Filtering SMS Spam Menggunakan Metode *Artificial Immune System (AIS)* dan Algoritma *Tokenization With Vectors*

SMS Spam Filtering Using Artificial Immune System (AIS) Method and Tokenization With Vectors Algorithm

Vero Arneal Octora[1], Shaufiah, S.T., M.T.[2], Moch. Arif Bijaksana, Ir., M.Tech.[3]

[1,2,3] Prodi S1 Teknik Informatika, Fakultas Informatika, Universitas Telkom Bandung

[1]arnealvero@student.telkomuniversity.ac.id, [2]shaufiah@telkomuniversity.ac.id, [3]arifbijaksana@telkomuniversity.ac.id

Abstrak

SMS merupakan layanan penting yang terdapat pada perangkat mobile disamping layanan panggilan suara. SMS Spam merupakan masalah yang sangat serius bagi hampir semua pengguna layanan SMS. Untuk mengatasi masalah spam ini dapat digunakan teknik klasifikasi yang dapat membedakan antara SMS spam dan ham (bukan spam) ketika suatu SMS masuk ke dalam perangkat mobile pengguna. Metode yang digunakan pada penelitian ini adalah metode Artificial Immune System (AIS) yang dikombinasikan dengan algoritma Tokenization With Vectors yang berfungsi sebagai preprocessing teks sebelum teks tersebut diklasifikasikan dengan metode AIS. Dari hasil ekperimen yang telah dilakukan, didapatkan hasil bahwa untuk pengujian cross validation 5-fold memiliki akurasi sebesar 89.26% pada penerapan metode AIS yang menambahkan algoritma Tokenization With Vectors dan 89.06% untuk metode AIS yang tidak menambahkan algoritma Tokenization With Vectors. Sedangkan Untuk pengujian cross validation dengan 10-fold memiliki akurasi sebesar 81.92% untuk penerapan metode AIS yang menambahkan algoritma Tokenization With Vectors dan 81.24% untuk metode AIS yang tidak menambahkan algoritma Tokenization With Vectors. Penggunaan algoritma Tokenization With Vectors memiliki rata-rata akurasi yang lebih tinggi daripada yang tidak menggunakan algoritma tersebut, tetapi selisih hasil rata-rata akurasi yang didapatkan tidak terlalu banyak.

Kata kunci : *Short Message Service, Artificial Immune System, filtering, spam, ham, Tokenization With Vectors*

Abstract

SMS is an important service found on mobile devices beside calling services. SMS spam is a very serious problem for almost all users of SMS services. To solve the problem of spam, filtering techniques can be used to distinguish between SMS spam and ham (not spam) when mobile device get incoming SMS. In this research will using Artificial Immune System (AIS) method and combine with Tokenization With Vectors algorithm that serves as the preprocessing of text before the text is classified by AIS method. From the experiment results, for testing with 5-fold cross validation, the result of average accuracy is 89.26% for the AIS method that adds Tokenization With Vectors algorithms and 89.06% for the AIS method that does not add Tokenization With Vectors algorithm. While for testing with 10-fold cross validation, the result of average accuracy is 81.92% for AIS method that adds Tokenization With Vectors algorithms and 81.24% for the AIS method that does not add Tokenization With Vectors algorithm. Implementing of Tokenization With Vectors algorithm have higher average accuracy than not implementing that algorithm, but the difference of both average accuracy does not significantly.

Keywords : *Short Message Service, Artificial Immune System, filtering, spam, ham, Tokenization With Vectors*

1. Pendahuluan

Short Message Service atau yang lebih dikenal dengan SMS merupakan sebuah fasilitas utama yang terdapat pada komunikasi mobile disamping fasilitas panggilan suara. Seiring dengan meningkatnya popularitas smart phone saat ini, SMS tidak hanya digunakan untuk saling berkomunikasi antara keluarga, sahabat, dan rekan kerja, melainkan dapat digunakan untuk tujuan yang lebih luas, seperti beriklan/berjualan bahkan dapat digunakan juga untuk hal yang negatif seperti penipuan. Beriklan dan melakukan penipuan melalui SMS adalah contoh dari SMS spam. SMS spam sendiri merupakan SMS yang tidak diharapkan dan tidak diinginkan yang masuk ke dalam perangkat mobile pengguna. SMS spam memiliki efek yang lebih besar kepada pengguna dibandingkan dengan email spam karena pengguna akan selalu melihat setiap SMS yang mereka terima, sehingga SMS spam dapat mempengaruhi pengguna secara langsung, terlebih lagi jika SMS spam yang mereka terima adalah SMS penipuan [6].

Untuk mengatasi masalah spam tersebut, salah satu cara yang dapat digunakan adalah dengan melakukan filtering. Pada dasarnya filtering ini akan membedakan antara SMS spam dan ham (bukan spam). Pada penelitian sebelumnya spam filtering lebih banyak diterapkan untuk pesan email dengan menggunakan beberapa algoritma klasifikasi teks seperti Naïve Bayes, Decision Tree, Support Vector Machine, dan Neural Network. Pada tugas akhir ini, penelitian tentang filtering spam akan difokuskan pada pesan SMS. SMS dan email memiliki beberapa perbedaan diantaranya ialah SMS hanya bisa mengirim dan menerima teks, sedangkan email bisa mengirim selain teks seperti gambar dan suara. SMS memiliki kapasitas teks yang terbatas jika dibandingkan dengan email. Selain itu pada SMS sering menggunakan singkatan-singkatan atau istilah-istilah sehingga dapat menyebabkan pengertian yang ambigu.

Metode yang akan digunakan dalam melakukan filtering SMS spam ini ialah Artificial Immune System (AIS). Metode ini pada dasarnya bekerja seperti system imun yang terdapat pada tubuh manusia yang bisa membedakan antara sel-sel di dalam tubuh dan sel-sel asing di luar tubuh yang berbahaya. Pada system filtering ini sel-sel di dalam tubuh dianalogikan sebagai ham dan sel-sel asing dari luar tubuh sebagai spam. Untuk mendapatkan akurasi yang lebih tinggi, metode AIS ini akan menambahkan algoritma Tokenization With Vectors yang akan diterapkan untuk dataset pada saat proses tokenisasi. Metode dan algoritma ini diyakini akan dapat menghasilkan persentase akurasi yang baik dalam memisahkan SMS spam dan juga ham.

2. Kajian Pustaka

2.1 Text Mining

Text mining merupakan variasi dari data mining yang berusaha untuk menemukan pola menarik dari sekumpulan data tekstual yang berjumlah besar [5]. Selain itu menurut situs Wikipedia menyebutkan bahwa text mining atau penambangan teks adalah proses ekstraksi pola berupa informasi dan pengetahuan yang berguna dari sejumlah besar sumber data teks, seperti dokumen Word, PDF, kutipan teks, dan sebagainya. Jenis masukan untuk penambangan teks ini disebut data tak terstruktur dan merupakan pembeda utama dengan penambangan data yang menggunakan data terstruktur atau basis data sebagai masukan. Penambangan teks dapat dianggap sebagai proses dua tahap yang diawali dengan penerapan struktur terhadap sumber data teks dan dilanjutkan dengan ekstraksi informasi dan pengetahuan yang relevan dari data teks terstruktur ini dengan menggunakan teknik dan alat yang sama dengan data mining atau penambangan data [10].

2.2 SMS Spam

Terdapat berbagai definisi tentang apa itu spam dan bagaimana membedakan antara spam dan ham. Secara garis besar spam merupakan suatu pesan yang tidak diminta atau tidak diinginkan. Sedangkan ham adalah kebalikan dari spam. Beberapa contoh dari SMS spam adalah SMS promosi, SMS ancaman, SMS penipuan, dan sebagainya. Mobile spam atau dikenal juga sebagai SMS spam adalah bagian dari spam yang mengandung pesan teks tentang iklan dan dikirim ke perangkat mobile melalui layanan SMS. Masalah SMS spam lebih serius dibandingkan dengan email spam. Hal ini dikarenakan perangkat mobile yang bersifat personal dan akan sangat terganggu jika SMS spam masuk ke dalam perangkat mobile setiap orang. Hal ini juga dikarenakan untuk setiap SMS yang masuk ke dalam perangkat mobile bersifat real-time dan akan memberikan pemberitahuan kepada pengguna secara real-time juga, sehingga pengguna akan merasa terganggun jika ada SMS spam yang masuk. Berbeda halnya dengan email spam yang bersifat tidak real-time dan setiap ada email yang masuk tidak akan mengganggu pengguna karena tidak ada pemberitahuan yang muncul secara real-time. Beberapa perbedaan antara email dan SMS terdapat pada tabel 2.1 di bawah ini:

Tabel 2.1 perbedaan antara email dan SMS

Fitur	Email	SMS
Panjang	Tidak terbatas	160 karakter Latin atau 70 karakter Arab dan China
Proses	Tidak real-time	Real-time
Isi	Teks, gambar, suara, dll	Hanya Teks

Dengan penyebaran SMS spam, beberapa operator jaringan mobile telah mengambil langkah-langkah untuk melawan spammer, dan mereka ingin mengurangi volume spam dan memuaskan para pengguna jasa layanan mereka. Beberapa cara yang dilakukan untuk mengurangi SMS spam yang ditawarkan oleh beberapa operator ialah dengan menggunakan alamat alias sebagai alamat pesan teks menggantikan nomor ponsel. Jadi, pesan yang disampaikan hanya yang dikirim ke alamat alias saja, sedangkan pesan yang dikirim ke nomor telepon akan dibuang. Solusi ini tidak praktis dan tidak berlaku karena solusi ini tidak dapat mengambil umpan balik pengguna dalam proses klasifikasi. Kekuatan komputasi perangkat mobile dan perangkat lain saat ini sedang meningkat, sehingga semakin mungkin untuk

melakukan penyaringan spam di perangkat mobile, sehingga dapat mengarah ke personalisasi dan efektivitas yang lebih baik [6]. Pada tugas akhir ini juga akan mencoba untuk mengimplementasikan salah satu teknik filtering yang dapat menyaring SMS spam yang masuk ke dalam perangkat mobile.

2.3 Artificial Immune System (AIS)

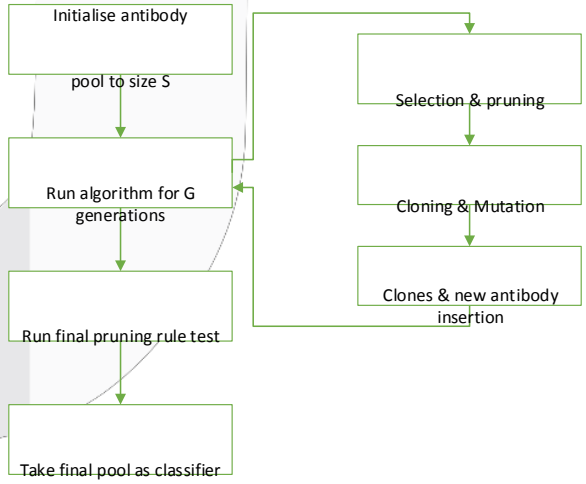
Artificial Immune System (AIS) merupakan sistem komputasi yang terinspirasi oleh teori imunologi pada Biological Immune System (BIS). Hal ini dapat diketahui dengan mengamati fungsi, prinsip, dan mekanisme kekebalan tubuh manusia yang diaplikasikan dalam pemecahan masalah. Untuk melakukan hal tersebut, sistem imun harus melakukan tugas pengenalan pola untuk membedakan molekul dan sel-sel yang terdapat di dalam tubuh dan sel-sel asing yang terdapat di luar tubuh [4].

Algoritma AIS berkaitan erat dengan cara kerja sel limfosit di dalam sistem kekebalan tubuh untuk mendeteksi masuknya benda asing atau *pathogen* ke dalam tubuh. *Pathogen* merupakan himpunan bagian dari antigen, dimana antigen ini bisa apa saja yang mengancam tubuh, seperti virus, bakteri, alergi atau molekul racun [8]. Namun, sistem kekebalan tubuh tidak bisa mendeteksi secara langsung adanya *pathogen* yang masuk ke dalam tubuh, melainkan dengan cara mendeteksi melalui bagian dari *pathogen* yang disebut dengan *antigen*. Berdasarkan mekanisme pendeteksian *pathogen* yang masuk, AIS terbagi menjadi 3 proses lagi, yaitu *clonal selection*, *negative selection*, dan *immune network theory*. Tujuan dikembangkannya algoritma ini adalah untuk memecahkan masalah di berbagai bidang seperti pengenalan pola (*pattern recognition*), dan menentukan solusi optimum dalam suatu persoalan optimasi (*optimization problem*).

Ketiga proses yang menjadi bagian dari AIS ini memiliki perbedaan dalam mekanismenya walaupun sama-sama melibatkan limfosit untuk mengenali *antigen*. *Clonal selection* merupakan teori di mana bagian limfosit yaitu sel B yang disebut antibodi bisa mengenali *antigen* asing (*nonself-antigen*). Jika sel B mampu mengikat *antigen* dengan nilai *affinity* yang besar, maka sel tersebut akan dipertahankan hidup dan diperbanyak melalui proses *mitosis*. Sedangkan sel-sel yang tidak mampu mengikat *antigen* akan mati. Berlawanan dengan *clonal selection*, proses *negative selection* membuat bagian limfosit yaitu sel T untuk beradaptasi agar bisa mengikat *antigen* yang terdapat di dalam tubuh (*self-antigen*). Mekanismenya merupakan kebalikan dari *clonal selection*, di mana sel T yang bisa mengikat *antigen* dengan nilai *affinity* yang besar akan dimusnahkan.

Dasgupta dan Forrest memperkenalkan aplikasi algoritma *negative selection* dari AIS untuk mendeteksi hal-hal baru pada *data time series*. Dasgupta dan Forrest juga memperkenalkan penggunaan AIS dalam mendeteksi kerusakan mesin. Metode ini diinspirasikan dari algoritma *negative selection* yang memungkinkan untuk membedakan sel-sel sendiri dan yang bukan. Pada studinya sel sendiri (*self*) didefinisikan untuk operasi pemotongan yang normal dan sel luar (*non-self*) adalah deviasi atau penyimpangan dari variasi kelonggaran dari kekuatan potong. De castro dan Timmis memperkenalkan penggunaan AIS dalam pengenalan pola (*pattern recognition*). De castro dan Von Zuben mengaplikasikan algoritma *clonal selection* untuk menyelesaikan masalah optimasi multimodal, penugasan *pattern recognotion* dan masalah perjalanan salesman. Hart dkk menggunakan pendekatan AIS untuk menyelesaikan masalah penjadwalan *job shop* [1] [6].

Pada tugas akhir ini akan menggunakan algoritma Clonal Selection Classification Algorithm (CSCA) yang merupakan salah satu bagian dari metode AIS sebagai proses clonal selection untuk melakukan klasifikasi SMS spam dan ham. Pada gambar 2.3 di bawah ini adalah alur proses dari CSCA yang akan digunakan untuk mengklasifikasikan SMS spam dan ham [3].



Gambar 2.1 Alur Proses CSCA

Penjelasan dari gambar 2.1 diatas adalah sebagai berikut:

1- **Initialization,** mengisi pool antibody dengan nilai S yang dipilih secara acak
2- **Loop,** menjalankan langkah utama dari algoritma untuk generasi G.
   a- **Selection & Pruning**, mengekspos seluruh populasi ke set antigen dan menghitung nilai fitness untuk setiap antibodi. Seluruh Populasi

ini kemudian dipilih. Aturan seleksi kemudian diterapkan dalam urutan sebagai berikut

i- Antibody dengan nilai klasifikasi salah = 0 akan dihapus dari set yang dipilih

ii- Antibody dengan klasifikasi benar = 0 dan klasifikasi salah > 0 maka nilai fitness akan dihitung ulang.

iii- Antibody dengan nilai fitness dibawah minimum fitness treshold (**ε**) akan dihapus dari set yang dipilih dan dari populasi antibody.

b- Cloning & Mutation, set yang dipilih kemudian dikloning dan bermutasi menggunakan nilai operasi fitness.

c- Insertion, hasil cloning tadi kemudian akan dimasukkan ke dalam populasi antibody yang utama. n antigen yang dipilih secara acak dari set antigen akan dimasukkan ke dalam populasi, dimana n adalah jumlah dari antibody pada set yang dipilih tahap 2.a.

3- Final pruning, mempersiapkan populasi antibody untuk produksi. Populasi diekspos ke populasi antigen, nilai fitness disiapkan, dan pemangkasan dilakukan seperti pada tahap 2.a.iii.

4- Classification, populasi antibody diambil sebagai satu set. Untuk sebuah unclassified data instance yang diberikan, maka akan di ekspos ke populasi. Affinity tertinggi dipilih, dan kelas mayoritas dari antigen diterapkan kepada unclassified instance.

2.4 Tokenizaton With Vectors

Proses tokenisasi yaitu proses pemisahan kata dari dokumen dengan menggunakan karakter spasi sebagai tanda pemisahnya. Proses ini diawali dari mengambil isi dokumen dari tabel *corpus* [9]. *Corpus* adalah kumpulan dari beberapa teks sebagai sumber penelitian bahasa dan sastra. Kumpulan teks disebut *corpus* jika kumpulan teks tersebut digunakan sebagai objek dari penelitian bahasa dan sastra. Teks yang digunakan sebagai *corpus* dikumpulkan dengan sistematis [2]. Terdapat beberapa algoritma untuk melakukan proses tokenisasi yaitu *Tokenization With Vectors* dan *Tokenization Without Vectors.* Pada dasarnya kedua algoritma tersebut digunakan untuk melakukan proses tokenisasi pada sistem *Information Retrieval* [7].

Pada Tugas Akhir ini akan menggunakan algoritma Tokenization With Vectors tersebut untuk melakukan proses tokenisasi sebagai pendukung dalam mengimplementasikan metode Artificial Immune System (AIS) untuk menyaring SMS spam. Algoritma Tokenization With Vectors dipilih sebagai pendukung untuk metode AIS ini karena berdasarkan performansi yang dihasilkan pada sistem Information Retrieval lebih efisien dibandingkan dengan tidak menggunakan algoritma tersebut. Pada algoritma Tokenization With Vectors terdapat empat fase yang digunakan untuk preprocessing, sehingga jika algoritma tersebut tidak digunakan, maka fase preprocessing tidak akan dijalankan, salah satu fase yang tidak dijalankan tersebut adalah fase stemming.

Pada algoritma Tokenization With Vectors terdapat empat fase utama, empat fase tersebut adalah sebagai berikut [7]:

- Fase 1 : Input file/dokumen

S.No.	Documents Contents
doc1	Military is a good option for a career builder for youngsters. Military is not covering only defense it also includes IT sector and its various forms are Army, Navy, and Air force. It satisfies the sacrifice need of youth for their country.

Gambar 2.2 Fase 1 proses *tokenization with vectors*

- Fase 2: memisahkan kata per kata

Name: doc1

[Military, is, a, good, option, for, a, career, builder, for, youngsters, Military, is, not, covering, only, defense, it, also, includes, IT, sector, and, its, various, forms, are, Army,, Navy,, and, Air, force., It, satisfies, the, sacrifice, need, of, youth, for, their, country.]

Gambar 2.3 Fase 2 proses *tokenization with vectors*

- Fase 3 dan 4: menghilangkan stop words dan melakukan stemmming

Name: doc1

[militari, good, option, for, career, builder, for, youngster, militari, not, cover, onli, defens, it, also, includ, it, sector, it, variou, form, ar, armi, navi, air, forc, it, satisfi, sacrific, need, youth, for, their, country]

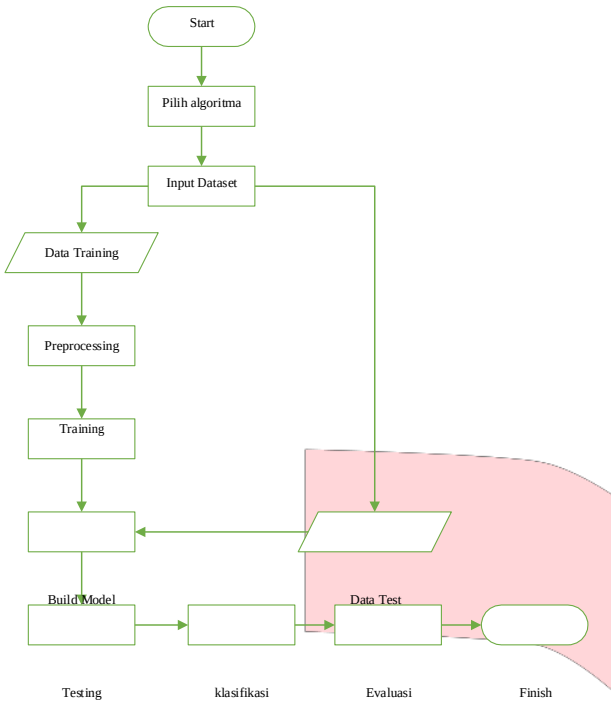
Gambar 2.4 Fase 3 dan 4 proses *tokenization with vectors*

Setelah semua fase tersebut dilakukan, maka langkah selanjutnya akan dilakukan klasifikasi dengan menggunakan metode AIS.

3. Analisis Kebutuhan

3.1 Desain umum sistem

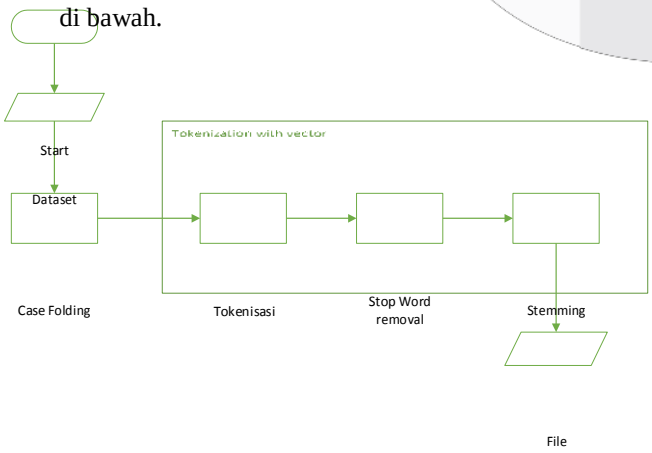
Sistem yang akan dibangun dalam penelitian tugas akhir ini meliputi proses training dan juga testing berdasarkan dataset yang digunakan. Pada gambar 3.1 di bawah ini merupakan desain sistem secara umum yang akan dibangun.



Gambar 3.1 Desain Umum Sistem

Proses pertama dari sistem ini ialah menginput dataset yang telah dibagi dua menjadi data training dan juga data testing dengan pembagian yang telah diatur berdasarkan skenario. Pada bagian data training akan dilakukan preprocessing untuk membangun model klasifikasi. Selanjutnya pada data testing akan dilakukan preprocessing juga dan sistem akan melakukan prediksi terhadap isi yang terdapat pada data testing apakah isi-isi tersebut termasuk ke dalam spam ataupun ham. Setelah sistem berhasil melakukan klasifikasi terhadap data testing, maka selanjutnya akan dilakukan proses evaluasi untuk menetukan akurasi yang dihasilkan oleh sistem.

Pada tahap preprocessing terdapat beberapa proses seperti case folding, tokenisasi, eliminasi stop word, dan stemming. Untuk lebih jelasnya tahapan-tahapan dari preprocessing akan digambarkan pada gambar 3.2 di bawah.



Gambar 3.2 Diagram Alir Preprocessing

3.2 Pengukuran performansi

Pengukuran performansi atau evaluasi dilakukan pada penelitian klasifikasi teks bertujuan untuk mengetahui apakah metode yang digunakan telah mengklasifikasi dengan benar atau tidak. Parameter evaluasi utama yang akan dianalisis pada penelitian ini mencakup recall, precision, F-Measure dan akurasi keseluruhan.

- Precision adalah jumlah jumlah dokumen yang diklasifikasikan dengan benar oleh sistem dibagi dengan jumlah keseluruhan klasifikasi yang dilakukan oleh sistem (Sunantyo, et al., 2013).

$$\text{Precision} = 100\% \times \frac{TP}{TP + FP}$$

- Recall adalah jumlah dokumen yang terklasifikasi dengan benar oleh sistem dibagi dengan jumlah dokumen yang seharusnya bisa dikenali sistem (Sunantyo, et al., 2013).

$$\text{Recall} = 100\% \times \frac{TP}{TP + FN}$$

- F-measure merupakan nilai yang mewakili kinerja keseluruhan sistem dan merupakan penggabungan nilai recall dan precision (Sunantyo, et al., 2013).

$$F\text{-Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- Akurasi keseluruhan adalah jumlah data yang diprediksi benar dibagi dengan jumlah keseluruhan data

$$\text{Akurasi} = \frac{TP + TN}{TP + FP + TN + FN}$$

Tabel 3.1 Confusion Matrix

	Ham	Spam
Ham	TP	FP
Spam	FN	TN

Keterangan:
TP atau True Positive adalah pesan ham yang benar terdeteksi sebagai ham.
FP atau False Positive adalah pesan ham yang terdeteksi sebagai spam
TN atau True Negative adalah pesan spam yang benar terdeteksi sebagai spam
FN atau False Negative adalah pesan spam yang terdeteksi sebagai ham

4. Pengujian

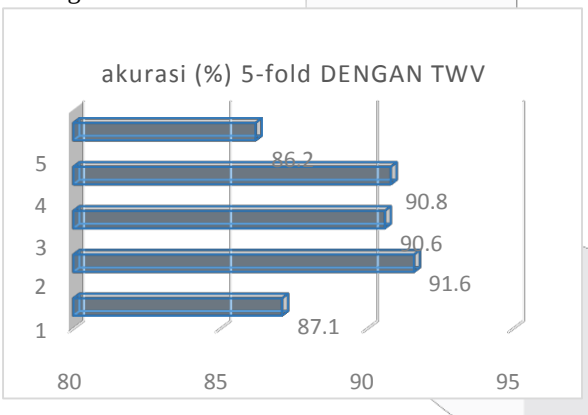
Skenario pengujian yang dilakukan pada penelitian adalah dengan menggunakan pengujian cross validation dengan nilai k=5-fold dan juga k=10-fold. Untuk membandingkan performansi yang dihasilkan, maka akan dilakukan empat pengujian sebagai berikut:

- k=5-fold untuk AIS dengan TWV
- k=10-fold untuk AIS dengan TWV
- k=5-fold untuk AIS tanpa TWV
- k=10-fold untuk AIS tanpa TWV

Adapun hasil yang didapatkan dari keempat skenario pengujian yang telah ditentukan tersebut adalah sebagai berikut:

- **k=5-fold untuk AIS dengan Tokenization With Vectors**

Pada skenario ini, didapatkan hasil yang bervariasi untuk setiap fold. Pada pengujian kali ini, rata-rata akurasi yang dihasilkan adalah sebesar 89.26%. Adapun untuk akurasi tertinggi terdapat pada fold ke-2, yaitu sebesar 91.6%. Dan akurasi terendah pada pengujian kali ini terdapat pada fold ke-5 yaitu hanya sebesar 86.2%. Pada gambar 4.1 di bawah ini akan ditampilkan akurasi yang dihasilkan dari masing-masing fold.



Gambar 4.1 Grafik rata-rata akurasi k=5 untuk AIS dengan TWV

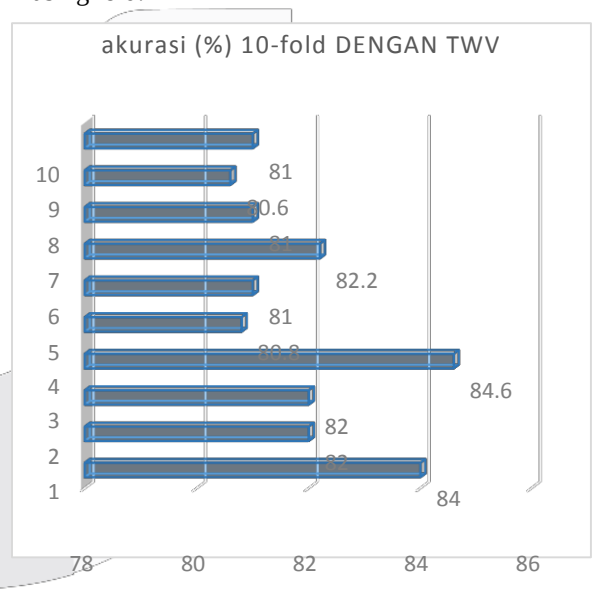
Di bawah ini adalah tabel yang menampilkan hasil confusion matrix dari pengujian cross validation 5-fold yang dilakukan dengan menggunakan metode AIS ditambah dengan algoritma Tokenization With Vectors. Pada tabel 4.1 di bawah ini akan terlihat seberapa banyak sistem dapat menentukan SMS dengan kategori spam dan ham yang benar.

Tabel 4.1 Confusion Matrix k=5 fold untuk AIS dengan TWV

K=5	Data Training	Data Testing	TP	FP	TN	FN
1	BCDE	A	775	25	96	104
2	ACDE	B	765	35	151	49
3	ABDE	C	765	35	141	59
4	ABCE	D	763	37	145	55
5	ABCD	E	787	13	75	125

- **k=10-fold untuk AIS dengan Tokenization With Vectors**

Pada skenario ini, didapatkan hasil yang bervariasi untuk setiap fold. Pada pengujian kali ini, rata-rata akurasi yang dihasilkan adalah sebesar 81.92%. Adapun untuk akurasi tertinggi terdapat pada fold ke-4, yaitu sebesar 84.6%. Dan akurasi terendah pada pengujian kali ini terdapat pada fold ke-9 yaitu hanya sebesar 80.6%. Pada gambar 4.2 di bawah ini akan ditampilkan akurasi yang dihasilkan dari masing-masing fold.



Gambar 4.2 Grafik rata-rata akurasi k=10 untuk AIS dengan TWV

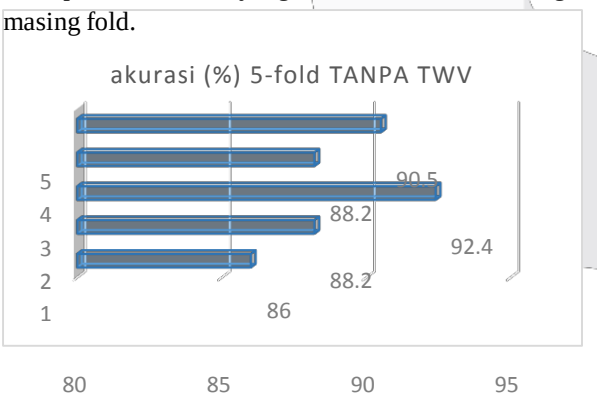
Di bawah ini adalah tabel yang menampilkan hasil confusion matrix dari pengujian cross validation 10-fold yang dilakukan dengan menggunakan metode AIS ditambah dengan algoritma Tokenization With Vectors. Pada tabel 4.2 di bawah ini akan terlihat seberapa banyak sistem dapat menentukan SMS dengan kategori spam dan ham yang benar.

Tabel 4.2 Confusion Matrix k=10 fold untuk AIS dengan TWV

K=10	Data Training	Data Testing	TP	FP	TN	FN
1	BCDEFGHIJ	A	400	0	20	80
2	ACDEFGHIJ	B	398	2	12	88
3	ABDEFGHIJ	C	400	0	10	90
4	ABCEFGHIJ	D	395	5	28	72
5	ABCDFGHIJ	E	397	3	7	93
6	ABCDEGHIJ	F	400	0	5	95
7	ABCDEFHIJ	G	400	0	11	89
8	ABCDEFGIJ	H	400	0	5	95
9	ABCDEFGHJ	I	400	0	3	97
10	ABCDEFGHI	J	400	0	5	95

- **k=5-fold untuk AIS tanpa Tokenization With Vectors**

Pada skenario ini, didapatkan hasil yang bervariasi untuk setiap fold. Pada pengujian kali ini, rata-rata akurasi yang dihasilkan adalah sebesar 89.06%. Hasil ini tidak berbeda jauh dengan yang dihasilkan dari pengujian dengan menggunakan algortima Tokenization With Vectors yaitu sebesar 89.26%. Adapun untuk akurasi tertinggi terdapat pada fold ke-3, yaitu sebesar 92.4%. Dan akurasi terendah pada pengujian kali ini terdapat pada fold ke-1 yaitu hanya sebesar 86%. Pada gambar 4.3 di bawah ini akan ditampilkan akurasi yang dihasilkan dari masing-masing fold.



Gambar 4.3 Grafik rata-rata akurasi k=5 untuk AIS tanpa TWV

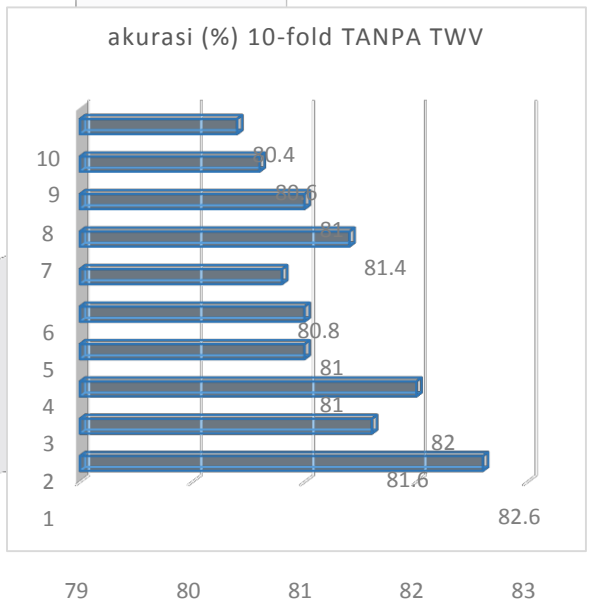
Di bawah ini adalah tabel yang menampilkan hasil confusion matrix dari pengujian cross validation 5-fold yang dilakukan dengan menggunakan metode AIS tanpa menambahkan algoritma Tokenization With Vectors. Pada tabel 4.3 di bawah ini akan terlihat seberapa banyak sistem dapat menentukan SMS dengan kategori spam dan ham yang benar.

Tabel 4.3 Confusion Matrix k=5 fold untuk AIS tanpa TWV

K=5	Data Training	Data Testing	TP	FP	TN	FN
1	BCDE	A	793	7	67	133
2	ACDE	B	789	11	93	107
3	ABDE	C	787	13	137	63
4	ABCE	D	766	34	116	84
5	ABCD	E	774	26	131	69

- **k=10-fold untuk AIS tanpa Tokenization With Vectors**

Pada skenario ini, didapatkan hasil yang bervariasi untuk setiap fold. Pada pengujian kali ini, rata-rata akurasi yang dihasilkan adalah sebesar 81.24%. Hasil ini tidak berbeda jauh dengan yang dihasilkan dari pengujian dengan menggunakan algortima Tokenization With Vectors yaitu sebesar 81.92%. Adapun untuk akurasi tertinggi terdapat pada fold ke-1, yaitu sebesar 82.6%. Dan akurasi terendah pada pengujian kali ini terdapat pada fold ke-10 yaitu hanya sebesar 80.4%. Pada gambar 4.4 di bawah ini akan ditampilkan akurasi yang dihasilkan dari masing-masing fold.



Gambar 4.4 Grafik rata-rata akurasi k=10 untuk AIS tanpa TWV

Di bawah ini adalah tabel yang menampilkan hasil confusion matrix dari pengujian cross validation 10-fold yang dilakukan dengan menggunakan metode AIS tanpa menambahkan algoritma Tokenization With Vectors. Pada tabel 4.4 di bawah ini akan terlihat seberapa banyak sistem dapat menentukan SMS dengan kategori spam dan ham yang benar.

Tabel 4.4 Confusion Matrix k=10 fold untuk AIS tanpa TWV

K=10	Data Training	Data Testing	TP	FP	TN	FN
1	BCDEFGHIJ	A	400	0	13	87
2	ACDEFGHIJ	B	400	0	8	92
3	ABDEFGHIJ	C	400	0	10	90
4	ABCEFGHIJ	D	400	0	5	95
5	ABCDFGHIJ	E	400	0	5	95
6	ABCDEGHIJ	F	400	0	4	96
7	ABCDEFHIJ	G	399	1	8	92
8	ABCDEFGIJ	H	400	0	5	95
9	ABCDEFGHJ	I	400	0	3	97
10	ABCDEFGHI	J	400	0	2	98

5. Kesimpulan

1- Rata-rata akurasi yang dihasilkan dari metode AIS yang dikombinasikan dengan algoritma Tokenization With Vectors ialah sebesar 89.26% untuk pengujian dengan cross validation 5-fold, dan 81.92% untuk pengujian dengan cross validation 10-fold.

2- Rata-rata akurasi yang dihasilkan dari metode AIS yang tidak dikombinasikan dengan algoritma Tokenization With Vectors ialah sebesar 89.06% untuk pengujian dengan cross validation 5-fold, dan 81.24% untuk pengujian dengan cross validation 10-fold.

3- Dengan menggunakan algoritma Tokenization With Vectors pada metode AIS menghasilkan rata-rata akurasi yang lebih tinggi dibandingkan dengan tidak menggunakan algoritma Tokenization With Vectors, tetapi selisih rata-rata akurasinya tidak terlalu banyak.

4- Penggunaan algoritma Tokenization With Vectors pada metode AIS tidak terlalu berpengaruh besar terhadap akurasi yang dihasilkan.

6. Saran

1- Untuk penelitian pada tahap selanjutnya dapat dikombinasikan dengan algoritma-algoritma lainnya untuk mendapatkan performansi akurasi yang lebih tinggi.

2- Untuk dataset training dan testing yang digunakan bisa mencoba dengan dataset yang berbahasa Indonesia, sehingga tidak hanya dataset berbahasa Inggris saja yang digunakan.

3- Menggunakan beberapa metode pengujian untuk membandingkan hasil yang didapatkan.

Daftar pustaka

[1] (Mulia), m., 2010. *Algoritma Apriori.* [Online] Available at: http://tugaskuliah-sabanamulia.blogspot.com/2010/06/tugas2.html [Accessed 9 11 2014].

[2] Abidin, Z., 2013. *Dasar-Dasar Korpus Dalam Ilmu Bahasa*. [Online] Available at: http://abidin.lecturer.uin-malang.ac.id/2013/10/dasar-dasar-korpus-dalam-ilmu-bahasa/ [Accessed 16 November 2014].

[3] Brownlee, J., 2005. *Clonal Selection Theory & Clonalg The Clonal Selection Classification Algorithm (CSCA)*, Melbourne: Swinburne University of Technology (SUT).

[4] Ginting, R. & S. Ginting, T. H., 2006. *Studi Aplikasi Metode Artificial Immune System Dalam Penjadwalan Flow Shop.* Medan: Departemen Teknik Industri dan Departemen Teknik Mesin Fakultas Teknik, Universitas Sumatera Utara.

[5] Kurniawan, B., Effendi, S. & Sitompul, S., 2012. Klasifikasi Konten Berita Dengan Metode Text Mining. *JURNAL DUNIA TEKNOLOGI INFORMASI*, Volume I, pp. 14-19.

[6] Mahmoud, T. M. & Mahfouz, A. M., 2012. SMS Spam Filtering Technique Based on Artificial Immune System. *International Journal of Computer Science Issues*, 9(2), pp. 589-597.

[7] Singh, V. & Saini, B., 2014. An Effective Tokenization Algorithm For Information Retrieval Systems. *CS & IT-CSCP*, pp. 109-119.

[8] Sridianti, 2014. *Peran Limfosit dalam Sistem Kekebalan Tubuh.* [Online] Available at: http://www.sridianti.com/peran-limfosit-dalam-sistem-kekebalan-tubuh.html [Accessed 9 11 2014].

[9] Wibowo, J. S. & Hartati, S., 2011. Text Document Retrieval In English Using Keywords of Indonesian Dictionary Based. *IJCCS*, Volume 5, pp. 26-32.

[10] Wikipedia bahasa Indonesia, e. b., 2013. *Penambangan Teks.* [Online] Available at: https://id.wikipedia.org/wiki/Penambangan_teks [Accessed 22 Juny 2015].