

2018

Identifikasi SMS SPAM Berbahasa Indonesia Menggunakan Algoritma Support Vector Machine

Vynaima, Rona I Dona

<http://repositori.usu.ac.id/handle/123456789/4333>

Downloaded from Repositori Institusi USU, Universitas Sumatera Utara

IDENTIFIKASI SMS *SPAM* BERBAHASA INDONESIA
MENGUNAKAN ALGORITMA *SUPPORT*
VECTOR MACHINE

SKRIPSI

RONA I DONA VYNAIMA S.
121402100



PROGRAM STUDI S1 TEKNOLOGI INFORMASI
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
UNIVERSITAS SUMATERA UTARA
MEDAN
2018

IDENTIFIKASI SMS *SPAM* BERBAHASA INDONESIA
MENGUNAKAN ALGORITMA *SUPPORT*
VECTOR MACHNIE

SKRIPSI

Diajukan untuk melengkapi tugas dan memenuhi syarat memperoleh ijazah Sarjana
Teknologi Informasi

RONA I DONA VYNAIMA S.

121402100



PROGRAM STUDI S1 TEKNOLOGI INFORMASI
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
UNIVERSITAS SUMATERA UTARA
MEDAN
2018

PERSETUJUAN

Judul : IDENTIFIKASI SMS *SPAM* BERBAHASA
INDONESIA MENGGUNAKAN ALGORITMA
SUPPORT VECTOR MACHINE

Kategori : SKRIPSI

Nama : RONA I DONA VYNAIMA S.

Nomor Induk Mahasiswa : 121402100

Program Studi : TEKNOLOGI INFORMASI

Fakultas : ILMU KOMPUTER DAN TEKNOLOGI
INFORMASI

UNIVERSITAS SUMATERA UTARA

Komisi Pembimbing :

Pembimbing 2

Pembimbing 1



Sarah Purnamawati, ST., M.Sc

NIP 19830226 201012 2 003



Dani Gunawan, ST., M.T

NIP. 19820915 201212 1 002

Diketahui/disetujui oleh

Program Studi S1 Teknologi Informasi

Ketua,



Romi Fadillah Rahmat, B.Comp.Sc., M.Sc.

NIP. 19860303 201012 1 004

PERNYATAAN

IDENTIFIKASI SMS *SPAM* BERBAHASA INDONESIA MENGUNAKAN ALGORITMA *SUPPORT* *VECTOR MACHINE*

SKRIPSI

Saya mengakui bahwa skripsi ini adalah hasil karya saya sendiri, kecuali beberapa kutipan dan ringkasan yang masing-masing telah disebutkan sumbernya.

Medan, 8 Pebruari 2018

Rona I Dona Vynaima S.
121402100

UCAPAN TERIMA KASIH

Diatas segalanya penulis mengucapkan puji dan syukur kepada Tuhan Yesus atas berkatNya yang melimpah sehingga penulis dapat menyelesaikan skripsi ini sebagai syarat kelulusan dan memperoleh gelar Sarjana dari Program Studi Teknologi Informasi Universitas Sumatera Utara.

Banyak rasa terima kasih yang ingin penulis ucapkan kepada seluruh pihak yang turut serta terlibat dalam masa perkuliahan dan masa pengerjaan skripsi ini:

1. Penulis mengucapkan terima kasih kepada orangtua penulis, Pernando Simbolon dan Risma Siahaan, atas cinta dan doa yang selalu diberikan. Juga bagi adik-adik penulis, Anggita Simbolon, Rosi Simbolon, Asimima Simbolon, Doly Simbolon yang selalu memberikan dukungan moral maupun materiil kepada penulis.
2. Penulis mengucapkan terima kasih kepada Bapak Dani Gunawan, ST., MT selaku dosen pembimbing pertama dan kepada Ibu Sarah Purnamawati, ST.,M.Sc selaku dosen pembimbing kedua. Terima kasih telah meluangkan waktu, ide dan tenaganya untuk membimbing penulis baik dalam pembuatan program maupun dalam penulisan skripsi.
3. Penulis mengucapkan terima kasih kepada Bapak Romi Fadillah Rahmat, B.Comp.Sc.,M.Sc selaku dosen pembimbing pertama dan Bapak Ivan Jaya, S.Si, M.Kom selaku dosen pembimbing kedua yang telah memberikan kritik dan saran yang bermanfaat.
4. Penulis mengucapkan terima kasih kepada Dekan, Wakil Dekan, Ketua Program Studi Teknologi Informasi, Sekretaris Program Studi Teknologi Informasi, seluruh dosen dan *staff* di Program Studi Teknologi Informasi USU yang telah mengajar, membimbing dan membantu penulis selama proses perkuliahan dan proses pengerjaan skripsi.
5. Penulis mengucapkan terima kasih kepada teman-teman angkatan 2012 yang telah menemani, memotivasi, memahami dan menerima sifat serta perilaku penulis, memberikan kritik dan saran yang baik selama proses perkuliahan

maupun dalam masa pengerjaan skripsi, terkhusus bagi Arsandi Saputra, Grace Lumanauw, Theresia Aruan, Harysa Octafine, Eric Suwarno, Siti Fatimah, Novia Elisa, Franco Bagio, Tito Pandiangan, Tika Hairani, Athmanathan, Ulfa Chairani, Siti Hasanah, Mafia Team dan Iqbal.

6. Penulis tentunya tak lupa mengucapkan terima kasih kepada seluruh abang, kakak, dan adik dari angkatan 2008, 2009, 2010, 2011, 2013, dan 2014 Teknologi Informasi USU, terkhusus bagi, Veronica, Tuti Simanjuntak, Livia Kemit, Odysius B. Anwar, Charlie, Anggi Nasution, Halimatusadiah, Fahrurrisa Khairani, Hans Noel A P, Sintong Siregar dan Reza Taqyuddin.
7. Penulis mengucapkan terima kasih kepada sahabat terbaik penulis, Stevany Manalu, Christine, Rena Tarigan dan Grup Indomie yang sudah memberikan motivasi dan dukungan kepada penulis.
8. Terima kasih penulis ucapkan untuk semua pihak yang telah terlibat dalam pengumpulan data maupun pengujian sistem.

Kiranya Tuhan memberkati kalian semua.

ABSTRAK

Semakin luas pengguna SMS pada masyarakat, banyak disalahgunakan oleh pihak-pihak yang tidak bertanggung jawab seperti melakukan tindak kejahatan dan dapat mengganggu penerimanya dengan menyebarkan SMS *spam* yang tidak diminta atau tidak diinginkan, diantaranya promosi, penipuan, pesan porno, dan lain sebagainya. Berdasarkan penelitian yang sudah ada, identifikasi SMS *spam* dapat dilakukan dengan dua cara yaitu dengan pembuatan daftar hitam dan pengklasifikasian teks. Pengklasifikasian teks merupakan cara yang lebih efisien dalam mengidentifikasi SMS *spam*. Pengembangan penelitian yang menggunakan pengklasifikasian teks masih terus dilakukan untuk mendapatkan metode yang tepat dan memiliki hasil akurasi yang lebih baik. Pada penelitian ini, dikembangkan sistem untuk identifikasi SMS *spam* berbahasa Indonesia dengan metode *Support Vector Machine*. Tahapan keseluruhan metode yang digunakan adalah *preprocessing* (*tokenizing*, *case folding*, *stopword removal*, *stemming*), dan identifikasi menggunakan *Support Vector Machine*. Penelitian ini membandingkan hasil *F-score* percobaan terhadap klasifikasi SMS menggunakan model yang telah di *cross-validation*. Hasil pengujian pada penelitian ini menunjukkan metode *Support Vector Machine* dapat mengidentifikasi SMS *spam* sangat baik dengan memperoleh tingkat akurasi sebesar 99,28%.

Kata kunci: identifikasi teks, SMS *spam*, *Support Vector Machine*

IDENTIFICATION OF INDONESIAN SPAM SMS USING SUPPORT VECTOR ALGORITHM MACHINE

ABSTRACT

The widespread use of SMS users in the community, many abused by irresponsible parties such as committing a crime and can interfere with the recipients by spreading unsolicited or unwanted spam SMS, including promotions, fraud, porn messages, and so forth. Based on existing research, spam SMS identification can be done in two ways, namely by making black list and classification of text. Text classification is a more efficient way of identifying spam SMS. The development of research using text classification is still being done to get the right method and have better accuracy result. In this study, developed a system for identification of SMS spam in Indonesian with Support Vector Machine (SVM) method. The overall stages of the method used are preprocessing (tokenizing, case folding, stopword removal, stemming), and identification using SVM. This study compared the results of F-score experiments on SMS classification using cross-validation model. The results of testing in this study shows SVM method can identify spam SMS very well by obtaining an accuracy of 99,28%.

Keywords : text identification, SMS *spam*, *Support Vector Machine*

DAFTAR ISI

	Hal.
PERSETUJUAN	ii
PERNYATAAN	iii
UCAPAN TERIMA KASIH	iv
ABSTRAK	vi
ABSTRACT	vii
DAFTAR ISI	viii
DAFTAR TABEL	xi
DAFTAR GAMBAR	xii
 BAB 1 PENDAHULUAN	 1
1.1. Latar Belakang	1
1.2. Rumusan Masalah	4
1.3. Tujuan Penelitian	4
1.4. Batasan Masalah	4
1.5. Manfaat Penelitian	5
1.6. Metodologi Penelitian	5
1.7. Sistematika Penulisan	6
 BAB 2 LANDASAN TEORI	 7
2.1. <i>Short Message Service (SMS)</i>	7
2.2. <i>Stemming</i>	8
2.3. <i>LIBSVM</i>	12
2.4. <i>Support Vector Machine (SVM)</i>	12

2.5. <i>Cross Validation</i>	16
2.6. Metode Evaluasi	17
2.7. Penelitian Terdahulu	17
 BAB 3 ANALISIS DAN PERANCANGAN SISTEM	 22
3.1. Data Penelitian	22
3.2. Analisis Sistem	24
3.2.1. <i>Input</i>	25
3.2.2. <i>Preprocessing</i>	25
3.2.3. Implementasi <i>Support Vector Machine</i>	27
3.3. Perancangan Sistem	28
3.3.1. Perancangan Tampilan Halaman Pelatihan	28
3.3.2. Perancangan Tampilan Halaman Model	29
3.3.3. Rancangan Tampilan Halaman Pengujian	30
3.4. Metode Evaluasi	31
 BAB 4 IMPLEMENTASI DAN PENGUJIAN SISTEM	 33
4.1. Implementasi Sistem	33
4.1.1. Spesifikasi Perangkat Keras dan Perangkat Lunak	33
4.1.2. Implementasi Perancangan Antarmuka	33
1. Tampilan Halaman Pelatihan	34
2. Tampilan Halaman Model	35
3. Tampilan Halaman Pengujian	35
4.2. Pengujian Sistem	37
4.2.1. Pelatihan Model	37
4.2.2. Pengujian Model	39
4.3. Hasil Pengujian Sistem	42
 BAB 5 KESIMPULAN DAN SARAN	 45
5.1. Kesimpulan	45
5.2. Saran	45

DAFTAR PUSTAKA

47

DAFTAR TABEL

	Hal.
Tabel 2.1. Aturan Peluruhan Kata Dasar (Adriani et al., 2007)	10
Tabel 2.2. Daftar Kemungkinan Perubahan Awalan (Adriani et al., 2007)	11
Tabel 2.3. Daftar Kombinasi Awalan dan Akhiran yang Tidak Diperbolehkan (Adriani et al., 2007)	11
Tabel 2.4 Variabel Perhitungan <i>F-score</i>	17
Tabel 2.5. Penelitian Terdahulu	20
Tabel 3.1. Rincian Tipe SMS <i>Spam</i>	23
Tabel 3.2. Pembagian Data Penelitian	23
Tabel 3.3 Tabel Contoh teks SMS Berbahasa Indonesia	25
Tabel 3.4 Tahapan <i>Tokenizing</i>	25
Tabel 3.5 Tahapan <i>Case Folding</i>	26
Tabel 3.6 Tahapan <i>Stopword Removal</i>	26
Tabel 3.7 Tahapan <i>Stemming</i>	27
Tabel 3.8 Contoh Perhitunagn Menggunakan <i>F-Score</i>	31
Tabel 4.1. Data Latih yang digunakan	38
Tabel 4.2. Data hasil pengujian	40
Tabel 4.3. Hasil Pengujian Sistem pada Model Pelatihan Sistem	42
Tabel 4.4. Rata-rata nilai <i>F-score</i> Pengujian Sistem (100%)	42

DAFTAR GAMBAR

	Hal.
Gambar 2.1. Alternatif bidang pemisah (kiri) dan bidang pemisah terbaik dengan margin (m) terbesar (kanan) (Krisantus, 2007).	13
Gambar 2.2. Contoh 3- <i>Fold Cross-Validation</i> (Refaeilzadeh et al., 2008)	16
Gambar 3.1. Arsitektur Umum	24
Gambar 3.2. Pseudo code untuk SVM	27
Gambar 3.3. Rancangan Tampilan Halaman Pelatihan	28
Gambar 3.4. Rancangan Tampilan Halaman Model	29
Gambar 3.5. Rancangan Tampilan Halaman Pengujian	30
Gambar 4.1. Tampilan Halaman Pelatihan	34
Gambar 4.2. Tampilan Halaman Proses Pelatihan	34
Gambar 4.3. Tampilan Halaman Model	35
Gambar 4.4. Tampilan Halaman Pengujian	36
Gambar 4.5. Tampilan Hasil Identifikasi sebagai SMS <i>spam</i>	36
Gambar 4.6. Tampilan Hasil Identifikasi sebagai SMS <i>ham</i>	37
Gambar 4.7. Grafik <i>F-score</i> Pengujian Sistem	43

BAB 1

PENDAHULUAN

1.1. Latar Belakang

SMS (*Short Message Service*) merupakan salah satu media komunikasi tanpa kabel yang masih digunakan masyarakat dengan biaya murah, yang memungkinkan pengirim SMS dan penerima SMS dapat dilakukan dengan cepat dan mudah. Pesatnya perkembangan teknologi SMS dipengaruhi oleh banyaknya *provider* penyedia jasa telekomunikasi yang menawarkan jasanya dengan harga yang cukup terjangkau oleh masyarakat luas. Semakin luas pengguna SMS pada masyarakat disalahgunakan oleh pihak yang tidak bertanggung jawab untuk melakukan tindak kejahatan dengan menyebarkan SMS *spam* yang tidak diminta dan tidak diinginkan, seperti promosi, penipuan, pesan porno, dan lain sebagainya (Ma *et al.*, 2016).

Ada dua cara untuk melakukan penyaringan SMS *spam*, yaitu dengan pembuatan daftar hitam dan pengklasifikasian teks (Taufiq *et al.*, 2010). Daftar Hitam adalah cara yang sederhana, dengan membandingkan kata-kata dalam isi SMS dan kata kunci yang ada pada daftar hitam. Namun cara ini kurang efisien, karena pengguna masih harus memilih dan memasukkan kata kunci dalam daftar hitam secara manual. Penerapan teknik ini sudah tersedia di banyak jenis telepon seluler seperti *SMS Spam Manajer* berjalan di *Symbian OS* dan *Spam SMS Blocker* berjalan di *Google Android*. Sebaliknya, pengklasifikasian teks menggunakan pengenalan pola teks seperti *Naïve Bayes*, *Support Vector Machine* (SVM), *Artificial Neural Network* (ANN), *Decision Tree*, *k-Nearest Neighbor* (kNN), dan *Hidden Markov Model* (HMM) (Nuruzzaman *et al.*, 2011).

Penelitian mengenai deteksi SMS *spam* telah banyak dilakukan dan dikembangkan dengan menggunakan berbagai macam metode pengklasifikasian baik dengan pendekatan pembelajaran mesin (*machine learning*) atau pendekatan

dari gabungan beberapa metode sehingga memperoleh hasil yang akurat. Beberapa penelitian yang terkait dengan deteksi SMS *spam* diantaranya Khemapatapan (2010) melakukan penelitian tentang penyaringan SMS *spam* dengan menggunakan algoritma *Support Vector Machine* (SVM) dan *Naive Bayesian* (NB). Pada penelitian tersebut menerapkan proses semantik untuk menganalisis atau mengoreksi kata dalam bahasa Thai. Salah satu hasil yang diperoleh adalah SVM memberikan akurasi yang lebih tinggi dari *Naive Bayesian*, akan tetapi waktu pemrosesan klasifikasi dengan menggunakan SVM membutuhkan waktu pemrosesan lebih lama dibandingkan dengan menggunakan *Naive Bayesian*. Kemudian Shahi & Yadav (2014) melakukan penelitian deteksi SMS *spam* untuk teks berbahasa Nepali dengan membandingkan algoritma *Naive Bayesian* dengan *Support Vector Machine*. Dari penelitian tersebut diperoleh hasil akurasi algoritma *Naive Bayesian* lebih tinggi daripada *Support Vector Machine*, dengan akurasi masing-masing berurutan 92,74% dan 87,15% dengan jumlah total data latih dan data uji sebanyak 150 SMS. Kemudian Arifin *et al.* (2016) mengusulkan untuk ditingkatkan SMS *spam* yang kinerja penyaringan dengan menggabungkan dua asosiasi tugas *data mining* dan klasifikasi. *FP-Growth* di asosiasi digunakan untuk *mining frequent pattern* pada SMS dan klasifikasi *Naive Bayes* digunakan untuk mengklasifikasikan apakah SMS *spam* atau *ham*. Peneliti menggunakan SMS sebesar 5.574 SMS yang terdiri dari 4.827 SMS *ham* dan 747 SMS *spam*. Pada penelitian tersebut mereka menggunakan kolaborasi *Naive Bayes* dan *FP-Growth* didapatkan hasil akurasi rata-rata tertinggi 98,506% dan 0,025% lebih baik daripada tanpa menggunakan *FP-Growth* untuk dataset SMS *Spam Koleksi v.1*, dan meningkatkan nilai presisi, maka akan didapatkan hasil klasifikasi yang lebih akurat. Kemudian Saputra (2017) melakukan penelitian untuk mendeteksi SMS *spam* bahasa Indonesia menggunakan algoritma *Twitter-LDA*. Tahapan keseluruhan metode yang digunakan pada penelitian tersebut ialah *preprocessing* (*case folding*, *punctuation removing*, tokenisasi, penanganan alamat URL dan nomor telepon, *stemming*, *filtering*, dan normalisasi), pemodelan topik dengan *Twitter-LDA*, dan klasifikasi SMS. Peneliti membandingkan hasil *F-score* percobaan terhadap klasifikasi SMS menggunakan model yang menerapkan penambahan *filtering* dan/atau normalisasi dengan model tanpa *filtering* dan/atau normalisasi. Dari penelitian tersebut, peneliti memperoleh hasil bahwa *Twitter-*

LDA dengan nilai F-score sebesar 96,24% menggunakan 774 SMS *spam* sebagai data latih dan 221 SMS *spam* dan *ham* sebagai data uji.

Beberapa penelitian terdahulu tersebut yaitu pengidentifikasian SMS *spam* menggunakan algoritma *Support Vector Machine* (SVM) dalam teks SMS. Berdasarkan penelitian yang dilakukan oleh Delany *et al.* (2012), diantara semua algoritma pengklasifikasian teks untuk mendeteksi SMS *spam*, SVM adalah yang paling baik. Hal ini dibuktikan dengan penelitian yang dilakukan Fernandes *et al.* (2015) untuk mendeteksi SMS *spam* berbahasa Inggris dengan menggunakan algoritma *Optimum-Path Forest* (OPF) yang kemudian membandingkan akurasi antara OPF dengan SVM, dari 747 SMS *spam* dan 4,827 SMS *ham*, dengan 1674 SMS untuk data latih dan 3900 SMS untuk data uji, diperoleh hasil bahwa SVM dengan akurasi 97,79% lebih akurat jika dibandingkan dengan OPF dengan akurasi 92,23%. Agarwal *et al.* (2015) melakukan penyaringan pesan mobile sebagai *Ham* atau *Spam* untuk pengguna masyarakat India dengan menganalisis perbedaan pengklasifikasi *machine learning* dalam *corpus* besar. Dari penelitian tersebut, *Support Vector Machine* dan *Multinomial Naive Bayes* terbukti menghasilkan pengklasifikasian yang terbaik untuk deteksi SMS *spam*. Pada penelitian tersebut mereka mengubah dataset yang sama untuk pasar India yang tersedia oleh peneliti-peneliti sebelumnya yang terdiri dari 4.827 pesan *ham* dan 747 *spam*, dengan menambahkan 439 pesan yang sah dan 748 *spam* dari perspektif India. Hasil terbaik diubah dari SMS *Spam* Pengumpulan Data Set termasuk konten India keluar menjadi 98,23% dari *Accuracy*, 92,88% dari *Spam Caught* dan 0,54% dari *Blocked ham* dengan SVM. Kemudian Ma *et al.* (2016) melakukan penelitian untuk mendeteksi SMS *spam* berbahasa Inggris menggunakan algoritma *Message Topic Model* (MTM) yang merupakan gabungan antara algoritma *K-Means* dan *Probabilistic Latent Semantic Analysis* (PLSA). Ide utama dari algoritma ini adalah penerapan modifikasi algoritma PLSA yang sering digunakan untuk pengklasifikasi teks dalam jumlah besar. PLSA bekerja baik pada banyak data, maka diperlukan suatu metode agar PLSA bisa bekerja baik juga pada data dalam jumlah kecil. Untuk mengatasi masalah tersebut, peneliti menambahkan algoritma *K-Means* pada awal pemrosesan. Dari penelitian tersebut, para peneliti memperoleh hasil bahwa MTM dengan persentase akurasi sebesar 97% menggunakan 1083 SMS *spam* sebagai data latih dan 770 SMS sebagai data uji.

Berdasarkan penelitian-penelitian yang telah dilakukan diatas, maka penulis mengusulkan menggunakan algoritma *Support Vector Machine*. Dengan harapan, penerapan metode tersebut dapat memberikan hasil maksimal untuk melakukan identifikasi SMS *spam* berbahasa Indonesia.

Berdasarkan latar belakang diatas, penulis mengajukan penelitian terhadap identifikasi SMS *spam* dengan judul “Identifikasi SMS *Spam* Berbahasa Indonesia Menggunakan *Support Vector Machine*”.

1.2. Rumusan Masalah

Semakin luas pengguna SMS pada masyarakat, banyak disalahgunakan oleh pihak-pihak yang tidak bertanggung jawab seperti melakukan tindak kejahatan dengan menyebarkan SMS *spam* yang tidak diminta atau tidak diinginkan, diantaranya promosi, penipuan, pesan porno, dan judi. Oleh karena itu, dibutuhkan pendekatan yang paling baik untuk mengidentifikasi SMS *spam*, khususnya SMS *spam* berbahasa Indonesia.

1.3. Tujuan Penelitian

Penelitian ini bertujuan untuk mengidentifikasi SMS yang mengandung unsur *spam* berbahasa Indonesia dengan menggunakan algoritma *Support Vector Machine (SVM)*.

1.4. Batasan Masalah

Dalam penelitian ini, peneliti memberikan beberapa batasan masalah yakni:

1. Data latih dan data uji teks SMS yang digunakan dalam sistem adalah teks SMS berbahasa Indonesia.
2. Identifikasi dilakukan secara *offline*.
3. Sistem yang dibangun mengidentifikasi SMS yang dinilai sebagai *spam*.
4. Jenis sms spam yang diproses adalah Promo/Iklan, Dana Tunai, Hadiah Lomba, Judi, Sex, dan Penipuan.

5. Dataset yang digunakan adalah file dokumen berekstensi *txt* yang mengandung teks SMS unik pada setiap barisnya.
6. Jumlah SMS yang diinput dalam pengujian maksimal satu text SMS untuk setiap pengujian.

1.5. Manfaat Penelitian

Manfaat dalam penelitian ini adalah sistem dapat memberikan hasil yang baik dalam melakukan identifikasi SMS yang mengandung unsur *spam* dalam bahasa Indonesia. Mengukur kinerja sistem identifikasi SMS *spam* dengan teks bahasa Indonesia menggunakan algoritma *Support Vector Machine (SVM)*.

1.6. Metodologi Penelitian

Tahapan-tahapan yang akan dilakukan pada pelaksanaan penelitian adalah sebagai berikut:

1. Studi Literatur

Pada tahap ini penulis mengumpulkan bahan referensi berkaitan dengan *Short Message Service (SMS)*, *Stemming*, *LIBSVM*, dan *Support Vector Machine (SVM)* dari berbagai jurnal, skripsi, buku, artikel dan berbagai sumber referensi lainnya.

2. Analisis Masalah

Pada tahap ini dilakukan analisis untuk setiap informasi yang telah di peroleh dari tahap sebelumnya agar mendapatkan pemahaman akan masalah dan metode yang akan digunakan untuk menyelesaikan permasalahan.

3. Perancangan Sistem

Pada tahap ini dilakukan perancangan sistem sesuai dengan hasil dari data, interface dan sistem keseluruhan.

4. Implementasi Program (*Coding*)

Pada tahap ini hasil dari analisis dan perancangan sistem akan di implementasikan ke dalam kode program perangkat lunak pengidentifikasian SMS *spam* menggunakan bahasa pemrograman Java.

5. Pengujian

Pada tahap ini dilakukan pengujian terhadap sintaksis pemrograman. Selain itu, pengujian juga dilakukan untuk memastikan bahwa proses mengidentifikasi SMS yang dilakukan sistem memberikan hasil yang terbaik.

6. Dokumentasi dan Penyusunan Laporan

Pada tahap terakhir membuat dokumentasi dan menyusun laporan hasil dari analisis dan implementasi dari penelitian yang dilakukan dalam bentuk skripsi.

1.7. Sistematika Penulisan

Sistematika penulisan dari skripsi ini terdiri dari lima bagian utama sebagai berikut.

Bab 1: Pendahuluan

Bab ini berisi penjelasan mengenai latar belakang pemilihan judul skripsi “Identifikasi SMS *Spam* Berbahasa Indonesia Menggunakan Algoritma *Support Vector Machine*”, rumusan masalah tujuan penelitian, batasan masalah, manfaat penelitian, metodologi penelitian dan sistematika penulisan.

Bab 2: Landasan Teori

Bab ini akan membahas teori-teori yang digunakan pada penelitian ini. Teori-teori yang berhubungan dengan *Short Message Service* (SMS), *Stemming*, *LIBSVM*, dan *Support Vector Machine* (SVM).

Bab 3: Analisis dan Perancangan

Bab ini terdiri dari metode yang digunakan, arsitektur umum, analisis kebutuhan perangkat lunak dan penerapan metode *Support Vector Machine* dan perancangan aplikasi untuk melakukan pengidentifikasi SMS *spam*.

Bab 4: Implementasi dan Pengujian

Bab ini berisi implementasi dari analisis dan perancangan pada bab 3 dan pengujian pada aplikasi yang berhasil di bangun.

Bab 5: Kesimpulan dan Saran

Bab ini berisi rangkuman dari hasil penelitian yang telah dilakukan dan saran-saran untuk pengembangan aplikasi atau penelitian selanjutnya.

BAB 2

LANDASAN TEORI

Pada bab ini akan dibahas mengenai teori-teori yang digunakan sebagai landasan dalam penyelesaian masalah dalam penelitian ini.

2.1. *Short Message Service (SMS)*

Short Message Service (SMS) merupakan sebuah layanan dasar yang memungkinkan pengguna bertukaran pesan teks dalam jumlah yang pendek. Pesan teks singkat diyakini telah dialihkan pada tahun 1992 lebih dari sinyal saluran dari jaringan GSM Eropa. Sejak uji coba yang sukses ini, penggunaan SMS telah menjadi perhatian pertumbuhan yang luar biasa. Mobile Data Association melaporkan bahwa jumlah total pesan teks orang-ke-orang yang dikirim ke empat jaringan GSM Inggris pada tahun 2003 berjumlah 20,5 miliar. Pesan singkat ini dapat dikirim dari perangkat mobile seluler GSM/UMTS tetapi bisa juga dikirim dari perangkat lain dengan cakupan yang lebih luas seperti *internet host*, *telex*, dan SMS faksimile. *Short Message Service (SMS)* ialah teknologi yang sangat maju dan 100% didukung oleh perangkat GSM dan sebagian besar jaringan GSM yang ada di dunia (Bodic, 2005).

Terdapat dua fitur paling dasar dalam sebuah SMS ialah mengirim dan menerima satu pesan singkat (Bodic, 2005). Pesan yang disampaikan dari satu *Mobile Station (MS)* ke satu *SMS Centre (SMSC)* adalah pesan yang dikirim oleh perangkat. Pesan tersebut dialamatkan ke *SME* yang lain seperti perangkat pengguna atau *Internet hosts* lainnya. Kemudian pesan yang dikirim oleh satu *SMSC* ke satu *MS* adalah pesan yang diterima oleh perangkat.

2.2. Stemming

Stemming merupakan proses pengubahan kata yang mengandung imbuhan menjadi kata dasar dengan menghapus awalan dan akhiran yang terdapat pada suatu kata, tujuannya untuk melakukan pengelompokan kata-kata yang diturunkan dari sebuah data *stem* yang umum dan kata dasar (Adriani et al., 2007).

Dalam kata bahasa Indonesia dikenal algoritma *stemming* diantaranya yaitu; algoritma Arifin dan Setiono, algoritma Vega, algoritma Nazief dan Adriani, serta algoritma Ahmad, Yussof, dan Sembok. Dari beberapa algoritma tersebut terdapat algoritma yang paling baik dalam melakukan stemming untuk kata bahasa Indonesia yaitu algoritma Nazief dan Adriani (Asian et al., 2005).

Algoritma Nazief dan Adriani memiliki konsep dasar bahwa sebuah kata dasar dapat ditambah pada imbuhan berupa *Derrivation Prefix* (DP) di awal dan/atau diakhiri secara berurutan oleh *Derrivation Suffix* (DS), *Possesive Pronoun* (PP) dan *Particle* (P) yang masing-masing bersifat opsional. Untuk lebih jelasnya, algoritma ini menggunakan aturan imbuhan sendiri dengan model sebagai berikut (Adriani et al., 2007) :

$$\left[\left[\left[DP + \right] DP + \right] DP + \right] \text{ kata dasar } \left[\left[+DS \right] \left[+PP \right] \left[+P \right] \right] \dots\dots\dots(2.1)$$

Dimana DP = awalan (*prefix*)

DS = akhiran (*suffix*)

PP = kata ganti kepemilikan (*possesive pronoun*)

P = partikel (*particle*).

Tanda kurung besar menandakan bahwa imbuhan adalah optional.

Algoritma *stemming* memiliki beberapa langkah yang diperlukan sebagai berikut (Adriani et al., 2007):

1. Lakukan pencarian kata di kamus. Jika kata tersebut ditemukan, maka dapat diasumsikan bahwa kata tersebut adalah kata dasar dan algoritma dihentikan.
2. Jika dari langkah sebelumnya tidak ditemukan, maka akan dilakukan apakah mengandung akhiran yang merupakan sebuah partikel (“-lah” atau “-kah”). Jika ada, partikel tersebut akan dihapus dari kata.
3. Kemudian, lanjutkan pengecekan apakah mengandung kata ganti milik (“-ku”,

“-mu”, “-nya”). Jika ada, maka hapus dari kata ganti tersebut.

4. Melakukan pengecekan akhiran (“-i”, “-an”). Jika ada, maka hapus akhiran tersebut. Dengan catatan; sampai pada langkah ini dibutuhkan ketelitian untuk mengecek apakah “-an” termasuk bagian dari akhiran “-kan” atau bukan, dan juga mengecek apakah partikel (“-lah”, “-kah”) dan kata ganti milik (“-ku”, “-mu”, “-nya”) yang telah dihapus padaa langkah 2 dan 3 bukan merupakan bagian dari kata dasar.
5. Melakukan pengecekan apakah kata mengandung awalan (“se-”, “ke-”, “di-”, “te-”, “be-”, “pe-”, “me-”). Jika ada, maka hapus awalan tersebut. Pengecekan dilakukan dengan berulang mengingat adanya kemungkinan *multi-prefix*. Dengan catatan; Langkah ini juga membutuhkan ketelitian untuk mengecek kemungkinan terjadinya peluruhan awalan berdasarkan Tabel 2.1, perubahan awalan yang disesuaikan dengan huruf-awal kata berdasarkan Tabel 2.2, dan aturan kombinasi awalan-akhiran yang tidak diperbolehkan berdasarkan tabel 2.3.

Pada tabel 2.1 dapat dilihat aturan-aturan peluruhan kata yang apabila digabungkan oleh awalan “me-“, “be-“, “te-“, “pe-“. Dimana pada kolom kedua dari tabel tersebut menjelaskan bentuk kata dasar yang digabungkan awalan “me-“, “be-“, “te-“, “pe-“, sedangkan pada kolom ketiga menjelaskan perubahan karakter pada kata dasar yang mungkin terjadi apabila algoritma telah menghilangkan awalan yang telah menggabungkan kata dasar tersebut. Huruf “V” pada tabel tersebut menunjukkan huruf hidup atau huruf vocal, huruf “C” menunjukkan huruf mati atau konsonan, huruf “A” menunjukkan huruf vocal atau huruf konsonan dan huruf “P” menunjukkan pecahan “er”.

6. Setelah semua langkah dilakukan dan diperoleh kata dasarnya, maka algoritma ini akan mengembalikan kata dasar yang dihasilkan tersebut. Jika tidak diperoleh, maka kata semula yang akan dikembalikan.

Tabel 2.1 Aturan Peluruhan Kata Dasar (Adriani et al., 2007)

Aturan	Bentuk Awalan	Peluruhan
1	berV...	ber-V... be-rV...
2	berCAP...	ber-CAP... dimana C!= 'r' dan P!= 'er'
3	berCAerV...	ber-CAerV... dimana C!= 'r'
4	belajar...	bel-ajar...
5	beC ₁ erC ₂ ...	be-C ₁ erC ₂ ...dimana C ₁ !={'r' 'l'}
6	terV...	ter-V... te-rV...
7	terCP...	ter-CP...dimana C!= 'r' dan P!= 'er'
8	terCer...	ter-Cer...dimana C!= 'r'
9	teC ₁ erC ₂	te-C ₁ erC ₂ ...dimana C ₁ != 'r'
10	me{l r w y}V...	me-{l r w y}V...
11	mem{b f v}...	mem-{b f v}...
12	mempe...	mem-pe...
13	mem{rV V}...	me-m{rV V}... me-p{rV V}...
14	men{c d j z}...	men-{c d j z}...
15	menV...	me-nV... me-tV...
16	meng{g h q k}...	meng-{g h q k}...
17	mengV...	meng-V... meng-kV...
18	meny...	meny-sV...
19	mempV...	mem-pV... dimana V!= 'e'
20	pe{w y}V...	pe-{w y}V...
21	perV...	per-V... pe-rV...
22	perCAP...	per-CAP... dimana C!= 'r' dan P!= 'er'
23	perCAerV...	per-CAerV... dimana C!= 'r'
24	pem{b f v}...	pem-{b f v}...
25	pem{rV V}...	pe-m{rV V}... pe-p{rV V}
26	pen{c d j z}...	pen-{c d j z}...

Tabel 2.1 Aturan Peluruhan Kata Dasar (Adriani et al., 2007)
(Lanjutan)

Aturan	Bentuk Awalan	Peluruhan
27	penV...	pe-nV... pe-tV...
28	peng{g h q}	peng-{g h q}
29	pengV...	peng-V... peng-kV...
30	penyV...	peny-sV...
31	pelV...	pe-lV...; kecuali untuk kata “pelajar” menjadi “ajar”
32	peCP...	pe-CP... dimana $C! = \{r w y l m n\}$ dan $P! = \text{'er'}$
33	perCerV...	per-CerV... dimana $C! = \{r w y l m n\}$

Tabel 2.2 Daftar Kemungkinan Perubahan Awalan (Adriani et al., 2007)

No	Awalan	Perubahan
1	di-	tidak berubah
2	ke-	tidak berubah
3	se-	tidak berubah
4	be-	ber-
5	me-	mem-, men-, meng-, meny-
6	pe-	per-, pen-, pem-, peng-
7	te-	ter-

Tabel 2.3 Kombinasi Awalan dan Akhiran yang Tidak Diperbolehkan
(Adriani et al., 2007)

No	Awalan (<i>Prefix</i>)	Akhiran (<i>Suffix</i>)
1	be-	-i
2	di-	-an

**Tabel 2.3 Kombinasi Awalan dan Akhiran yang Tidak Diperbolehkan
(Adriani et al., 2007) (Lanjutan)**

No	Awalan (<i>Prefix</i>)	Akhiran (<i>Suffix</i>)
3	ke-	-i-an
4	me-	-an
5	pe-	-i-kan
6	se-	-i-kan
7	te-	-an

2.3. LIBSVM

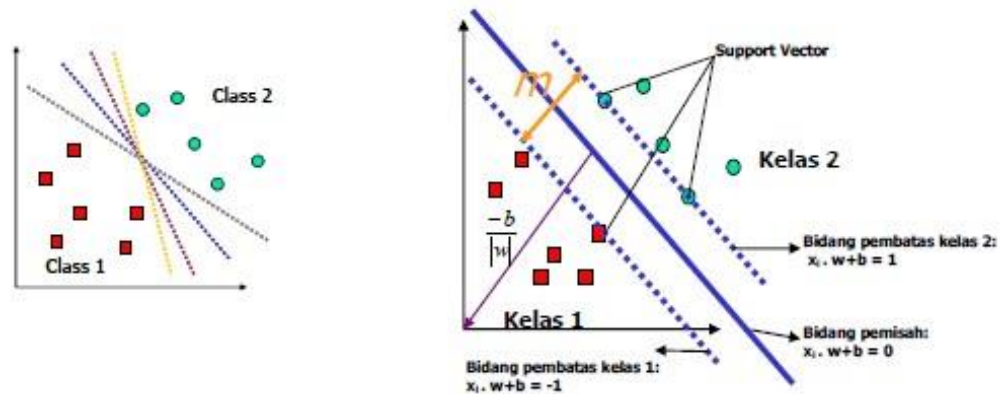
LIBSVM merupakan *library* implementasi *Support Vector Machine*. LIBSVM saat ini adalah salah satu perangkat lunak *Support Vector Machine* yang paling banyak digunakan dan bagian dari metode *decomposition* (Lin, 2005). Metode ini bekerja berdasarkan prinsip '*working set*' dengan mengubah beberapa multiplier α_i dalam jumlah tertentu pada setiap iterasi, sementara nilai yang lain bernilai tetap. *Working set* merupakan kumpulan variabel yang sedang dioptimasi pada *current iteration*. Secara umum, prinsip kerja *decomposition* pada LIBSVM adalah mengoptimasi masalah global dengan hanya menggunakan sebagian kecil data pada satu saat (Krisantus, 2007).

2.4. Support Vector Machine (SVM)

Support Vector Machine (SVM) memiliki dasar teoritis yang kuat dan keberhasilan empiris yang sangat baik. *Support Vector Machine* (SVM) telah diterapkan untuk tugas-tugas seperti *handwritten digit recognition* (pengenalan tulisan tangan digit), *object recognition* (pengenalan obyek) dan *text classification* (klasifikasi teks) (Vapnik, 1982).

Support Vector Machine (SVM) adalah sistem pembelajaran yang menggunakan hipotesis fungsi linear dalam ruang berdimensi tinggi dan dilatih dengan algoritma berdasarkan teori optimasi dengan menerapkan *learning bias* yang berasal dari teori statistik (Christianini & Shawe-Taylor, 2000). Support

Vector Machine merupakan salah satu teknik yang relatif baru dibandingkan dengan teknik lain yang memiliki performansi yang lebih baik di berbagai bidang seperti bioinformatics, pengenalan tulisan tangan, klasifikasi teks dan lain sebagainya (Krisantus, 2007). Tujuan utama dari metode ini adalah untuk membangun OSH (*Optimal Separating Hyperplane*), yang membuat fungsi pemisahan optimum yang dapat digunakan untuk klasifikasi.



Gambar 2.1 Alternatif bidang pemisah (kiri) dan bidang pemisah terbaik dengan margin (m) terbesar (kanan) (Krisantus, 2007).

Menurut Osuna et al (1997) data dikatakan *linearly separable* jika permasalahan tersebut dapat dicari pasangan (w, b) . *Linearly separable data* merupakan data yang dapat dipisahkan secara linier. *Support Vector Machine* menggunakan model linear sebagai *decision boundary* dengan bentuk umum sebagai berikut :

$$y(x) = w \cdot x + b$$

Dimana w = parameter bobot (normal bidang)

x = vektor input (label kelas)

b = bias.

Misalnya $\{x_1, \dots, x_n\}$ adalah dataset dan $y_i \in \{+1, -1\}$ adalah label kelas dari data x_i . Pada gambar 2.2 dapat dilihat berbagai alternatif bidang pemisah yang dapat memisahkan semua data set sesuai dengan kelasnya. Namun, bidang pemisah terbaik tidak hanya dapat memisahkan data tetapi juga memiliki margin paling besar.

Data yang berada pada bidang pembatas disebut dengan *support vector*.

Dalam gambar 2.2, dua kelas dapat dipisahkan oleh sepasang bidang pembatas yang sejajar. $\frac{|b|}{||w||}$ merupakan jarak bidang pemisah yang tegak lurus dari titik pusat koordinat dan $||w||$ adalah jarak *euclidean* dari w . Bidang pembatas pertama membatasi kelas pertama sedangkan bidang pembatas kedua membatasi kelas kedua, sehingga diperoleh:

$$\begin{aligned} x_i \cdot w + b &\geq +1 \text{ for } y_i = +1 \\ x_i \cdot w + b &\leq -1 \text{ for } y_i = -1 \end{aligned} \quad \dots\dots\dots (2.2)$$

w adalah normal bidang dan b adalah posisi bidang alternatif terhadap pusat koordinat. Nilai margin (jarak) antara bidang pembatas (berdasarkan rumus jarak garis ke titik pusat) adalah $\frac{1-b-(-1-b)}{||w||} = \frac{2}{||w||}$. Nilai margin ini dimaksimalkan dengan tetap memenuhi persamaan 2.2. Dengan mengalikan b dan w dengan sebuah konstanta, akan dihasilkan nilai margin yang dikalikan dengan konstanta yang sama. Oleh karena itu, *constraint* pada persamaan 2.2 merupakan *scaling constraint* yang dapat dipenuhi dengan *rescaling* b dan w . Selain itu karena memaksimalkan $\frac{1}{|w|}$ sama dengan meminimumkan $|w|^2$. Jika kedua bidang pembatas pada persamaan 2.2 direpresentasikan dalam pertidaksamaan,

$$y_j (x_i w + b) - 1 \geq 0 \quad \dots\dots\dots (2.3)$$

maka pencarian bidang pemisah terbaik dengan nilai margin terbesar dapat dirumuskan menjadi masalah optimasi konstrain, seperti terlihat pada persamaan 2.4 :

$$\begin{aligned} \min \frac{1}{2} ||w||^2 \\ y_i (x_i w + b) - 1 \geq 0 \end{aligned} \quad \dots\dots\dots (2.4)$$

dimana x_i merupakan data masukan dan y_i merupakan keluaran, sedangkan w dan b merupakan parameter yang kita cari nilainya.

Persoalan ini akan lebih mudah diselesaikan jika diubah ke dalam formula *lagrangian* yang menggunakan *lagrange multiplier*. Dengan demikian permasalahan optimasi konstrain dapat diubah menjadi:

$$\min_{w,b} L_p(w, b, \alpha) \equiv \frac{1}{2} |w|^2 - \sum_{i=1}^n \alpha_i y_i (x_i \cdot w + b) + \sum_{i=1}^n \alpha_i \dots\dots\dots (2.5)$$

dengan tambahan konstrain, $\alpha_i \geq 0$ (nilai dari koefisien *lagrange*). Dengan meminimumkan L_p terhadap w dan b , maka dari $\frac{\partial}{\partial b} L_p(w, b, \alpha) = 0$ diperoleh (3.2) dan $\frac{\partial}{\partial w} L_p(w, b, \alpha) = 0$ dari diperoleh (3.3).

$$\sum_{i=1}^n \alpha_i y_i = 0 \quad \dots\dots\dots (2.6)$$

$$w = \sum_{i=1}^n \alpha_i y_i x_i \quad \dots\dots\dots (2.7)$$

Vektor w sering kali bernilai besar (mungkin tak terhingga), tetapi nilai α_i terhingga. Untuk itu, formula *lagrangian* L_p (primal problem) diubah kedalam *dual problem* L_D . Dengan mensubstitusikan persamaan (3.4) ke L_p diperoleh *dual problem* L_D dengan konstrain berbeda.

$$L_D(\alpha) \equiv \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1, j=1}^n \alpha_i \alpha_j y_i y_j x_i x_j$$

$$\min_{w, b} L_p = \max_{\alpha} L_D \quad \dots\dots\dots (2.8)$$

Jadi persoalan pencarian bidang pemisah terbaik dapat dirumuskan sebagai berikut:

$$L_D(\alpha) \equiv \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1, j=1}^n \alpha_i \alpha_j y_i y_j x_i x_j$$

$$s. t. \sum_{i=1}^n \alpha_i y_i = 0, \alpha_i \geq 0 \quad \dots\dots\dots (2.9)$$

Dengan demikian, dapat diperoleh nilai α_i yang nantinya digunakan untuk menemukan w . Terdapat nilai α_i untuk setiap data pelatihan. Data pelatihan yang memiliki nilai $\alpha_i > 0$ adalah *support vector* sedangkan sisanya memiliki nilai $\alpha_i = 0$. Dengan demikian fungsi keputusan yang dihasilkan hanya dipengaruhi oleh *support vector*.

Formula pencarian bidang pemisah terbaik ini adalah permasalahan *quadratic programming*, sehingga nilai maksimum global dari α_i selalu dapat ditemukan. Setelah solusi permasalahan *quadratic programming* ditemukan (nilai α_i), maka kelas dari pengujian x dapat ditentukan berdasarkan nilai dari fungsi keputusan:

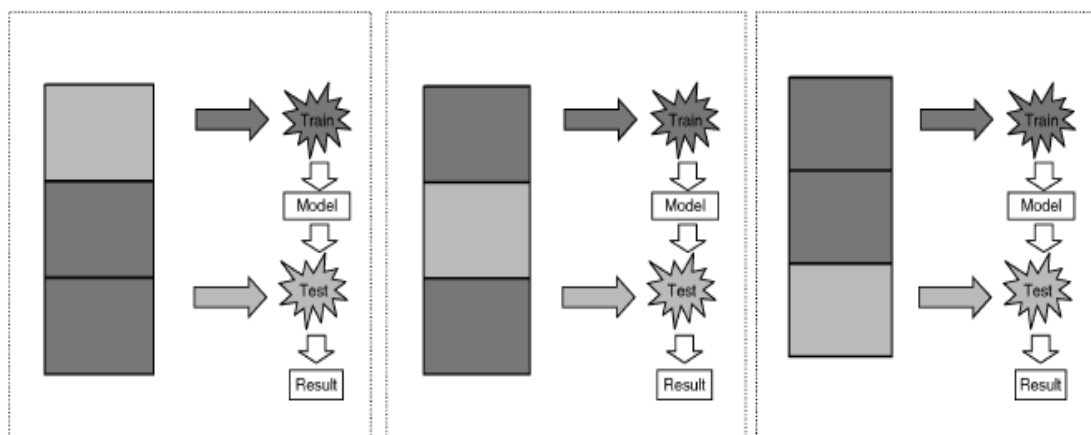
$$f(x_d) = \sum_{i=1}^{n_s} \alpha_i y_i x_i \cdot x_d + b \quad \dots\dots\dots (2.10)$$

x_i adalah support vector, ns = jumlah support vector dan x_d adalah data yang akan diklasifikasikan.

2.5. *Cross-Validation*

Menurut Refaeilzadeh et al. (2008), *Cross-Validation* adalah sebuah metode statistika untuk mengevaluasi dan membandingkan algoritma pembelajaran atau pelatihan dengan cara membagi data menjadi dua segmen; satu segmen digunakan untuk sebuah model dan satu segmen lainnya digunakan untuk proses validasi model.

Pada *k-fold cross-validation*, pertama dataset dibagi menjadi k bagian/segmen. Kemudian dilakukan perulangan sebanyak k kali untuk menjalankan proses pelatihan dan validasi, dimana di setiap perulangannya satu segmen yang berbeda dijadikan sebagai bagian untuk validasi, sedangkan sisanya sebanyak $k-1$ bagian dijadikan sebagai bahan untuk pelatihan (Refaeilzadeh et al., 2008). Sebagai contoh untuk *3-fold cross-validation* dapat dilihat pada Gambar 2.3.



Gambar 2.2 Contoh 3-Fold Cross-Validation (Refaeilzadeh et al., 2008)

Pada gambar 2.3 terlihat bahwa pada *k-fold cross-validation*, data dibagi menjadi 3 bagian/segmen, kemudian dilakukan perulangan sebanyak 3 kali dimana setiap satu bagian yang berbeda digunakan sebagai bagian untuk validasi (dalam penelitian ini, dijadikan data uji), sedangkan dua bagian lain digunakan sebagai bagian untuk data pelatihan.

2.6. Metode Evaluasi

Tahapan evaluasi merupakan perhitungan seberapa baik sistem dalam mengidentifikasi SMS *spam* dan *ham* yang diujikan dalam sistem. Tahapan ini menerapkan aturan variabel pada tabel 2.4 dengan menggunakan rumus umum perhitungan *precision*, *recall* dan *F-score* yang diperkenalkan oleh Baeza-Yates & Ribeiro-Neto (1999), dapat dilihat pada persamaan dibawah ini.

Tabel 2.4 Variabel Perhitungan *F-score*

		Label Manual SMS	
		<i>Spam</i>	<i>Ham</i>
Hasil Identifikasi	<i>Spam</i>	<i>True Positive</i> (TP)	<i>False Positive</i> (FP)
	<i>Ham</i>	<i>False Negative</i> (FN)	<i>True Negative</i> (TN)

$$Recall = \frac{TP}{TP+FN} \dots\dots\dots (2.11)$$

$$Precision = \frac{TP}{TP+FP} \dots\dots\dots (2.12)$$

$$F - score = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \dots\dots\dots (2.13)$$

Keterangan:

- True Positive (TP) adalah kondisi dimana SMS dengan label *spam* berhasil diidentifikasi oleh sistem sebagai *spam*.
- False Positive (FP) adalah kondisi dimana SMS dengan label *ham* gagal diidentifikasi oleh sistem sebagai *ham*.
- True Negative (TN) adalah kondisi dimana SMS dengan label *ham* berhasil diidentifikasi oleh sistem sebagai *ham*.
- False Negative (FN) adalah kondisi dimana SMS dengan label *spam* gagal diidentifikasi oleh sistem sebagai *spam*.

2.7. Penelitian Terdahulu

Penelitian mengenai identifikasi SMS *spam* sudah banyak dilakukan, diantaranya Khemapatapan (2010) melakukan penelitian tentang penyaringan SMS *spam*

dengan menggunakan algoritma *Support Vector Machine* (SVM) dan *Naive Bayesian* (NB). Proses analisis semantik diterapkan pada penelitian ini untuk menganalisis atau mengoreksi kata dalam bahasa Thai. Salah satu hasil yang diperoleh dengan menggunakan *filtering method*#2 yaitu SVM memberikan akurasi 95,6968% yang lebih tinggi dari *Naive Bayesian* dengan 83,3109%, akan tetapi waktu pemrosesan klasifikasi dengan menggunakan SVM membutuhkan waktu pemrosesan lebih lama dibandingkan dengan menggunakan *Naive Bayesian*.

Shahi & Yadav (2014) melakukan penelitian deteksi SMS *spam* untuk teks berbahasa Nepali dengan membandingkan algoritma *Naive Bayesian* dengan *Support Vector Machine* (SVM). Pada penelitian tersebut skema TF-IDF digunakan untuk membentuk vektor ciri yang tidak mempertimbangkan kata di dalam SMS, melainkan hanya mempertimbangkan bobot dari masing-masing kata. Dari penelitian tersebut diperoleh hasil bahwa untuk SMS dengan teks berbahasa Nepali, akurasi algoritma *Naive Bayesian* lebih tinggi daripada SVM, dengan akurasi masing-masing berurutan 92,74% dan 87,15% dengan jumlah total data latih dan data uji sebanyak 150 SMS.

Agarwal *et al.* (2015) melakukan penyaringan pesan *mobile* sebagai *Ham* atau *Spam* untuk pengguna India dengan menambahkan pesan India ke *dataset* SMS yang tersedia di seluruh dunia, dengan menganalisis perbedaan pengklasifikasi *machine learning classifiers* dalam corpus besar pada pesan SMS untuk orang-orang India. Dari penelitian tersebut, menunjukkan bahwa *Support Vector Machine* dan *Multinomial Naive Bayes* didapat pengklasifikasian yang terbaik untuk deteksi SMS *spam*. Pengklasifikasi SVM dengan *Linear Kernel* memiliki akurasi terbaik tetapi waktu yang dibutuhkan untuk proses yang tinggi. Di sisi lain *Multinomial Naive Bayes* dengan *Laplace smoothing* juga memiliki akurasi yang sangat dekat dengan SVM, namun waktu yang dibutuhkan oleh *Multinomial Naive Bayes* jauh lebih rendah daripada SVM. Pada penelitian tersebut mereka mencoba untuk mengubah *dataset* yang sama untuk pasar India yang tersedia sebelumnya oleh peneliti-peneliti sebelumnya yang terdiri dari 4.827 pesan *ham* dan 747 *spam*, dengan menambahkan 439 pesan yang sah dan 748 *spam* dari perspektif India. Hasil terbaik dari Diubah SMS Spam Pengumpulan Data Set termasuk konten India keluar menjadi 98,23% dari *Accuracy*, 92,88% dari *Spam*

Caught dan 0,54% dari *Blocked ham* dengan SVM.

Fernandes *et al.* (2015) juga melakukan penelitian untuk mendeteksi SMS *spam* berbahasa Inggris dengan menggunakan algoritma *Optimum-Path Forest* (OPF). Ide dasar metode OPF adalah untuk memodelkan masalah pengenalan pola dalam bentuk graf. OPF akan mengatur sebuah proses untuk persaingan di antara *node* yang ada di dalam graf untuk menarik *node* lain ke dalam kumpulan *node* masing-masing. Pada penelitian tersebut, mereka membandingkan akurasi antara OPF dengan SVM, dari 747 SMS *spam* dan 4.827 SMS *ham*, dengan 1674 SMS untuk data latih dan 3900 SMS untuk data uji, diperoleh hasil bahwa SVM dengan akurasi 97,79% lebih akurat jika dibandingkan dengan OPF dengan akurasi 92,23%.

Arifin *et al.* (2016) mengusulkan untuk ditingkatkan SMS *spam* yang kinerja penyaringan dengan menggabungkan dua asosiasi tugas *data mining* dan klasifikasi. *FP-Growth* diasosiasi digunakan untuk *mining frequent pattern* pada SMS dan klasifikasi *Naive Bayes* digunakan untuk mengklasifikasikan apakah SMS *spam* atau *ham*. Pelatihan data menggunakan koleksi SMS *spam* dari penelitian sebelumnya yang diambil dari Dt.fee.unicamp.br "*YouTube Spam Collection*" dengan SMS sebesar 5.574 SMS yang terdiri dari 4.827 SMS *ham* dan 747 SMS *spam*. Pada penelitian tersebut mereka menggunakan kolaborasi *Naive Bayes* dan *FP-Growth* didapatkan hasil akurasi rata-rata tertinggi 98,506% dan 0,025% lebih baik daripada tanpa menggunakan *FP-Growth* untuk dataset SMS *Spam Koleksi v.1*, dan meningkatkan nilai presisi. Dengan demikian, didapatkan hasil klasifikasi yang lebih akurat.

Ma *et al.* (2016) melakukan penelitian untuk mendeteksi SMS *spam* berbahasa Inggris menggunakan algoritma *Message Topic Model* (MTM) yang merupakan gabungan antara algoritma *K-Means* dan *Probabilistic Latent Semantic Analysis* (PLSA). Ide utama dari algoritma ini adalah penerapan modifikasi algoritma PLSA yang sering digunakan untuk pengklasifikasi teks dalam jumlah besar. PLSA bekerja baik pada banyak data, maka diperlukan suatu metode agar PLSA bisa bekerja baik juga pada data dalam jumlah kecil. Untuk mengatasi masalah tersebut, peneliti menambahkan algoritma *K-Means* pada awal pemrosesan. Dari penelitian tersebut, para peneliti memperoleh hasil bahwa

MTM dengan persentasi akurasi sebesar 97% menggunakan 1083 SMS *spam* sebagai data latih dan 770 SMS sebagai data uji.

Saputra (2017) melakukan penelitian untuk mendeteksi SMS *spam* bahasa Indonesia menggunakan algoritma *Twitter-LDA*. Tahapan keseluruhan metode yang digunakan pada penelitian tersebut ialah *preprocessing* (*case folding*, *punctuation removing*, tokenisasi, penanganan alamat URL dan nomor telepon, *stemming*, *filtering*, dan normalisasi), pemodelan topik dengan *Twitter-LDA*, dan klasifikasi SMS. Peneliti membandingkan hasil *F-score* percobaan terhadap klasifikasi SMS menggunakan model yang menerapkan penambahan *filtering* dan/atau normalisasi dengan model tanpa *filtering* dan/atau normalisasi. Dari penelitian tersebut, peneliti memperoleh hasil bahwa *Twitter-LDA* dengan nilai *F-score* sebesar 96,24% menggunakan 774 SMS *spam* sebagai data latih dan 221 SMS *spam* dan *ham* sebagai data uji. Untuk lebih jelasnya, ringkasan mengenai penelitian terdahulu dapat dilihat pada Tabel 2.4.

Tabel 2.5. Penelitian Terdahulu

No	Nama Peneliti	Tahun	Metode Penelitian	Keterangan
1	Khemapatapan	2010	<i>Support Vector Machine</i> (SVM) dan <i>Naive Bayesian</i> (NB)	Total data latih dan data uji : 25860 SMS Akurasi SVM : 95,6968% Akurasi NB : 83,3109%
2	Shahi & Yadav	2014	<i>Naive Bayesian</i> (NB), <i>Support Vector Machine</i> (SVM)	Total data latih dan data uji : 150 SMS Akurasi NB : 92,74% Akurasi SVM : 87,15%
3	Agarwal <i>et al.</i>	2015	<i>Support Vector Machine</i> (SVM)	Total data latih dan data uji : 6761 SMS Akurasi : 98,23%

Tabel 2.5. Penelitian Terdahulu (Lanjutan)

No	Nama Peneliti	Tahun	Metode Penelitian	Keterangan
4	Fernandes <i>et al.</i>	2015	<i>Optimum-Path Forest</i> (OPF)	Total data latih dan data uji : 5574 SMS Akurasi : 92,23%
5	Arifin <i>et al.</i>	2016	<i>FP-Growth</i> dan <i>Naive Bayes</i> (NB)	Total data latih dan data uji : 5574 SMS Akurasi <i>FP-Growth</i> dan <i>Naive Bayes</i> (NB) : 98,506 %
6	Ma <i>et al.</i>	2016	<i>Message Topic Model</i> (MTM)	Total data latih dan data uji : 1853 SMS Akurasi : 97%
7	Saputra	2017	<i>Twitter-LDA</i>	Total data latih dan data uji : 985 SMS <i>F-score</i> : 96,24%

BAB 3

ANALISIS DAN PERANCANGAN SISTEM

Bab ini membahas tentang analisis dan implementasi metode yang digunakan dalam sistem identifikasi SMS *spam* berbahasa Indonesia menggunakan Algoritma *Support Vector Machine* (SVM). Pada bab ini juga akan di bahas tentang data yang digunakan, perancangan sistem serta tahap analisis dan perancangan sistem.

3.1. Data Penelitian

Dalam penelitian ini menggunakan data SMS berbahasa Indonesia. Data SMS *spam* diperoleh dari hasil kuesioner yang terbuka untuk umum yang didapat dari SMS *spam* yang diterima pada perangkat seluler pengguna. Sedangkan untuk data SMS *ham* diperoleh dari beberapa sumber. Dalam kuisisioner pengambilan data SMS *spam*, telah disediakan form pengisian Nama, Usia, Teks SMS *spam* beserta jenis SMS *spam*-nya. Tersedia 10 kolom isian Teks SMS *spam* dalam memudahkan pengisian SMS *spam* yang banyak diterima pengguna. Dengan jenis SMS *spam* yang disediakan, diantaranya: Iklan/Promo, Hadiah Lomba, Dana Tunai, Judi, Sex, Penipuan, dan lainnya. Data yang diterima dari kuisisioner akan di simpan dalam bentuk file *.txt*.

Setelah didapatkan hasil kuisisioner, maka akan dilakukan pembersihan data dengan cara menghapus data yang tidak layak secara manual, pembersihan ini dilakukan untuk menghindari adanya duplikasi data. Adapun diperoleh jumlah keseluruhan data yang layak adalah sebanyak 690 SMS. Rincian data yang diperoleh dari hasil kuisisioner berdasarkan tipe sms dapat dilihat pada Tabel 3.1.

Tabel 3.1. Rincian Tipe SMS Spam

No	Tipe SMS Spam	Jumlah
1	Iklan/Promo	115
2	Hadiah Lomba	115
3	Dana Tunai	115
4	Judi	115
5	Sex	115
6	Penipuan	115
Jumlah Keseluruhan		690

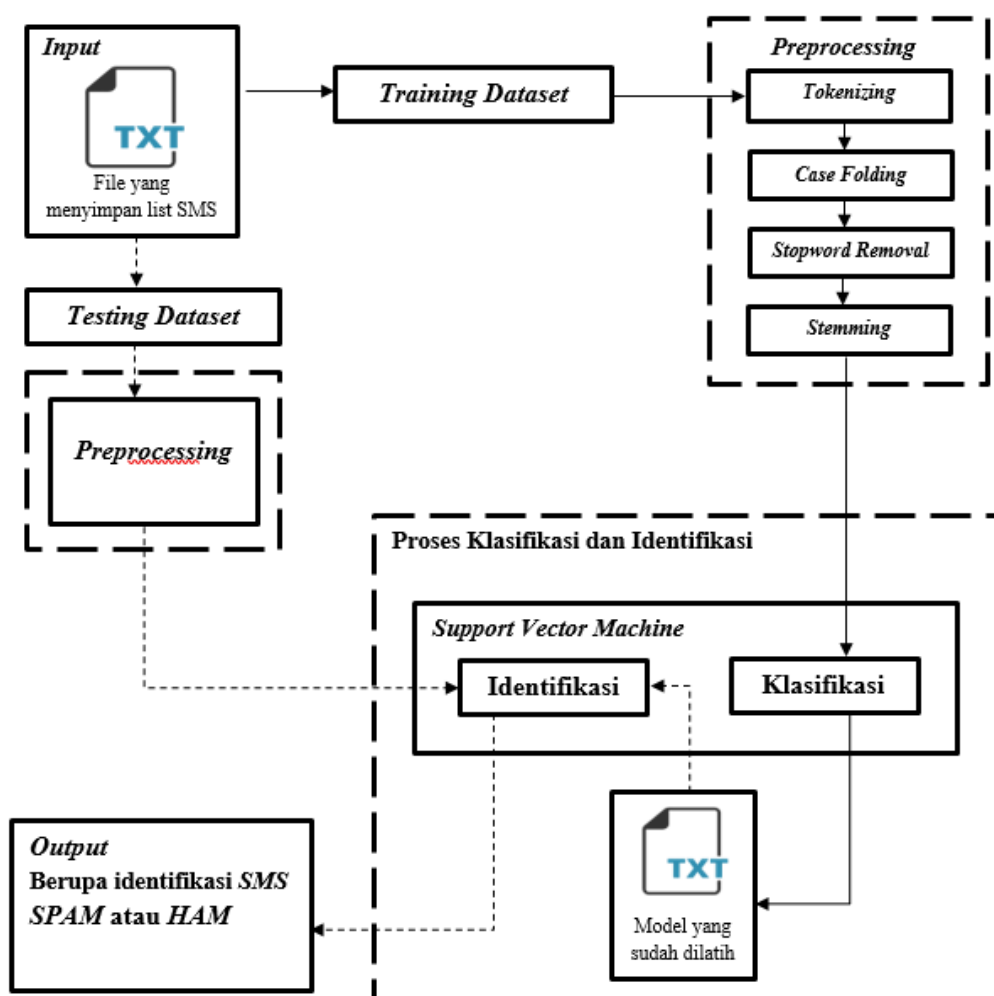
Data yang digunakan pada penelitian ini tidak hanya SMS *spam* tetapi juga dibutuhkan SMS *ham*. SMS *ham* ialah SMS normal atau SMS yang sah yang diinginkan dan tidak mengandung unsur *spam*. Data pembanding dari SMS *spam* yang akan digunakan ialah SMS *ham* yang diperoleh berbagai sumber berjumlah 460 teks sms, yang akan diuji apakah sistem identifikasi SMS *spam* dapat membedakan antara SMS *spam* atau *ham*. Data SMS *spam* dan *ham* selanjutnya digabung untuk menjadi dataset penelitian. Data yang digunakan pada penelitian ini berjumlah 1150 yang terbagi menjadi dua bagian yaitu bagian data latih dalam proses pelatihan model dan bagian data uji dalam proses pengujian untuk menguji model yang sudah dilatih sebelum pada proses pelatihan. Pembagian data dalam penelitian ini dapat dilihat pada Tabel 3.2.

Tabel 3.2. Pembagian Data Penelitian

Dataset	Spam	Ham	Jumlah
Data Pelatihan	621	414	1035
Data Pengujian	69	46	115
Jumlah Data Keseluruhan	690	460	1150

3.2. Analisis Sistem

Metode untuk mengidentifikasi SMS *spam* berbahasa Indonesia pada penelitian ini terdiri dari tahapan-tahapan utama yaitu: *input* penelitian ialah file dataset berekstensi *txt* yang berisi SMS *spam* dan *ham* yang masing-masing akan dipisahkan menjadi dataset pelatihan dan dataset pengujian. Kemudian dataset pelatihan (*training dataset*) dan dataset pengujian (*testing dataset*) akan melalui tahapan *preprocessing* sebagai berikut; (*tokenizing*), (*case folding*), (*stopword removal*), (*stemming*), dan klasifikasi menggunakan *Support Vector Machine*. Setelah tahapan tersebut dilakukan, maka data *testing* akan diidentifikasi menggunakan metode *Support Vector Machine*, dengan *output* berupa hasil dari teks sms yang berhasil dikenali. Untuk lebih jelasnya, metode penelitian ini dapat dilihat pada arsitektur umum penelitian yang ditunjukkan pada gambar 3.1.



Gambar 3.1. Arsitektur Umum

3.2.1. Input

Input pada penelitian ini adalah teks SMS dengan bahasa Indonesia. *Input* terdiri dari dua bagian yaitu data latih dan data uji. Data latih adalah data (SMS) yang telah didefinisikan sebagai SMS *spam*. Data uji adalah data (SMS) yang masuk ke perangkat pengguna yang akan diuji untuk membuktikan data (SMS) bernilai *spam* atau *ham* (pesan yang sah). Adapun contoh untuk teks SMS yang akan di proses dapat dilihat pada Tabel 3.3.

Tabel 3.3 Contoh teks SMS Berbahasa Indonesia

Input
HANYA BAYAR 6.600! Karaoke 1 Jam. Tukarkan segera sms ini di Happy Puppy / Happup Jl. Teuku Umar. Berlaku setiap hari. SKB. Promo *606#

3.2.2. Preprocessing

Teks yang telah dideteksi akan masuk ke pemrosesan selanjutnya, yaitu:

1) *Tokenizing*

Proses tokenizing berfungsi melakukan pemotongan input teks menjadi bagian terkecil untuk setiap kata berupa rangkaian angka dan rangkaian angka dengan huruf yang memiliki makna tertentu, yang terdapat dalam SMS. Karakter selain huruf, rangkaian angka, atau rangkaian angka dengan huruf akan dihilangkan. Setiap kata, rangkaian angka, maupun rangkaian angka dengan huruf disebut sebagai *token*. Berdasarkan hasil dari teks SMS pada tabel 3.3, maka hasil dari tahapan *tokenizing* dapat dilihat pada Tabel 3.4.

Tabel 3.4 Tahapan *Tokenizing*

HANYA	BAYAR	Karaoke	Jam
Tukarkan	Segera	Sms	Ini
Di	Happy	Puppy	Happup
Jl	Teuku	Umar	Berlaku
Setiap	Hari	SKB	Promo

2) *Case Folding*

Proses *Case Folding* merupakan tahapan mengubah setiap huruf pada kata yang akan menjadi huruf kecil atau huruf besar sehingga jenis huruf yang akan diproses menjadi seragam dan dapat mempermudah tahapan selanjutnya. Pada penelitian ini penulis mengubahnya menjadi huruf kecil. Hasil dari tahapan *case folding* dari *tokenizing* teks SMS pada Tabel 3.4 dapat dilihat pada Tabel 3.5.

Tabel 3.5 Tahapan *Case Folding*

hanya	bayar	karaoke	jam
tukarkan	segera	sms	ini
di	happy	puppy	happup
jl	teuku	umar	berlaku
setiap	hari	skb	promo

3) *Stopword Removal*

Proses *Stopwords Removal* merupakan proses penghapusan kata yang termasuk di dalam daftar *stopwords* yang ada didalam file *.txt* yang dianggap tidak berpengaruh dalam kalimat. Beberapa kata yang termasuk dalam daftar *stopwords* adalah yang, di, ke, dari, adalah, dan, atau, dan lain sebagainya. Hasil dari tahap ini dapat dilihat pada Tabel 3.6.

Tabel 3.6 Tahapan *Stopword Removal*

bayar	karaoke	jam	tukarkan
sms	happy	puppy	happup
jl	teuku	umar	berlaku
skb	promo		

4) *Stemming*

Stemming merupakan proses pengubahan kata yang mengandung imbuhan menjadi kata dasar dengan menghapus awalan dan akhiran yang terdapat pada suatu kata, tujuannya untuk melakukan pengelompokan kata-kata yang diturunkan dari sebuah data *stem* yang umum dan kata

dasar. Hasil dari tahap ini dapat dilihat pada Tabel 3.7.

Tabel 3.7 Tahapan *Stemming*

bayar	karaoke	jam	tukar
sms	happy	puppy	happup
jl	teuku	umar	laku
skb	promo		

3.2.3. Implementasi *Support Vector Machine*

Dataset Pelatihan yang telah melalui tahapan preprocessing akan dijadikan *input* pada proses pelatihan dengan menerapkan algoritma *Support Vector Machine*. Klasifikasi teks dilakukan untuk mengklasifikasikan data ke dalam kelas yang telah ditentukan sebelumnya, yaitu data SMS *spam* sebagai data normal dan *ham* sebagai data tidak normal. Langkah pertama dalam klasifikasi teks adalah mengubah data teks SMS menjadi format yang sesuai untuk algoritma pembelajaran *Support Vector Machine*, kemudian akan dilakukan pemodelan terhadap dataset SMS tersebut.

Setelah tahapan pemodelan dilakukan, berarti model hasil pelatihan sudah selesai dibuat. Model yang dihasilkan pada penelitian ini pada dasarnya menyimpan nilai-nilai distribusi dari rumus *Support Vector Machine* yang digunakan untuk tahapan selanjutnya, yaitu identifikasi teks pada data Pengujian. Dimana dataset uji akan diuji kecocokan pola yang dihasilkan model yang telah dilatih sebelumnya. Adapun pseudo code SVM pada gambar 3.2.

```

Langkah 1 Input dataset, baca data dalam bentuk file
Langkah 2 Preparing data, yaitu ubah dataset menjadi angka
Langkah 3 Inisialisasi data pelatihan ( $x_i, y_i$ ) untuk  $x_i = 1 \dots N$  dan  $y_i \in \{+1, -1\}$ 
Langkah 4 Tentukan  $w$  = parameter bobot(normal bidang),  $x$  = vektor input(label kelas) dan  $b$  = bias (posisi bidang relatif)

$$y(x) = w \cdot x + b$$

Langkah 5 Latih data menggunakan svmtrain
Langkah 6 Buat kelompok untuk pelatihan yang ditetapkan seperti itu.
svmstruct = svmtrain (set pelatihan, kelompok/kelas)
Langkah 7 Temukan support vector dengan menggunakan svmclassify, seperti svmstruct, testdata, kelompok/kelas
Langkah 8 selesai (end)

```

Gambar 3.2 Pseudo code untuk SVM

3.3. Perancangan Sistem

Perancangan sistem dibuat sebagai alat untuk menjalankan sistem yang akan dibangun dengan tujuan untuk memudahkan pengguna untuk menjalankan sistem yang nantinya akan dibangun. Rancangan ini terdiri dari beberapa bagian halaman, diantaranya: halaman Latih (untuk proses pelatihan) dan halaman Uji (untuk proses identifikasi teks menggunakan model yang sudah dilatih). Adapun untuk penjelasan rancangan desain yang akan diterapkan tersebut pada sistem adalah sebagai berikut.

3.3.1 Perancangan Tampilan Halaman Pelatihan

Pada rancangan halaman pelatihan ini pengguna dapat mengupload file yang akan dilatih dengan model yang sudah ada, kemudian disimpan untuk proses pelatihan. Rancangan tampilan Halaman Pengambilan Data Pelatihan dapat dilihat pada Gambar 3.3.

Gambar 3.3. Rancangan Tampilan Halaman Pelatihan

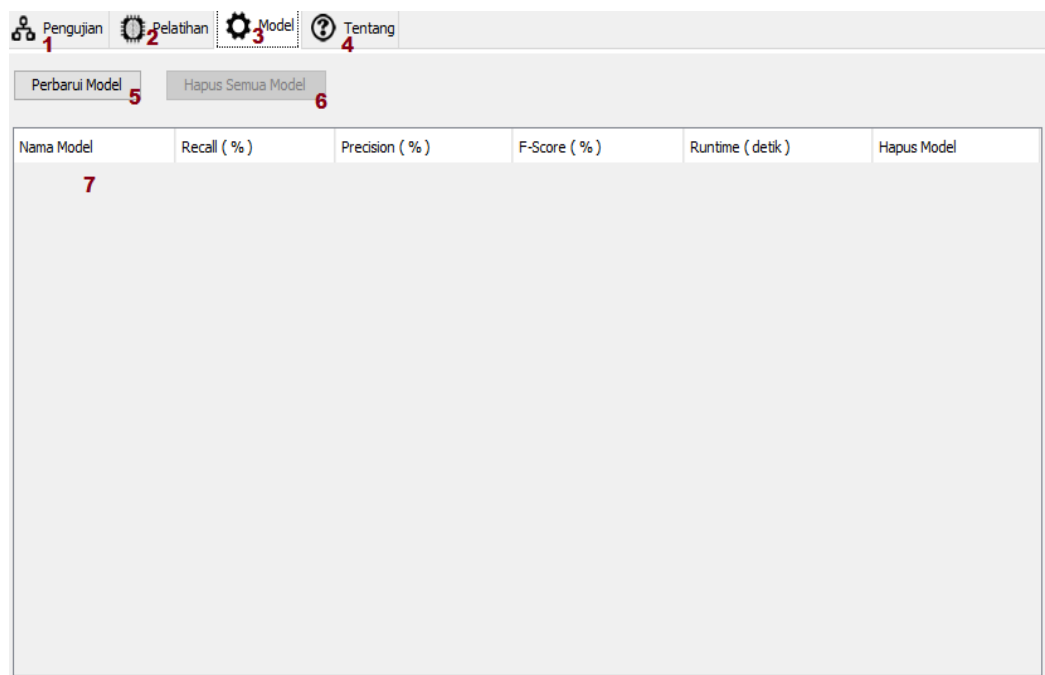
Adapun keterangan rancangan tersebut adalah sebagai berikut:

- 1) Menuju halaman Pengujian.
- 2) Halaman yang saat ini sedang aktif yaitu halaman Pelatihan.
- 3) Menuju halaman Model.

- 4) Menuju halaman Tentang.
- 5) Tombol Pilih untuk memilih file yang akan digunakan.
- 6) Tombol untuk memilih file dataset Spam.
- 7) Tombol untuk memilih file dataset Spam.
- 8) Tombol untuk menjalankan proses pelatihan dengan dataset SMS
- 9) Area untuk menampilkan kondisi dan hasil proses pelatihan dengan model yang baru digunakan atau telah ada sebelumnya.
- 10) Memilih model yang telah dilatih sebelumnya, bila pengguna ingin menggunakan model lain untuk dilatih kembali.
- 11) Tombol untuk memperbarui list model yang telah dilatih sebelumnya.

3.3.2 Perancangan Tampilan Halaman Model

Pada rancangan halaman model ini pengguna dapat melihat nilai *F-score* model dari proses Pelatihan dan dapat memilih model yang sudah dilatih sebelumnya untuk dihapus. Rancangan tampilan Halaman Pengujian dapat dilihat pada Gambar 3.4.



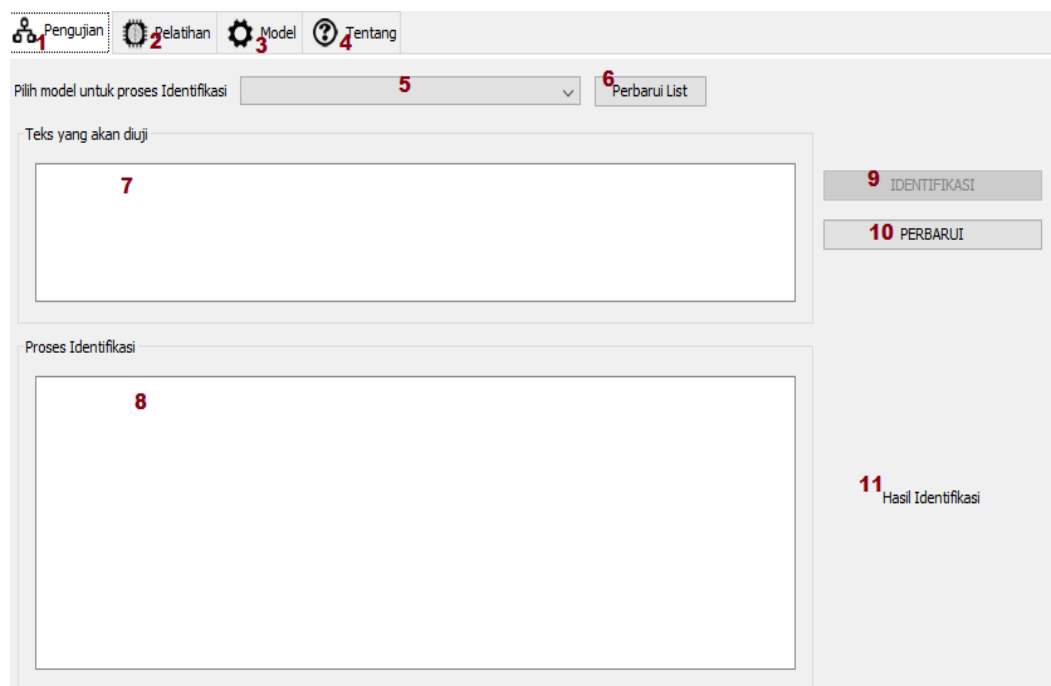
Gambar 3.4. Rancangan Tampilan Halaman Model

Adapun keterangan rancangan tersebut adalah sebagai berikut:

- 1) Menuju halaman Pengujian.
- 2) Menuju halaman Pelatihan.
- 3) Halaman yang saat ini sedang aktif yaitu halaman Model.
- 4) Menuju halaman Tentang.
- 5) Pilihan yang berisi nama model yang sudah dilatih sebelumnya.
- 6) Tombol untuk menghapus model yang tidak diinginkan.
- 7) Tombol untuk menghapus semua model yang telah dilatih dan dievaluasi.
- 8) Area untuk menampilkan hasil dari proses Pelatihan, yang berisi nama model, nilai *recall*, *precision*, *f-score* dari model tersebut, waktu yang digunakan pada saat Pelatihan dan kolom hapus model.

3.3.3 Perancangan Tampilan Halaman Pengujian

Pada rancangan halaman pengujian ini pengguna dapat memilih model yang sudah dilatih sebelumnya, kemudian akan diuji dengan teks SMS yang akan diidentifikasi. Rancangan tampilan Halaman Pengujian dapat dilihat pada Gambar 3.5.



Gambar 3.5. Rancangan Tampilan Halaman Pengujian

Adapun keterangan rancangan tersebut adalah sebagai berikut:

- 1) Halaman yang saat ini sedang aktif yaitu halaman Pengujian.
- 2) Menuju halaman Pelatihan.
- 3) Menuju halaman Model.
- 4) Menuju halaman Tentang.
- 5) Tombol Pilih untuk memilih model yang akan digunakan.
- 6) Tombol untuk memperbarui list model identifikasi.
- 7) Area untuk menuliskan teks SMS yang akan diuji dengan model yang telah dilatih sebelumnya.
- 8) Area untuk menampilkan kondisi dan hasil proses pengujian dengan model yang digunakan.
- 9) Tombol untuk menjalankan proses pengujian dengan teks SMS.
- 10) Tombol untuk memperbarui halaman Pengujian kembali awal.
- 11) Tampilan untuk hasil dari proses pengujian yang telah dilakukan.

3.4. Metode Evaluasi

Tahapan evaluasi berfungsi untuk melakukan perhitungan seberapa baik sistem dalam mengidentifikasi SMS *spam* dan *ham* yang diujikan dalam sistem. Tahapan ini menerapkan aturan variabel pada tabel 2.4. Sebagai contoh untuk tahapan evaluasi, digunakan data sebagai berikut pada tabel 3.8 .

Tabel 3.8 Contoh perhitungan menggunakan *F-score*

No.	Variabel	Jumlah
1	<i>True Positive</i> (TP)	68
2	<i>False Negative</i> (FN)	1
3	<i>False Positive</i> (FP)	1
4	<i>True Negative</i> (TN)	45
	Total	115

Dari Tabel 3.8 dapat dihitung nilai *recall*, *precesion*, dan *F-score* yang secara berurutan menggunakan 2. 11, 2.12 dan 2.13 sebagai berikut :

- **Recall** = $\frac{TP}{TP+FN} = \frac{68}{68+1} = \frac{68}{69} = 0,98$
- **Precision** = $\frac{TP}{TP+FP} = \frac{68}{68+1} = \frac{68}{69} = 0,98$
- **F – score** = $\frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} = \frac{2 \times (0,98 \times 0,98)}{0,98+0,98} = \frac{1,92}{1,96} = 0,979$

BAB 4

IMPLEMENTASI DAN PENGUJIAN SISTEM

Bab ini akan membahas tentang hasil yang didapat dari implementasi metode *Support Vector Machine* untuk mengidentifikasi SMS *Spam* berbahasa Indonesia, sesuai dengan analisis dan perancangan sistem yang telah dibahas pada Bab 3.

4.1. Implementasi Sistem

Pada tahapan ini metode *Support Vector Machine* (SVM) akan diimplementasikan ke dalam system menggunakan bahasa pemrograman Java sesuai dengan perancangan yang dilakukan.

4.1.1. Spesifikasi Perangkat Keras dan Perangkat Lunak

Spesifikasi perangkat keras yang digunakan dalam pembangunan sistem ini adalah:

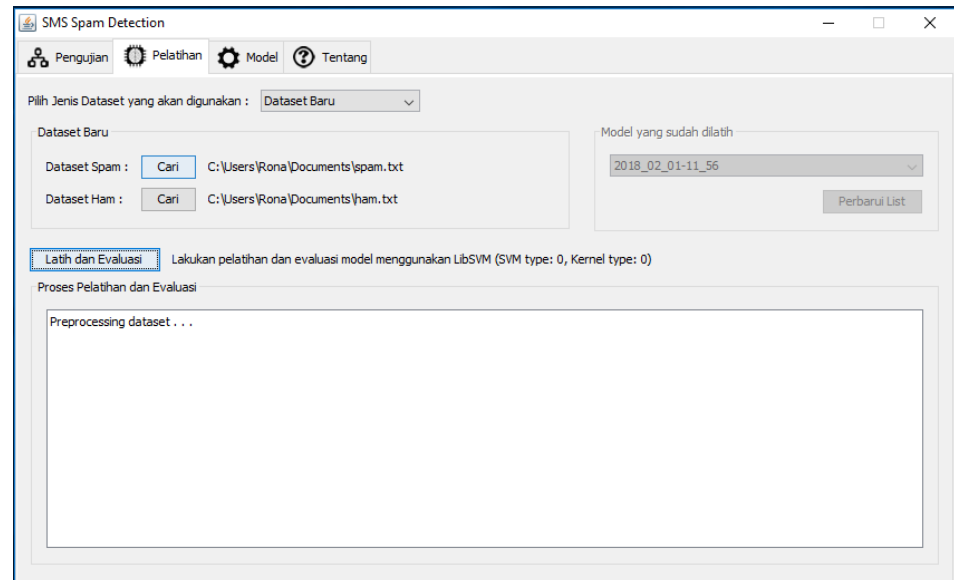
1. Laptop *ASUS A450L*
2. Sistem Operasi *Windows 10 Pro 64 Bit*
3. Prosesor *Intel(R) Core(TM) i5-4200 CPU @ 1.60GHz 2.30GHz*
4. Kapasitas *hard disk 500GB HDD*
5. *Memory RAM 4GB*
6. *Netbeans IDE 8.2*
7. *Java SE Development Kit 8*

4.1.2. Implementasi Perancangan Antarmuka

Implementasi perancangan antarmuka yang telah dilakukan pada Bab 3 adalah sebagai berikut:

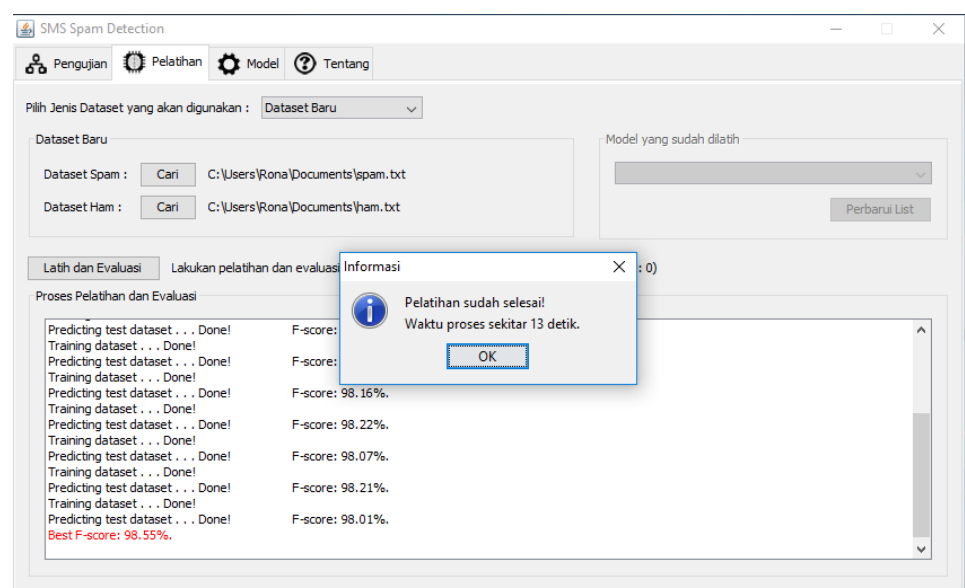
1. Tampilan Halaman Pelatihan

Halaman pelatihan merupakan halaman untuk mengupload file yang akan dilatih dengan model yang sudah ada, kemudian disimpan untuk proses pelatihan. Halaman Pelatihan dapat dilihat pada Gambar 4.1.



Gambar 4.1. Tampilan Halaman Pelatihan

Tampilan ini ialah tampilan awal dari *tab* Pelatihan. Tampilan ini akan menampilkan proses yang sedang berjalan pada area informasi dimana menampilkan kondisi yang diperoleh akurasi dari proses yang dilakukan serta ketika proses sudah mencapai akhir. Dapat dilihat pada gambar 4.2.

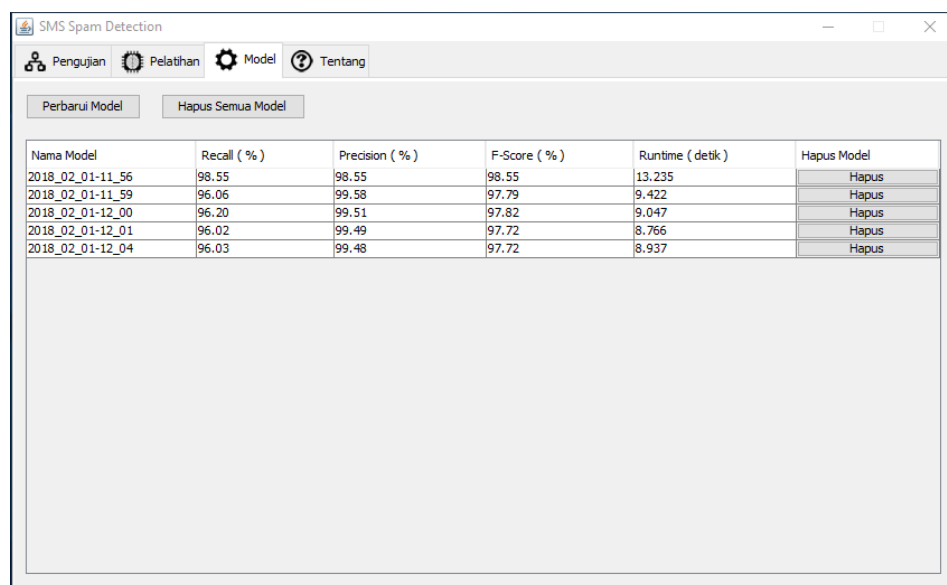


Gambar 4.2. Tampilan Halaman Proses Pelatihan

Ketika proses pelatihan telah mencapai akhir, maka akan ditampilkan informasi detail model yang diperoleh pada area informasi dan sebuah pemberitahuan bahwa proses telah selesai.

2. *Tampilan Halaman Model*

Tampilan ini memuat model hasil dari proses pelatihan yang dilakukan sebelumnya dan disimpan di dalam system. Tampilan ini dapat dilihat pada Gambar 4.3.



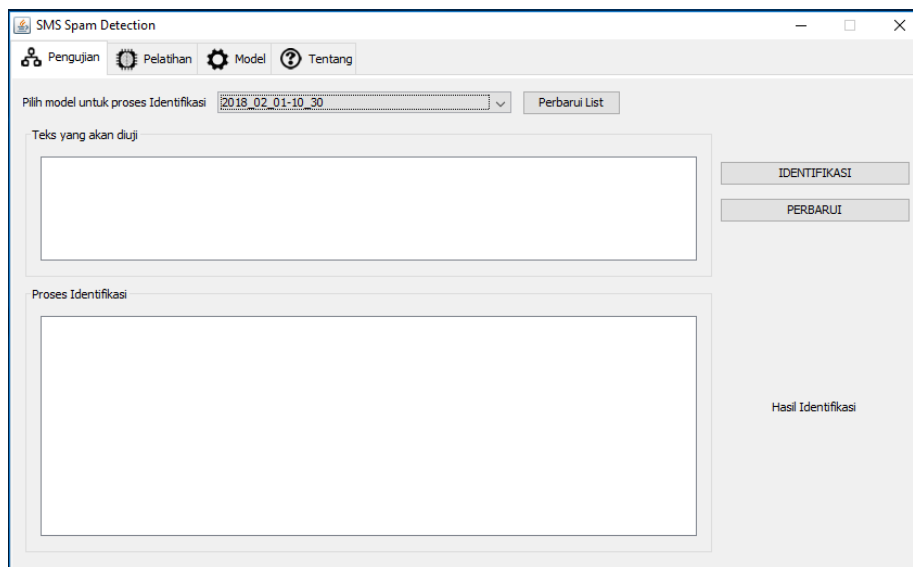
Nama Model	Recall (%)	Precision (%)	F-Score (%)	Runtime (detik)	Hapus Model
2018_02_01-11_56	98.55	98.55	98.55	13.235	Hapus
2018_02_01-11_59	96.06	99.58	97.79	9.422	Hapus
2018_02_01-12_00	96.20	99.51	97.82	9.047	Hapus
2018_02_01-12_01	96.02	99.49	97.72	8.766	Hapus
2018_02_01-12_04	96.03	99.48	97.72	8.937	Hapus

Gambar 4.3 Tampilan Halaman Model

Pada tampilan ini memaparkan detail dari model-model yang telah dilatih sebelumnya dan diberi akses untuk menghapus model yang diinginkan dengan menekan tombol pada kolom Hapus Model atau dengan menghapus semua model yang telah dilatih sebelumnya dengan menekan tombol Hapus Semua Model.

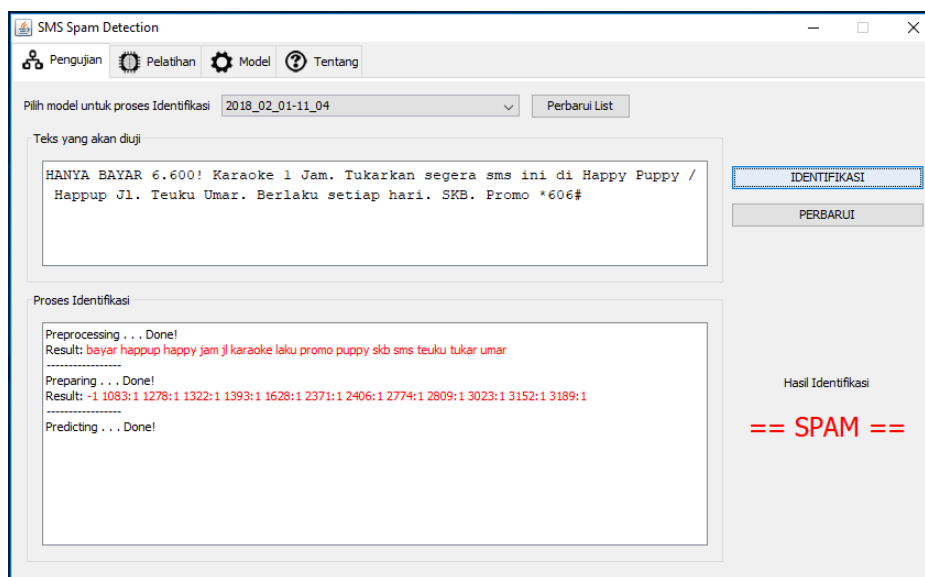
3. *Tampilan Halaman Pengujian*

Halaman pengujian merupakan tampilan yang dimana dapat memilih model yang sudah dilatih sebelumnya, kemudian akan diuji dengan teks SMS yang akan diidentifikasi. Rancangan tampilan Halaman Pengujian dapat dilihat pada Gambar 4.4.



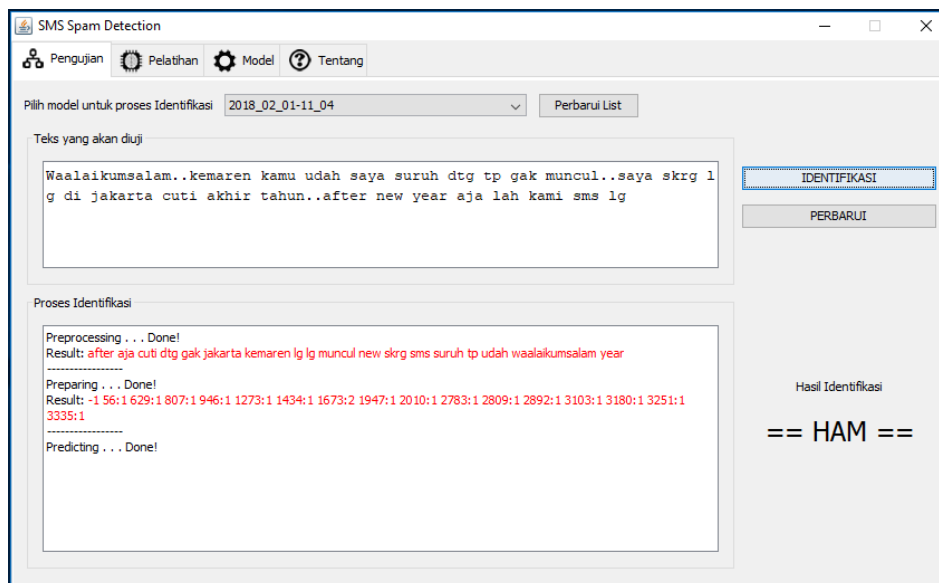
Gambar 4.4. Tampilan Halaman Pengujian

Sebagai contoh proses testing pada halaman pengujian, dapat dilihat pada Gambar 4.4 dan 4.5 untuk hasil dari identifikasi sebagai SMS *spam* dan *ham*.



Gambar 4.5. Tampilan Hasil Identifikasi sebagai SMS *Spam*

Untuk memudahkan pengguna dalam mengetahui hasil identifikasi teks yang mereka masukkan, pada Gambar 4.4 dapat dilihat bahwa ketika diuji dan diidentifikasi sebagai SMS *spam*, maka akan tampil teks “SPAM” berwarna merah. Sedangkan jika teks yang diuji diidentifikasi sebagai SMS *ham*, maka akan tampil teks “SPAM” berwarna hitam, yang dapat dilihat pada gambar 4.5.



Gambar 4.6. Tampilan Hasil Identifikasi sebagai SMS *Ham*

4.2. Pengujian Sistem

Untuk melakukan pemeriksaan terhadap sistem yang sudah dibangun, maka dilakukan pengujian sistem yang terdiri proses pelatihan dan pengujian model sesuai dengan metode pengujian yang sudah dijelaskan sebelumnya pada bagian 3.3.

4.2.1 Pelatihan Model

Pada pelatihan model adalah proses yang diterapkan untuk mencari model yang terbaik. Sebelum memasuki proses pelatihan model, data akan dijalankan pada *preproccesing* seperti yang sudah dijelaskan pada bab 3. Dimana tahapan yang dilakukan adalah *tokenizing*, *case folding*, *stopword removal*, dan *stemming*. Pelatihan dengan jumlah model tertentu dilakukan dengan validasi dengan metode *k-fold cross validation*, dengan $k=10$, dimana dataset SMS *spam* akan dibagi menjadi 10 bagian secara acak, 9/10 dari dataset SMS *spam* akan dipilih untuk menjadi dataset pelatihan, begitu juga dengan dataset SMS *ham* akan dibagi menjadi 10 bagian secara acak, 9/10 dari dataset SMS *ham* akan dipilih untuk menjadi dataset pelatihan. Kemudian sisanya 1/10 dari dataset SMS *spam* dan *ham* akan digunakan sebagai dataset pengujian. Adapun beberapa data Latih yang digunakan dalam Pelatihan, yang dapat dilihat pada Tabel 4.1.

Tabel 4.1 Data Latih yang digunakan

No	Teks SMS	Jenis SMS
1	butuh dana cepat ?. jaminkan bpkb mobil/truk. rate rendah,proses cepat,finance resmi. bpkb aman. bisa take over yang masih berjalan. hub : chandra 082217285115	Spam
2	dan@ exprezz.... jminan bpkbm0bil thn 2003 keatas... . prses 2hr cair...bs takeover . hub.yeni 0813 6764 6129	Spam
3	butuh tambahan modal usaha? kami solusinya!!! fundsupermart group bergerak dibidang penanaman modal usaha.info lebih lanjut hubungi marketing: 082343604357.	Spam
4	berkah"isi ulang. pulsa m-kios. no,anda t"pilih. sbgai pemenang. men-dpt hadiah. rp.77jt.dri. pt.m-kios. pin;5n939f7. info klik;. www.berkah-mkios.com	Spam
5	maaf kami sudah menghubungi tapi tidak tersambung anda meraih hadiah ke-iv dari pt.m-kios pin:hdr7jn7 cek pin anda di www.m-kios53.blogspot.com	Spam
6	nasabah bri yth! anda men-dptkn hadiah cek 27jt dri undian bri kode triplle cek anda (02599875) u/info klik www.daftar-pemenang2017.blogspot.co.id	Spam
7	no undian kejutan sms banking telkomsel anda gfuqa10. .menangkan liburan ke hongkong,spd motor,logam mulia&lainnya! skb.info tsel.me/kejutansmsbanking	Spam
8	promo terbaik www.rgocasino10.com bonus deposit 10% pendaftaran member baru & bonus reload hingga 2,5 juta!! situs live casino online no 1 di indonesia	Spam
9	poker online terpercaya! double jackpot ratusan juta, dp 25rb + mega referral sebesar 12%, mainkan pada pc dan juga android anda hanya di www.rgopoker2.com	Spam
10	aq akan memberikan yg gk akan pernah kmu lupain, telp aja di 08091401077 djamin gk bkalan nyesel dech .. buruan (no sex & sara)	Spam
11	Mak, isi pulsa im# 10 ewu neng Nomor. 085756595570 iki saiki...	Spam
12	assalamua'alaikum bg.. bg kalo jumpa jam 12 aja bisa? kegiatan yg d poltekkes selesai jam 11 irun	Ham

Tabel 4.1 Data Latih yang digunakan (Lanjutan)

No	Teks SMS	Jenis SMS
13	Mbak RONA saya mau infoin PROMO BELANJA hari ini di oriflame. Tiap belanja 399rb harga katalog BERHADIAH Lipstik seharga 129rb (desi)	Ham
14	assalamu'alaikum bang, selesai liqo' ke btbs kan? soalnya alita mau izin pulkam heehe	Ham
15	klo di daerah sini blum coba tanya2 sih. Oiya coba tanya si abdur, soalnya dkosnya itu ada kawannya yg udah wisuda, tpi gatau apakah di Injut ato gak lagi.	Ham
16	assalamualaikum pak, pak masih bisa mesan mie? indomaret sumber pak, indomaret setelah amik . kalo bisa kami mau pesan 5 porsi pak, mie aceh basah 2, ifumie 2, nasgor 1	Ham
17	Qan bsk cancel ya saya jatuh pula jadi agak terkilir susah jalan..Minggu ku hubungi lg ya	Ham
18	udah ada tanya pak dhani blum? coba tanya ran siang ini bisa gak kira2 kita bimbingan.	Ham
19	kunci rumah titipkan ke kakek ya, aku balek, Cuma gak tau jam berapa	Ham
20	tlg kirim aplikasi untuk menginstal printer canon pixma ip 2770	Ham
21	iya bg..nanti kalo udah d simpang pos irun sms	Ham
22	pas testing itu yang diproses data testingnya, trus hasil proses data testingnya itu dibandingin sama hasil proses trainingnya	Ham

4.2.2 Pengujian Model

Setelah melakukan pelatihan terhadap dataset dan menghasilkan model, maka akan dilakukan pengujian model yang dihasilkan melalui proses pengujian terhadap data uji (*testing data*) dengan menerapkan persamaan yang telah dibahas pada bab 3 yaitu menggunakan persamaan 3.7 implementasi *Support Vector Machine* untuk klasifikasi teks. Untuk melakukan proses pengujian digunakan beberapa contoh teks yang dapat dilihat pada Tabel 4.2.

Tabel 4.2 Data hasil Pengujian

No	Teks SMS	Jenis SMS	Hasil Uji
1	saya sepakat dengan penawaran kita kemarin. mohon hubungi saya di no. 081356890065	Spam	Spam
2	dan@ exprezz.... jminan bpkbm0bil thn 2003 keatas... . prses 2hr cair...bs takeover . hub.yeni 0813 6764 6129	Spam	Spam
3	butuh tambahan modal usaha? kami solusinya!!! fundsupermart group bergerak dibidang penanaman modal usaha.info lebih lanjut hubungi marketing: 082343604357.	Spam	Spam
4	berkah"isi ulang. pulsa m-kios. no,anda t"pilih. sbgai pemenang. men-dpt hadiah. rp.77jt.dri. pt.m-kios. pin;5n939f7. info klik;. www.berkah-mkios.com	Spam	Spam
5	mengenai Rumah bpk/ibu yg mau di jual sd di surpai dan sy berminat sy tlpn tdk bisa tembus masalah harga hub suami sy 085399700354 Dr.H.IRAWAN	Spam	Spam
6	nasabah bri yth! anda men-dptkn hadiah cek 27jt dri undian bri kode triplle cek anda (02599875) u/info klik www.daftar-pemenang2017.blogspot.co.id	Spam	Spam
7	no undian kejutan sms banking telkomsel anda gfuqa10. .menangkan liburan ke hongkong,spd motor,logam mulia&lainnya! skb.info tsel.me/kejutansmsbanking	Spam	Spam
8	INFO TOGEL... Ingin sio & angka jitu nomor GHOIB di jamin 100% tembus (jika anda berminat hub:ki-darjo 082326111133 untuk singapura togel)	Spam	Spam
9	poker online terpercaya! double jackpot ratusan juta, dp 25rb + mega referral sebesar 12%, mainkan pada pc dan juga android anda hanya di www.rgopoker2.com	Spam	Spam
10	aq akan memberikan yg gk akan pernah kmu lupain, telp aja di 08091401077 djamin gk bkalan nyesel dech .. buruan (no sex & sara)	Spam	Spam
11	Mak, isi pulsa im# 10 ewu neng Nomor. 085756595570 iki saiki...	Spam	Spam

Tabel 4.2 Data hasil Pengujian (Lanjutan)

No	Teks SMS	Jenis SMS	Hasil Uji
12	assalamua'alaikum bg.. bg kalo jumpa jam 12 aja bisa? kegiatan yg d poltekkes selesai jam 11 irun	Ham	Ham
13	Mbak RONA saya mau infoin PROMO BELANJA hari ini di oriflame. Tiap belanja 399rb harga katalog BERHADIAH Lipstik seharga 129rb (desi)	Ham	Ham
14	Maaf bg, boleh minta no WhatsApp abg? Coz sy hp baru, jadi no abg ga ada d hp baru ini	Ham	Ham
15	klo di daerah sini blum coba tanya2 sih. Oiya coba tanya si abdur, soalnya dkosnya itu ada kawannya yg udah wisuda	Ham	Ham
16	Blum kutanya Tp dah dpt tempatnya Rencana abis hari raya kutanya Klo misalnya dia blg paling dikit 5 buah atau lebih kita tawarin ke yang lain ya	Ham	Ham
17	Qan bsk cancel ya saya jatuh pula jadi agak terkilir susah jalan..Minggu ku hubungi lg ya	Ham	Ham
18	udah ada tanya pak dhani blum? coba tanya ran siang ini bisa gak kira2 kita bimbingan.	Ham	Ham
19	kunci rumah titipkan ke kakek ya, aku balek, Cuma gak tau jam berapa	Ham	Ham
17	Qan bsk cancel ya saya jatuh pula jadi agak terkilir susah jalan..Minggu ku hubungi lg ya	Ham	Ham
18	udah ada tanya pak dhani blum? coba tanya ran siang ini bisa gak kira2 kita bimbingan.	Ham	Ham
19	kunci rumah titipkan ke kakek ya, aku balek, Cuma gak tau jam berapa	Ham	Ham
20	tlg kirim aplikasi untuk menginstal printer canon pixma ip 2770	Ham	Ham
21	iya bg..nanti kalo udah d simpang pos irun sms	Ham	Ham
22	Kami mau tanya nomor Hp tulang pernando atau No. hp tulang Jopan. Kami dpt nomor ini dr Ito Aminton Sinaga. Sy anak opung simanjuntak br simbolon di samarinda	Ham	Spam

4.3. Hasil Pengujian Sistem

Hasil pengujian sistem dilakukan dengan 5 kali pelatihan model dengan dataset yang sama dan acak untuk mencari rata-rata dari nilai *F-score* dari pelatihan dengan menggunakan dataset yang digunakan dalam penelitian. Selain untuk mencari rata-rata 5 pelatihan untuk mencari model yang mempunyai nilai *F-score* paling tinggi, dapat dilihat pada tabel 4.3.

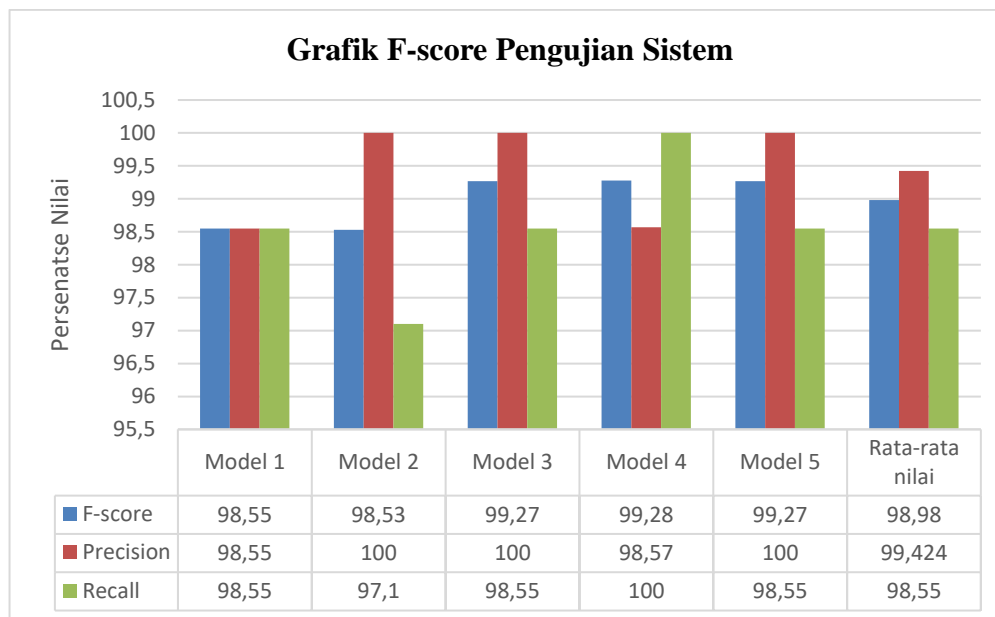
Tabel 4.3 Hasil Pengujian Sistem pada model Pelatihan Sistem (100%)

Hasil Model	TP	FN	FP	TN	Precision	Recall
Model 1	68	1	1	45	98.55	98.55
Model 2	67	2	0	46	100	97.10
Model 3	68	1	0	46	100	98.55
Model 4	69	0	1	45	98.57	100
Model 5	68	1	0	46	100	98.55

Pada tabel 4.3 dipaparkan nilai *precision* dan *recall* pada setiap model. Berdasarkan metode evaluasi yang telah dijelaskan pada bab 3 bagian 3.4 sebelumnya, untuk mencari nilai *F-score* dapat menggunakan nilai *precision* dan *recall* yang telah didapat pada tabel 4.3. Adapun nilai *F-score* pada masing-masing model, dapat dilihat pada tabel 4.4 dan Gambar 4.6.

Tabel 4.4 Rata-rata nilai *F-score* Pengujian Sistem (100%)

Model Pelatihan	F-score
Model 1	98.55
Model 2	98.53
Model 3	99.27
Model 4	99.28
Model 5	99.27
Rata-rata nilai <i>F-score</i>	98.98



Gambar 4.6 Grafik F-Score Pengujian

Pada Tabel 4.4 dapat dilihat bahwa secara umum model memiliki rata-rata nilai *F-score* yaitu 98,98%, dengan nilai yang paling rendah, yaitu 98,53%, sedangkan nilai *F-score* paling tinggi, yaitu 99,28%. Hal ini disebabkan oleh susunan data pelatihan menggunakan metode *cross-validation* yang dilakukan secara acak sebelum melakukan pelatihan dan perulangan pada proses model yang dilatih. Setelah diperoleh model dengan nilai *F-score* tertinggi dalam hal ini Model 4, maka dilakukan pengujian model tersebut untuk beberapa sampel data SMS *spam* dan *ham* sebagai berikut:

1. Teks : Blum kutanya Tp dah dpt tempatnya Rencana abis hari raya kutanya Klo misalnya dia blg paling dikit 5 buah atau lebih kita tawarin ke yang lain ya
 Label Manual : *Ham*
 Hasil Identifikasi: *Ham*
2. Teks : kunci rumah titipkan ke kakek ya, aku balek, Cuma gak tau jam berapa
 Label Manual : *Ham*
 Hasil Identifikasi: *Ham*
3. Teks : Kami mau tanya nomor Hp tulang pernando atau No. hp tulang Jopan. Kami dpt nomor ini dr Ito Aminton Sinaga. Sy anak opung

simanjuntak br simbolon di samarinda

Label Manual : *Ham*

Hasil Identifikasi: *Spam*

4. Teks : INFO TOGEL... Ingin sio & angka jitu nomor GHOIB di jamin 100% tembus (jika anda berminat hub:ki-darjo 082326111133 untuk singapura togel)

Label Manual : *Spam*

Hasil Identifikasi: *Spam*

5. Teks : mengenai Rumah bpk/ibu yg mau di jual sd di surpai dan sy berminat sy tlpn tdk bisa tembus masalah harga hub suami sy 085399700354 Dr.H.IRAWAN

Label Manual : *Spam*

Hasil Identifikasi: *Spam*

Dari proses identifikasi SMS sampel diatas dapat dilihat dari 5 sampel SMS terdapat kesalahan dalam mengidentifikasi, yaitu pada contoh sampel ke-3 dapat dilihat bahwa label manual SMS tersebut ialah *ham*, ketika diidentifikasi pada sistem hasilnya SMS *spam*. Ini dikarenakan pada proses Pengujian menggunakan model 4, SMS *ham* yang digunakan sebagai data uji memiliki pola yang sama dengan SMS *spam* yang digunakan pada model 4. Secara umum bahwa identifikasi untuk satu SMS dilakukan dengan minimal kesalahan dalam mengidentifikasi. Hal tersebut menunjukkan bahwa sistem identifikasi telah layak untuk diimplementasikan pada perangkat seluler.

BAB 5

KESIMPULAN DAN SARAN

Bab ini akan membahas tentang kesimpulan dari implementasi metode yang diajukan untuk identifikasi SMS *spam* berbahasa Indonesia, serta saran-saran yang dapat dilakukan pada penelitian selanjutnya.

5.1. Kesimpulan

Berdasarkan pengujian sistem menggunakan metode *Support Vector Machine* (SVM) pada identifikasi SMS *spam* berbahasa Indonesia didapatkan kesimpulan sebagai berikut:

1. Metode *Support Vector Machine* mampu melakukan identifikasi SMS *spam* berbahasa Indonesia dengan hasil pengujian menghasilkan nilai *F-score* sebesar 99,28%. Nilai akurasi yang dihasilkan saat pengujian data dipengaruhi oleh data yang unik.
2. Pada pelatihan data dengan menggunakan *Support Vector Machine* linear nilai akurasi yang dihasilkan sebesar 100% dengan jumlah data yang pada Pelatihan sebanyak 1035 data. Jumlah dataset *spam* dan *ham* yang digunakan mempengaruhi nilai akurasi pelatihan, semakin banyak data yang digunakan akan semakin hasilnya.

5.2. Saran

Saran penulis untuk pengembangan dalam penelitian selanjutnya adalah sebagai berikut:

1. Jumlah data *spam* dan data *ham* ditambah untuk menambah kemampuan sistem dalam mengidentifikasi SMS *spam* berbahasa Indonesia.

2. Implementasi dengan metode machine learning untuk dapat meningkatkan nilai *F-score* pada penelitian selanjutnya.
3. Pada penelitian ini hanya sebatas sistem yang dapat dijalankan di desktop, kedepannya diharapkan dapat diimplementasi pada sistem android.

DAFTAR PUSTAKA

- Adriani, M., Asian, J., Nazief, B., Tahaghoghi, S.M.M. & Williams, H.E. 2007. Stemming Indonesian: A Confix-Stripping Approach. *ACM Transactions on Asian Language Information Processing* 6(4): 1-33.
- Agarwal, S., Kaur, S. & Garhwal, S. 2015. SMS Spam Detection for Indian Messages. *Proceedings of 1st International Conference on Next Generation Computing Technologies (NGCT)*, pp.634-638.
- Arifin, D. D., Shaufiah & Bijaksana, M. A. 2016. Enhancing Spam Detection on Mobile Phone Short Message Service (SMS) Performance using FP-Growth and Naive Bayes Classifier. *IEEE Asia Pacific Conference on Wireless and Mobile*, pp. 80-84.
- Asian, J., Williams, H.E. & Tahaghoghi, S.M.M. 2005. Stemming Indonesian. *Proceedings of the 28th Australasian conference on Computer Science (ASC'05)* 38:307-314.
- Baeza-Yates, R.A. & Ribeiro-Neto, B. 1999. *Modern Information Retrieval*. Addison-Wesley Longman Publishing: Boston.
- Bodic G.L. 2005. *Mobile Messaging Technologies and Services SMS, EMS and MMS*. 2nd Edition. Wiley: England.
- Chang, Chih-Chung., Lin, Chih-Jen., 2011. LIBSVM: A Library for Support Vector Machines. Department of Computer Science, Nasional Taiwan University, Taipei, Taiwan.
- Christianini, N. & Shawe-Taylor, J. 2000. *An Introduction to Support Vector Machines*. UK, Cambridge: Cambridge University.
- Delany, S. J., Buckley, M. & Greene, D. 2012. SMS Spam Filtering: Methods and Data. *Proceedings of Expert Systems with Applications*, pp. 9899-9908.
- Fernandes, D., Costa, K.A.P.da., Almeida, T.A. & Papa, J.P. 2015. SMS Spam Filtering Thorough Optimum-path Forest-based Classifiers. *IEEE 14th International Conference on Machine Learning and Application*, pp. 133-137.
- Khemapatapan, C. 2010. Thai-English Spam SMS Filtering. *Proceedings of 16th Asia-Pacific Conference on Communications (APCC)*, pp. 226-230.

- Krisantus. 2007. Penerapan Teknik Support Vector Machine untuk Pendeteksian Intrusi Pada Jaringan. Skripsi. Institut Teknologi Bandung.
- Lin, Chih-Jen. 2005. Optimization, Support Vector Machine, and Learning Machine. Department of Computer Science, National Taiwan University. Taipei, Taiwan.
- Ma, J., Zhang, Y., Liu, J., Yu, K. & Wang, X.A. 2016. Intelligent Spam Filtering Using Topic Model. *2016 International Conference on Intelligent Networking and Collaborative Systems*, pp. 380-383.
- Nuruzzaman, M.T., Lee, C. & Choi D. 2011. Independent and Personal SMS Spam Filtering. *11th IEEE International Conference on Computer and Information Technology*, pp. 429-435.
- Osuna EE, Freund R, Girosi F. 1997. Support Vector Machines: Training and Applications. AI Memo 1602, Massachusetts Institute of Technology.
- Refaeilzadeh, P., Tang, L. & Liu, H. 2009. Cross-Validation. *Encyclopedia of Database Systems*, pp. 532-533.
- Shahi, T.B. & Yadav, A. 2014. Mobile SMS Spam Filtering for Nepali Text Using Naïve Bayesian and Support Vector Machine. *International Journal of Intelligence Science*, pp. 24-28.
- Taufiq, M., Abdullah, M.F.A. Kang, K. & Choi, D. 2010. A Survey of Preventing, Blocking and Filtering Short Message Services (SMS) Spam. *Proceedings of International Conference on Computer and Electrical Engineering IACSIT Vol. 1*, pp. 462-466.
- Vapnik, V. 1982. *Estimation of Dependences Based on Empirical Data*. Springer Verlag.