Optimasi *Stemming* Porter KBBI dan *Cross Validation Naïve Bayes* untuk Klasifikasi Topik Soal UN (Ujian Nasional) Bahasa Indonesia

A. Yudi Permana, Ismasari dan M.Makmun Effendi

Jurusan Teknik Informatika STT PELITA BANGSA
Jl. A. Yani, Bekasi Barat, Indonesia
E-mail: {yudi,ismasari.n,effendiyan}@pelitabangsa.ac.id

Abstrak

Penelitian ini dimaksudkan untuk mencari nilai akurasi klasifikasi topik soal UN bahasa Indonesia yang terdiri dari 12 kategori topik soal UN (ujian nasional) bahasa indonesia untuk tingkat SMK. Tujuan utama dari penelitian ini adalah mencari klasifikasi dari pada topik soal UN (ujian nasional) Bahasa indoensia yang kemudian di evaluasi hasil akurasinya dengan beberapa pendekatan optimasi algoritma. Metode penelitian yang digunakan adalah dengan terlebih dahulu melakukan Preprocessing dan optimasi stemming porter KBBI pada data soal UN (ujian nasional) bahasa indonesia, Dengan jumlah data training 350 soal dan data testing 150 soal. Proses Preprocessing merupakan proses melakukan Case Folding, Stopword Removal dan menerapkan optimasi algoritma Stemming porter KBBI. Setelah Preprocessing data kemudian dicari nilai frekuensi kata pada tiap soal sehingga setiap kata dari dokumen soal UN (ujian nasional) Bahasa indoensia tersebut mempunyai nilai. Kemudian diberikan label klasifikasi secara manual dan pada tahapan akhir metode dilakukan klasifikasi dengan algoritma Naïve Bayes. Dari hasil penelitian akurasi klasifikasi dengan optimasi pendekatan stemming porter KBBI dan pendekatan metode klasifiksai cross validation Naïve Bayes, dengan hasil akurasi data training 94,34% dan data testing cross validation 72.67%.

Kata Kunci : stemming, klasifikasi, preprocessing, cross validation, naïve bayes.

Pendahuluan

UN (ujian nasional) merupakan syarat utama kelulusan siswa di tingkat SMK(sekolah menengah kejuruan), salah satu pelajaran yang menjadi bagian dari indikasi kelulusan adalah nilai bahasa Indonesia. Kurangnya pemahaman siswa dan guru tentang beberapa topik yang dibahas dalam soal bahasa Indonesia, menyebabkan tidak efektifnya belajar menjelang UN (ujian nasional) karna semua bahasan dipelajari . Soal UN (ujian nasional) bahasa indonesia pada umumnya memiliki 12 topik soal. Topik soal tersebut adalah fakta, opini, kalimat, pargraf, frasa, gagasan utama, puisi, karya sastra judul, kutipan, artikel dan tabel. Penelitian ini dilakukan untuk mencari optimasi dari pada nilai akurasi sehingga hasil dari klasifikasi topik soal ujian nasional bahasa Indonesia memliki akurasi yang baik dengan standar kelulusan nilai adalah 70%. Sehingga guru dan siswa dengan penelitian ini lebih mudah belajar sesuai hasil topik soal yang diklasifikasikan dan lebih bisa belajar dengan efektif.

Metode penelitian yang digunakan adalah dengan terlebih dahulu melakukan Preprocessing pada data soal UN (ujian nasional) bahasa indonesia yang sebelumnya sudah dipisahkan antara data Training dan data testing, Dengan jumlah data training 350 soal dan data testing 150 soal. Preprocessing merupakan proses melakukan Case Folding, Stopword Removal dan optimasi algoritma Stemming porter KBBI. Setelah Preprocessing data kemudian setiap kata dicari nilai frekuensi katanya masing masing dalam tiap dokumen lalu diberikan label secara manual dan pada tahapan akhir metode dilakukan optimasi pendekatan klasifikasi dengan algoritma Naïve Bayes untuk mengetahui berapa tingkat akurasi klasifikasi topik soal UN bahasa indonesia.

Metode Pemrosesan Teks

Imbuhan Bahasa Indonesia

Bahasa Indonesia merupakan bahasa yang memiliki morfologi yang berbeda dan unik dengan bahasa lainnya. Bahasa Indonesia memiliki imbuhan-imbuhan yang beraneka ragam dan masuk sebagai kata serapan atau lainnya. Sering kali sebuah kata dasar atau bentuk dasar perlu diberi imbuhan untuk dapat digunakan dalam pertuturan [1]. Imbuhan ini dapat mengubah makna, jenis dan fungsi sebuah kata dasar atau bentuk dasar menjadi kata lain yang fungsinya berbeda dengan kata dasar atau bentuk dasarnya.

Data Mining

Data Mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar [2]. Terdapat beberapa istilah lain yang memiliki makna sama dengan data mining, yaitu Knowledge discovery in databases (KDD), ekstraksi pengetahuan (knowledge extraction), Analisa data/pola (data/pattern analysis), kecerdasan bisnis (business intelligence) dan data archaeology dan data dredging [3].

Teks Mining

Salah satu bagian dari Data Mining yang cukup menarik adalah Text Mining. Teks mining dapat didefinisikan sebagai "penemuan informasi baru dan tidak diketahui sebelumnya oleh komputer, dengan secara otomatis mengekstrak informasi dari sumber – sumber teks tak terstruktur yang berbeda" [4]. Dalam melakukan implementasi text mining terdiri dari dua tahap besar yaitu pre-processing dan processing. Tahap preprocessing adalah tahap di mana aplikasi melakukan seleksi data yang akan diproses pada setiap dokumen. Setiap kata akan dipecah-pecah menjadi struktur bagian kecil yang nantinya akan mempunyai makna sempit. Ada beberapa hal yang perlu dilakukan pada tahap pre-processing ini, yaitu: Casefolding, Tokenizing, Filtering, dan Stemming [5].

Preprocessing

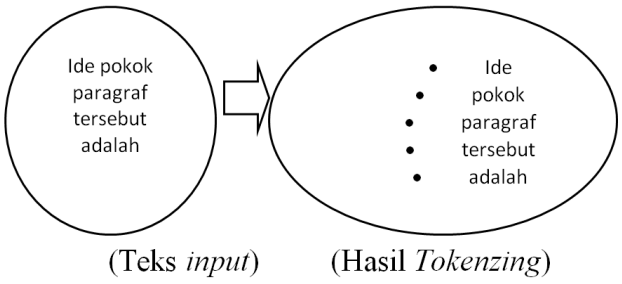
Prosess preprocessing dilakukan pada tahapan awal penelitian sebelum melakukan stemming dan labelisasi serta klasifikasi topik soal UN (ujian nasional) bahasa indonesia. Ada beberapa langkah dari preprocessing diantaranya adalah:

Case Folding

Pada proses case folding data set soal yang memiliki karakter, angka dan tanda baca dihilangkan sehingga tanda baca tidak akan muncul pada saat pelabelan klasifikasi kategori dan terdapat beberapa proses yang harus dilakukan adalah diantaranya mengubah semua huruf besar menjadi huruf kecil (text to lowercase).

Tokenizing

Pada tahapan tokenizing setiap kalimat dirubah menjadi suku kata dengan karakter dan tanda baca yang sudah dihilangkan.



Gambar 1: Contoh dari tahapan *tokenizing*

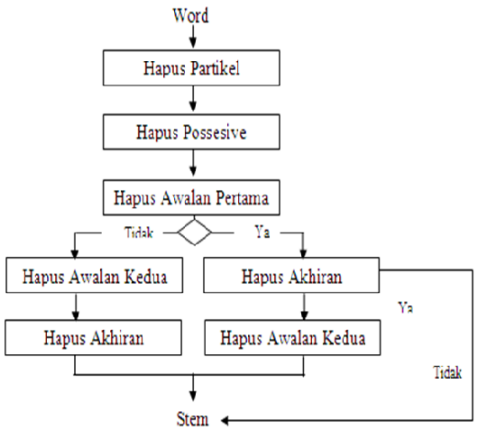
Stopword Removal

Kemudian selanjutnya yaitu proses memeriksa stop word list, stopword list adalah daftar kata-kata penghubung antar kalimat yang semestinya dihilangkan.

Word frequency training (Stemming Porter KBBI)

Tahapan stemming adalah tahap mencari root kata atau kata dasar dari tiap kata hasil filtering. Implementasi porter stemmer bahasa indonesia berdasarkan English Porter Stemmer telah dikembangkan oleh W.B. Frakes pada tahun 1992.[6]

Tahapan stemming porter dapat dilihat pada gambar 2.



Gambar 2: Tahapan Algoritma stemming Porter [7]

Frekuensi kata

Term frekuensi merupakan total kemunculan kata yang akan diproses didalam sebuah dokumen. Tahapan tf merupakan salah satu tahapan dari labelisasi dimana setiap kata memeliki kemungkinan muncul di dokumen secara berulang. TF (t,d) : total kemunculan term t pada dokumen d.

Proses training dokumen kategori (label)

Proses training dokumen kategori atau labelisasi dilakukan secara manual pada tiap dokumen soal ujian nasional bahasa indonesia untuk menentukan dan sekaligus acuan untuk proses klasifikasi dokumen, kategori klasifikasi topik soal ujian nasional bahasa indonesia menjadi 12 kelas topik yaitu: gagasan utama, tabel, kalimat, fakta, opini, paragraf, kutipan, puisi, judul, karya sastra, frasa dan artikel.

Klasifikasi Dokumen Naïve bayes

Metode naïve bayes mempunyai 2 tahapan ketika proses klasifikasi teks, yaitu proses pelatihan dan proses klasifikasi. Algoritma klasifikasi naïve bayes bertujuan untuk mencari klasifikasi dari data yang akan diujikan dengan mencari nilai probabilitas tertinggi dalam pengujian data. Maka untuk tahapan diatas dibutuhkan dokumen yang akan di training dan dokumen yang akan di testing.

1. Dokumen training

   Dokumen training dibutuhkan untuk pembentukan kelas dan mempermudah proses klasifikasi dokumen dengan membentuk model klasifikasi, dalam hal ini penelitian menggunakan bank data soal ujian nasional bahasa Indonesia dengan 350 soal data training.

2. Dokumen testing

   Dokumen testing dalam penelitian ini menggunakan dokumen dengan extension CSV, dokumen percobaan sejumlah 150 dokumen testing soal ujian bahasa Indonesia. Dalam algoritma klasifikasi naïve bayes, setiap kumpulan dokumen diuraikan dengan pasangan atribut x_1 x_2 x_3. . . sampai dengan x_n dimana [(x)]_adalah kata awal dan seterusnya, sedang V merupakan himpunan topik soal. Pada saat tahapan pengujian naïve bayes akan mencari nilai probabilitas tertinggi dari semua dokumen yang akan diujikan [8]. Persamaannya disajikan pada persamaan 1.

$$V_{map} = \underset{Vj \in V}{argmax} \left(\frac{(P(x_1, x_2, x_3 ... x_n \mid Vj) P(Vj)}{P(x_1, x_2, x_3 ... x_n)} \right) \quad (1)$$

Untuk $P(x_1 x_2 x_3 \ldots x_n)$ nilainya konstan untuk semua Kategori (Vj) sehingga persamaan dapat ditulis sebagai persamaan 2 berikut :

$$V_{map} = \underset{Vj \in V}{argmax} \; (P(x_1, x_2, x_3 ... x_n \mid Vj) P(Vj) \quad (2)$$

Persamaan 2 dapat disederhanakan menjadi persamaan 3 sebagai berikut:

$$V_{map} = \underset{Vj \in V}{argmax} \prod_{i=1}^{n} (P(x_i \mid Vj) P(Vj) \quad (3)$$

Keterangan :
Vj: Kategori soal j1,2,3. . . .n
dimana :
J1=Artikel
J2=topik soal Fakta
J3= topik soal Frasa
J4= topik soal Gagasan utama
J5= topik soal kalimat
J6 topik soal Judul
J7= topik soal karya sastra
J8=topik soal kutipan
J9=topik soal Opini
J10=topik soal
Paragraf J11=topik soal karya Puisi J12=topik soal
Tabel $P(Xi|Vj)$: Probabilitas Xi pada $VjP(Vj)$: Probabilitas dari Vj

Untuk$P(Vj)$ dan $P(Xi|Vj)$ dihitung pada saat pelatihan dengan persamaan 4 dan 5 sebagai berikut:

$$P(Vj) = \frac{|\ docs\, j\ |}{|\ contoh\ |} \quad (4)$$

$$(P(x_i \mid Vj) = \frac{n_{k+1}}{n+ \mid kosakata \mid} \quad (5)$$

Keterangan:

$|docsj|$**:** jumlah dokumen setiap kategori j

$|contoh|$**:** jumlah dokumen dari semua kategori

n_k : jumlah frekuensi kemunculan setiap kata

n**:** jumlah frekuensi kemunculan kata dari setiap kategori

$|kosakata|$**:** jumlah semua kata dari semua kategori

Cross Validation

CV (Cross-validation) adalah metode statistik yang dapat digunakan untuk mengevaluasi kinerja model atau algoritma dimana data dipisahkan menjadi dua subset yaitu data proses pembelajaran dan data validasi / evaluasi. Model atau algoritma dilatih oleh subset pembelajaran dan divalidasi oleh subset validasi. Selanjutnya pemilihan jenis CV (Cross-validation) dapat didasarkan pada ukuran dataset. Biasanya CV (Cross-validation) K-fold digunakan karena dapat mengurangi waktu komputasi dengan tetap menjaga keakuratan estimasi.

Akurasi

Akurasi diperlukan untuk evaluasi dan mengukur keakuratan dari hasil klasifikasi, semakain besar nilai akurasi maka semakin baik tingkat klasifikasinya:

$$Accuracy = \frac{jumlah\, dokumen\, yang\, terklasifikasi}{jumlah\, dokumen\, keseluruhan}\, x\, 100\,\%$$

Metodologi Penelitian

Tahapan dalam penelitian disajikan pada gambar 3 dan meliputi:

1. Mengumpulkan data soal ujian nasional bahsa indonesia sebanyak 500 data sample dan membaginya menjadi menjadi data training dan data testing.

2. Metode usulan yaitu Impelementasi dan perancangan sistem berbasis web dengan php mysql untuk uji sample data training dan data testing soal ujian nasional bahasa indonesia dan klasifikasi dengan tool weka.

3. Lakukan proses preprocessing pada data sample yang sudah dibagi antara data training dan data testing, dengan data training sebanyak 350 soal dan data testing 150 soal dan melakukan preprocessing data dengan tahapan case folding, tokenizing, stopword removal sebagai berikut:

   (a) Case folding melakukan proses penghapusan karakter dan angka pada kalimat soal dan merubah huruf besar menjadi huruf kecil.

   (b) Tokenizing melakukan proses pemisahan struktur kalimat menjadi struktur kata-kata.

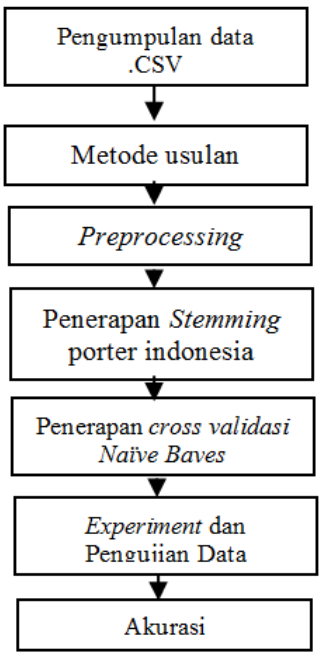
   (c) Stopword removal melakukan penghapusan pada kata penghubung yang dianggap tidak mempunyai bobot kata dalam kalimat.

4. Lakukan proses optimasi stemming pada kata berimbuhan yang sudah melalui proses preprocessing dengan stemming porter KBBI(kamus besar bahasa Indonesia).

5. Lakukan proses penerapan algoritma naive bayes untuk mengelompokkan kategori soal ujian nasional bahasa indonesia dan melakukan analisa penyimpangan matrik pada kategori soal yang tidak terklasifikasi dengan baik.

6. Lakukan eksperimen dan pengujian data manual klasifikasi dengan data klasifikasi terkomputerisasi.

7. Menghitung nilai akurasi dari masing masing stemming dengan algoritma cross validation naive bayes.



Gambar 3: Metode penelitian.

Hasil dan Pembahasan

Koleksi dokumen yang digunakan untuk pengujian adalah dokumen training sebanyak 350 soal untuk data set training dan 150 soal testing dan kemudian soal berformat CSV, diambil dari soal ujian nasional. Contoh soal dapat dilihat pada tabel 1 dan tabel 2 adalah koleksi data soal UN.

Table 1: Koleksi Dokumen

No	Soal UN
1.	apa minta kornea donor di indonesia dapat tuh
2.	transplantasi ginjal mudah laku selain faktor biaya donor cocok temu kalimat fakta paragraph

Hasil Klasifikasi Dengan Naïve Bayes

Dilakukan data training dengan stemming porter KBBI, dan eksperimen data testing dengan stemming Porter KBBI, tabel 3 dan 4 adalah hasil dari data training dengan stemming porter KBBI.

Table 2: Koleksi data soal UN bahasa Indonesia berdasarkan topiknya

No	Kelas topik un	Soal sampel	Soal training	Soal testing
1.	Artikel	2	1	1
2.	Fakta	23	12	3
3.	Frasa	18	16	9
4.	Gagasan Utama	31	16	4
5.	Judul	8	4	2
6.	Kalimat	258	180	84
7.	Karya Sastra	24	12	2
8.	Kutipan	20	10	2
9.	Opini	17	13	4
10.	Paragraf	63	56	22
11.	Puisi	34	18	13
12.	Tabel	14	12	4
	Total Soal	500	350	150

Table 3: Contoh hasil training preprocessing dan stemming Porter KBBI perkata

No	Term	Doc_id	Count
1.	Apa	1	1
2.	Minta	1	1
3.	Kornea	1	1
4.	Donor	1	1
5.	Di	1	2

Table 4: Contoh data soal training sebelum preprocessing

No	Soal UN
1.	Mengapa permintaan kornea donor di Indonesia tidak dapat di penuhi?
2.	Transplantasi ginjal tidak mudah untuk dilakukan. Selain faktor biaya, donor yang cocok juga tidak mudah untuk ditemukan. Kalimat fakta dalam paragraf tersebut adalah
3.	Lanskap budaya subak di Bali telah ditetapkan sebagai situs warisan dunia. Kalimat fakta paragraf tersebut terdapat pada nomor

Hasil evaluasi dataset soal UN (ujian nasional) bahasa indoensia dengan menggunakan algoritma naive bayes:

1) Training data

Evaluasi klasifikasi data soal UN bahasa indonesia dengan pilihan use training set untuk untuk mencari hasil akurasi dengan menggunakan training data yang berjumlah 350 dokumen dan meghasilkan menghasilkan akurasi dan mean absolute error data pada tabel 5.

Table 5: Hasil Training Set Naive Bayes soal UN bahasa Indonesia

No	Klasifikasi	*Preprocessing* dan *stemming* porter KBBI		Akurasi
		Terklasi-fikasi	Tidak terk-lasi-fikasi	
1.	Artikel	1	0	0.29
2.	Fakta	12	0	3.43
3.	Frasa	14	2	4.00
4.	Gagasan Utama	14	2	4.00
5.	Judul	4	0	1.14
6.	Kalimat	163	17	46.57
7.	Karya Sastra	12	0	3.43
8.	Kutipan	10	2	2.86
9.	Opini	13	0	3.71
10.	Paragraf	55	1	15.71
11.	Puisi	18	0	5.14
12.	Tabel	12	0	3.43
	Presentase Akurasi	318	22	93.7143%

Tabel 5 menunjukkan akurasi yang diperoleh adalah 93.7143 % dengan hasil klasifikasi benar sebanyak 328 dokumen, jumlah data training tidak terklasifikasi sebanyak 22 atau 6.287%. Dan tabel 6 menunjukan hasil dari pada akurasi dengan confussion matrix, untuk melihat data asal dan data prediksi yang kemungkinan besar ada perpindahan dari kelas asal ke kelas prediksi sehingga ada hasil akurasinya. Tabel 6 menunjukan Hasil dari data training dengan menggunakan stemming porter KBBI sebanyak 350 soal menghasilkan data yang terklasifikasi sebanyak 328 soal terklasifikasi dengan baik sesuai kelas kategori masing-masing dengan hasil akurasi 93,7143% dan 22 soal tidak terklasifikasi dengan baik.

Table 6: Hasil akurasi data training soal UN bahasa Indonesia

No	Hasil Training Akurasi	Presentase
1.	*Correctly Classified Instances*	328 atau 93.7143 %
2.	*Incorrectly Classified Instances*	22 atau 6.2857%
3.	*Kappa statistic*	0.912
4.	*Mean absolute error*	0.0105
5.	*Root mean squared error*	0.0965
6.	*Relative absolute error*	8.9797 %
7.	*Root relative squared error*	40.0172 %
8.	*Total Number of Instances*	350

2) Testing data / evaluasi cross validation 10 fold

Klasifikasi yang telah terbentuk pada tahap training selanjutnya diuji dengan menggunakan data testing evaluasi cross validation 10 fold dengan data testing sebanyak 150 soal data testing menghasilkan data pada Tabel 5.

Table 7: Hasil Evaluasi Test Set Naive Bayes cross validation 10 fold soal UN bahasa indoenesia

No	Klasifikasi	*Preprocessing* dan *stemming* porter KBBI		Akurasi
		Terklasi-fikasi	Tidak terk-lasi-fikasi	
1.	Artikel	0	1	0
2.	Fakta	1	2	0.67
3.	Frasa	0	9	0
4.	Gagasan Utama	0	4	0
5.	Judul	2	0	1.33
6.	Kalimat	84	0	0
7.	Karya Sastra	0	2	0
8.	Kutipan	0	2	0
9.	Opini	0	4	1.33
10.	Paragraf	14	8	5.33
11.	Puisi	8	5	5.33
12.	Tabel		4	0
	Presentase Akurasi	109	41	72.66%

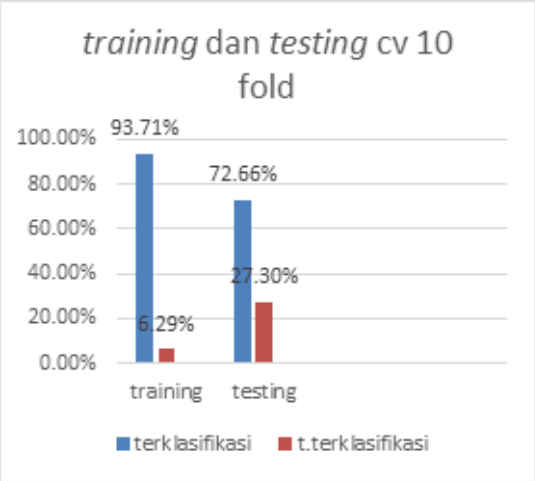
Tabel 7 menunjukkan akurasi yang diper-

oleh adalah 72.66% dengan test record yang diklasifikasi secara benar sebanyak 109, jumlah test record yang diklasifikasi secara tidak benar sebanyak 41 atau 27.3%, dengan hasil mean absolute error adalah 0,046%.

Table 8: Hasil akurasi data testing cross validation Naive Bayes 10 fold soal UN bahasa Indoenesia

No	Hasil Training Akurasi	Presentase
1.	*Correctly Classified Instances*	109 atau 72.66 %
2.	*Incorrectly Classified Instances*	41 atau 27.3 %
3.	*Kappa statistic*	0.52
4.	*Mean absolute error*	0.046
5.	*Root mean squared error*	0.20
6.	*Relative absolute error*	41.45 %
7.	*Root relative squared error*	86.14 %
8.	*Total Number of Instances*	150

Tabel 8 menunjukan Hasil dari data testing cross validation 10 fold dengan menggunakan stemming porter KBBI sebanyak 150 soal menghasilkan data yang terklasifikasi sebanyak 109 soal terklasifikasi dengan baik sesuai kelas kategori masing-masing dengan hasil akurasi 72,66% dan 41 soal tidak terklasifikasi dengan baik. Kemudian hasil dari grafik training datanya sesuai grafik dibawah ini.



Gambar 4: Hasil evaluasi data terklasifikasi 10 fold.

Dari hasil grafik pada gambar 4 menunjukan hasil training 93.71% dan testing cross validation 10 fold 72.66%, sehingga selisih akurasi data training yaitu 21.11%,sedangkan untuk data tidak terklasifikasi menunjukan hasil training 6.29% dan testing cross validation 10 fold 27.30%, sehingga selisih akurasi data training yaitu 21.1%. 3) Testing data / evaluasi cross validation 8 fold Klasifikasi yang telah terbentuk pada tahap training selanjutnya diuji dengan menggunakan data testing evaluasi cross validation 8 fold dengan data testing sebanyak 150 soal data testing menghasilkan data pada Tabel 9.

Table 9: Hasil Evaluasi Test Set Naive Bayes cross validation 8 fold soal UN bahasa Indonesia

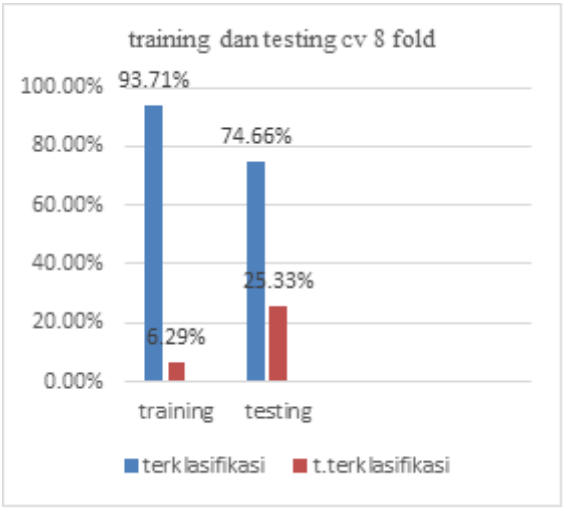
No	Klasifikasi	*Preprocessing* dan *stemming* porter KBBI		Akurasi
		Terklasifikasi	Tidak terklasifikasi	
1.	Artikel	1	0	0.67
2.	Fakta	1	2	0.67
3.	Frasa	0	9	0
4.	Gagasan Utama	0	4	0
5.	Judul	2	0	1.33
6.	Kalimat	83	1	55.33
7.	Karya Sastra	0	2	0
8.	Kutipan	0	2	0
9.	Opini	1	3	0.67
10.	Paragraf	15	7	10
11.	Puisi	9	4	6
12.	Tabel	0	4	0
	Presentase Akurasi	112	38	74.67%

Tabel 9 menunjukkan akurasi yang diperoleh adalah 74.66% dengan test record yang diklasifikasi secara benar sebanyak 112, jumlah test record yang diklasifikasi secara tidak benar sebanyak 38 atau 25.33%, dengan hasil mean absolute error adalah 0,045%.

Tabel 10 menunjukan Hasil dari data testing cross validation 8 fold dengan menggunakan stemming porter KBBI sebanyak 150 soal menghasilkan data yang terklasifikasi sebanyak 112 soal terklasifikasi dengan baik sesuai kelas kategori masing-masing dengan hasil akurasi 74,66% dan 38 soal tidak terklasifikasi dengan baik.

Table 10: Hasil akurasi data testing cross validation 8 fold Naive Bayes soal UN bahasa Indonesia

No	Hasil Training Akurasi	Presentase
1.	*Correctly Classified Instances*	112 atau 74.66 %
2.	*Incorrectly Classified Instances*	38 atau 25.33 %
3.	*Kappa statistic*	0.55
4.	*Mean absolute error*	0.045
5.	*Root mean squared error*	0.20
6.	*Relative absolute error*	40.11 %
7.	*Root relative squared error*	85.68 %
8.	*Total Number of Instances*	150



Gambar 5: Hasil evaluasi data terklasifikasi 8 fold

Dari hasil grafik pada gambar 5 menunjukan hasil training 93.71% dan testing cross validation 8 fold 74.66%, sehingga selisih akurasi data training yaitu 19.04%,sedangkan untuk data tidak terklasifikasi menunjukan hasil training 6.29% dan testing cross validation 8 fold 25.33%, sehingga selisih akurasi data training yaitu 19.16%.

4) Testing data / evaluasi cross validation 6 fold Klasifikasi yang telah terbentuk pada tahap training selanjutnya diuji dengan menggunakan data testing evaluasi cross validation 6 fold dengan data testing sebanyak 150 soal data testing menghasilkan data pada Tabel 11, pendekatan ini dilakukan untuk mencari nilai akurasi dari optimasi beberapa pendekatan cross validasi.

Table 11: Hasil Evaluasi Test Set Naive Bayes cross validation 6 fold soal UN bahasa Indonesia

No	Klasifikasi	*Preprocessing* dan *stemming* porter KBBI		Akurasi
		Terklasifikasi	Tidak terklasifikasi	
1.	Artikel	1	0	0.67
2.	Fakta	1	2	0.67
3.	Frasa	0	9	0
4.	Gagasan Utama	0	4	0
5.	Judul	2	0	1.33
6.	Kalimat	83	1	55.33
7.	Karya Sastra	0	2	0
8.	Kutipan	0	2	0
9.	Opini	0	4	0
10.	Paragraf	15	7	10
11.	Puisi	9	4	6
12.	Tabel	0	4	0
	Presentase Akurasi	111	39	74%

Tabel 11 menunjukkan akurasi yang diperoleh adalah 74% dengan test record yang diklasifikasi secara benar sebanyak 111, jumlah test record yang diklasifikasi secara tidak benar sebanyak 39 atau 26%, dengan hasil mean absolute error adalah 0,044%. Dari hasil testing mulai dari 10 fold,8 fold dan 6 fold testing yang memunyai otimasi yang baik yaitu yang menggunakan pendekatan cross validasi 8 fold. Dengan demikian hasil akurasi dari testing cross valiadi 8 fold lebih unggul dibandingkan dengan 10 cross valiasi 10 fold dan cross validasi 6 fold. Dan menempati akurasi terbaik yaitu mulai dari 8 fold cross validasi kemudian diikuti cross validasi 6 fold dan berikutnya yaitu cross validasi 6 fold. Hasil akurasi 8 fold lebih tinggi dibandingan dengan hasil akurasi 10 fold cross validasi.

Tabel 12 menunjukan Hasil dari data testing cross validation 8 fold dengan menggunakan stemming porter KBBI sebanyak 150 soal menghasilkan data yang terklasifikasi sebanyak 111 soal terklasifikasi dengan baik sesuai kelas kategori masing-masing dengan hasil akurasi 74% dan 39 soal tidak terklasifikasi dengan baik.

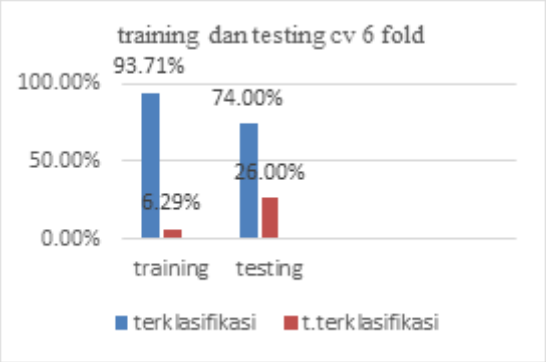
Table 12: Hasil akurasi data testing cross validation 6 fold Naive Bayes soal UN bahasa Indonesia

No	Hasil Training Akurasi	Presentase
1.	*Correctly Classified Instances*	111 atau 74 %
2.	*Incorrectly Classified Instances*	39 atau 26 %
3.	*Kappa statistic*	0.54
4.	*Mean absolute error*	0.044
5.	*Root mean squared error*	0.198
6.	*Relative absolute error*	39.74 %
7.	*Root relative squared error*	85.13 %
8.	*Total Number of Instances*	150

Dari hasil grafik pada gambar 6 menunjukan hasil training 93.71% dan testing cross validation 8 fold 74.66%, sehingga selisih akurasi data training yaitu 19.04%,sedangkan untuk data tidak terklasifikasi menunjukan hasil training 6.29% dan testing cross validation 8 fold 25.33%, sehingga selisih akurasi data training yaitu 19.16%. 5) Testing data / evaluasi cross validation 12 fold Klasifikasi yang telah terbentuk pada tahap training selanjutnya diuji dengan menggunakan data testing evaluasi cross validation 12 fold dengan data testing sebanyak 150 soal data testing menghasilkan data pada Tabel 10.

Table 13: Hasil akurasi data testing cross validation 12 fold Naive Bayes soal UN bahasa Indonesia

No	Hasil Training Akurasi	Presentase
1.	*Correctly Classified Instances*	110 atau 73.3 %
2.	*Incorrectly Classified Instances*	40 atau 26.67 %
3.	*Kappa statistic*	0.529
4.	*Mean absolute error*	0.0449
5.	*Root mean squared error*	0.199
6.	*Relative absolute error*	40.00 %
7.	*Root relative squared error*	85.30 %
8.	*Total Number of Instances*	150



Gambar 6: Hasil evaluasi data terklasifikasi 8 fold.

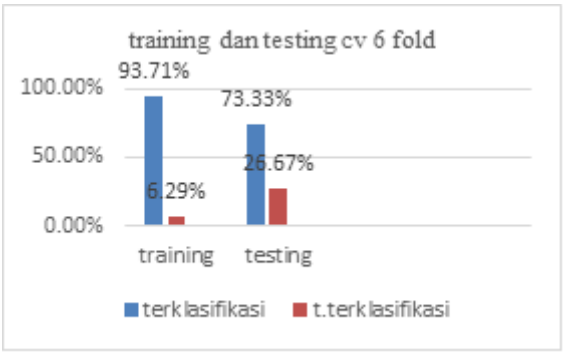
Tabel 13 menunjukkan akurasi yang diperoleh adalah 73.33% dengan test record yang diklasifikasi secara benar sebanyak 110, jumlah test record yang diklasifikasi secara tidak benar sebanyak 40 atau 26.67%, dengan hasil mean absolute error adalah 0,0449%. Dari hasil testing mulai dari 10 fold,8 fold, 6 fold dan 12 fold testing yang memunyai otimasi yang baik yaitu yang menggunakan pendekatan cross validasi 8 fold. Dengan demikian hasil akurasi dari testing cross valiadi 8 fold lebih unggul dibandingkan dengan 10 cross valiasi 10 fold dan cross validasi 6 fold. Dan menempati akurasi terbaik yaitu mulai dari 8 fold cross validasi kemudian diikuti cross validasi 6 fold dan berikutnya yaitu cross validasi 12 fold dan terakhir dengan nilai akurasi terendah yaitu dengan pedekatan 10 fold. Dengan demikian hasil dari pada pendekatan 8 fold validasi lebih optimal mneghasilkan nilai akurasi dari hasil klasifikasi topik soal ujian nasional bahasa Indonesia dengan pendekatan algortma naïve bayes.

Tabel 14. menunjukan Hasil dari data testing cross validation 12 fold dengan menggunakan stemming porter KBBI sebanyak 150 soal menghasilkan data yang terklasifikasi sebanyak 110 soal terklasifikasi dengan baik sesuai kelas kategori masing-masing dengan hasil akurasi 73.33% dan 40 soal tidak terklasifikasi dengan baik. Pendekatan 12 fold cross validasi nilai akurasinya lebih baik dibandingkan nilai akurasi 10 fold cross validasi yang menghasilkan nilai akurasi testing cross validasi yaitu 72.6% nilai daripada akuras akhirnya.

.

Table 14: Hasil Evaluasi Test Set Naive Bayes cross validation 12 fold soal UN bahasa Indonesia

No	Klasifikasi	*Preprocessing* dan *stemming* porter KBBI		Akurasi
		Terklasi-fikasi	Tidak terk-lasi-fikasi	
1.	Artikel	0	1	0
2.	Fakta	1	2	0.67
3.	Frasa	0	9	0
4.	Gagasan Utama	0	4	0
5.	Judul	2	0	1.33
6.	Kalimat	83	1	55.33
7.	Karya Sastra	0	2	0
8.	Kutipan	0	2	0
9.	Opini	0	4	0
10.	Paragraf	15	7	10
11.	Puisi	9	4	6
12.	Tabel	0	4	0
	Presentase Akurasi	110	40	73.33%

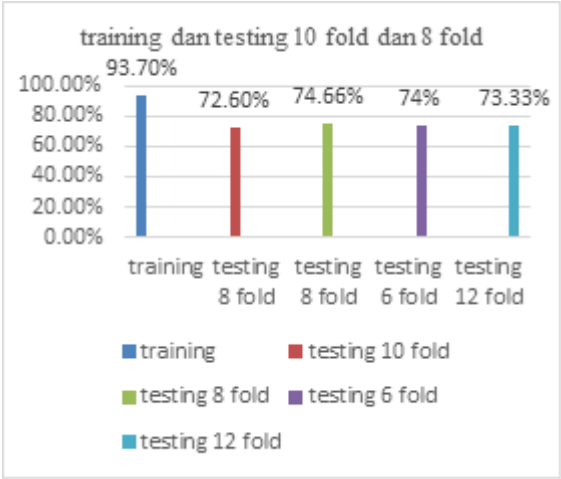


Gambar 7: hasil evaluasi data terklasifikasi 12 fold.

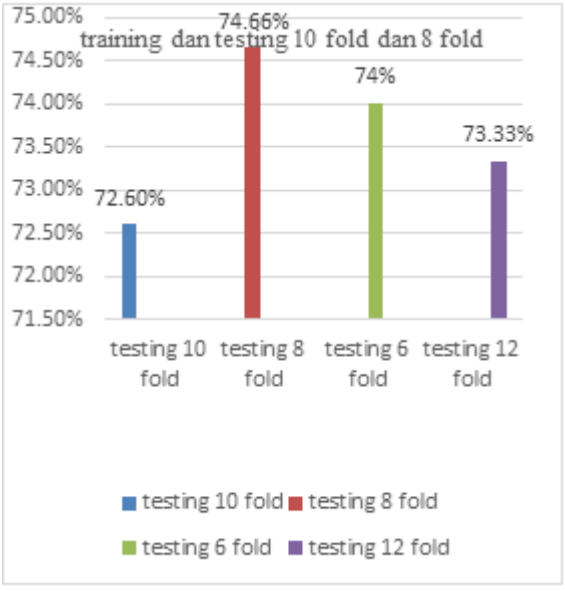
Dari hasil grafik pada gambar 7 menunjukan hasil training 93.71% dan testing cross validation 12 fold 73.33%, sehingga selisih akurasi data training yaitu 19.04%,sedangkan untuk data tidak terklasifikasi menunjukan hasil training 6.29% dan testing cross validation 12 fold 25.33%, sehingga selisih akurasi data training yaitu 19.16%

Hasil presentase akurasi training dan testing stemming porter KBBI dan Naïve Bayes

Gambar 8 ini hasil oprtimasi perbandingan nilai akurasi data training dan testing stemming porter KBBI dan juga cross validasi naïve bayes yang sudah dilakukan eksperimen.



Gambar 8: Hasil presentase akurasi nilai training dan testing porter KBBI dan cross validasi 10 dan 8 fold Naïve Bayes.



Gambar 9: Hasil presentase akurasi nilai testing porter KBBI dan cross validasi 10 dan 8 fold Naïve Bayes.

Dari Grafik pada gambar 8 menjelaskan bahwa tingkat akurasi data training dan testing stemming porter KBBI dan naïve bayes memeberikan hasil optimasi akurasi training data 93,7%, dan testing data cross validasi

10 fold 86%, kemudian testing cross validasi 8 fold dengan akurasi 74.66% sehingga stemmer porter KBBI memberikan optimasi akurasi pada klaifikasi topik soal UN bahas Indonesia karna stemming porter KBBI mempunyai hasil kata dasar yang baik dan mempengaruhi hasil dari pada akurasi klasifikasinya Dan testing cross validasi memberikan pengaruh optimasi dalam evaluasi akurasi hasil akhir. Evaluasi tahap akhir diperlukan untuk menjelaskan dan mengevaluasi hasil data testing sehingga bisa mengetahui soal mana yang tetap pada kelas asalnya dan soal yang keluar dari kelas asalnya.

Analisa Hasil Stemming

Untuk mengetahui ketepatan hasil stemming perlu dilakukan analisa secara manual. Mengingat jumlah kata unique yang cukup banyak (1414 kata) pada data training yang nantinya kata ini menjadi fitur untuk dijadikan atribut klasifikasi, pengamatan mencakup sebagian saja. Kesalahan hasil stemming pada algoritma porter KBBI adalah apabila kata tidak ditemukan di kamus database dan kemudian dianggap kata dasar. Tabel 15 adalah kesalahan hasil stemming pada algoritma porter KBBI terhadap kata berimbuhan.

Table 15: Analisa hasil stemming pada algoritma porterKBBI

Contoh	Hasil *stemming*	Seharusnya
Dipenuhi	Tuh	Penuh
Kendaraan	Ndara	Kendara
Berupa	Upa	Rupa
Pemenggalan	Nggal	Penggal
Diperlukan	Lu	Perlu

Analisa Kesalahan Hasil Pengetikan Kata

Evaluasi dan analisis kesalahan stemming dilakukan untuk mengetahui keselahan dari data sample , baik itu kesalahan stemming dan kesalahan pengetikan serta karakterisitik masing masing algoritma stemming porter KBBI. Tabel 16 menunjukkan kesalahan hasil stemming karena pengetikan kata yang salah dalam dokumen.

Table 16: Analisa hasil stemming karena kesalahan pengetikan kata

Contoh	Seharusnya	Hasil *stemming*	Kata Dasar
Pura i	Pura 1	Pura	Pura
Multikhasiat	Multi khasiat	Multikhasiat	Multi khasiat
Sdah	Sudah	Sdah	sudah
parag;raf	Paragraf	Paragraf	Paragraf
Nierupakan	Merupakan	Rupa	Rupa

Kesalahan stemming dari kesalahan pengetikan kata akan memepengaruhi hasil frekuensi kata dalam dokumen. Berikut kesalahan hasil stemming terhadap nama orang, nama tempat dan istilah dapat dilihat pada Tabel 17.

Table 17: Analisa hasil stemming terhadap nama orang, tempat dan istilah

Contoh	Hasil *stemming*	Seharusnya
Bali	Bal	Bali
Manjadi	Manjad	Manjadi
Kartini	Kart	Kartini

Table 18: Kesalahan hasil Klasifikasi dengan stemming porter KBBI pada algoritma cross validasi Naïve Bayes

No.	Soal	Hasil Topik	Seharusnya
1.	apa manfaat terong kering	Kalimat	Artikel
2.	frase makna ganda dapat kalimat	Kalimat	Frasa

Analisa hasil klasifkasi topik soal UN (ujian nasional) pada algoritma Naïve Bayes

Untuk mengetahui apakah hasil klasifikasi topik soal ujian nasional bahasa indonesia terklasifikasi baik atau tidak maka dilakukan proses analisa kesalahan pada klasifikasi topik soal ujian nasional dengan algoritma naïve bayes, sebagai bahan evaluasi dalam penentuan apakah akurasinya baik atau tidak. Tabel 18 adalah hasil analisa kesalahan untuk data set training soal dengan preprocessing dan stemming porter KBBI dengan kesalahan tidak

terkoreksi dengan baik atau tidak terklasifkasi dengan baik sesuai kategori topik soalnya.

Penutup

Kesimpulan dari hasil penelitian sebagai berikut: 1. Penggunaan stemmer porter KBBI pada proses preprocessing bisa membantu optimasi akurasi klasifikasi pada data testing. 2. Penggunaan stemmer porter KBBI dengan kata dasar yang baik mempengaruhi Hasil akurasi klasifikasi kategori soal UN nya. 3. Hasil akurasi klasifikasi dengan penerapan optimasi cross validasi naïve bayes membantu data tidak terstruktur untuk menghasilkan akurasi yang baik.

Daftar Pustaka

[1] Abdul Chaer, "Tata Bahasa Praktis Bahasa Indonesia", Rineka Cipta Jakarta. 194-197, 2011.

[2] E. Turban, "Decision Support Systems and Intelligent Systems", Edisi Bahasa Indonesia Jilid 1. Andi: Yogyakarta, 2005..

[3] Daniel T. Larose, "Discovering Knowledge in Data : An Introduction to Data Mining", John Willey & Sons, Inc, 2005.

[4] A. Tan, "Text Mining: The state of the art and the challenges", In Proc of the Pacific Asia Conf on Knowledge Discovery and Data Mining PAKDD'99 workshop on Knowledge Discovery from Advanced Databases, 1999.

[5] Chandra Triawati, "Metode Pembobotan Statistical Concept Based untuk Klastering dan Kategorisasi Dokumen Berbahasa Indonesia", Institut Teknologi Telkom, Bandung, 2009.

[6] W. Frakes, "Stemming algorithms, in W. Frakes & R. Baeza-Yates, eds, 'Information Retrieval: Data Structures and Algorithms",Prentice-Hall, chapter 8, pp. 131{160}, 1992.

[7] Ledy Agusta, "Perbandingan Algoritma Stemming Porter dengan Algoritma Nazief & Adriani untuk Stemming Dokumen Teks Bahasa Indonesia", Konferensi Nasional Sistem dan Informatika. KNS&I09-036, 2009.

[8] A.Yudi Permana, "Perbandingan stemming Porter KBBI dengan Tala untuk mencari akurasi klasifikasi topik soal UN Bahasa indoensia dengan algoritma naïve bayes", SNTI 2018, 2018.