KLASIFIKASI TOPIK SOAL UN BHS.INDONESIA DENGAN PENDEKATAN STEMMING PORTER KBBI DAN NAÏVE BAYES

A.Yudi permana[1], Arif Siswandi[2].

STT Pelita Bangsa[1,2]
yudi@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 bahasa indonesia. Metode penelitian yang digunakan adalah dengan tahapan awal Preprocessing pada soal UN bahasa indonesia dengan jumlah data sampel sebanyak 500 soal, dengan model data training 350 soal dan data testing 150 soal. Tahapan awal preprocessing dilakukan proses Case Folding, Stopword Removal dan menerapkan algoritma Stemming porter KBBI, kemudian dilakuakan tahapan klasifikasi dengan pendekatan algoritma naive bayes. Dari hasil penelitian stemming porter KBBI menghasilkan nilai akurasi data training 93,71% sedangkan data testing dengan akurasi 86%. Stemming porter KBBI mempengaruhi hasil dari nilai akurasi klasifikasi topik soal UN dengan ketepatan kata dasar pada proses stemmingnya.

Kata kunci: stemming, naive bayes, klasifikasi.

ABSTRACT

This research is intended to find the accuracy of classification of topic about the UN Indonesian language which consists of 12 categories of topics about the UN language of Indonesia. The research method used is the initial stages of Preprocessing on the problem of Indonesian language with the amount of sample data as much as 500 questions, with the model data 350 training questions and 150 question data testing. Initial stages of preprocessing are Case Folding, Stopword Removal process and apply Stemming KBBI Stemming algorithm, then do classification stages with naive bayes algorithm approach. From result of research of KBBI porter stemming result accuracy value of training data 93,71% while data testing with 86% accuracy. Stemming the KBBI porter affects the result of the accuracy of the UN topic classification accuracy with the precision of the word base on the stemming process..

Keywords: *stemming, naïve bayes,classification.*

PENDAHULUAN

Dokumen soal UN bahasa indonesia pada tingkat SMA dan SMK memiliki 12 topik dan kategori soal. Kategori soal tersebut adalah fakta, opini, kalimat, pargraf, frasa, gagasan utama, puisi, karya sastra judul, kutipan, artikel dan tabel. Penelitian ini dilakukan untuk mempermudah proses klasifikasi topik soal ujian nasional bahasa indonesia dengan komputerisasi dan mencari nilai akurasi klasifikasinya.

Teknologi 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.

Adapun langkah-langkah algoritma pada algoritma *porter* adalah sebagai berikut (Agusta, 2009):

1. Hapus *Particle*.
2. Hapus Possesive Pronoun.
3. Hapus awalan pertama. Jika tidak ada lanjutkan ke langkah 4a, jika ada cari maka lanjutkan ke langkah 4b.
4. a.Hapus awalan kedua, lanjutkan ke langkah 5a.
   b.Hapus akhiran, jika tidak ditemukan maka kata tersebut diasumsikan sebagai

root word. Jika ditemukan maka lanjutkan ke langkah 5b.

5. a.Hapus akhiran. Kemudian kata akhir diasumsikan sebagai *root word*
   b.Hapus awalan kedua. Kemudian kata akhir diasumsikan sebagai *rootword*.

Penelitian yang dilakukan adalah dengan terlebih dahulu melakukan Preprocessing pada soal UN bahasa indonesia yang sebelumnya sudah dipisahkan antara data Training dan data testing, Dengan jumlah data training 350 soal dan data testing 150 soal. Preprocessing melakukakan tahapan diantaranya Case Folding, Stopword Removal dan menerapkan algoritma Stemming porter KBBI.setelah proses preprocessing, data sample akan dilakukan proses labelisasi dan selanjutnya dilakukan proses klasifikasi dengan algoritma naïve bayes. Pada saat tahapan pengujian naïve bayes akan mencari nilai probabilitas tertinggi dari semua dokumen yang akan diujikan (amir hamzah:2012).
Persamaan klasifikasi bayesian sebagai berikut:

$$V_{map} = \underset{Vj \in V}{\arg max} \left(\frac{P(x_1 x_2 x_3 \dots x_n|Vj) P(Vj)}{P(x_1 x_2 x_3 \dots x_n)} \right) \quad [2.1]$$

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

$$V_{map} = \underset{Vj \in V}{\arg max} \left(P(x_1 x_2 x_3 \dots x_n|Vj) P(Vj) \right) \quad [2.2]$$

Persamaan di atas dapat disederhanakan menjadi sebagai berikut:

$$V_{map} = \underset{Vj \in V}{\arg max} \prod_{i=1}^{n} (P(x_1|Vj) P(Vj)) \quad [2.3]$$

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 Vj

$P(Vj)$: Probabilitas dari Vj

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

$$P(Vj) = \frac{|docs\ j|}{|contoh|} \quad [2.4]$$

$$P(Xi|Vj) = \frac{n_{k+1}}{n+|kosakata|} \quad [2.5]$$

Keterangan:

|*docs j*| : 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

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

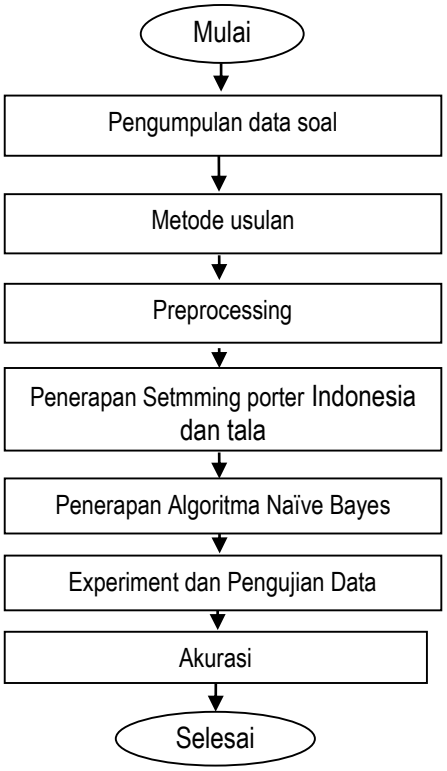
$$Accuracy = \left(\frac{jumlah\ dokumen\ yang\ terklasifikasi}{jumlah\ dokumen\ keseluruhan} X\ 100\ \% \right)$$

METODE

Tahapan dalam penelitian meliputi:

1. Mengumpulkan data soal ujian nasional bahsa indonesia sebanyak 500 data sample dan membaginya menjadi menjadi data training dan data testing.
2. 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.

3. metode usulan Lakukan proses stemming pada kata berimbuhan yang sudah melalui proses preprocessing dengan stemming porter KBBI
4. 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.
5. Menghitung nilai akurasi dengan algoritma naive bayes.



Gambar 1. Diagram Alur Penelitian

HASIL

Koleksi Dokumen

Koleksi dokumen yang digunakan untuk pengujian adalah dokumen training sebanyak 350 soal dan 150 soal testing.

Tabel 1. Koleksi dokumen soal UN Bahasa indonesia

No	topik	Soal sample	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

dilakukan 2 eksperimen yaitu menguji data training dan testing. Berikut hasil dari data training dengan stemming porter KBBI.

Tabel 2. 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

Table 2 di atas merupakan data training yang akan diujikan dengan stemming porter KBBI dengan dokumen training 350 soal.

Table 3. hasil stemming porter KBBI per kata

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

Hasil kata dasar tidak selalu baik sesuai kata dasar dalam KBBI, setelah tahapan stemming kemudian dialkukan tahapan klasifikasi dengan algoritma naïve bayes.

Tabel 4.hasil akurasi *training preprocessing* dan *stemming* porter KBBI

No	Klasifikasi	*Preprocessing* dan *stemming* porter KBBI		Akurasi
		Terklasifikasi	Tidak terklasifikasi	
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	0	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
	Persentase akurasi	328	22	93.71%

Hasil dari pengujian data training dengan menggunakan stemming porter KBBI sebanyak 350 soal menghasilkan data yang terklasifikasi sebanyak 328 soal sesuai kelas kategori masing-masing dan 22 soal tidak terklasifikasi dengan baik dengan hasil akurasi 93,71%.

Tabel 5.hasil akurasi *testing preprocessing* dan *stemming* porter KBBI

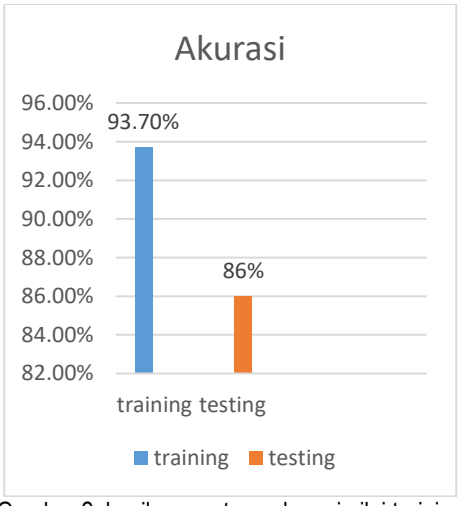
No	Klasifikasi	*Preprocessing* dan *stemming* porter *KBBI*		Akurasi
		Terklasifikasi	Tidak terklasifikasi	
1	Artikel	0	1	0.00
2	Fakta	3	0	2.00
3	Frasa	8	1	5.33
4	Gagasan utama	4	0	2.67
5	Judul	2	0	1.33
6	Kalimat	67	17	44.67
7	Karya sastra	2	0	1.33
8	Kutipan	2	0	1.33
9	Opini	4	0	2.67
10	Paragraf	20	2	13.33
11	Puisi	13	0	8.67
12	Tabel	4	0	2.67
	Persentase	129	21	86%

Hasil dari data testing sebanyak 150 soal menghasilkan data yang terklasifikasi yaitu sebanyak 129 soal sesuai kelas kategori masing-masing dan 21 soal tidak terklasifikasi dengan baik.

Tabel 6. Uji validasi data testing klasifikasi

Kelas label	Kelas prediksi											
	a	b	c	d	e	f	G	h	i	j	k	l
Artikel	0	0	0	0	0	1	0	0	0	0	0	0
Fakta	0	3	0	0	0	0	0	0	0	0	0	0
Frasa	0	0	8	0	0	0	0	0	0	1	0	0
Gagasan utama	0	0	0	4	0	0	0	0	0	0	0	0
Judul	0	0	0	0	2	0	0	0	0	0	0	0
Kalimat	0	0	3	1	0	67	2	0	2	3	0	1
Karya sastra	0	0	0	0	0	0	2	0	0	0	0	0
Kutipan	0	0	0	0	0	0	0	2	0	0	0	0
Opini	0	0	0	0	0	0	0	0	4	0	0	0
Paragraf	0	1	0	1	0	0	0	0	0	20	0	0
Puisi	0	0	0	0	0	0	0	0	0	0	13	0
Tabel	0	0	0	0	0	0	0	0	0	0	0	4

Hasil dari data testing klasifikasi menunjukan ada data yang keluar dari data kelas aslinya sehingga mengisi data kelas prediksi yang salah dan tidak terklasifikasi dengan baik.



Gambar 2. hasil presentase akurasi nilai training dan testing porter KBBI

Gambar 2 menunjukan hasil pencapaian data testing yang mendekati hasil data asli trainingnya.

Analisa kesalahan 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. Berikut kesalahan hasil *stemming* pada algoritma *porter* KBBI terhadap kata berimbuhan.

Tabel 7 Kesalahan hasil *stemming* pada algoritma porter KBBI

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

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.

Berikut adalah hasil analisa kesalahan untuk data set *training* soal dengan *preprocessing* dan *stemming porter* KBBI dan *stemming* TALA dengan kesalahan tidak terkoreksi dengan baik atau tidak terklasifkasi dengan baik sesuai kategori topik soalnya.

Tabel 8. Kesalahan hasil Klasifikasi dengan *stemming* porter KBBI pada algoritma *Naïve Bayes*

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

Pada table 8 hasil klasifikasi yang tidak terklasifkasi dengan baik, dikarenakan komputerisasai menganggap bahwa soal yang tidak terklasifkasi dengan baik jauh dari kelas asalnya, sehingga hasil klasifikasi topik soalnya keluar dari kelas asalnya dan masuk kelas prediksi yang salah.

SIMPULAN

Penggunaan stemmer dengan kata dasar yang baik mempengaruhi Hasil akurasi klasifikasi kategori soal UN nya, hasil stemmer kata dasar yang baik akan memepengaruhi hasil probabilitas pada saat testing klasifikasi topik soal UN nasional Bahasa Indonesia yang akan terklasifikasi dengan baik dan klasifikasi yang baik menghasilkan kelas asal matrik sama dengan kelas prediksi.

DAFTAR RUJUKAN

Agusta, L. (2009). Perbandingan algoritma stemming Porter dengan Algoritma Nazief & Adriani untuk stemming dokumen teks bahasa indonesia. konferensi nasional sistem dan informatika. KNS&I09-036.Gender issues accros the life cycle (pp. 107-123). New York:Springer.

Amir, H. (2012). Klasifikasi teks dengan naïve bayes classifier (NBC) untuk mengelompokkan teks berita dan abstract akademis, Seminar Nasional Aplikasi Sains & Teknologi (SNAST).

Arifin,A.Z.&Setiono,A.N.(2002),Classification of Event News Documents in Indonesian Language Using Single Pass Clustering Algorithm, in 'Proceedings of the Seminar on Intelligent Technology and its Applications (SITIA)', Teknik Elektro, Sepuluh Nopember Institut of Technology, Surabaya, Indonesia.

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

Novitasari, D. (2016). Perbandingan Algoritma Stemming Porter Dengan arifin Setiono Untuk Menentukan Tingkat Ketepatan Kata Dasar.

Samodra, J. (2009). Klasifikasi dokumen teks berbahasa Indonesia dengan menggunakan Naive Bayes. Seminar nasional electrical, informatics, and it's education.

Wen, Z.A. (2010). comparative study of TFIDF, LSI and multi-words for text classification,School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1-1 Ashahidai, Nomi, Ishikawa 923-1292, Japan.