Filtering SMS *Spam* Berdasarkan *Naive Bayes Classifier* dan *Apriori Algorithm Frequent Itemset*

Fahrizal Masyhur[1], **Shaufiah, S.T.,M.T.**[2], **M. Arif Bijaksana, Ir.,M.Tech** [3]
Departemen Teknik Informatika Universitas Telkom, Bandung
[1]**fahrizalmasyhur@gmail.com**, [2]**shaufiah@gmail.com**, [3]**arifbijaksana@gmail.com**

Abstrak :

SMS masih menjadi salah satu pelayanan terpenting dalam media komunikasi. Namun karena SMS murah dan banyak digunakan, maka banyak muncul SMS spam. Untuk menanggulanginya, dalam tugas akhir ini penulis menggunakan Naive Bayes Classifier dan Apriori Algorithm Frequent Itemset. Penulis memilih Naive Bayes Classifier dikarenakan Naive Bayes Classifier dianggap sebagai salah satu algoritma learning yang efektif. Sedangkan Apriori Algorithm Frequent Itemset merupakan algoritma yang cocok untuk menanggulangi data dan transaksi yang banyak. Dalam kasus klasifikasi SMS spam menggunakan Naive Bayes Classifier, setiap kata yang di anggap sebagai data dan setiap sms dianggap transaksi. Hasilnya, dengan menggabungkan Apriori Algorithm Frequent Itemset pada Naive Bayes Classifier, terdapat peningkatan daripada menggunakan klasik Naive Bayes Classifier pada data SMS Corpus v.0.1 Big. Akurasi rata-rata Naive Bayes Classifier sebesar 97.22 sedangkan akurasi rata-rata Naive Bayes Classifier dan Apriori Algorithm Frequent Itemset mengalami peningkatan menjadi 97.33
Kata Kunci: SMS, Naive Bayes classifier, Apriori frequent itemset, spam

Abstract :

SMS is still one of the most important services in the communication media. However, because SMS is cheap and widely used, it appears many spam SMS. To overcome, in this thesis the author uses Naive Bayes classifier and Apriori Algorithm frequent itemset. The author chose Naive Bayes classifier because Naive Bayes classifier is regarded as one of the effective learning algorithms. While the frequent itemset Apriori Algorithm is an algorithm that is suitable to cope with a lot of data and transactions. In the case of SMS spam classification using Naive Bayes classifier, every word that is considered as the data and every sms is considered a transaction. The result, by combining Apriori Algorithm frequent itemset on Naive Bayes classifier, there is increasing rather than using the classic Naive Bayes classifier on the data SMS Corpus v.0.1 Big. Average accuracy Naive Bayes classifier at 97.22 while the average accuracy Naive Bayes classifier and Apriori Algorithm frequent itemset increased to 97.33.
Keywords*: SMS, Naive Bayes classifier, Apriori frequent itemset algoritm, spam*

1. Pendahuluan

Seiring dengan pasar mobile phone yang makin meningkat dan ketergantungan masyarakat akan cell phone, SMS telah menjadi salah satu pelayanan terpenting dalam media komunikasi[13]. Bagi beberapa pengguna cell phone, khususnya di Indonesia, SMS spam dianggap remeh, namun di US terdapat 1.1 milyar SMS spam dan di China setiap user menerima 8,29 SMS spam setiap minggunya[20]. Peningkatan SMS spam sangat signifikan, pada tahun 2013, di region Asia SMS spam meningkat sebesar 30%.

Untuk menanggulangi masalah SMS spamming ini adalah melakukan adalah melakukan filtering SMS dengan klasifikasi teks. Beberapa teknik yang populer untuk klasifikasi teks diantaranya decision trees, Naive Bayes, rule induction, neural network, nearest neighbors, dan Support Vector Machine. Namun dalam klasifikasi SMS ini berbeda dengan klasifikasi pada teks dokumen biasa atau email dikarenakan teks pada SMS sangat pendek (maksimal 160 7-bit karakter), banyak terdapat teks yang disingkat, dan cenderung tidak formal [12].

Penulis memilih Naive Bayes Classifier dikarenakan Naive Bayes Classifier dianggap sebagai salah satu algoritma learning yang efektif. Sedangkan Apriori Algorithm Frequent Itemset merupakan algoritma yang cocok untuk menanggulangi data dan transaksi yang banyak. Dalam kasus klasifikasi SMS spam menggunakan Naive Bayes Classifier, setiap kata yang di anggap sebagai data dan setiap sms dianggap transaksi.

Naïve bayes classifier menggunakan teorema bayes untuk menghitung probabilitas kategori berdasarkan dokumen yang telah diketahui[17]. Apriori algoritm frequent itemset adalah algoritma association rule mining yang akan di integrasikan ke naïve bayes classifier yang bertujuan untuk meningkatkan performansi.

2. SMS Spam dan Ham

Menurut Androulidakis (2012) sebenarnya tidak ada definisi yang baku secara international dan konstitusi untuk SMS Spam [1]. Mardiani dan Tanaliah (2013) mengatakan beberapa negara memiliki definisi sendiri mengenai arti dari SMS spam [1].

Namun intinya, SMS spam adalah sebuah pesan teks (SMS) yang tidak diminta atau tidak diinginkan oleh pengguna yang dikirim ke perangkat seluler, yang biasanya mengandung materi pemasaran (promosi), penipuan, dan lain-lain, seperti email spam [1] . Orang yang melakukan spam disebut spammer. Tindakan spam dikenal dengan nama spamming [1].

Spam dikirimkan melalui pesan teks dengan biaya operasional yang sangat rendah untuk mencapai para pelanggan-pelanggan yang diinginkan. Karena hambatan masuk yang rendah, maka banyak spammers yang muncul dan jumlah pesan yang tidak diminta menjadi sangat tinggi. Akibatnya, banyak pihak yang dirugikan [8].

Sedangkan SMS ham merupakan kebalikan dari SMS spam, yaitu sebuah pesan teks (SMS) yang bisa saja diharapkan dan diinginkan oleh pengguna atau dengan kata lain merupakan sms legal. SMS ham biasanya dikirimkan dari pengguna lain yang dikenal baik melalui content yang dikenal maupun berdasarkan database kontak ponsel pengguna dan tidak menimbulkan gangguan atau kerugian.

3. Naïve Bayes Clasifier

Naïve Bayes Classifier merupakan sebuah metoda klasifikasi yang berakar pada teorema Bayes. Ciri utama dari Naïve Bayes Classifier ini adalah asumsi yang sangat kuat (naïf) akan independensi dari masing-masing kondisi/kejadian. Nilai asumsi ini secara dramatis menyederhanakan dan mengurangi kompleksitasnya[2]. Sebelum menjelaskan Naïve Bayes Classifier ini, akan dijelaskan terlebih dahulu Teorema Bayes yang menjadi dasar dari metoda tersebut.

Pada teorema Bayes, bila terdapat dua kejadian yang terpisah (misalkan A dan B), maka teorema Bayes dirumuskan sebagai berikut:

$$P(A|B) = \frac{P(A)}{P(B)} P(B|A)$$

Untuk keperluan classifier, maka teorema bayes dapat di kembangkan menjadi sebagai berikut :

$$P(Y|X1 \dots Xn) = P(Y) \prod_{i=1}^{n} P(Xi|Y)$$

Klasifikasi adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu obyek[3]. Oleh karena itu, kelas yang ada

tentulah lebih dari satu. Penentuan kelas dari suatu dokumen dilakukan dengan cara membandingkan nilai probabilitas suatu sampel berada di kelas yang satu dengan nilai probabilitas suatu sampel berada di kelas yang lain.

Penentuan kelas yang cocok bagi suatu sampel dilakukan dengan cara membandingkan nilai Posterior. untuk masing-masing kelas, dan mengambil kelas dengan nilai Posterior yang tinggi. Secara matematis klasifikasi dirumuskan sebagai berikut:

$$C_{nb} = argmax_{c \epsilon C} P(Y) \prod_{i=1}^{n} P(Xi|Y)$$

dengan c yaitu variabel kelas yang tergabung dalam suatu himpunan kelas C.

4. Apriori Algorithm

Apriori adalah suatu algoritma yang sudah sangat dikenal dalam menentukan pencarian freuquent item set dengan menggunakan teknik association rule [10]. Algoritma ini pertama kali diperkenalkan oleh Agrawal dan Srikant pada tahun 1994 untuk penentuan frequent itemset untuk aturan asosiasi boolean [16].

Analisa asosiasi atau association rule mining adalah teknik data mning untuk menemukan aturan suatu kombinasi item [16]. 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 [16] 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 asosiasi [16].

Erwin (2009) menambahkan bahwa algoritma ini menggunakan pengetahuan mengenai frequent item set yang telah diketahui sebelumnya untuk memproses informasi selanjutnya [3]. Pada algoritma apriori untuk menentukan kandidat-kandidat yang mungkin muncul akan dilakukan dengan cara memperhatikan minimum support [3].

Ada dua proses utama yang dilakukan dalam algoritma apriori [4], 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.

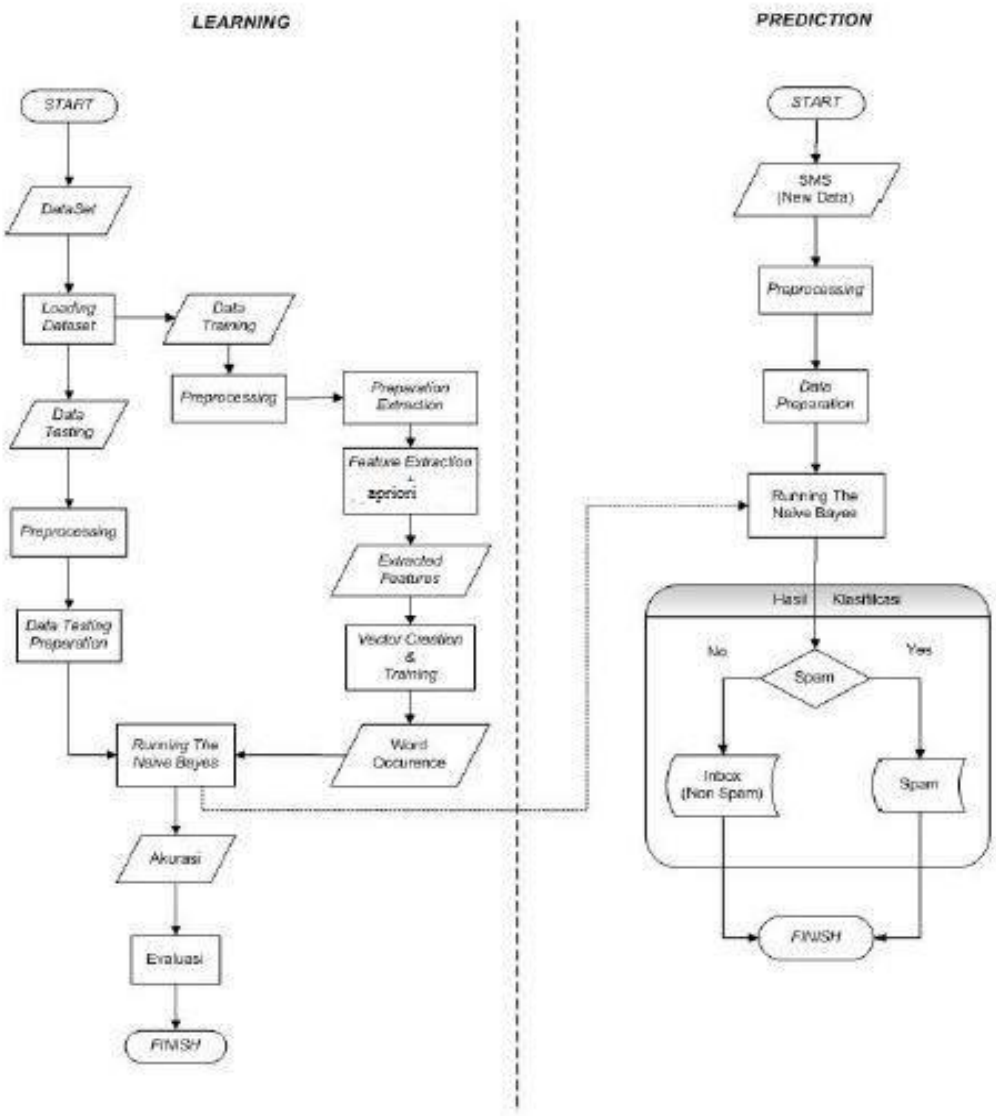
5. Perancangan Sistem

Dalam penelitian ini, peneliti akan membangun sistem untuk memfilter spam SMS dengan menggunakan algoritma Naive Bayes Classifier dan Apriori Algorithm Frequent Itemset.

Pada penelitian ini, proses akan dilakukan dalam dua tahap yaitu Learning dan Prediction. Learning adalah proses pembentukan model klasifikasi dan Prediction adalah proses implementasi model yang dilakukan setelah model diperoleh dengan akurasi yang baik.

Tahap pertama yang akan dilakukan adalah proses Learning. Pada tahap ini, data set awalnya dibagi dulu menjadi dua jenis data, yaitu data latih (data training) dan data uji (data testing) dengan proporsi yang akan disesuaikan pada tahap pengujian nantinya. Kemudian data set dalam format ".txt" disimpan dalam sebuah folder program dan diambil secara otomatis dengan sistem yang dibuat. Dalam tahap awal ini akan dilakukan dalam dua proses yaitu, proses training dan proses testing.

Proses training adalah pengolahan data uji (data training) untuk dijadikan model klasifikasi. Pada tahap proses ini akan dilakukan preprocessing guna mendapatkan data yang bersih, mengurangi volume kosa kata, dam membuat data lebih terstruktur, sehingga data lebih mudah untuk diolah pada sistem nantinya. Setelah itu harus dilakukan feature extraction dengan menggunakan algoritma Apriori dalam menentukan frequent itemset untuk menambahkan item set dan dilanjukan dengan vector creation dan training dengan menghitung frekuensi seluruh kata, termasuk item set baru yang di hasilkan oleh frequent set, dari masing-masing SMS berdasarkan kelasnya dan hasilnya akan ditampilkan dalam table word occurance atau tabel kejadian yang menampilkan jumlah kemunculan setiap kata yang telah diekstrak pada masing-masing kelas. Proses selanjutnya yaitu proses running the naïve bayes system dengan menggunakan data dari table word occurance. Tahap awal dari proses running the naïve bayes system adalah menghitung prior probability yaitu dengan menghitung nilai probabilitas yang diyakini benar sebelum dilakukannya ujicoba. Hasil perhitungan tersebut merupakan model yang kemudian dilakukan pengujian dengan melibatkan data testing yang telah di pre-processing terlebih dahulu. Apabila didapatkan hasil akurasi yang baik, maka tahapan selanjutnya, yaitu proses prediction dapat dijalankan guna melihat hasil klasifikasi yang didapat apabila terdapat sms baru. Secara sistematis, proses diatas dapat dilihat dalam skema pada gambar 3.1 dibawah ini:



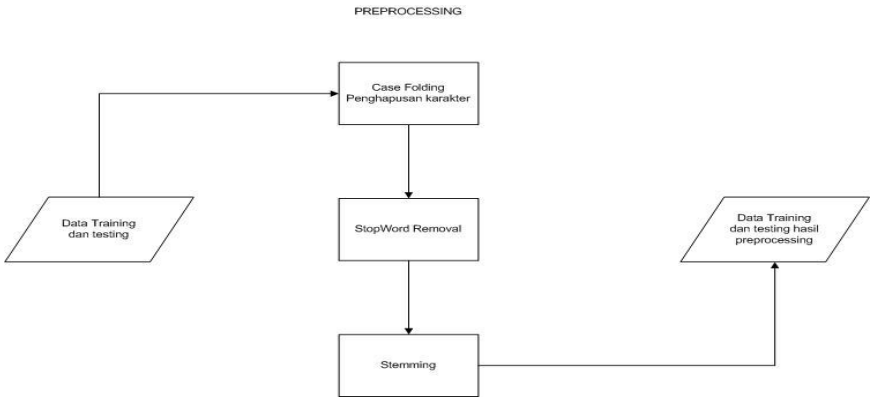
Gambar 3-1 Gambaran Umum Sistem

5.1. Proses Learning

Secara garis besar terdapat dua tahap pemrosesan dalam proses ini, yaitu proses training dan proses testing. Proses training ditujukan untuk pelatihan data agar membentuk model klasifikasi sedangkan proses testing ditujukan untuk menguji hasil klasifikasi berdasarkan model yang telah di proses.

5.1.1. Preprocessing

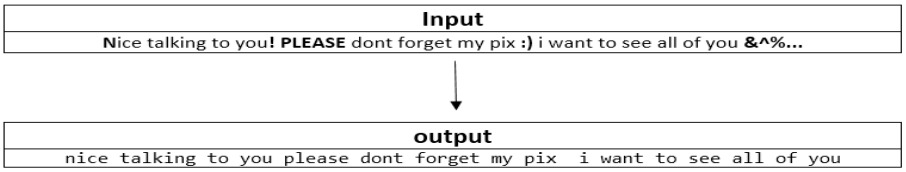
Prepocessing pada masing-masing tipe data, training dan testing, dilakukan secara terpisah untuk memudahkan proses selanjutnya. Secara sistematis, skema prepocessing dapat dilihat pada gambar 5-1 dibawah ini.



Gambar 5-1 Skema *Prepocessing*

a. Case Folding + Hapus Karakter

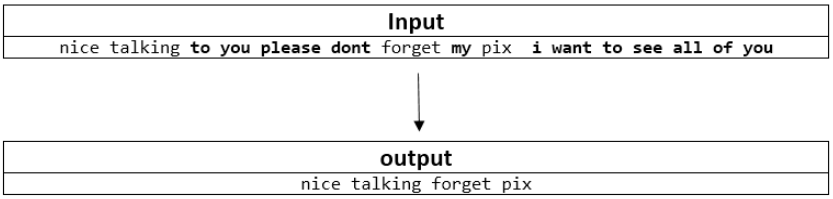
Pada tahap ini seluruh teks diubah menjadi huruf kecil untuk menyeragamkan dara serta menghapus karakter selain huruf, angka, dan juga menghapus tanda baca.



Gambar 5-2 Proses *Case Folding* + Hapus Karakter

b. Stopward Removal

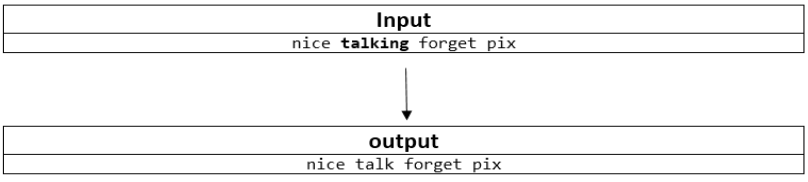
Proses ini ditujukan untuk menghapus kata-kata yang termasuk dalam kamus kata *stopwords.* Berikut ini gambaran proses *stopwords.*



Gambar 5-3 Proses *Stopword Removal*

c. *Stemming*

Pada proses ini akan dilakukan eliminasi kata imbuhan yang terdapat pada dataset sehingga kata kembali ke bentuk dasarnya. Hal ini dimaksudkan agar dapat mengurangi variasi kata yang seharusnya memiliki arti sama namun memiliki bentuk imbuhan yang berbeda.



Gambar 5-4 Proses *Stemming*

5.1.2. Feature Extraction

Pada data training dilakukan feature extraction dengan melibatkan algoritma Apriori untuk mendapatkan frequent itemset. Pertama data dikelompokkan berdasarkan kelasnya, spam atau ham. Setelah itu sistem akan membentuk setiap kata kata unik dari masing masing kelas untuk dijadikan parameter perhitungan probabiliti naive bayes yang dimasukkan dalam parameter "key". Lalu dengan menjalankan apriori, akan menghasiskan frequent item set yang akan dianggap sebagai kata unik baru dan akan di tambahkan juga ke dalam parameter "key".

5.1.3. Vector Creation and Training

Pada proses ini dilakukan perhitungan untuk setiap kata yang telah diekstrak pada masing-masing kelas. Untuk mempermudah perhitungan, maka word occurrence table. Word occurrence table merupakan tabel yang berisi kemunculan setiap kata di spam dan di ham. 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 . Sebagai contoh :

word	love	play	bx420	vary	video	hope	dosomething	uup	ave	wid
ham	1	0	0	0	0	1	1	1	0	0
spam	0	1	1	1	1	0	0	0	1	1

Gambar 5-5 Proses *word occurrence table*

5.1.4. Running the Naïve Bayes Sistem

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 pada proses preparation untuk dilakukan klasifikasi.

5.2. Proses Prediction

Setelah didapatkan akurasi yang baik berdasarkan proses learning, maka selanjutnya model (hasil perhitungan dari word occurrence) dapat digunakan untuk proses prediksi klasifikasi SMS baru. Pada proses prediksi ini data SMS baru tetap melewati tahap preprocessing dan preparation seperti yang diterapkan pada data testing untuk kemudian dilakukan klasifikasi seperti pada sub bab 3.1.1.4 apakah SMS tersebut termasuk spam atau ham.

6. Pengujian dan Analisis

Untuk mengidentifikasi performansi dari sistem SMS filtering yang dibangun pada tugas akhir ini, digunakan dataset yang berasal 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.

6.1. Analisis Pengujuian pertama

Pengujian pertama dilakukan dengan hanya menerapkan algoritma Naive Bayes tanpa melibatkan algoritma Apriori untuk feature extraction. Hal ini dilakukan untuk membuktikan bahwa Naive Bayes merupakan salah satu algoritma terbaik untuk klasifikasi teks dan selanjutnya dapat digunakan sebagai pembanding untuk proses pengujian dengan melibatkan algoritma Apriori.

Tabel 6-1. Hasil Pengujian Pertama

NO.	SMS *SPAM* CORPUS BIG V.0.1		Naive Bayes			
	Data Training	Data Testing	Akurasi			
			Precision	*Recall*	*Fmeasure*	*Accuracy (%)*
1	BCDEFGHIJ	A	0,884	0,766	0,821	92,3
2	ACDEFGHIJ	B	1	0,933	0,965	98,46
3	ABDEFGHIJ	C	1	0,933	0,965	98,46
4	ABCEFGHIJ	D	0,967	1	0,983	99,23
5	ABCDFGHIJ	E	0,967	1	0,983	99,23
6	ABCDEGHIJ	F	0,928	0,866	0,896	95,38
7	ABCDEFHIJ	G	0,967	1	0,983	99,23
8	ABCDEFGIJ	H	0,933	0,933	0,933	96,92
9	ABCDEFGHJ	I	0,964	0,9	0,931	96,92
10	ABCDEFGHI	J	0,931	0,9	0,915	96,15
		Average:	0,954	0,923	0,937	97,22

Berdasarkan hasil pengujian terbukti bahwa Naive Bayes merupakan salah satu algoritma terbaik untuk klasifikasi teks dengan menunjukan akurasi rata-rata sebesar 97,22%. Hasil algoritma Naive Bayes sangat dipengaruhi dengan jumlah data training yang digunakan, karena algoritma ini menggunakan prinsip probabilitas sehingga jumlah dokumen dan jumlah fitur sangat diperhitungkan.

6.2. Hasil Pengujian Kedua

Pengujian kedua naïve bayes dikolaborasikan dengan algoritma Apriori untuk membantu dalam proses feature extraction. Pada pengujian ini digunakan 3 parameter minimum support yaitu 4%, 6%, dan 8%.

NO.	SMS *SPAM* CORPUS BIG V.0.1		Naive Bayes + FP Growth				
	Data Training	Data Testing	*minsup*	Akurasi			
				Precision	*Recall*	*Fmeasure*	*Accuracy (%)*
1	BCDEFGHIJ	A	8%	0,884	0,766	0,821	92,3
			6%	0,884	0,766	0,821	92,3
			4%	0,884	0,766	0,821	92,3
2	ACDEFGHIJ	B	8%	1	0,933	0,965	98,46
			6%	1	0,933	0,965	98,46

			4%	1	0,933	0,965	98,46
3	ABDEFGHIJ	C	8%	1	0,933	0,965	98,46
			6%	1	0,933	0,965	98,46
			4%	1	0,933	0,965	98,46
4	ABCEFGHIJ	D	8%	0,967	1	0,983	99,23
			6%	0,967	1	0,983	99,23
			4%	0,967	1	0,983	99,23
5	ABCDFGHIJ	E	8%	0,967	1	0,983	99,23
			6%	0,967	1	0,983	99,23
			4%	0,967	1	0,983	99,23
6	ABCDEGHIJ	F	8%	0,928	0,866	0,896	95,38
			6%	0,928	0,866	0,896	95,38
			4%	0,928	0,866	0,896	95,38
7	ABCDEFHIJ	G	8%	1	1	1	100
			6%	1	1	1	100
			4%	1	1	1	100
8	ABCDEFGIJ	H	8%	0,933	0,933	0,933	96,92
			6%	0,933	0,933	0,933	96,92
			4%	0,933	0,933	0,933	96,92
9	ABCDEFGHJ	I	8%	0,964	0,9	0,931	96,92
			6%	0,964	0,9	0,931	96,92
			4%	0,964	0,9	0,931	96,92
10	ABCDEFGHI	J	8%	0,931	0,9	0,915	96,15
			6%	0,931	0,9	0,915	96,15
			4%	0,964	0,9	0,931	96,92
		Average:		0,958	0,923	0,939	97,33

Berdasarkan pengujian kedua, terlihat rata-rata semua paramater uji yang digunakan mengalami kenaikan di bandingkan dengan hasil pengujian pertama. Precision mengalami kenaikan dari 0,954 menjadi 0,958. Recall tidak mengalami kenaikan, tetap di angka 0,923. F-measure mengalami kenaikan dari 0,937 menjadi 0,939. Akurasi yang dihasilkan pun membaik dari 97,22% menjadi 97,33%.

Kenaikan parameter uji tersebut merupakan hasil dari penerapan Apriori Algorithm Frequent Itemset. Itemset yang dihasilkan oleh Apriori Algorithm Frequent Itemset merupakan Itemset dari data set sms spam. Sedangkan data set sms ham tidak di hasilkan itemset baru. Data set sms ham baru menghasilkan itemset baru jika penulis menerap Apriori Algorithm Frequent Itemset dengan minimum support

7. Kesimpulan dan Saran

Berdasarkan analisis terhadap pengujian yang dilakukan dalam tugas akhir ini, maka dapat disimpulkan bahwa :

1. Dari kedua metode yang digunakan, performansi kedua metode sama baiknya untuk pengklasifikasian sms dengan akurasi rata-rata diatas 90%. Penerapan *apriori frequent itemset algorithm* ke naïve bayes *classifier* berhasil menambah rata-rata akurasi sebesar 0,11%, *precision* sebesar 0,004, dan *f-measure* sebesar 0.002.

2. Perubahan minimum suport pada apriori berpengaruh pada *frequent item* set baru yang di hasilkan apriori, makin kecil *minimum support*-nya makin banyak *itemset* yang di *generate*.

3. Akurasi rata-rata terbaik didapat saat menggunakan apriori dengan minimum support 4% dengan rata-rata akurasi sebesar 97,38%.

Saran:

Beberapa saran dari penulis untuk pengembangan lebih lanjut adalah sebagai berikut :

1. Menggunakan *preprocessing* yang lebih baik dalam mengatasi kata baku dikarenakan dalam kasus *dataset* 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.

8. Daftar Pustaka

[1] A.W, E., Mardiani, & Tinaliah. (2013). Penerapan Metode Naive Bayes Untuk Sistem Klasifikasi SMS Pada Smartphone Android

[2] Ahmed, Guan, dan Chung. (2014). SMS Classification Based on Naive Bayes Classifie and Apriori Algorithm Frequent Itemset. *International Journal of Machine Learning and Computing*. Vol 4: No. 2.

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

[4] Han Jiawei, and M. Kamber. (2006). *Data Mining: Concepts and Techniques, Morgan Kaufmann,* USA.

[5] Ibrahim. (2011). Pengembangan sistem infomasi monitoring TA berbasis SMS di Fasilkom UNSRI. *JUSI*. Vol 1. No. 2.

[6] Khang, B. 2002. *Trik pemrograman aplikasi berbasis SMS,* Jakarta: Elex Media Komputindo

[7] M. Taufiq, M. F. A. Abdullah, K. Kang, and D. Choi. 2010. A survey of preventing, blocking And filtering short message services (SMS) Spam. *International Conference on Computer and Electrical Engineering (IACSIT).* Vol. 1, Hal: 462-466.

[8] Makhtidi, K. 2012. *Sistem Spam Detector Untuk SMS Berbahasa Indonesia Pada Smartphone Android. Bogor*: Departemen Ilmu Komputer,Fakultas matematika dan Ilmu Pengetahuan Alam Institut Pertanian Bogor

[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 Apotek. Jurusan Ilmu Komputer, Universitas Katolik Parahyangan, Bandung.

[11] Motoda, H., & Liu, H. (t.thn.). Feature Selection, Extraction and Construction.

[12] Natalius Samuel. (2011). Metoda Naïve Bayes Classifier dan Penggunaannya pada Klasifikasi Dokumen. .

[13] P. J. Denning. (1982). Electronic Junk. *ACM Communications.* Vol. 25, no. 3, Mar. 1982 Hal: 163–165.

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

[15] Rish, I., (2006). *An empirical study of The Naive Bayes Classifier, International Joint Conference on Artificial Intelligence, California*.

[16] Sari, Eka Novita. 2013. Analisa Algoritma Apriori untuk Menentukan Merek Pakaian yang Paling Diminati pada Mode Fashion Group Medan. *Pelita Informatika Budi Darma*. Vol. IV No. 3.

[17] Shirani-Mehr, H. (t.thn.). SMS Spam Detetection using Machine Learning Approach.

[18] Sunardi, M., & Listiyono, H. (2009). Aplikasi SMS Gateway. *Jurnal Teknologi Informasi DINAMIK*, XIV(1). Hal: 30-34

[19] Tan, P. N., Steinbach, M., & Kumar, V., *2006*, Introduction to Data Mining, Pearson Education, Boston.

[20] W. Qian, H. Xue, and W. Xiayou/ (2009). *Studying of Classifying Junk Messages Based on The Data Mining*. International Conference on Management and Service Science, IEEE. Hal: 1-4.

[21] Witten, I.H., Don, K.J., Dewsnip, M. and Tablan, V. (2004) *Text mining in a digital library. Journal of Digital Libraries.* Volume 4(1). Hal:. 56-59.