• Stemming</li> </ul>	Confix-stripping stemmer	Multinomial Naive Bayes Classifier	✓	Recall, Precision, Accuracy, F1-measure	Surat keluar dinas 4 kategori	Java dan MySQL

**2.2 Klasifikasi Surat**

Surat adalah suatu media komunikasi yang berisi pernyataan tertulis mengenai data atau informasi yang ingin disampaikan atau ditanyakan kepada penerima surat [3]. Dalam klasifikasi surat suatu instansi, aturan sistem klasifikasi yang digunakan adalah *Dewey Decimal Classification* (DDC).

DDC diciptakan oleh Melvil Dewey pada tahun 1873 dan diterbitkan pada tahun 1876. DDC membagi ilmu pengetahuan manusia menjadi 10 kelas utama, masing-masing kelas utama dibagi menjadi 10 divisi dan masing-masing divisi dibagi lagi menjadi 10 seksi, sehingga DDC mempunyai 10

kelas, 100 divisi, dan 1000 seksi, dan masih ada kemungkinan pembagian lebih lanjut.

**2.3 Text Mining**

Menurut Ronen Feldman dan James Sanger (2007), Text mining dapat didefinisikan sebagai proses pengetahuan intensif di mana pengguna berhubungan sekumpulan dokumen dari waktu ke waktu dengan menggunakan seperangkat *tool* analisis [4]. Tujuan utama *text mining* adalah untuk menganalisis dan menemukan pola yang menarik termasuk tren, *outlier* dan gagasan *query* yang tidak memiliki kandungan atau bahkan yang relevan [5].

Proses *Text Mining* dibagi menjadi 3 tahap utama, diantaranya : [6]

- a) *Text Preprocessing* merupakan tahapan yang dilakukan proses *tokenizing*, dan *case folding*.
- b) *Text Transformation* disebut juga proses *filtering*. Pada tahap ini tindakan yang dilakukan adalah menghilangkan *stopword* (*stopword removal*) dan *stemming* terhadap kata yang berimbuhan.
- c) *Pattern Discovery* merupakan tahapan berusaha menemukan pola atau pengetahuan dari keseluruhan teks.

**2.4 Confix Stripping Stemmer**

Algoritma *confix-stripping stemmer* adalah algoritma yang digunakan untuk melakukan proses *stemming* terhadap kata-kata berimbuhan [7]. Algoritma *Confix-stripping stemmer* mempunyai aturan imbuhan sendiri dengan model sebagai berikut :

<p>[[[AW + ] AW +] AW +] Kata-Dasar [[+AK][+KK][+P]]</p>
--

Gambar 1. Aturan *Algoritma Confix-stripping stemmer* [8]

Keterangan :  
 AW : Awalan  
 AK : Akhiran  
 KK : Kata ganti kepunyaan  
 P : Partikel

Tidak semua kombinasi dari prefiks dan sufiks dapat bergabung bersama untuk membentuk konfiks a. Ada beberapa kombinasi awalan dan akhiran yang tidak diizinkan. Algoritma *Confix-stripping stemmer* biasa dilakukan dengan penghapusan imbuhan awalan. Namun pada kombinasi tersebut merupakan pengecualian terhadap prefiks dan sufiks tertentu untuk menghapus akhiran terlebih dahulu.

Tabel 2. Kombinasi Prefiks dan Sufiks yang tidak diperbolehkan [7]

Awalan ( Prefiks)	Akhiran (Suffiks)
be-	-i
di-	-an
ke-	-i –kan
me-	-an
se-	-i –kan
te-	-an

Berikut merupakan tabel aturan peluruhan imbuhan pada algoritma *confix-stripping stemmer*.

Tabel 3. Aturan peluruhan kata dasar [7]

Aturan	Bentuk Awalan	Peluruhan
1	berV...	ber-V...   be-rV...
2	belajar...	bel-ajar
3	beC1erC2...	be-C1erC2 ... dimana C1!={r l}
4	terV...	ter-V... te-rV...
5	terCer...	ter-Cer ... dimana C!='r'
6	teC1erC2	te-C1erC2... dimana C1!='r'
7	me{l r w y}V...	me-{l r w y}V...
8	mem{b f v}...	mem-{b f v}...
9	mempe...	mem-pe...
10	mem{rV V}...	me-m{rV V} ...  me-p{rV V}...
11	men{c d j z}...	men-{c d j z}...
12	menV...	me-nV... me-tV...
13	meng{g h q k} ...	meng-{g h q k}...
14	mengV...	meng-V... meng-kV...
15	mengeC	menge-C
16	menyV...	me-ny...  meny-sV...
17	mempV...	mem-pV...
18	pe{w y}V...	pe-{w y}V...
19	perV...	per-V... pe-rV...
20	pem{b f v}...	pem-{b f v}...
21	pem{rV V}...	pe-m{rV V}... pe-p{rV V}
22	pen{c d j z}...	pen-{c d j z}...
23	penV...	pe-nV... pe-tV...
24	peng{g h q}	peng-{g h q}
25	pengV	peng-V peng-kV
26	penyV...	pe-nyal peny-sV
27	peIV..	pe-IV...; kecuali untuk kata "pelajar" menjadi "ajar"
28	peCP	pe-CP...dimana C!={r w y l m n} dan P!='er'
29	perCerV	Per-CerV... dimana C!={r w y l m n}

**2.5 Term Weighting**

Algoritma TF-DF adalah suatu algoritma yang berdasarkan nilai statistik menunjukkan kemunculan suatu kata di dalam dokumen [2]. TF (*Term Frequency*)

menyatakan banyaknya suatu kata muncul dalam sebuah dokumen. *DF (Document Frequency)* menyatakan banyaknya dokumen yang mengandung suatu kata dalam satu segmen publikasi.

$$IDF = \log\left(\frac{N}{DF(w)}\right)$$

$$TF - IDF(w, d) = TF(w, d) \times IDF(w)$$

Keterangan:

- TF-IDF(w,d) : bobot suatu kata dalam keseluruhan dokumen
- w : suatu kata (*word*)
- d : suatu dokumen (*document*)
- TF(w,d) : frekuensi kemunculan sebuah kata w dalam dokumen d
- IDF(w) : *inverse DF* dari kata w
- N : jumlah keseluruhan dokumen
- DF(w) : jumlah dokumen yang mengandung kata w

### 2.6 Multinomial Naive Bayes Classifier

*Multinomial NBC* merupakan model pengembangan dari algoritma *bayes* yang cocok dalam pengklasifikasian teks atau dokumen. Pada formula *Multinomial Naive Bayes Classifier*, kelas dokumen tidak hanya ditentukan dengan kata yang muncul tetapi juga jumlah kemunculannya [9].

$$C_{MAP} = \arg \max_{c \in \{c_l, c_s\}} P(c|d)$$

$$= \arg \max_{c \in \{c_l, c_s\}} P(c) \prod_{k=1}^m P(t_k|c)$$

Parameter  $P(t_k|c)$  (*probability likelihood*) diestimasi dengan menghitung kejadian  $t_k$  pada semua dokumen training di  $c$ , menggunakan *Laplacean prior*: [10]

$$P(t_k|c) = \frac{1+N_k}{|V|+N}$$

dimana  $N_k$  adalah jumlah kemunculan  $t_k$  dalam dokumen pelatihan di  $c$  dan  $N$  adalah jumlah total kejadian kata dalam  $c$ .

### 2.7 Confusion Matrix

*Confusion matrix* berisikan informasi mengenai aktual dan prediksi yang diberikan oleh *classifier*. [11]

Tabel 4. *Confusion matrix dari classifier* [11]

	Actual	Positive	Negative
Classified			
Positive		TP	FN
Negative		FP	TN

where

- TP: the number of correct classifications of the positive examples (**true positive**)
- FN: the number of incorrect classifications of positive examples (**false negative**)
- FP: the number of incorrect classifications of negative examples (**false positive**)
- TN: the number of correct classifications of negative examples (**true negative**)

Pengukuran utama adalah klasifikasi *accuracy*, yang merupakan jumlah kasus diklasifikasikan dengan benar di *test set* dibagi dengan jumlah total kasus dalam *test set*. *Precision* dan *recall* mengukur seberapa tepat dan seberapa lengkap klasifikasi ini pada kelas yang positif. [11]

$$r = \frac{TP}{TP + FN} \quad p = \frac{TP}{TP + FP}$$

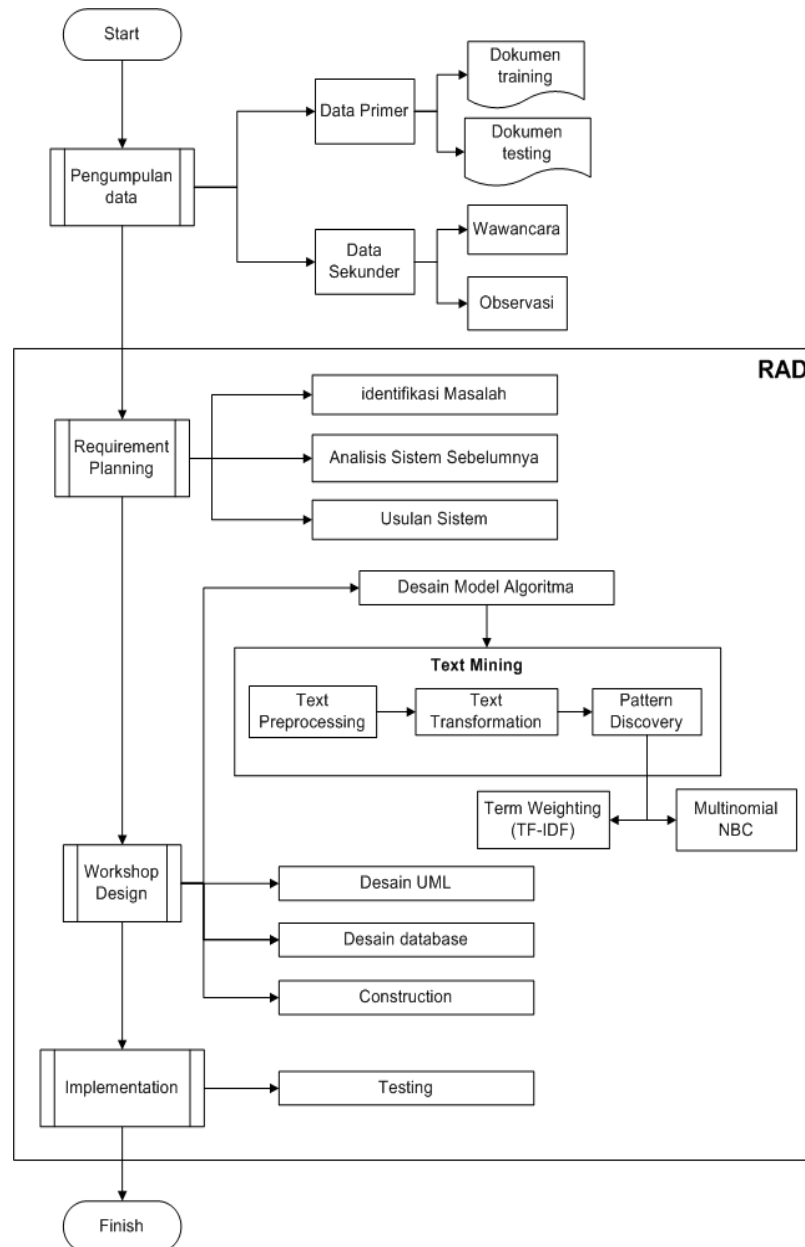
$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

*Precision* ( $p$ ) adalah rasio terjadinya secara aktual diklasifikasikan sebagai positif untuk semua ketentuan yang diklasifikasikan sebagai positif. *Recall* ( $r$ ) adalah rasio terjadinya secara aktual diklasifikasikan sebagai positif untuk semua ketentuan positif. *F-measure* yang didefinisikan sebagai rata-rata harmonik dari *precision* dan *recall*. [12]

$$F - measure = \frac{2 \times precision \times recall}{precision + recall}$$

## III. PEMBAHASAN

Penelitian ini menggunakan pendekatan *Rapid Application Development (RAD)* sebagai metode pengembangan sistem. Dalam pengembangan sistem, metode ini dapat mempersingkat waktu pemrosesan dan menghasilkan sistem aplikasi secara akurat dan cepat. Proses RAD hanya melibatkan tiga tahap, meliputi: *requirement planning*, *workshop design*, dan implementasi.



Gambar 2. Kerangka penelitian

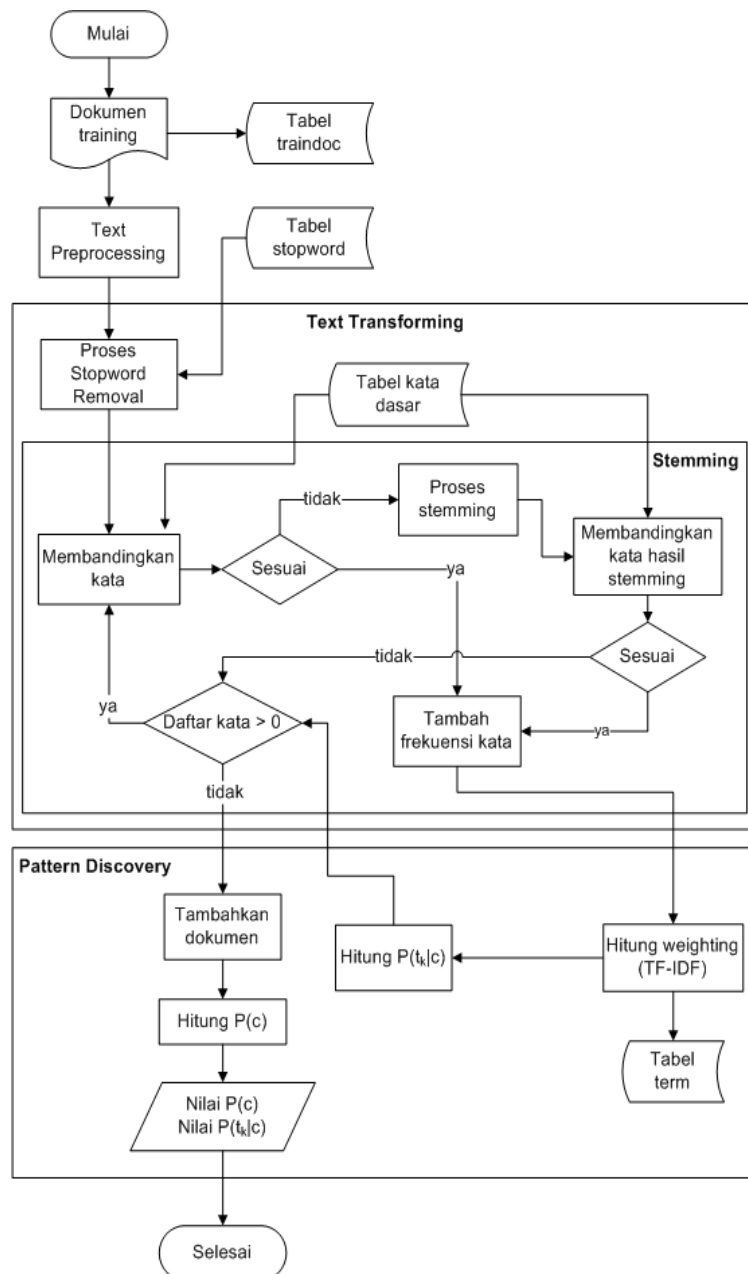
### 3.1 Pengumpulan Data

Dokumen yang dijadikan sampel penelitian yaitu surat keluar dinas dengan kategori undangan, cuti, perintah, dan pinjam ruang. Surat yang terkumpul sejumlah 6 sampel untuk masing-masing kategori sehingga total keseluruhan yaitu 24 dokumen surat dan ditambah 1 dokumen surat untuk implementasi.

### 3.2 Requirement Planning

Proses klasifikasi dilakukan melalui 2 tahap, di antaranya tahap *training* dan *testing*.

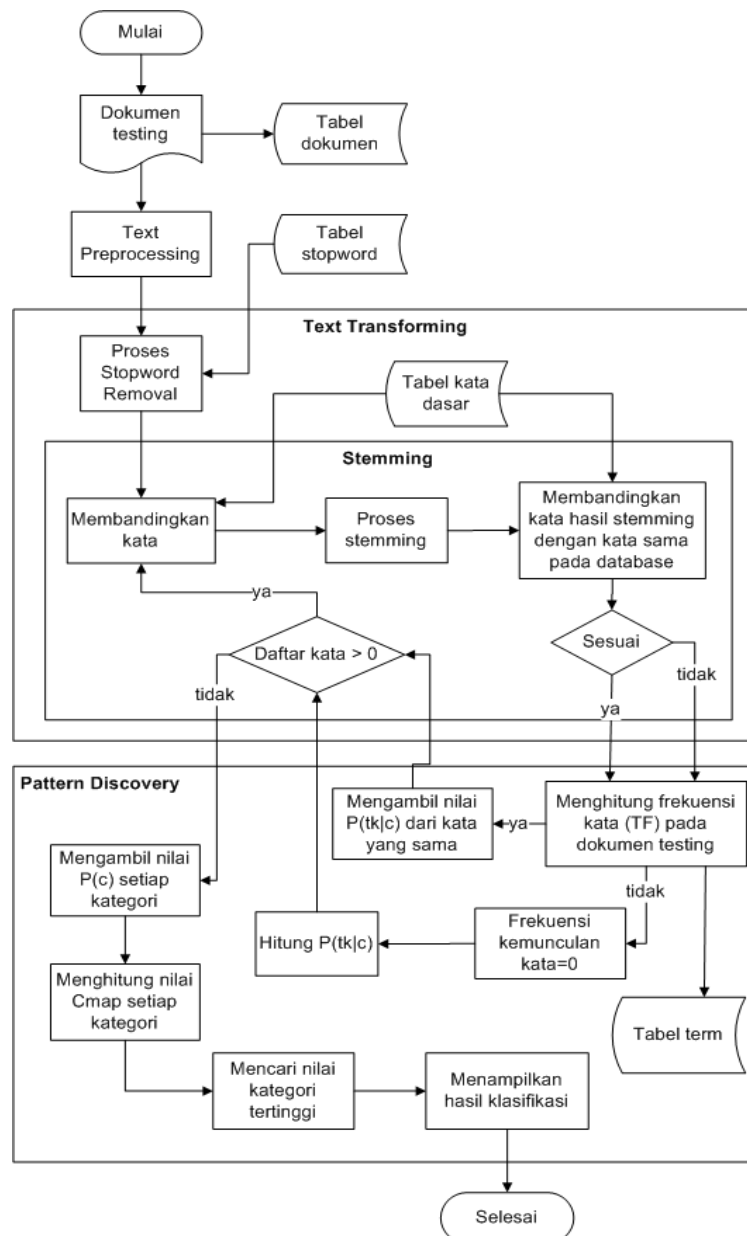
Tahapan pembelajaran atau *training* merupakan proses mengekstraksi surat yang telah diketahui kategorinya. Proses ini dilakukan dengan penentuan dan pembentukan *bag of word* pada setiap dokumen *training*. Setiap kata yang disimpan dianalisis nilai *weighting* terhadap semua dokumen pada setiap klasifikasi sehingga dapat menghitung probabilitas kata yang nantinya akan digunakan pada tahap *testing*.



Gambar 3. Alur proses *training*

Tahapan *testing* merupakan proses pengklasifikasian terhadap surat yang belum diketahui kategorinya. Perhitungan probabilitas sebuah dokumen terhadap

kategori dengan mengimplementasikan algoritma *Multinomial Naive Bayes Classifier*.

Gambar 4. Alur proses *testing*

### 3.3 Workshop Design

Sistem desain dilaksanakan dengan menggunakan metode *object-oriented* dengan UML (*Unified Modeling Language*). Sistem desain dibuat dengan beberapa diagram pemodelan termasuk *use case diagram*, *class diagram*, *sequence diagram*, dan *activity diagram*.

### 3.4 Pengujian

Pengujian ini dilakukan pada salah satu dokumen *testing* dan sisanya sebagai dokumen *training*. Oleh karena itu, pengujian 24 dokumen yang tersedia masing-masing diuji dengan dokumen *training* dari 23

dokumen lainnya. Berikut adalah hasil klasifikasi dari setiap dokumen *testing*.



Tabel 5. Hasil klasifikasi setiap dokumen

No.	Dokumen	Hasil Klasifikasi
1.	Undangan1	Cuti
2.	Undangan2	Pinjam Ruang
3.	Undangan3	Undangan
4.	Undangan4	Pinjam Ruang
5.	Undangan5	Undangan
6.	Undangan6	Pinjam Ruang
7.	Perintah1	Perintah
8.	Perintah2	Perintah
9.	Perintah3	Perintah
10.	Perintah4	Perintah
11.	Perintah5	Perintah
12.	Perintah6	Perintah
13.	Cuti1	Cuti
14.	Cuti2	Cuti
15.	Cuti3	Cuti
16.	Cuti4	Cuti
17.	Cuti5	Cuti
18.	Cuti6	Cuti
19.	PinjamRuang1	Pinjam Ruang
20.	PinjamRuang2	Cuti
21.	PinjamRuang3	Pinjam Ruang
22.	PinjamRuang4	Pinjam Ruang
23.	PinjamRuang5	Pinjam Ruang
24.	PinjamRuang6	Pinjam Ruang

Berikut hasil analisa pengujian sistem secara keseluruhan.

Tabel 6. Rata-rata tingkat recall, precision, accuracy, dan f-measure

Klasifikasi	Accuracy	Precision	Recall	F-measure
Undangan	83,33%	33,33%	100%	50%
Perintah	100%	100%	100%	100%
Cuti	91,67%	100%	75%	85,71%
Pinjam Ruang	83,33%	83,33%	62,5%	71,43%
Rata-rata	89,58%	79,17%	84,38%	76,79%

#### IV. KESIMPULAN

Sistem klasifikasi surat yang dibentuk berbasis *desktop* dengan bahasa pemrograman *Java* dan basis data *mySQL* dengan pembatasan data input yaitu berupa surat keluar dengan format *word* berekstensi (.docx), (.doc), dan (.txt). Sistem klasifikasi surat keluar didukung dengan penerapan *Confix-Stripping Stemmer* pada proses *stemming* dan *TF-IDF* untuk pembobotan setiap kata yang terkandung pada dokumen (*term weighting*).

Hasil evaluasi sistem klasifikasi surat keluar didapatkan nilai *Accuracy*, *Precision*,

*Recall*, dan *F-measure* secara berurutan sebesar 89,58%, 79,17%, 78,72%, dan 77,05%. Untuk mengembangkan sistem ini pada penelitian lebih lanjut, dapat dilakukan dengan memperluas objek klasifikasi, baik surat keluar maupun surat masuk. Sistem juga perlu menambahkan lebih banyak *bag of words* karena semakin banyak kamus kata, semakin lebih tepat proses *stemming* dilakukan. Sistem dapat juga diimplementasikan menggunakan metode lainnya.

#### DAFTAR PUSTAKA

- [1] C. Bridge. 2011. "Unstructured Data and the 80 Percent Rule.," [Online]. Available: <https://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/>. [Accessed 20 April 2016].
- [2] A. P. Wijaya. 2016. "Klasifikasi Dokumen dengan Naive Bayes Classifier (NBC) untuk Mengetahui Konten E-Goverment," *Journal of Applied Intelligent System, Vol.1, No. 1*, pp. 48-55,
- [3] S. Wijaya. 2009. Surat-Surat Kesekretariatan, Jakarta: Pustaka Grahatama
- [4] R. Feldman and J. Sanger. 2007. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, New York: Cambridge University Press.
- [5] C. C. Aggarwal and C. Zhai. 2012. An Introduction to text mining, New York: Springer
- [6] S. Andini. 2013. "Klasifikasi Dokumen Teks menggunakan Algoritma Naive Bayes dengan Bahasa Pemograman Java," *Jurnal Teknologi Informasi dan Pendidikan, Vol 6 no.2*, pp. 140 - 147,
- [7] M. Adriani, J. Asian, B. Nazief, S. Tahaghoghi and H. Williams. 2007. "Stemming Indonesian : A Confix-Stripping Approach.," *Transactions on Asian Language Information Processing, Vol. 6, No.4*
- [8] B. Kurniawan, S. Effendi and O. S. Sitompul. 2012. "Klasifikasi Konten Berita dengan Metode Text Mining," *Jurnal Dunia Teknologi Informasi, Vol. 1*, pp. 14-19
- [9] I. H. Witten, F. Eibe and M. A. Hall. 2011. Data mining : Practical Machine Learning

- Tools and Techniques. Third Edition, USA: Elsevier
- [10] M. W. Berry and J. Kogan. 2010. Text Mining Application and Theory, United Kingdom: John Wiley and Sons
- [11] B. Liu. 2011. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, Second Edition, New York: Springer
- [12] P. Cichosz. 2014. Data Mining Algorithms: Explained Using R, Chichester: John Wiley & Sons