PERBANDINGAN KINERJA VARIASI NAÏVE BAYES MULTIVARIATE BERNOULLI DAN NAÏVE BAYES MULTINOMIAL DALAM PENGKLASIFIKASIAN DOKUMEN TEKS

Widyawati[1], Sutanto[2]
Program Studi Teknik Informatika, Fakultas Ilmu Komputer
Universitas Banten Jaya
Jl. Ciwaru II No. 73, Kota Serang–Banten 42117, Indonesia
email: widyawati.astrabuwono@gmail.com[1], sutanto@unbaja.ac.id[2]

ABSTRACT

Classification of text documents with large amounts will be a job that requires a lot of time, effort and cost of having to read text documents and then categorize them manually, therefore the automatic classification of text documents is needed. The algorithm developed is K-Nearest-Neighbor (KNN), Naïve Bayes, Support Vector Machine (SVM), Decision Tree (DT), Neural Network (NN) and Maximum Entropy. The algorithm used as the object of research is a variation of the Naïve Bayes algorithm, the Naïve Bayes Multivariate Bernoulli and the Naïve Bayes Multinomial. This study discusses whether there are differences between the Algorithms. The Naïve Bayes Multivariate Bernoulli algorithm and the Naïve Bayes Multinomial can be seen from the value of the agreement and the speed of the process of classifying text documents, as well as more information about the process of processing requests that are getting more and more requested. While the highest value using the non-stemming Naïve Bayes Bernoulli method is 71.33%, and the fastest processing time is required using the non-stemming Naïve Bayes method which requires 0.12 seconds processing time.

Keywords: Text Classification, Multinomial, Multivariate, Bernoulli, Naïve Bayes

Pendahuluan

Seiring dengan bertambahnya jumlah informasi berbasis teks yang berasal dari dokumen teks elektronik dan *world wide web* maka semakin meningkatnya kebutuhan akan pencarian dokumen teks yang efektif dan efisien. Salah satu cara yang dapat dilakukan adalah dengan mengklasifikan dokumen teks kedalam kelas-kelas yang sudah ditetapkan berdasarkan pola kata atau kalimat yang terdapat dalam dokumen tersebut ataupun juga berdasarkan atribut lainnya seperti jenis dokumen, pengarang dan tahun terbit.

Namun pengklasifikasian dokumen teks dengan jumlah besar akan menjadi pekerjaan yang menghabiskan banyak waktu, tenaga serta biaya bagi manusia jika

harus membaca keseluruhan dokumen teks kemudian mengkategorikanya secara manual. Untuk menjawab permasalahan tersebut dibutuhkan pengklasifikasian dokumen teks secara otomatis (*automatic document classification*).

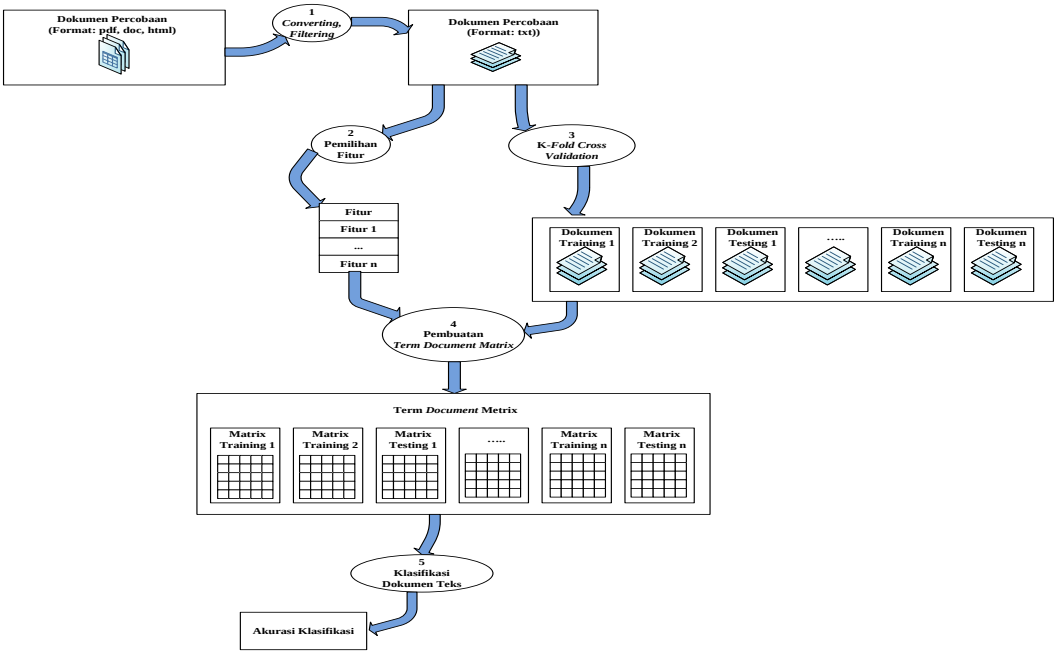
Dalam melakukan pengklasifikasian dokumen teks secara otomatis terdapat banyak algoritma yang dikembangkan. Menurut Ting, beberapa algoritma yang dikembangkan untuk menyelesaikan permasalahan klasifikasi dokumen adalah k-nearest-neighbor (KNN), naïve bayes, Support Vector Machine (SVM), desicion tree (DT), neural network (NN) dan maximum entropy. Dari algoritma tersebut peneliti menemukan permasalahan dalam memilih algoritma untuk mendapatkan kinerja yang lebih baik pada pengklasifikasian dokumen teks secara otomatis. Sehingga dalam penelitian ini, peneliti memilih algoritma yang akan dijadikan objek penelitian adalah algoritma Naïve Bayes. Menurut Wu, algoritma Naïve Bayes ini memiliki kelebihan karena cukup handal serta sederhana sehingga mudah ketika diaplikasikan di *dataset* (bahasa Indonesia:"set data") yang berukuran besar, selain itu algoritma ini termasuk kedalam kumpulan algoritma yang sering digunakan dalam *data mining*. Menurut McCallum dan Nigam, terdapat beberapa variasi dalam algoritma Naïve Bayes yang bisa digunakan dalam pengklasifikasian dokumen teks yaitu Naïve Bayes Multivariate Bernoulli dan Naïve Bayes Multinomial. Penelitian ini dimaksudkan untuk melakukan perbandingan kinerja antara Naïve Bayes Multivariate Bernoulli dan Naïve Bayes Multinomial.

Metode Penelitian

Metode penelitian ini dilakukan dengan beberapa tahap seperti Penetapan gambaran umum proses pengklasifikasian dokumen teks, persiapan data, pengolahan data, dan analisis hasil.

1. Gambaran Umum Pengklasifikasian Dokumen Teks

Berikut ini merupakan kerangka pemikiran yang dilakukan dalam penelitian:



Gambar 1. Gambaran Umum Proses Klasifikasi Dokumen Teks

2. Persiapan Data

Data yang didapatkan tidak bisa langsung dijadikan bahan percobaan, dibutuhkan beberapa proses yang diperlukan agar data tersebut siap uji. Proses tersebut meliputi *converting* dan *filtering*, dimana dalam proses *converting* dilakukan proses perubahanekstensi dokumen teksyang awalnya berekstensi pdf, doc ataupun html menjadi berekstensi txt, hal ini dilakukan agar memudahkan perangkat lunak yang digunakan (Weka) dalam membaca dokumen teks tersebut. Proses selanjutnya adalah proses *filtering*, yaitu proses penghilangan *stopwords*dan tanda baca. *Stopwords* adalah kata-kata yang tidak dipakai dalam pemrosesan bahasa alami, daftar *stopwords* yang digunakan pada penelitian ini bisa dilihat pada

lampiran. Dalam proses *filtering*, selain dilakukan penghilangan *stopwords* dan tanda baca, dilakukan juga proses *stemming* yang berupa perubahan kata-kata yang terdapat pada dokumen teks menjadi kata dasar (contohnya adalah *testing* → *test* dan *eats* → *eat*) hal ini dilakukan agar tidak terjadi pengulangan kata yang mempunyai arti yang sebenarnya sama (contoh *testing*, *tester* dan *tested*→*test dan eating, eats, eater* → *eat.*

3. Metode Pengolahan Data

a. Pemilihan Fitur

Pemilihan fitur dilakukan dengan menghitung terlebih dahulu kata-kata yang muncul pada data dokumen teks. Setelah itu, data tersebut akan diurutkan sesuai dengan frekuensi kemunculan data dimana kata yang diurutkan hanya merupakan kata yang lolos nilai minimum frekuensi, hal ini ditujukan untuk mengurangi kata-kata yang jarang muncul dan tidak memiliki makna yang berarti sehingga bisa menghemat waktu dalam proses.

b. K-*fold Cross Validation*

Pada penelitian ini dalam proses verifikasinya menggunakan metode k-*fold cross validation*, metode ini bertujuan agar data yang dijadikan bahan dalam proses pengklasifikasian dokumen teks tidak bias. Metode ini membagi dokumen kedalam k bagian. Dengan menggunakan metode ini akan dilakukan percobaan sebanyak k buah dimana dalam tiap percobaan akan menggunakan satu buah data testing dan k-1 bagian menjadi data training, selain itu dalam tiap proses percobaan, data *testing* akan ditukar dengan satu data *training* yang ada sehingga untuk dalam setiap percobaan akan menggunakan data *testing* yang berbeda.

c. *Term Document Matrix*

Term document matrix merupakan respresentasi kumpulan dokumen yang akan digunakan untuk melakukan proses klasifikasi dokumen teks. Pada term document matrix sebuah dokumen direspresentasikan sebagai kumpulan fitur dan dapat diilustrasikan sebagai $d_i = [f_{i1}, f_{12}, ..., f_{ij}]$ dengan d_i merupakan

dokumen ke-i dan f_{ij} merupakan nilai kemunculan fitur ke-j pada dokumen d_i. Matriks ini akan berisi nilai fitur dimana fitur yang digunakan adalah *frequency* dan *presence*. Baris pada *term document matrix* merupakan data dokumen, sedangkan kolom dari *term document matrix* merupakan fitur yang digunakan.

4. Algoritma Klasifikasi

Naïve Bayes merupakan salah satu algoritma *Machine Learning* yang menggunakan konsep probabilitas. Algoritma ini melakukan klasifikasi dengan menghitung nilai probabilitas p(h|x), yaitu probabilitas kelas h jika diketahui suatu b, berdasarkan algoritma NaïveBayes.

Proses klasifikasi dapat dilakukan dengan menentukan nilai suatu kelas $h \in H$ dari suatu dokumen $x \in X$ dengan $H = \{h_1, h_2, h_3, \cdots, h_p\}$ dan $X = \{x_1, x_2, x_3, \cdots, x_q\}$. Penentuan kelas dalam klasifikasi dokumen tersebut dilakukan dengan cara memilih nilai maximum dari p(h|x), berdasarkan distribusi probabilitas $P = \{p(h|x) \mid h \in H \text{ dan } x \in X\}$. Suatu dokumen x ke i dapat dipresentasikan sebagai vector dan nilai-nilai fitur yang ada pada dokumen tersebut sehingga nilai-nilai fitur yang ada pada dokumen tersebut sehingga $x = \{f_{i1}, f_{i2}, f_{i3}, \cdots, f_{in}\}$. Nilai dari elemen tiap vektor merupakan nilai untuk fitur f_j pada himpunan fitur $F = \{f_1, f_2, f_3, \cdots, f_n\}$ dengan f_{ij} merupakan nilai dari fitur ke j pada dokumen x ke i. Berdasarkan algoritma NaïveBayes berikut ini merupakan persamaan perhitungan nilai dari probabilitas p(h|x) :

$$\text{Posterior} = \frac{\text{Likelihood . Prior}}{\text{Evidence}}$$

$$p(h|x) = \frac{p(x|h) \,.\, p(h)}{p(x)}$$

Keterangan:

p(h|x) = Nilai posterior atau probabilitas kata h dari kelas x

p(x|h) = Nilai likelihood atau probabilitas kemunculan kelas x untuk kata h

p(h) = Nilai prior atau probabilitas kata h

p(x) = Nilai evidence atau probabilitas kelas x

Menurut McCallum dan Nigam, terdapat beberapa variasi dalam algoritma Naïve Bayes yang bisa digunakan dalam pengklasifikasian dokumen teks yaitu Naïve Bayes Multivariate Bernoulli dan Naïve Bayes Multinomial yang akan dijadikan sebagai algoritma penelitian ini. Yang akan dijelaskan pada tahapan diskusi.

Diskusi

1) Naïve Bayes Multivariat Bernoulli

   a) Membuat tabel kelas untuk keperluan dokumen *testing*

   Berikut ini merupakan proses penentuan kelas untuk dokumen *testing*:

Tabel 1. Tabel Keperluan Dokumen *Testing*

Dokumen	**Kelas**	**Fitur (*frequency*)**
Dokumen 1	Electronic	Wireless (2), Comunication (3), Signal (4), Broken (1)
Dokumen 2	Medical	Aches (3), Aspirin (1), Surface (3), Head (2)
Dokumen 3	Medical?	Broken (4), Surface (2), Arm (2), Accident (1)

   pada tabel 1 akan dicari penyesuaian klasifikasi dokumen teks dari dokumen 3.

   b) Membuat *Term Dokumen Matriks*

   Dibawah ini merupakan tabel *Term Dokumen Matriks,* perbedaan antara naïve beyes multinomial dengan model naïve bayes terletak pada proses *Term Document Matrix* dengan memberikan nilai "0" dan "1". Dimana nilai "0" diberikan jika tidak terdapat fitur dalam dokumen tersebut begitupun sebaliknya,

Tabel 2. Tabel *Term Document Matriks*

	wireless	Comunication	Signal	Broken	Aches	Aspirin	Surface	Head	Arm	accident
dokumen 1	1	1	1	1	0	0	0	0	0	0
dokumen 2	0	0	0	0	1	1	1	1	0	0
dokumen 3	0	0	0	1	0	0	1	0	1	1

c) Membuat tabel probabilitas

Berikut ini merupakan model probabilitas yang terbentuk :

Tabel 3. Model Probabilitas yang Terbentuk

Kelas	p(xi)	p(hkj|xi)									
		wireless	Comunication	Signal	Broken	Aches	Aspirin	Surface	Head	Arm	accident
Electronic	0.5	0.100	0.100	0.100	0.100	0.050	0.050	0.050	0.050	0.050	0.050
Medical	0.5	0.053	0.053	0.053	0.053	0.105	0.105	0.105	0.105	0.053	0.053

d) Penentuan kategori untuk dokumen *testing*

$$a^* = \arg \max_{x_i \in X} \prod_k p(h_{kj}|x_i).p(x_i)$$

P(“Electronic”|“dokumen 3”) = 0.00000625

P(“Medical”|“dokumen 3”) = 0.00000767

e) *Verifikasi* Hasil

Berdasarkan hasil perhitungan diatas didapatkan nilai P(“Electronic”|“dokumen 3”) sebesar 0.00000625 sedangkan P(“Medical”|“dokumen 3”) sebesar 0.00000767. Karena nilai dari P(“Medical”|“dokumen 3”)>P(“Electronic”| “dokumen 3”) maka dokumen 3 masuk kedalam kelas Medical.

2) Naïve Bayes Multinomial

a). Membuat tabel kelas untuk keperluan dokumen *testing*

Berikut ini merupakan proses penentuan kelas untuk dokumen *testing*:

Tabel 4. Tabel Keperluan Dokumen *Testing*

Dokumen	Kelas	Fitur (*frequency*)
Dokumen 1	Electronic	Wireless (2), Comunication (3), Signal (4), Broken (1)
Dokumen 2	Medical	Aches (3), Aspirin (1), Surface (3), Head (2)
Dokumen 3	Medical?	Broken (4), Surface (2), Arm (2), Accident (1)

pada tabel 4 akan dicari penyesuaian klasifikasi dokumen teks dari dokumen 3.

b). Membuat *Term Dokumen Matriks*

Dibawah ini merupakan tabel *Term Dokumen Matriks*:

Tabel 5. Tabel *Term Document Matriks*

	wireless	Comunication	Signal	Broken	Aches	Aspirin	Surface	Head	Arm	accident
dokumen 1	2	3	4	1	0	0	0	0	0	0
dokumen 2	0	0	0	0	3	1	3	2	0	0
dokumen 3	0	0	0	4	0	0	2	0	2	1

c). Membuat tabel probabilitas

Berikut ini merupakan model probabilitas yang terbentuk:

Tabel 6. Model Probabilitas yang Terbentuk

Kelas	p(xi)	p(hkj|xi)									
		wireless	Comunication	Signal	Broken	Aches	Aspirin	Surface	Head	Arm	accident
Electronic	0.5	0.150	0.200	0.250	0.100	0.050	0.050	0.050	0.050	0.050	0.050
Medical	0.5	0.053	0.053	0.053	0.053	0.211	0.105	0.211	0.158	0.053	0.053

d). Penentuan kategori untuk dokumen *testing*

$$a^* = \arg \max_{x_i \in X} \prod_k p(h_{kj}|x_i).p(x_i)$$

P(“Electronic”|“dokumen 3”) = 0.00000625

P(“Medical”|“dokumen 3”) = 0.00001535

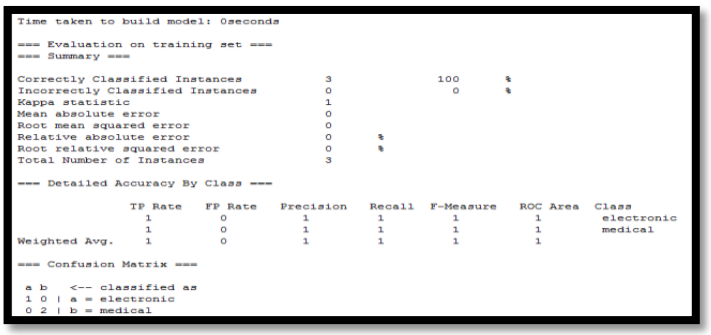
e). *Verifikasi* Hasil

Berdasarkan hasil perhitungan diatas didapatkan nilaiP(“Electronic”|“dokumen 3”) sebesar 0.00000625 sedangkan P(“Medical”|“dokumen 3”) sebesar 0.00001535. Karena nilai dari P(“Medical”|“dokumen 3”)>P(“Electronic”|“dokumen 3”) maka dokumen 3 masuk kedalam kelas Medical.

3) Pengujian Hasil Menggunakan Perangkat Lunak Weka

Dengan menggunakan data yang sama, dilakukan percobaan untuk membuktikan kebenaran dari verifikasi hasil perhitungan Algoritma Naïve Bayes Multivariate Bernoulli dan Naïve Bayes Multinomial perangkat lunak Weka 3.6.5. Hasil dari pengujian menggunakan perangkat luna Weka dapat dilihat pada Gambar 2 dan Gambar 3.

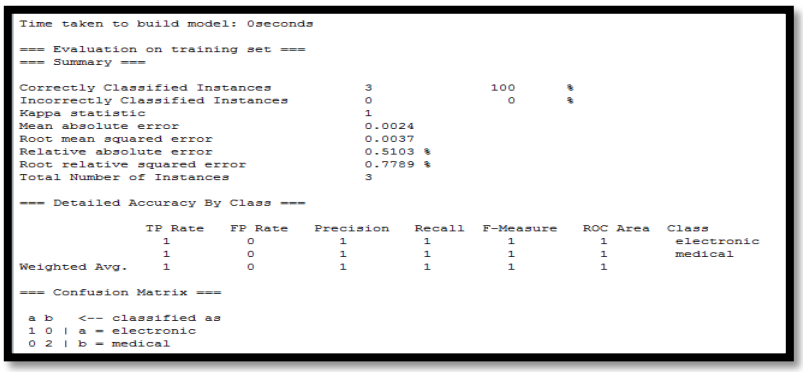
Presence:



```
Time taken to build model: 0seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances           3               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances                3

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1         1         1          1        electronic
                 1         0          1         1         1          1        medical
Weighted Avg.    1         0          1         1         1          1

=== Confusion Matrix ===

 a b   <-- classified as
 1 0 | a = electronic
 0 2 | b = medical
```

Gambar 2. Hail Pengujian Naïve Bayes Multivariate Bernoulli

***Frequency*:**



```
Time taken to build model: 0seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances           3               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0024
Root mean squared error                  0.0037
Relative absolute error                  0.5103 %
Root relative squared error              0.7789 %
Total Number of Instances                3

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1         1         1          1        electronic
                 1         0          1         1         1          1        medical
Weighted Avg.    1         0          1         1         1          1

=== Confusion Matrix ===

 a b   <-- classified as
 1 0 | a = electronic
 0 2 | b = medical
```

Gambar 3. Hasil Pengujian Naïve Bayes Multivariate Bernoulli

Berdasarkan hasil pengujian pada Gambar 2 dan Gambar 3, dapat diketahui bahwa hasil pengujian menggunakan perhitungan manual ataupu n menggunakan perangkat lunak Weka menghasilkan kesimpulan yang sama bahwa dalam data percobaan tersebut terbagi menjadi dua dokumen teks dengan kelas *medical* dan satu dokumen teks dengan kelas *electronic*.

Bab ini merupakan bab yang menjelaskan mengenai pembahasan dari hasil penelitian mengenai implementasi perbandingan kinerja variasi Algoritma Naïve Bayes Multivariate Bernoulli dan Naïve Bayes Multinomial dalam melakukan pengklasifikasian dokumen teks.

4) Hasil Penelitian Berdasarkan Aspek *Stemming*

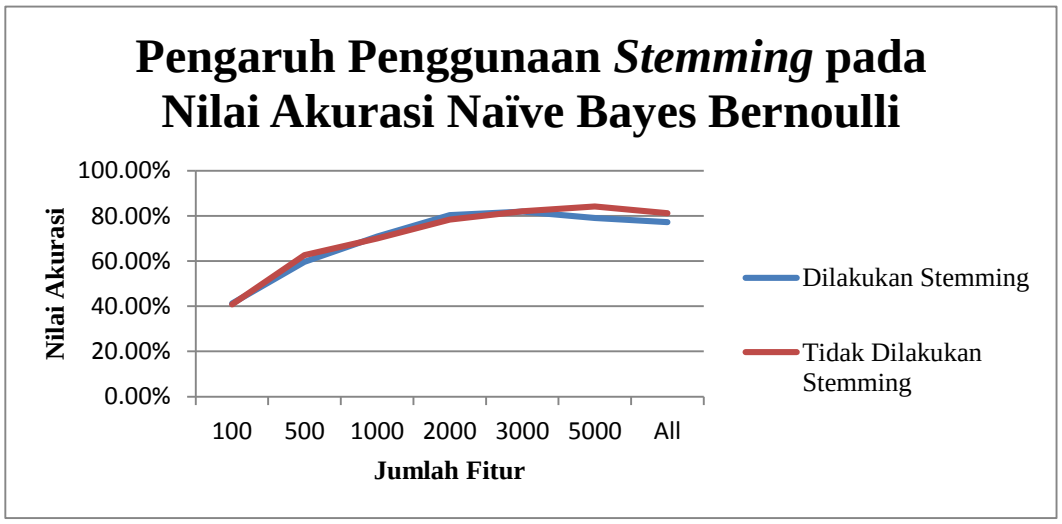
Tabel 7 berikut ini merupakan hasil penelitian dalam bentuk persentase berdasarkan aspek dilakukannya atau tidak dilakukannya proses *Stemming* dari tiap metode serta fitur yang digunakan (pengaruh pengunaan atau tidak menggunakan proses *Stemming* pada nilai akurasi setiap metode).

Tabel 7. Hasil Penelitian Berdasarkan Aspek *Stemming*

Metode	Fitur							Persentase Rata-rata
	100	**500**	**1000**	**2000**	**3000**	**5000**	***All***	
Naïve Bayes Bernoulli *with Stemming*	**41.27%**	59.72%	**70.91%**	**80.32%**	81.94%	79.08%	**77.29%**	70.08%
Naïve Bayes Bernoulli *Non Stemming*	40.84%	**62.66%**	69.98%	78.31%	**82.07%**	**84.24%**	81.22%	**71.33%**
Naïve Bayes Multinomial *with Stemming*	36.50%	53.87%	67.72%	78.42%	81.90%	82.69%	82.89%	69.14%
Naïve Bayes Multinomial *Non Stemming*	36.85%	56.79%	68.20%	77.21%	81.87%	83.51%	**84.06%**	69.78%

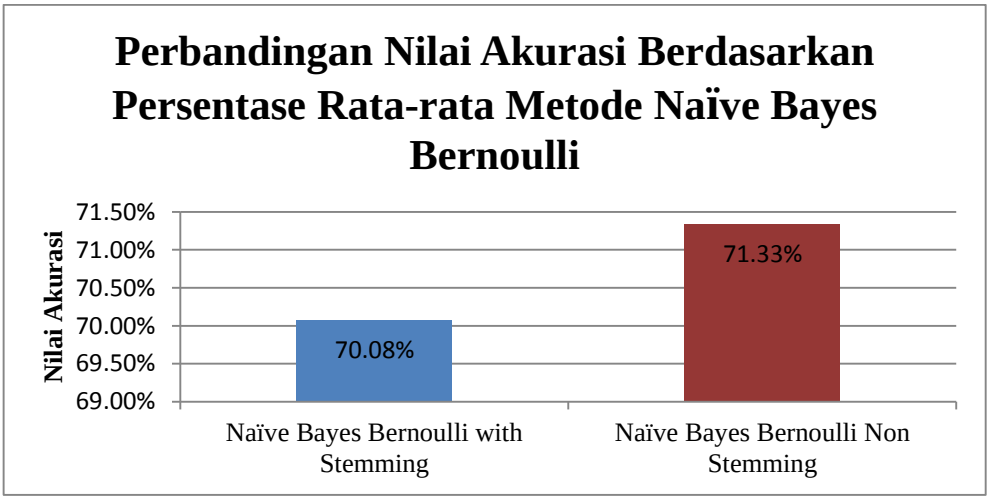
Berdasarkan Tabel 7 dilihat bahwa hasil penelitian berdasarkan aspek *Stemming* tiap fitur yang digunakan memberikan pengaruh yang berbeda dari setiap metode. Dari Tabel 7 menunjukan bahwa keseluruhan variasi metode yang digunakan bahwa metode yang memiliki nilai akurasi terbaik adalah menggunakan

metode Naïve Bayes Bernoulli *non stemming* jika dilihat dari persentase tertinggi nilai akurasi setiap fitur yang digunakan.



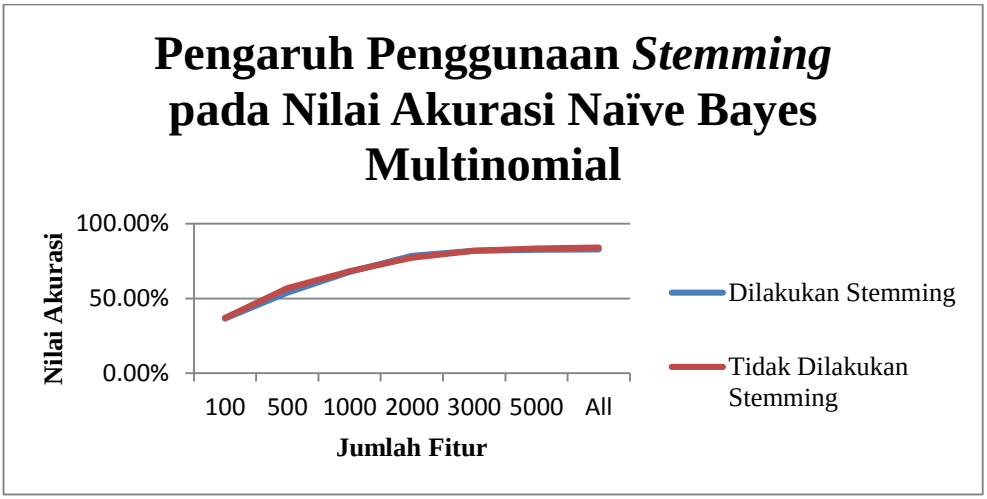
Gambar 4. Pengaruh Penggunaan *Stemming* pada Nilai Akurasi Naïve Bayes Bernoulli

Berdasarkan Gambar 4 menunjukan bahwa ada pengaruh dari penggunaan proses *Stemming* pada nilai akurasi Naïve Bayes Bernoulli setiap fitur yang digunakan, dilihat dari bentuk grafik yang semakin meningkat dari setiap perubahan dalam menggunakan fitur yang berbeda dan akan mengalami penurunan pada saat penggunaan *all* fitur.



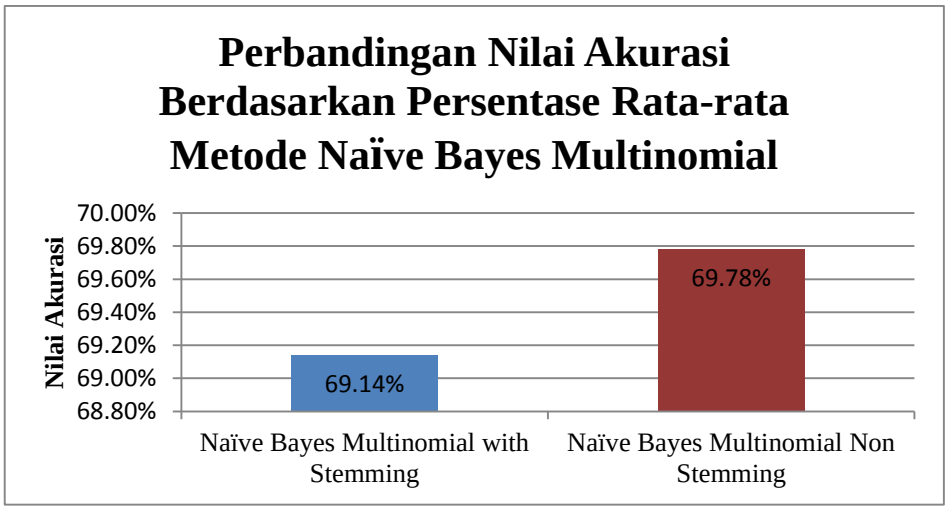
Gambar 5. Perbandingan Nilai Akurasi Berdasarkan Persentase Rata-rata Metode Naïve Bayes Bernoulli

Terdapat perbedaan antara penggunaan atau tidak menggunakan proses *Stemming* dalam pengklasifikasian dokumen teks. Gambar 5 menunjukan bahwa tidak melakukan prosess*Stemming* dalam metode Naïve Bayes Bernoulli memiliki nilai akurasi yang lebih besar yaitu sebesar 71.33% jika dibandingkan dengan menggunakan proses *stemming* dalam pengklasifikasian dokumen teks yaitu sebesar 70.08%.



Gambar 6. Pengaruh Penggunaan *Stemming* pada Nilai Akurasi Naïve Bayes Multinomial

Berdasarkan Gambar 6 menunjukan bahwa ada pengaruh dari penggunaan proses s*temming* pada nilai akurasi Naïve Bayes Multinomial setiap fitur yang digunakan, dilihat dari bentuk grafik yang semakin meningkat dari setiap perubahan dalam menggunakan fitur yang berbeda. Semakin besar jumlah fitur yang digunakan maka nilai akurasi semakin meningkat.



Gambar 7. Perbandingan Nilai Akurasi Berdasarkan Persentase Rata-rata Metode Naïve Bayes Multinomial

Terdapat perbedaan antara penggunaan atau tidak menggunakan proses s*temming* dalam pengklasifikasian dokumen teks. Gambar 7menunjukan bahwa tidak menggunakan proses s*temming* dalam metode Naïve Bayes Multinomial memiliki nilai akurasi yang lebih besar yaitu sebesar 69.78% jika dibandingkan dengan menggunakan proses s*temming* dalam pengklasifikasian dokumen teks yaitu sebesar 69.14%.

5) Hasil Penelitian Berdasarkan Aspek Metode

Tabel 8. Hasil Penelitian Berdasarkan Aspek Metode

Metode	Fitur							Persentase Rata-rata
	100	500	1000	2000	3000	5000	All	
Naïve Bayes Bernoulli	41.06%	61.19%	70.45%	79.32%	**82.01%**	81.66%	79.26%	**70.70%**
Naïve Bayes Multinomial	36.68%	55.33%	67.96%	77.82%	81.89%	83.10%	**83.48%**	69.46%

Berdasarkan Tabel 8 menunjukan bahwa metode terbaik berdasarkan persentase rata-rata nilai akurasi adalah metode Naïve Bayes Bernoulli yaitu sebesar 70.70%. Apabila berdasarkan jumlah fitur maka metode Naïve Bayes Bernoulli dengan jumlah fitur 3000 merupakan metode yang memiliki nilai akurasi paling besar yaitu sebesar 82.01%. Sedangkan berdasarkan metode Naïve Bayes Multinomial dengan jumlah fitur *all* memiliki nilai akurasi paling besar yaitu sebesar 83.48%.

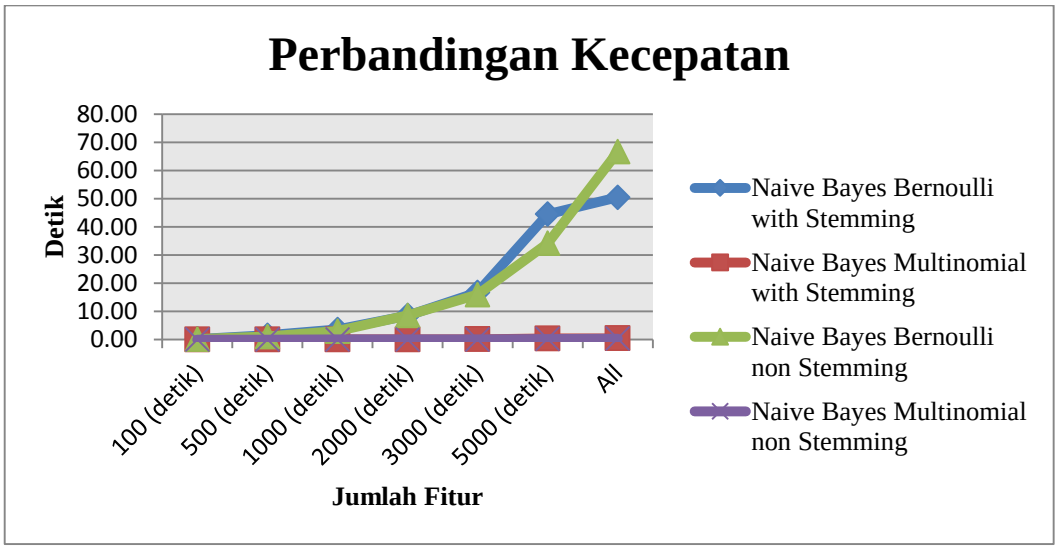
6) Hasil Penelitian Berdasarkan Aspek Waktu

Tabel 9. Hasil Penelitian Berdasarkan Aspek Waktu

Metode	Fitur							Waktu Rata-rata
	100 (detik)	500 (detik)	1000 (detik)	2000 (detik)	3000 (detik)	5000 (detik)	*All* (detik)	
Naïve Bayes Bernoulli *with Stemming*	0.27	1.56	3.59	8.52	16.74	44.52	**50.51**	17.96
Naïve Bayes Multinomial *with Stemming*	**0.00**	0.05	0.06	0.08	0.13	0.42	**0.47**	0.17
Naïve Bayes Bernoulli *non Stemming*	0.20	1.11	3.03	8.60	15.96	34.20	**66.67**	**18.54**
Naïve Bayes Multinomial *non Stemming*	0.02	0.02	0.05	0.06	0.09	0.23	**0.37**	0.12

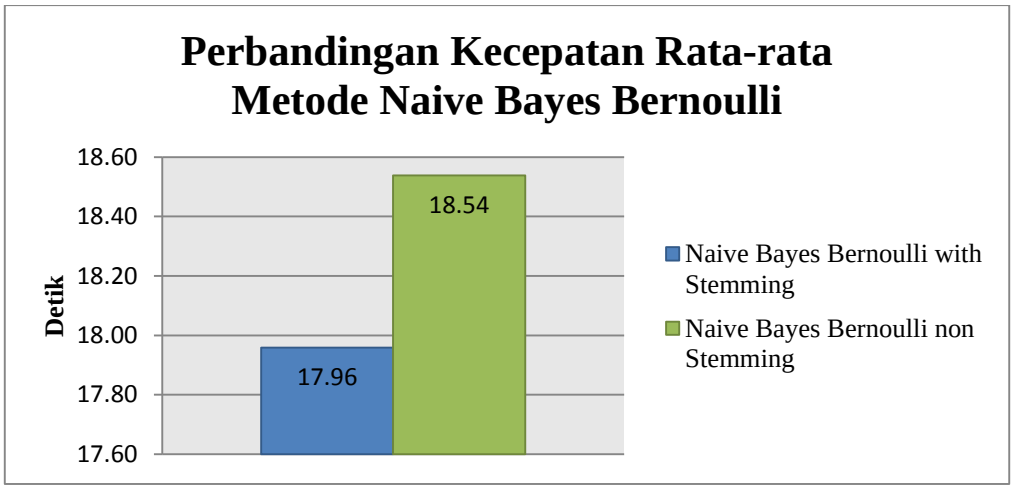
Berdasarkan Tabel 9 menunjukan bahwa metode yang memiliki waktu tercepat dalam pemrosesan adalah metode Naïve Bayes Multinomial *with stemming* dengan jumlah fitur 100 selama 0.00 detik, sedangkan metode yang memiliki waktu terlama

dalam pemrosesan adalah metode Naïve Bayes Bernoulli *non stemming* dengan jumlah fitur *all* yaitu selama 66.67 detik, dan berdasarkan waktu proses rata-rata selama 18.54 detik. Hal tersebut disebabkan oleh semakin banyak jumlah fitur yang digunakan maka semakin lama waktu proses yang dibutuhkan begitupun sebaliknya.



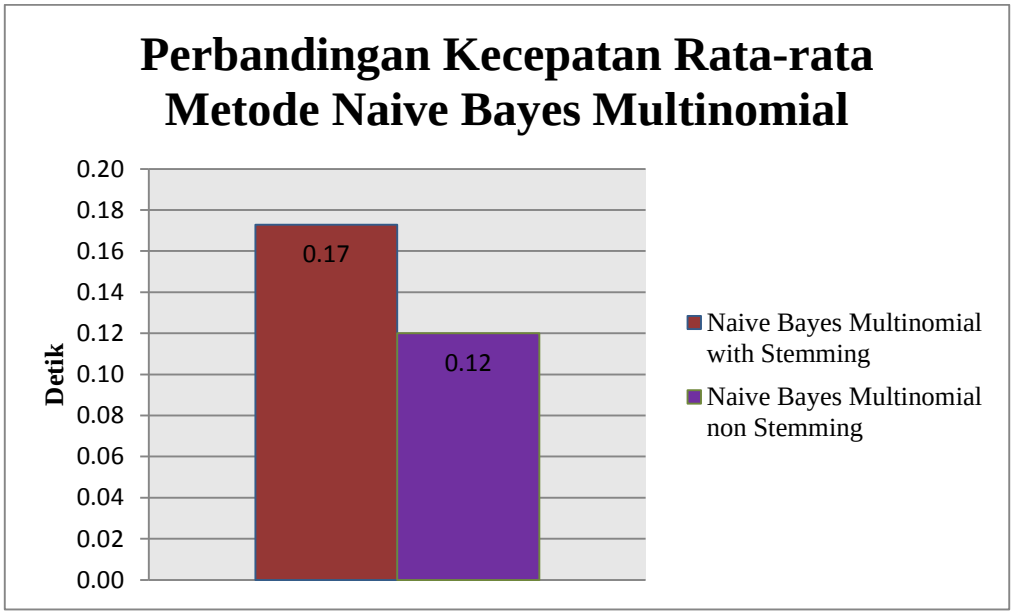
Gambar 8. Perbandingan Kecepatan Waktu Proses Berdasarkan Metode dan Jumlah Fitur yang Digunakan

Gambar 8 menunjukan bahwa metode Naïve Bayes Bernoulli membutuhkan waktu yang lebih lama dalam melakukan pemrosesan data (pengklasifikasian dokumen teks), jika dibandingkan dengan metode Naïve Bayes Multinomial. Hal tersebut dikarenakan dalam melakukan pengklasifikasian dokumen teks menggunakan metode Naïve Bayes Bernoulli perlu melakukan proses perubahan data menjadi data binary yaitu "0" dan "1", sehingga membutuhkan waktu yang lebih lama jika dibandingkan dengan metode Naïve Bayes Multinomial.



Gambar 9. Perbandingan Kecepatan Rata-rata Waktu Proses Metode Naïve Bayes Bernoulli

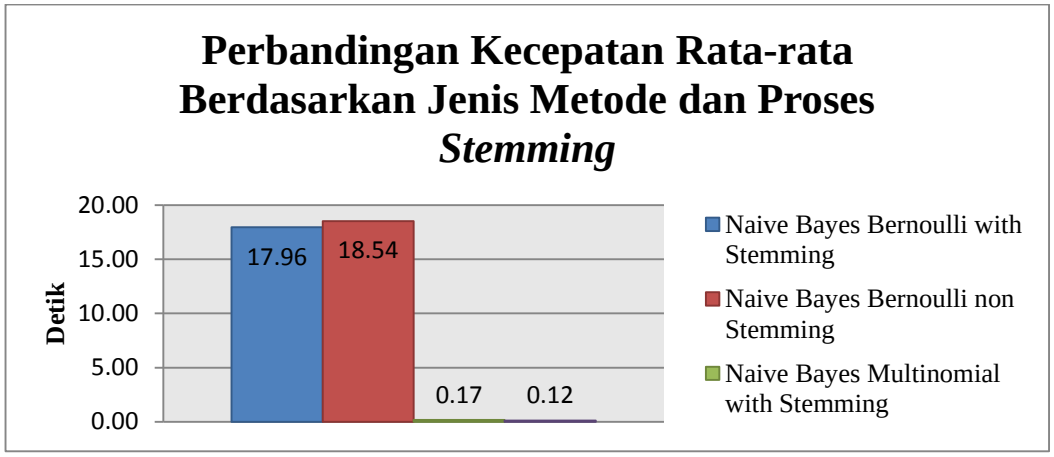
Gambar 9 menunjukan bahwa metode Naïve Bayes Bernoulli yang melalui proses *stemming* membutuhkan waktu yang lebih cepat dalam melakukan pemrosesan data (pengklasifikasian dokumen teks) jika dibandingkan dengan metode Naïve Bayes Bernoulli yang tidak melewati proses *stemming* terlebih dahulu.



Gambar 10. Perbandingan Kecepatan Rata-rata Waktu Proses Metode Naïve Bayes Bernoulli

Gambar 10 menunjukan bahwa metode Naïve Bayes Multinomial yang melalui proses *stemming* membutuhkan waktu yang lebih lama dalam melakukan pemrosesan data (pengklasifikasian dokumen teks), jika dibandingkan dengan

metode Naïve Bayes Multinomial yang tidak melewati proses *stemming* terlebih dahulu.



Gambar 11. Perbandingan Kecepatan Rata-rata Berdasarkan Jenis Metode dan Proses *Stemming*

Gambar 11 menunjukan bahwa metode Naïve Bayes Bernoulli *non stemming* membutuhkan waktu yang lebih lama dalam melakukan pemrosesan data (pengklasifikasian dokumen teks), jika dibandingkan dengan metode lainnya dan metode Naïve Bayes Multinomial yang tidak melalui proses *stemming* membutuhkan waktu proses yang lebih cepat. Oleh karena itu sebaiknya gunakan metode Naïve Bayes Multinomial yang tidak perlu melalui proses *stemming*, namun metode ini memiliki nilai akurasi yang lebih rendah dibandingkan dengan metode Naïve Bayes Bernoulli yang melewati proses *stemming* terlebih dahulu dan begitu pula sebaliknya.

7) Hasil Penelitian

Tabel 10. Hasil Penelitian

Metode	Akurasi	Waktu Proses
Naïve Bayes Bernoulli *with stemming*	70.08%	17.96
Naïve Bayes Bernoulli *non stemming*	**71.33%**	18.54
Naïve Bayes Multinomial *with stemming*	69.14%	0.17
Naïve Bayes Multinomial *non stemming*	69.78%	**0.12**

Berdasarkan Tabel 10 didapatkan hasil bahwa semakin membutuhkan tingkat akurasi yang tinggi maka membutuhkan waktu proses yang lebih lama, begitupun sebaliknya.

Kesimpulan

Adapun kesimpulan didalam penelitian ini adalah terdapat perbedaan kinerja dari penerapan Algoritma Naïve Bayes Multivariate Bernoulli dan Naïve Bayes Multinomial dalam melakukan klasifikasi dokumen teks dilihat dari nilai persentase akurasi dan kecepatan waktu proses dalam melakukan pengklasifikasian dokumen teks, dimana semakin tinggi persentase akurasi maka semakin membutuhkan waktu proses yang lebih lama begitupun sebaliknya. Adapun nilai akurasi tertinggi apabila menggunakan metode Naïve Bayes Bernoulli *non stemming* sebesar 71.33%, dan waktu proses tercepat apabila menggunakan metode Naïve Bayes Multinomial *non stemming* yaitu membutuhkan waktu proses 0.12 detik

Referensi

Agastya, M. (2018). Pengaruh Stemmer Bahasa Indonesia Terhadap Performa Analisis Sentimen Terjemahan Ulasan FIlm. *Jurnal TEKNOKOMPAK, Vol. 12, No. 1, 2018, 18-23. ISSN 1412-9663 (print)*, 18-23.

Ibid. (n.d.).

Juang, D. (2016). Analisis Spam dengan menggunakan Naive Bayes . *Jurnal Teknovasi Volume 03, Nomor 2, 2016, 51 – 57 ISSN : 2355-701X* , 51-57.

Ma, J., Zhang, Y., Liu, J., & Yu, K. (2016). Intelligent SMS Spam Filtering Using Topic Model. *2016 International Conference on Intelligent Networking and Collaborative Systems*, 380-383.

Pratiwi, S., & Ulama , B. (2016). Klasifikasi Email Spam dengan Menggunakan Metode Support Vector Machine dan k-Nearest Neighbor. *JURNAL SAINS DAN SENI ITS Vol. 5 No. 2 (2016) 2337-3520 (2301-928X Print)* , D-344 - D-349.

Rahmayani, I. (2019, Juli Sabtu). *https://kominfo.go.id/content/detail/6095*. Retrieved from https://kominfo.go.id: https://kominfo.go.id/content/detail/6095/indonesia-raksasa-teknologi-digital-asia/0/sorotan_media

Rahmi, F., & Wibisono, Y. (2016, Juli Sabtu). *Aplikasi SMS Spam Filtering pada Android menggunakan Naive Bayes, Unpublished manuscript*. Retrieved from http://nlp.yuliadi.pro: http://nlp.yuliadi.pro/dataset

Raschka, S. (2014, Juli Sabtu). *https://sebastianraschka.com/Articles*. Retrieved from https://sebastianraschka.com: https://sebastianraschka.com/Articles/2014_naive_bayes_1.html

Santosa, B. (2007). *Data Mining Teknik Pemanfaatan Data Untuk Keperluan Bisnis.* Yogyakarta: Graha Ilmu.

Ting, S., Ip, W., & Tsang , A. (3, July, 2011). Is Naïve Bayes a Good Classifier for Document Classification? *International Journal of Software Engineering and Its Applications Vol. 5, No.* , 37-46.

.