SMS Classification Deteksi Spam dengan menggunakan Algoritma Artificial Immune System dan Apriori Frequent Itemset

Firizqy Ramadhana Nasution[1], Shaufiah, S.T, M.T[2], Moch. Arif Bijaksana, ir. MTECH[3]

[1]Firizqyramadhana@yahoo.com [2]shaufiah@gmail.com

Departemen Teknik Informatika Universitas Telkom, Bandung

Abstrak

Saat ini mobile Phone sudah menjadi salah satu kebutuhan penting masyarakat dan salah satu fiture yang sangat di andalkan pada *mobile phone* adalah *short message service* (SMS). SMS saat berguna untuk menyampaikan suatu pesan singkat padat jelas serta hemat biaya pemakaian, data yang disampaikan pun berupa *text* sehingga data bisa di simpan dan di gunakan untuk keperluan lain. Tetapi dalam kasus yang sama, terjadi penurunan harga dari SMS, hal ini memicu meningkatnya *unsolicited commercial advertisements* (Spam). Spam sangat mengguntungkan bagi pengirim pesan namun saat mengganggu bagi penerima pesan dikarenakan pesan yang di terima bersifat memaksa dengan output yang sangat besar, peningkatan SMS spam sangat signifikan, pada tahun 2013 di region Asia SMS spam meningkat sebesar 30%. Pada penelitian tugas akhir ini dilakukan analisis dan implementasi Spam detection untuk melakukan filtering pada SMS spam dengan algoritma *Artificial Immune System*(AIS), sebuah algoritma pengelompokan yang memakai ide dari sistem kekebalan tubuh manusia dengan tamabahan algoritma apriori untuk menghasilkan *frequent itemset*. Sebagai hasilnya algoritma *Artificial Immune System* dapat meningkatkan performansi dari sistem *sms filtering* sebesar 5% dan nilai akurasi dari sistem di atas angka 95%, karena seperti kekebalan tubuh manusia sistem dapat membuat *antibody* baru yang dapat menanggulangi masalah sms spam dan penggabungan dari kombinasi baru data set yang dihasilkan *frequent itemset* menambah nilai dari performa sistem.

Kata kunci : sms,spam , Artificial Immune System, apriori

Abstract

Currently mobile Phone has become one of the essential needs of the community and one of the very Fiture count on the mobile phone is the short message service (SMS). SMS while useful to convey a clear concise message and cost-effective use of, any data presented in the form of text so that the data can be saved and used for other purposes. But in the same case, a decline in the price of SMS, this triggers increased unsolicited commercial advertisements (Spam). Spam is very mengguntungkan for the sender but annoying for the recipient when the message because the message received is forced to output a very large, very significant increase in spam SMS, in 2013 in the Asia region SMS spam increased by 30% .In this research is done Spam detection analysis and implementation to perform spam filtering on SMS with Algorithms Artificial Immune System (AIS), a grouping algorithm that uses the idea of the human immune system with supplemental priori algorithm to generate frequent itemset.as the result Artificial Immune System algorytm can increase the work for sms filtering system 5% and have an accuracy above 95% , because like immune system in human body it can make new antybodies that can solve a problem about SMS spam dan a new combination of dataset from frequent itemset increase the result of the system.

Keywords: sms**, spam, Artificial Immune System, a priori**

1. Pendahuluan

Dalam era teknologi seperti sekarang Mobile Phone adalah hal yang sangat Lumrah dan hampir dimiliki oleh semua golongan masyarakat dari bawah , menengah hingga atas. Bahkan dalam Praktiknya , Mobile Phone sudah menjadi kebutuhan sekunder oleh beberapa golongan masyarakat. Short Message Service (Sms) adalah suatu fitur dari mobile phone yang sudah sangat mendasar dan dibutuhkan dalam suatu perangkat Mobile Phone. Fitur ini sudah dianggap sebagai salah satu jenis pelayanan yang fundamental dan primitif karena pelayananya yang murah, kemudahaan penggunaanya dan dokumentasi untuk setiap pengguna Mobile Phone

Namun Dari semua keunggulan sms terdapat sebuah aktifitas yang sangat menggangu yang di sebut unsolicited commercial advertisements(spam). **Spam** adalah penggunaan perangkat elektronik untuk mengirimkan pesan secara bertubi-tubi atau pesan yang mengganggu dan tidak penting tanpa dikehendaki oleh penerimanya. Orang yang melakukan spam disebut spammer. Tindakan spam dikenal dengan nama spamming. Spam dikirimkan oleh pengiklan dengan biaya operasional yang sangat rendah, maka banyak spammers yang muncul dan jumlah pesan yang tidak diminta menjadi sangat tinggi. Akibatnya, banyak pihak yang dirugikan dan merasa tidak nyaman akan hal ini. Untuk menanggulangi masalah SMS spamming ini penulis akan melakukan penilitian mengenai sms classification untuk membedakan dan menggolongkan sms spam dan sms non spam (ham). Serta metode filtering yang akan di gunakan untuk menanggulangi masalah SMS spam tersebut .algoritma yang akan penulis gunakan adalah artificial immune system dan apriori pada frequent itemset.
Algoritma Artificial Immune System adalah sebuah algoritma yang dikembangkan dengan ide dari Biological Immnue System(BIS). Ide dasarnya diambil dengan prinsip immune pada manusia yang mempertahankan dan melawan virus penyakit dan infeksi yang menyerang tubuh. Dan algoritma AIS menghasilkan banyak ide baru dalam dunia Computer Science terlebih lagi dalam ranah Security Area. Gambaran kinerja algoritma ini adalah menganalisis dan mengklasifikasikan data yang masuk. Bila data baru dan merupakan spam maka sistem akan membuang data tersebut dan membentuk sistem kekebalan terhadap data spam yang terdeteksi tersebut. Algoritma AIS ini memiliki akurasi dan rasio deteksi yang sangat tinggi di atas angka 85% sehingga akan efisien untuk menanggulangi masalah sms Spamming

Algoritma Apriori adalah algoritma paling terkenal untuk menemukan pola frekuensi tinggi. Pola frekuensi tinggi adalah pola-pola item di dalam suatu database yang memiliki frekuensi atau support di atas ambang batas tertentu yang disebut dengan istilah minimum support. Pola frekuensi tinggi ini digunakan untuk menyusun aturan assosiatif dan juga beberapa teknik data mining lainnya .

Dalam tugas akhir kali ini penulis akan menguji performansi algoritma AIS dan apriori pada data. Karena dengan pola frekuensi data yang sudah ditemukan oleh algorithma apriori akan menambah performansi dari algoritma AIS baik dalam segi akurasi dan kecepatan proses kerja sistem.

2. Artificial immune system

Adalah algoritma yang didasarkan pada prinsip kekebalan tubuh manusia, yang melindungi tubuh terhadap berbahaya penyakit dan infeksi. Untuk melakukan hal ini,harus melakukan pola rekognisi untuk membedakan moleku ldan sel-sel tubuh(diri) dari sel sel asing (non-self).Biological Immune System (BIS) berbasis di satu set sel kekebalan yang disebut limfosit.terdiri dari sel B dan T. Pada permukaan setiap limfosit terdapat reseptor dan pengikat pada reseptor ini oleh interaksi kimia akan membentuk sebuah pola pada antigen yang dapat mengaktifkan kekebalan sel. Himpunan bagian dari antigen adalah patogen, sebuah agen biologis yang mampu merugikan tuan rumah(misalnya bakteri). Limfosit dibuat di sumsum tulang belakang dan bentuk reseptor ditentukan dengan menggunakan libraries dari gen.peran utama dari limfositd di AIS adalah encoding dan menyimpan titik diruang solusi. Persaingan antara sebuah reseptor dan antigen , proses pencocokan antara reseptor dan antigen nmungkin tidak terlalu tepat dan teliti, jadi ketika proses pengikatan berlangsung ia melakukannya dengan kekuatan yang disebut afinitas.

Seleksi klonal dan ekspansi adalah teori yang paling diterima digunakan untuk menjelaskan bagaimana sistem kekebalan tubuh berupaya dengan antigen. Singkatnya ,klona lteori seleksi menyatakan bahwa ketika antigen menyerang organisme, subset dari kekebalan Sel-sel yang mampu mengenali anti gen tersebut berkembang biak dan berdiferensiasi menjadi sel-sel yang aktif atau memori. Klon fittesta dalah mereka, yang menghasilkan antibodi yang mengikat antigen terbaik (dengan afinita stertinggi).

3. Algoritma ariori

Apriori adalah suatu algoritma yang sudah sangat dikenal dalam menentukan pencarian freuquent item set dengan menggunakan teknik association rule. Algoritma ini pertama kali diperkenalkan oleh Agrawal dan Srikant pada tahun 1994 untuk penentuan frequent itemset untuk aturan asosiasi boolean .

Analisa asosiasi atau association rule mining adalah teknik data mning untuk menemukan aturan suatu kombinasi item . Salah satu tahap analisas asosiasi yang menarik perhatian banyak peneliti adalah kemampuannya dalam menghasilkan algoritma yang efisien melalui analisis pola frekuensi tinggi atau frequent pattern mining Penting tidaknya suatu asosiasi dapat diketahui dua tolak ukur, yaitu support dan confidence. Support adalah prosentase kombinasi item tersebut dalam database, sedangkan confidence adalah kuatnya hubungan antar item dalam aturan

Erwin (2009) menambahkan bahwa algoritma ini menggunakan pengetahuan mengenai frequent item set yang telah diketahui sebelumnya untuk memproses informasi selanjutnya . Pada algoritma apriori untuk

menentukan kandidat-kandidat yang mungkin muncul akan dilakukan dengan cara memperhatikan minimum suppor.

Ada dua proses utama yang dilakukan dalam algoritma apriori, yaitu :

1. Join (Penggabungan)
   Pada proses ini setiap item dikombinasikan dengan item lainnya sampai tidak terbentuk kombinasi lagi.
2. Prune (Pemangkasan)
   Pada proses ini, hasil dari item yang telah dikombinasikan tadi lalu dipangkas dengan menggunakan minimum support yang telah ditentukan oleh user.

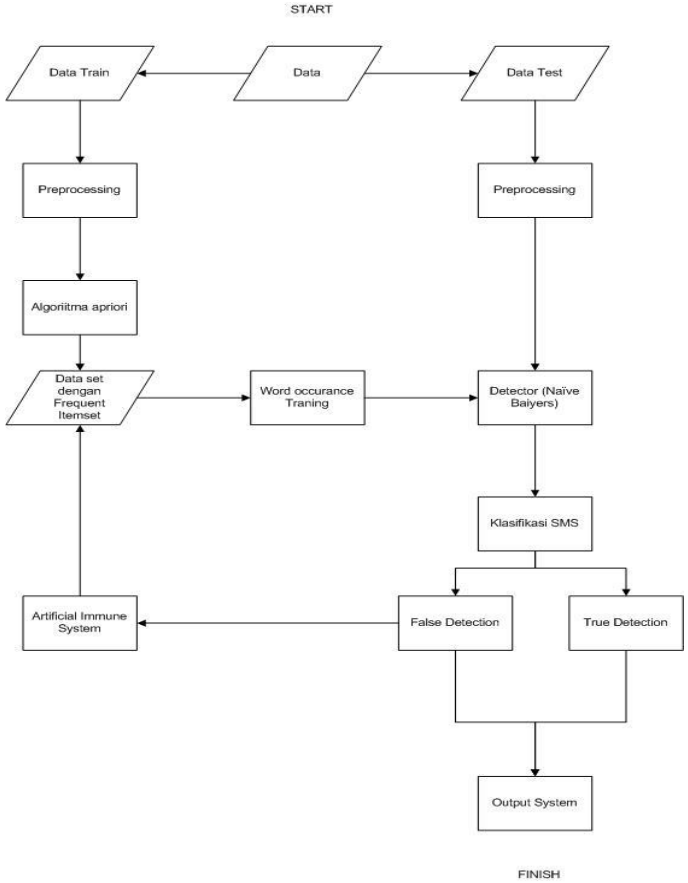
4. Perancangan Sistem

Sistem adalah sebuah classification data dan filtering data untuk data berupa text sms. Dan dapat membedakan data yang berupa spam dan ham, kemudian sistem mendeteksipola data berupa spam. Bila terdapat kesalahan dalam proses pendeteksian maka sistem akan secara otomatis mengikat pola data tersebut ,mentrainingnya kembali sehingga probabilitas dari kesalahan data menjadi lebih kecil dan pendeteksian data menjadi lebih akurat.

Alur kerja sistem dibagi menjadi 2 tahap, Training data dan Testing data .Training data adalah proses pembentukan model klasifikasi, dalam tahap ini penulis menggunakan algoritma apriori untuk membuat suatu frequent item set, sehingga pembentukan model training mempunyai probabilitas lebih tinggi untuk keakurasian sistem. Testing Data adalah proses implementasi model yang dilakukan setelah didapat model dengan akurasi yang baik, untuk mendeteksi suatu Ham atau Spam sistem mempunyai sebuah detector(antibody) dalam hal ini penulis menggunakan algorithma naïve baiyess sebagai detectornya. Output dari tahap pendeteksian adalah suatu nilai false dan true .semua data output dengan nilai false akan di masukan ke dalam Immune system dan d training kembali dan dimasukan kembali ke dalam data set , sehingga akurasi pendeteksian akan semakin akurat.

Namun sebelum memasuki 2 tahap ini dataset akan melalui proses preprocessing terlebih dahulu agar data yang didapat lebih bersih, terstruktur,dan berkulang volume kosakatanya. Hal ini sangat mempengaruhi kemudahan sistem untuk mengolah data.

Untuk mencapai kebutuhan sistem diatas, secara garis besar rancangan sistem digambarkan dengandiagram sebagai berikut:



4.1. Prerocessing

Preprocessing terbagi dalam 3 tahap yaitu, Case Folding , Stop World Removal dan Stemming. 3 proses tersebut berfungsi untuk memudahkan sistem untuk mengolah data set , berikut penjelasan 3 tahap pererocessing :

1.Case Folding + penghapusan karakter :

Menyeragam kan data sms menjadi format text huruf kecil dan menghapuskan tanda baca serta membuang semua karakter selain huruf dan angka:

2. Stopword Removal :

Dalam proses ini system akan mengeliminasi semua kata Stopword. Tahap ini Menggunakan library Stopword yang sudah dicocokan ke dalam sistem. Library tersebut di ambil dari situs http://www.ranks.nl/stopwords.

3.Stemming :

Tahap ini berfungsi untuk mengeleminasi semua kata yang memiliki imbuhan, sehingga dapat mengurangi kata yang memiliki variasi berbeda namun pengartianya sama. seharusnya memiliki arti sama namun memiliki bentuk imbuhan yang berbeda. Pada proses ini digunakan Snowball Tartarus library yang menerapkan algoritma Porter Stemmer, diambil dari situs snowball.tartarus.org.

4.2. Feature Extraction

Dalam *data Training* dilakukan *Feature Extraction* dengan menggunakan algoritma apriosi untuk mendapatkan suatu frequent item set. Sebelum diproses, data isi SMS dalam format kata tersebut diubah ke dalam format angka melalui proses *preparation extraction*.

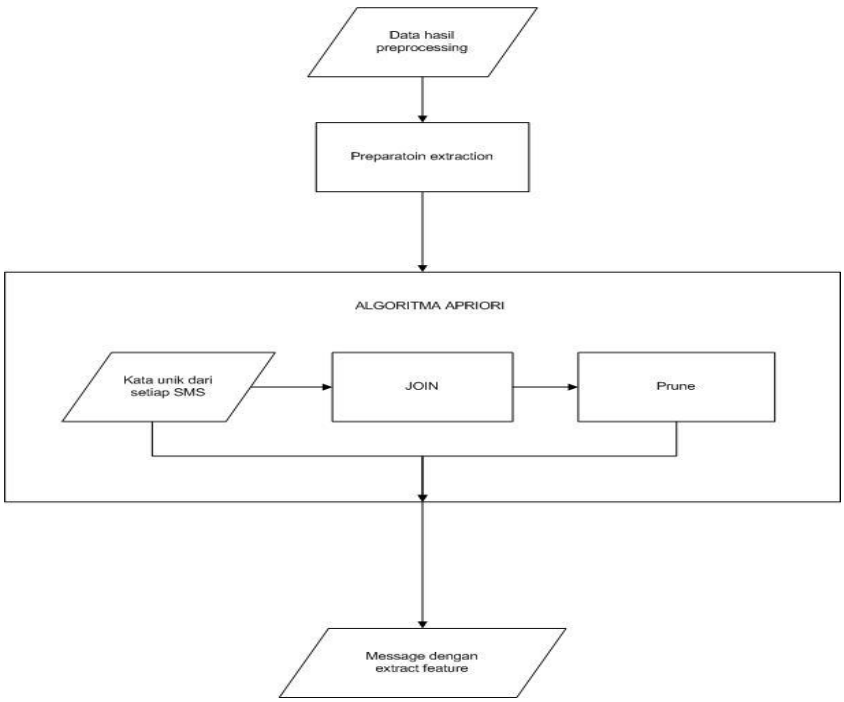
Lalu jalankan algoritma apriori. Dalam algoritma apriori sendiri terdapat 2 proses utama yaitu *join* dan *prune* :

4.2.1 *join*

Dalam tahap ini semua item pada data dkombinasikan dengan item lainya sehingga menghasilkan suatu kombinasi baru proses ini dilakukan hingga tidak ada lagi kombinasi baru yang terbuat.

4.2.1 *Prune*

Dalam tahap ini data di pangkas berdasakan minimum support yang sudah ditentukan misalkan minimum support yang di tentukan adalah 50%, maka data di bawah 50% di pangkas dan data di atas kemunculan 50% di bawa oleh sistem ke tahap selanjutnya.



4.3. Word Occurence Training

Tahap Selanjutnya adalah melakukan perhitungan untuk masing-masing kata yang telah diekstrak pada masing-masing kelas., dibuat dalam format word occurrence table.

seluruh kata tersebut akan dihitung kemunculannya dari masing-masing dokumen SMS dalam format word occurrence atau tabel kejadian yang menggambarkan perhitungan seluruh kata dari setiap kelas SMS.

4.4 Detector Naïve Baiyers

Pada tahap ini, mulai dilakukan proses klasifikasi dengan Algoritma Naive Bayes. Kata-kata yang sudah dihitung pada word occurence table selanjutnya dilakukan perhitungan total dan prior probability terhadap masing-masing kelas (spam dan ham). Kemudian memasukkan data testing yang telah dilakukan proses preparation untuk dilakukan klasifikasi. Pada tahap klasifikasi, diterapkan perhitungan dengan laplace estimator atau laplace smoothing untuk menghindari nilai probabilitas 0.

Contoh :

Hitung berdasarkan hasil pada word occurrence table.

- Prior probability of ham P(ham) = Jumlah sms Ham : Total Sms = 3/5 = 0,6
- Prior probability of spam P(spam) = = Jumlah sms Spam : Total Sms = 2/5 = 0,4
- Total number of vocabulary |v| = 29
- Total number of ham words Nham = 14
- Total number of spam words Nspam = 20

Hasil perhitungan diatas merupakan model yang selanjutnya digunakan untuk proses testingdan klasifikasi SMS baru.Untuk menghasilkan akurasi yang lebih baik, maka diterapkan Laplace estimator untuk menghindari kemungkinan nol pada SMS. Di awal telah didapatkan hasil prior probability untuk spam dan ham, selanjutnya tiap kata dihitung nilai peluangnya dengan rumus sebagai berikut :

P (word | class) = (Nword + 1) / (Nclass + |v|)

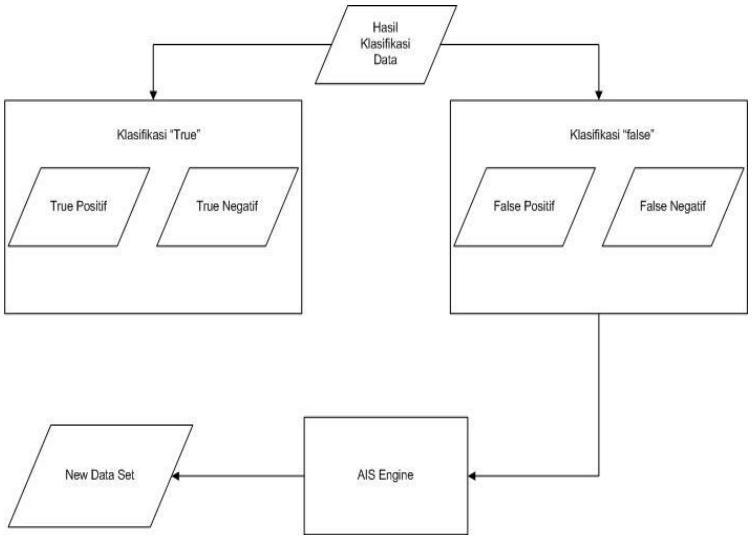
4.5 AIS Engine And User Feedback

Dalam Tahap ini output dalam pendeteksian akan di klasifikasi, dalam tahap ini terdapat 2 jenis eror dalam sistem klasifikasi :

1.Spam diklarifikasikan sebagai Ham (False Negatif)

2.Ham diklarifikasikan sebagai Spam (False Positif)

Kedua kesalahan klarifikasi di atas akan segera siap di training kembali ke dalam data set oleh sistem sehingga sistem akan mempunyai tambahan data set baru yang bias menambah keakurasian sistem untuk mendeteksi sms .



5. Pengujian Sistem

5.1. Skenario Pengujian

Untuk mengidentifikasi performansi dari sistem SMS filtering yang dibangun pada tugas akhir ini, digunakan dataset yang berasal dari SMS Corpus dari SMS Spam Corpus v.0.1 Big dengan SMS berjumlah 1324 yang terdiri dari 1002 SMS ham dan 322 SMS spam.

Dalam pengujian ini digunakan teknik 10-fold cross validation, yaitu teknik pembagian data training dan data testing dimana data dibagi menjadi 10 kelompok dan proses pelatihan data akan dilakukan sebanyak 10 kali. Sehingga setiap kelompok data menjadi data latih sebanyak 9 kali dan menjadi data uji sebanyak 1 kali. Hal ini dilakukan untuk mendapatkan rata-rata akurasi yang akurat terhadap keseluruhan dataset.

5.2. Proses Pengujian

Proses pengujian dalam tugas akhir ini dilakukan dalam beberapa skenario.

1.Pengujian Naive Bayes dengan Apriori menggunakan dataset SMS Spam Corpus v.0.1 Big

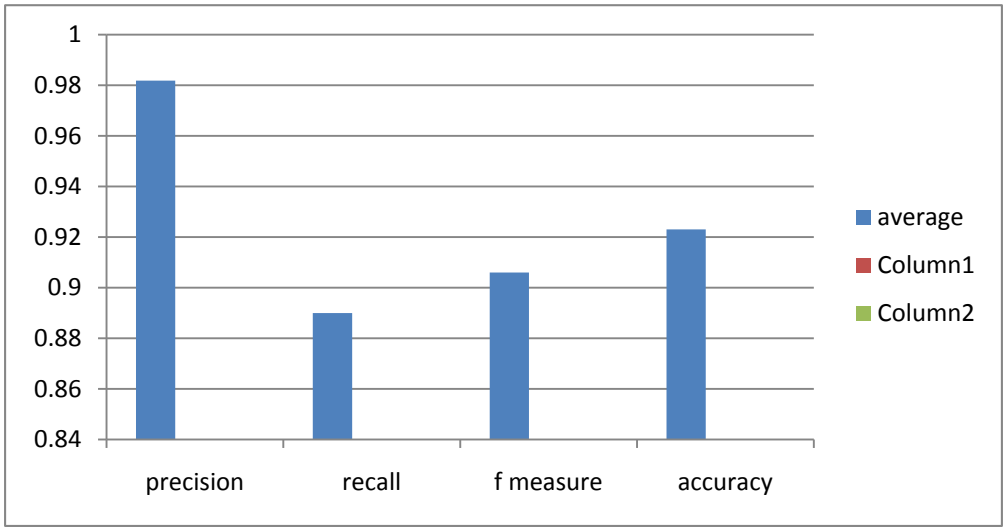
2.Pengujian Naive Bayes dengan Apriori menggunakan dataset SMS Spam Collection v.1dan setelah dilakukan proses Immunologi dari algorythma AIS.

3.Pegujian minimum support pada apriori dan system tanpa minum support

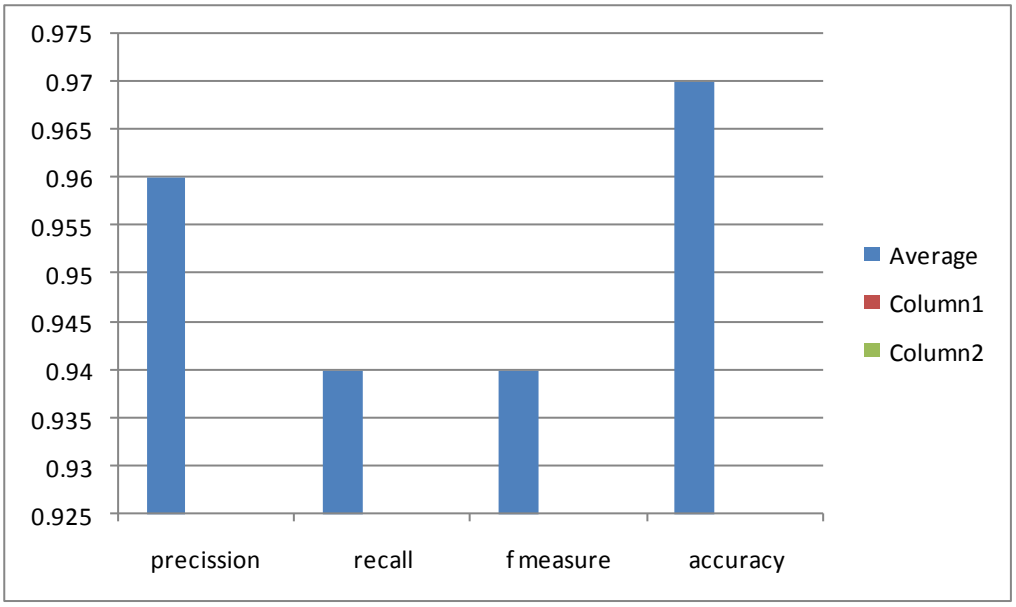
5.3Hasil Pengujian

5.3.1Hasil Perbandingkan dengan dan tanpa AIS

Berikut adalah nilai akurasi dari hasil pengujian pada kedua metode (dengan AIS- dan tanpa AIS) akan dibandingkan dengan menggunakan nilai evaluasi rata-rata.



Gambarl 1. Grafik Pengujian pertama

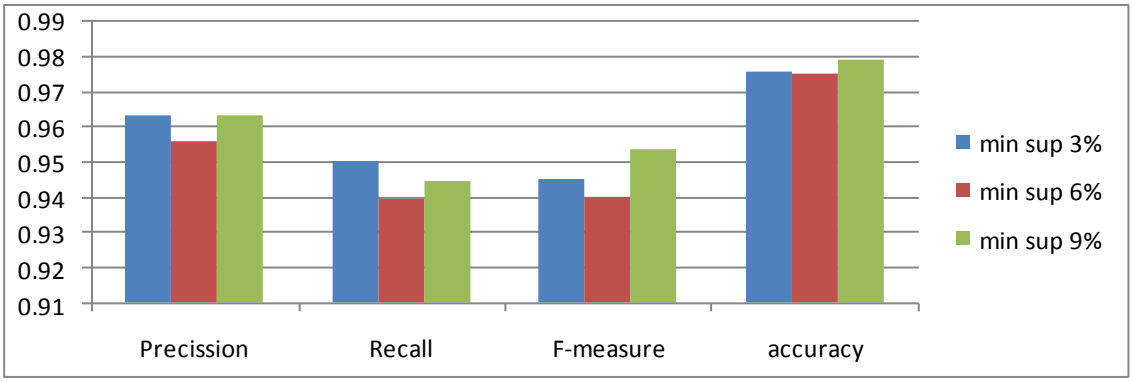


Gambarl 2. Grafik Pengujian kedua

Dari grafik di atas nilai *precission* mempunyai hasil yang lebih tinggi pada pengujian pertama namun *niali recall, Fmeasure dan accuracy* mengalami peningkatan pada pengujian kedua. Hasil peningkatan ini disebabkan karena immune system dari algoritma AIS, yang bila terdapat spam yang tidak terdeteksi maka system akan mentrainingnya kembali , sehingga nilai dari akurasinya pun bertambah.

5.3.2Hasil Perbandingkan Minimum support

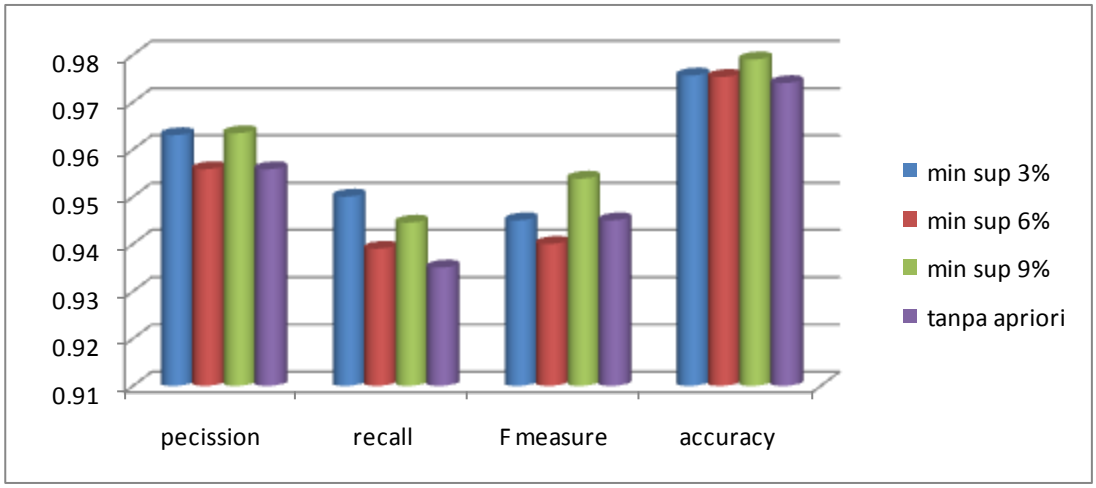
Pengaruh minimum support dengan parameter 3%, 6% , dan 9% dari pengujian system bisa dilihat dalam grafik sebagai berikut:



Gambar 3. Grafik pengaruh minimum support

Dari gambar 4-3 terlihat perbedaan *minimum support* 3% ke 6% mengalami penurunan pada smua parameter penilaian. Hal ini dikarenakan data frequent item set yang dihasilkan lebih sedikit saat minimum support di angka 6% dan mengalami peningkatan kembali *di mimimum support* 9% di karenakan data spam dari frequent itemset dengan minimum support 9 % memiliki tingkat kemunculan spam yang lebih tinggi, sehingga frequent item set baru ini lebih kuat dari dua parameter minimum support sebelumnya.\

5.3.2Analisis Proses Pengujian Tanpa Apriori



Gambarl 4-4. Grafik Pengujian tanpa dan dengan apriori

Dari grafik di atas terlihat apriori *frequent item* set meningkatkan performa system ,terlebih lagi di *minnum support* 9%. Namun peningkatan performansi system ini tidak terlalu siknifikan . dikarenakan performansi dari algoritma artificial immune system dan Naïve baiyes sebagai detector sudah sangat baik. Hampir mendekati angka 100% untuk akurasi. Dan immune buatan pada artificial immune system lebih berpengaruh jelas pada *sms filtering.*

6.1 Kesimpulan dan Saran

6.1 Kesimpulan

1. algoritma *artificial immune system* dapat memaksimalkan sistem SMS *filtering classification*.dikarenakan *detector* mendapat data set baru yang lebih akurat dan kuat , sehingga probabilitas dari *detecto*r untuk menidentifikasi data SMS meningkat lebih baik . dari hasil pengujian di atas dalam teknik pengujian *10-fold validation* ataupun *5-fold validation* . performa sistem selalu menghasilkan angka lebih saat algoritma *Artificial Immune System* di gunakan.
2. Minimum Support pada algoritma apriori tidak terlalu memberi pengaruh besar pada sistem.
3. Dengan menerapkan Apriori untuk feature extraction dapat meningkatkan nilai precision terlebih lagi dengan *minimum support* 3% dan 9%, sehingga sistem lebih tepat dalam memberikan jawaban terhadap informasi yang diminta pengguna untuk klasifikasi SMS. Namun secara keseluruhan algoritma *apriori frequent itemset* tidak membei dampak yang sangat besar pada algoritma *artificial immune system* perbedaan performansi ini masih d bawah angka 1 %, dikarenakan *antibody* yang di buat oleh sistem lebih kuat untuk mendeteksi spam dibandingkan dengan frequent itemset itu sendiri.

6.2 Saran

1. Menggunakan preprocessing yang lebih baik dalam mengatasi kata baku dikarenakan dalam kasus datset berbahasa inggris banyak kata-kata yang telah menjadi kata-kata tidak baku yang hanya muncul dalam SMS

2. Selanjutnya dapat digunakan dataset berbahasa Indonesia agar dapat lebih berguna khususnya untuk menanggulangi masalah SMS spam pada ruang lingkup di Indonesia.

3. Dilakukan analisis *data set* lebih jauh lagi untuk meneliti data data seperti apa yang lebih mudah di olah dan di klasifikasikan sistem .

Daftar Pustaka

[1]. BHATTARAI. 2012.Self-supervised Approach to Comment Spam Detection based on Content Analysis, Department of Computer Science, University of Memphis, Memphis, TN, USA.

[2]. Cormack et al., "Spam filtering for short messages," in Proc. The Sixteenth ACM Conference on Conference on Information And Knowledge Management, November 06-10, 2007,

[3] Erwin. 2009. Analisi Market Basket dengan Algoritma Apriori dan FP-Growth. Jurnal Generic Universitas Sriwijaya. Vol.4 No. 2.

[4] Fayyad,Usama., Piatetsky-Shapiro, Gregory dan Smyth, Padhraic. (1997). From Data Mining to Knowledge Discovery in Databases

[5] Han Jiawei, and M. Kamber. 2006. Data Mining: Concepts and Techniques, Morgan Kaufmann, USA.

[6] Han, J. dan M. Kamber, (2000). Data Mining: Concepts and Techniques.Data Mining: Concepts and Techniques.

[7] Ingo Feinerer, Kurt Hornik, and David Meyer. Text mining infrastructure in R. Journal of Statistical Software, 25(5):1-54, March 2008.

[8] Ishtiaq Ahmed, Donghai Guan, and Tae Choong Chung "SMS Classification Based on Naïve Bayes Classifier and Apriori Algorithm Frequent Itemset"

[9] Manning, C. D., Raghavan, P., & Schütze, H., 2008, Introduction to Information Retrieval, Cambridge University Press, Cambridge.

[10] Moertini, Veronika dan Marsela Yulita. 2007. Analisis Keranjang Pasar Dengan Algoritma Hash-Based Pada Data Transaksi Penjualan

[11] P. Madadi, "Text Categorization based on apriori algorithm's frequent itemsets," MSc. thesis, School of Computer Sceince., Howard R. Hughes College of Engineering, University of Nevada, Las Vegas, 2009.

[12] Raschka, S. (October 14, 2014). Naive Bayes and Text Classification I Introduction and Theory.

[13] Rosso, Paolo. dan Balaguer, Enrique Vallés. (2011) Detection of Near-duplicate User Generated Contents: The SMS Spam Collection

[14] Sari, Eka Novita. 2013. Analisa Algoritma Apriori untuk Menentukan Merek Pakaian yang Paling Diminati pada Mode Fashion Group Medan. ISSN. Vol. IV No. 3

[15] Secker Andrew , "an Artificial Immune System for E-mail Classification. "Computing LaboratoryUniversity of KentCanterbury'