SIGMA
JURNAL TEKNOLOGI PELITA BANGSA

Vol. 8 No. 3 September 2017 ISSN: 2407-3903

Diterima, 15 Agustus 2017 | Direvisi, 21 Agustus 2017 | Dipublikasikan, 28 September 2017

IMPLEMENTASI STEMMING PORTER KBBI UNTUK KLASIFIKASI TOPIK SOAL UJIAN NASIONAL BAHASA INDONESIA MENGGUNAKAN ALGORITMA NAIVE BAYES

A. Yudi Permana

Program Studi Teknik Informatika Sekolah Tinggi Teknologi Pelita Bangsa
yudi@pelitabangsa.ac.id

Abstraksi

Klasifikasi adalah pembagian sesuatu menurut kelas - kelas dan kategori kelasnya sudah ditentukan sebelumnya. Dalam hal ini soal ujian nasional akan diklasifikasikan dan dikelompokkan berdasarkan kategorinya sendiri secara otomatis. Soal ujian nasional bahasa indonesia secara manual dikelompokkan kedalam beberapa kategori topik. Pada penelitian ini akan ditentukan metode untuk *preprocessing*, *stemming* KBBI dan klasifikasi menggunakan algoritma Naive Bayes. Pengujian dilakukan menggunakan 805 soal ujian nasional bahasa indonesia yang sudah ditentukan sebelumnya. Dari 805 data set kemudian dibagi 2 bagian 600 soal untuk data set training dan 205 untuk soal testing. Hasil dari pengujian akhir penelitian tesis yang dilakukan menunjukkan bahwa dengan adanya proses *case folding*, *tokenizing*, *stopword* dan *stemming* porter bahasa indonesia dengan menentukan hasil akhir kata dasar yang sesuai dengan KBBI, sangat membantu dan menentukan proses klasifikasi soal ujian nasional dengan tingkat akurasi yang baik. Hasil training dengan metode *preprocessing* (*case folding*, *tokenizing*, *stopword*) dan *stemming* KBBI menghasilkan tingkat akurasi 95,5% dan hasil data *testing* menghasilkan tingkat akurasi 89,27%.

Kata kunci : *Preprocessing*, *Case Folding*, *Tokenizing*, *StopWord*, *Stemming* Porter Indonesia, *Naïve Bayes*, Klasifikasi.

Abstract

Classification is the division of something by class - the class and the class category is predetermined. In this case the national exam will be classified and grouped according to their own category automatically. Indonesian national exam to manually grouped into several categories topics. In this study will be determined the method for preprocessing, stemming KBBI and classification using Naive Bayes algorithm. Tests carried out using a 805 Indonesian national exam that is predetermined. Of the 805 data sets and then divided into 2 sections of 600 questions for 205 data sets for the matter of training and testing. Results of final testing thesis research carried out showed that the presence of prosescase folding, tokenizing, stopword and porter stemming Indonesian to determine the final results are in accordance with the basic words KBBI, very helpful and determines the classification process national exam with a good degree of accuracy. The result of training with preprocessing methods (case folding, tokenizing, stopword) and stemming KBBI produce accuracy rate of 95.5% and the results of testing the data generating 89.27% accuracy rate.

Keywords*: preprocessing, Case Folding, tokenizing, Stopword, Porter Stemming Indonesia, Naïve Bayes, classification.*

1. Pendahuluan

Soal ujian nasional bahasa indonesia memiliki beberapa topik dan kategori. Silabus bahasa indonesia umum memiliki 12 kategori topik

soal ujian nasional. Untuk memudahkan kategorisasi topik soal ujian nasional secara otomatis maka dengan ini peneliti akan melakukan penelitian terfokus pada klasifikasi teks soal UN bahasa Indonesia. Peneliti dalam hal ini akan melakukan penelitian dengan topik" klasifikasi topik soal ujian nasional Bahasa Indonesia menggunakan algoritma naïve bayes". Kemudian penelitian ini nantinya bertujuan untuk mengelompokkan (klasifikasi) soal ujian nasional bahasa indonesia dalam kategori atau topik-topik bahasan yang sudah ditentukan sebelumnya. Diantara pendekatan yang ada klasifikasi naïve bayes telah banyak digunakan karena motode naïve bayes memiliki kesederhanaan baik dalam tahap preprocessing teks dan klasifikasi teks itu sendiri. Meskipun ada kekurangan dari metode ini, dipenelitian sebelumnya metode ini sudah bisa membuktikan cukup efektif dalam pengklasifikasian yang mempunyai banyak kelasnya.

2. Tinjauan Pustaka

2.1. Dokumen *Preprocessing*

Ada beberapa langkah dari preprocessing pada penelitian ini diantaranya adalah:

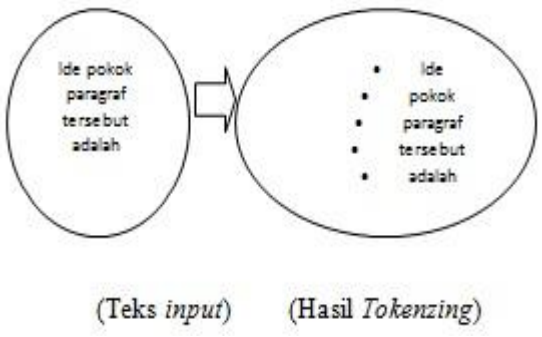
1.Case Folding

Pada proses case folding data set soal yang memiliki karakter dan tanda baca dihilangkan sehingga tanda baca tidak akan muncul pada saat pelabelan klasifikasi kategori.

2. Tokenizing

Pada tokenizing terdapat beberapa proses yang harus dilakukan adalah mengubah semua huruf besar menjadi kecil (text to lowercase). Proses selanjutnya adalah penguraian, proses penguraian yang dimaksud adalah membagi teks menjadi kumpulan kata tanpa memperhatikan keterhubungan diantara kata satu dengan yang lain serta peran dan posisinya pada kalimat, karakter diterima dalam kumpulan kata menurut abjad.

Contoh dari tahap ini seperti pada Gambar 2.1 berikut:



Gambar 1. Contoh dari tahapan *tokenizing*

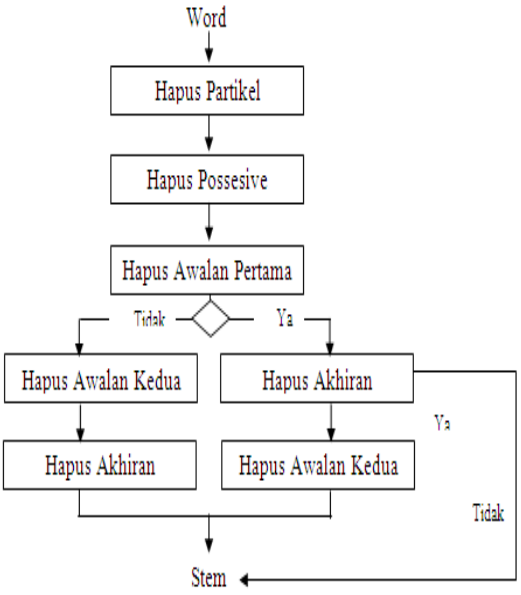
3. Stopword Removal

Kemudian proses selanjutnya yaitu memeriksa stop word list, stopword list adalah daftar kata-kata yang semestinya dihilangkan, jika kata pada dataset terdapat pada stop word list maka kata akan dihilangkan. Tetapi jika tidak terdapat di dalamnya maka proses akan berlanjut tanpa menghilangkan kata pada dokumen.

2.2. *Word frequency training* (*Stemming* Porter Bahasa Indonesia)

Tahapan stemming adalah tahap mencari root kata dari tiap kata hasil filtering. Kata-kata yang telah selesai dilakukan proses stemming kemudian disimpan sebagai data percobaan, setiap kata pada data percobaan dirubah menjadi format yang tidak diketahui oleh metode untuk selanjutnya dijadikan sebagai data masukan untuk prosses pembelajaran dengan metode terkait.

Tahapan stemming porter dapat dilihat pada gambar berikut ini :



Gambar 2. Tahapan Algoritma *stemming* Porter (Agusta:2009)

2.3. Proses *training* dokumen kategori (label)

Proses penentuan kategori dilakukan secara manual pada tiap dokumen soal ujian nasional bahasa indonesia unutk memepermudah 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.

2.4. 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 memepermudah proses klasifikasi dokumen dengan membentuk model klasifikasi, dalam hal ini penelitian menggunakan bank data soal ujian nasional bahasa Indonesia dengan 600 soal data training.

2.Dokumen testing

Dokumen testing dalam penelitian ini menggunakan dokumen dengan extension CSV, dokumen percobaan sejumlah 205 dokumen testing soal ujian bahasa Indonesia.

Dalam algoritma klasifikasi naïve bayes, setiap dokumen diuraikan dengan pasangan atribut x_1 x_2 x_3…sampai dengan x_ndimana 〖 x〗 _adalah kata pertama dan seterusnya, sedang V adalah himpunan topik soal. Pada saat tahapan pengujian naïve bayes akan mencari nilai probabilitas tertinggi dari semua dokumen yang akan diujikan. Persamaannya sebagai berikut:

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

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

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

Persamaan diatas dapat disederhanakan menjadi sebagai berikut:

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

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:

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

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

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

2.5 Akurasi

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

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

.

3. Metodologi

Metode penelitian yang dilakukan dalam penelitian ini. Pertama dari objek penelitian , desain penelitian dan teknik pengumpulan data yang dilakukan dari berbagai sumber penelitian, kemudian selanjutnya metode yang diusulkan.dan melakukan preprosesing data dan penerapan algoritma. Langkah selanjutnya yang akan dilakukan adalah melakukan eksperimen/pengujian terhadap data berupa soal ujian nasional dalam bentuk teks bahas indonesia dengan preprocessing, stemming porter indonesia dan klasifikasi teks menggunakan naïve bayes.

3.1 Desain Penelitian

Desain penelitian melalui tahapan eksperimen data teks soal UN bahasa indonesia dengan desain prosedur terlihat seperti pada Gambar 3.1

1.Objek penelitian

Objek penelitian yang dibahas dalam tesis ini adalah soal ujian nasional bahasa Indonesia yang akan dikategorisasi dalam bentuk soal.

2.Pengumpulan Data

Pada bagian pengumpulan data, akan dijelaskan lebih rinci bagaimana dan darimana data soal UN digunakan untuk penelitian ini didapatkan, kemudian data dalam penelitian ini diambil dari berbagai sumber dan dijadikan sebagai dokumen uji dengan ekstensi csv

3.Metode yang Diusulkan

Pada bagian ini akan dijelaskan oleh peneliti tentang metode yang akan diusulkan untuk melakukan stemming dengan porter Indonesia dan metode klasifikasi dengan naïve bayes, dengan porter stemming menggunakan kamus KBBI.

4.Preprocessing (Case folding, Tokenizing dan Stop word Removal)

Pada tahap ini akan dijelaskan tentang tahapan paling awal untuk melakukan ekstraksi sebuah dokumen yang terdiri dari case folding, tokenizing dan stopword removal sebelum dilakukan stemming, yaitu dengan menghilangkan tanda spasi, karakter dan angka dalam dokumen.

Selanjutnya akan dilakukan pemisahan kalimat menjadi per-kata dalam dokumen atau semua proses ini disebut preprocessing.

5.Penerapan Algoritma Stemming

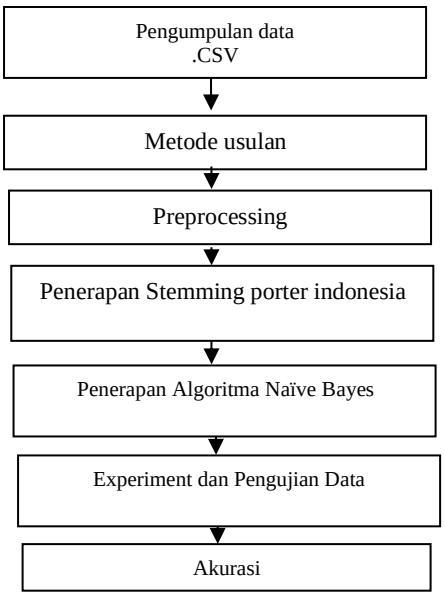
Pada tahapan ini diterapkan algoritma stemming porter indonesia, untuk membuktikan bahwa pemilihan metode ini bisa digunakan untuk menghasilkan kata dasar dalam tiap soal

6.Penerapan Algoritma Naïve Bayes

Pada tahapan ini tiap soal yang sudah diketahui kata – kata dasarnya, kemudian akan diklasifikasi berdasarkan kategori topik soal yang sudah ditentukan atau disebut pemilihan label.

7.Eksperimen dan Pengujian dataset

Pada tahapan ini langkah-langkah eksperimen yang meliputi cara desain eksperimen sampai memperoleh hasil stemming kata dari asal kalimat soal, sampai tahapan stemming penelitian sudah dan pendekatan ini sangat tepat, selanjutnya untuk tahapan penelitian klasifikasi dilakukan sesuai dengan tahapan metode penelitian pada gambar 3.1 berikut.



Gambar 3. Metode penelitian

3.2 Pengumpulan data

Data yang akan mendukung dalam penelitian ini yaitu meliputi data primer dan data sekunder. Data primer yaitu data yang diperoleh langsung dari kamus besar bahasa Indonesia. Data sekunder yaitu data yang diperoleh dengan membaca dan mempelajari beberapa referensi maupun dokumen yang berhubungan langsung dengan permasalahan dalam penelitian ini.

3.2 Metode yang diusulkan

Metode yang diusulkan untuk penelitian ini menggunakan pendekatan dengan preprocessing (case Folding, tokenisasi dan stop word removal) dengan algoritma stemming porter KBBI agusta dan klasifikasi naïve bayes.

3.4 Tahapan *Preprocessing*

Pada tahapan *preprocessing* ini ada beberapa proses sebelum dilakukan proses *stemming.* Langkah *preprocessing* ini sebagai tahapan awal dimana ada proses *case folding* untuk merubah hurup dalam soal menjadi hurup kecil. Untuk menghilangkan karakter karakter teks yang tidak diperlukan, kemudian selanjutnya ada proses *tokenizing* dimana *tokenizing* berfungsi membagi soal UN yang awal dalam bentuk kalimat menjadi kata per kata, kemudian proses selanjutnya ada tahapan *stop word removal* dimana kata kata yang merupakan penghubung dengan kata lainnya dalam kalimat soal dihilangkan akan tetapi *stopword removal* pada penelitian ini tidak berdasarkan *stopword removal* dengan basis data dan terbatas.

Proses *preprocessing* ini dibutuhkan sebagai parameter untuk melakukan *stemming* pada soal UN bahasa Indonesia.

Case Folding, Tokenizing dan Stop Word Removal

Case folding merupakan proses mengubah kalimat menjadi huruf kecil, akan tetapi yang dirubah hanya hurup saja. Selain itu *case folding* juga merupakan proses penghapusan spasi, angka dan tanda baca pada kalimat.

3.3 Algoritma Stemming Porter Indonesia

Proses *stemming porter* Indonesia akan menghasilkan kata dasar dengan mengacu kamus data dan imbuhan yang akan dihilangkan sesuai aturan algoritma *stemming porter* indonesia. Pada proses ini akan dihasilkan kata dasar sesuai *database*.

3.6 Label Klasfikasi

Setelah proses *stemming* maka selanjutnya akan dilakukan pelabelan untuk menentukan atribut dari data set training yang akan dilakukan proses klasifikasi dengan *naïve bayes.*

3.7 Klasifikasi *Naïve Bayes*

Setelah proses *stemming* maka selanjutnya adalah proses klasifikasi topik soal ujian nasional dengan klasifikasi *naïve bayes* dan akan menentukan nilai akurasinya. Pada proses ini soal ujian nasional yang sudah di *stemming* kemudian di *training* untuk menghasilkan klasifikasi topik soal ujian bahasa indoensia.

3.8 Eksperimen dan pengujian dataset

Sebelum diterapkan dan diujikan pada proses klasifikasi maka data soal UN akan melalui tahapan *preprocessing* dan algoritma *stemming*nya terlebih dahulu.

Setelah proses *preprocessing* dan *stemming* kemudian data untuk *training* dan *testing* klasifikasi dibuat label untuk mengetahui klasifikasi topik soal ujian nasional secara manual (logika Manusia).

4. Pembahasan dan Hasil

4.1. Pembahasan

Dalam pembahasan tentang hasil percobaan yang dilakukan mulai dari koleksi soal UN, kamus kata dasar untuk proses stemming, kemudian data training serta data testing yang akan diujikan.

Pengujian dilakukan menggunakan 805 soal ujian nasional bahasa indonesia yang sudah ditentukan sebelumnya.dari 805 data set ini di bagi 2 bagian 600 soal untuk data set training dengan variabel soal label klasifikasi acak dan soal pertanyaan dengan soal tunggal dan majemuk artinya soal pertanyaan ada yang terdiri dari konten soalnya ada yang tidak mengikutsertakan konten soalnya ada juga dari variable soal keduanya, 205 soal testing dengan variabel soal label klasifikasi acak dan soal pertanyaan dengan soal tunggal dan majemuk artinya soal pertanyaan ada yang terdiri dari konten soalnya ada yang tidak mengikutsertakan konten soalnya ada juga dari variable soal keduanya.

Hasil dari pengujian akhir dari penelitian tesis yang dilakukan menunjukkan bahwa dengan adanya proses case folding, stopword dan stemming porter bahasa indonesia dengan menentukan hasil akhir kata dasar yang sesuai dengan KBBI, dari masing-masing kata dasar mempunyai frekwensi katanya. Dengan adanya proses preprocessing sangat membantu dan menentukan proses klasifikasi soal ujian nasional dengan tingkat akurasi yang baik.

Agar metode usulan degan metode preprocessing dan stemming dinyatakan berhasil maka peneliti melakukan penelitian dengan melakukan 2 penelitian yang terdahulu dengan data set soal UN. Dari hasil penelitian yang dilakukan dengan 4 eksperimen percobaan terhadap data set training dan testing membuktikan bahwa klasifikasi topik soal ujian nasional dengan ditambahkan proses preprocessing dan stemming porter Indonesia KBBI menghasilkan klasifikasi untuk eksperimen 1 (case folding, stop word, dan stemming porter) dengan hasil training akurasi klasifikasi sebesar 95,5% dari 600 soal training, eksperimen 2 (case folding dan stemming porter) dengan hasil training akurasi klasifikasi sebesar 94,8% dari 600 soal training, ekperimen 3 (case folding dan stop word) dengan hasil training akurasi klasifikasi sebesar 90,8% dari 600 soal training, dan ekperimen 4 (case folding tanpa stopword dan stemming) dengan hasil training akurasi klasifikasi sebesar 90% dari 600 soal training.

Kemudian dari data testing soal dengan eksperimen 1 (case folding, stopword, dan stemming porter) dengan akurasi klasifikasi sebesar 89,27% dari 205 soal testing, eksperimen 2 (case folding dan stemming porter) dengan akurasi klasifikasi sebesar 87,32% dari 205 soal testing, eksperimen 3 (case folding dan stopword) dengan akurasi klasifikasi sebesar 77,56% dari 205 soal testing, dan eksperimen 4 (case folding tanpa stopword dan stemming porter) dengan akurasi klasifikasi sebesar 72,68% dari 205 soal testing.

4.2. Hasil Pengujian

4.2.2. Koleksi Dokumen

Koleksi dokumen yang digunakan untuk pengujian adalah dokumen training sebanyak 600 soal untuk data set training dan 205 soal testing dan kemudian soal berformat CSV, diambil dari soal ujian nasional.

Tabel 1. Koleksi data soal UN bahasa Indonesia berdasarkan topiknya

NO	KELAS TOPIK UN	SOAL SAMPEL	SOAL TRAINING	SOAL TESTING
1	Artikel	7	6	1
2	Fakta	78	44	34
3	Frasa	24	22	2
4	Gagasan utama	46	41	5
5	Kalimat	7	6	1
6	Judul	314	210	104
7	Karya sastra	108	90	18
8	Kutipan	51	40	11
9	Opini	46	41	5
10	Paragraf	32	27	5
11	Puisi	60	44	16
12	Tabel	32	29	3
	Total soal	805	600	205

Tabel 2. Contoh hasil akurasi training stop word dan stemming

No	Klasifikasi	*Case Folding, Tokenisasi, Stopword* dan *stemming*		Akurasi
		Terklasifikasi	Tidak terklasifikasi	
1	Artikel	6	0	1.000
2	Fakta	41	3	6.833
3	Frasa	21	1	3.500
4	Gagasan utama	40	1	6.667
5	Kalimat	6	0	1.000
6	Judul	199	11	33.167
7	Karya sastra	86	4	14.333
8	Kutipan	38	2	6.333
9	Opini	39	2	6.500
10	Paragraf	24	3	4.000
11	Puisi	44	0	7.333
12	Tabel	29	0	4.833
	Persentase	573	27	95,5%

Hasil dari eksperimen1 data training sebanyak 600 soal menghasilkan data yang terklasifikasi hampir sama yaitu sebanyak 573 soal terklasifikasi dengan baik sesuai kelas kategori masing-masing dan 27 soal tidak terklasifikasi dengan baik.

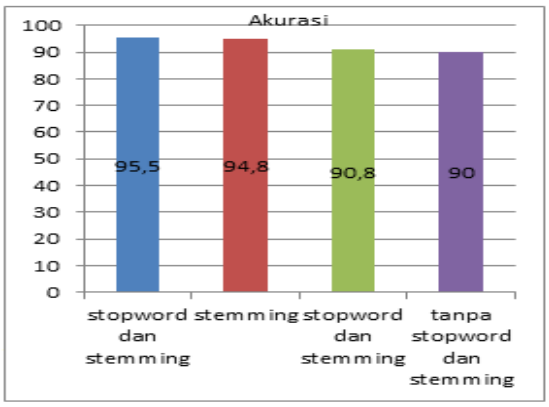
Tabel 3. Contoh Hasil akurasi *testing* dengan *stop word* dan *stemming*

No	Klasifikasi	*Case Folding, Tokenisasi, Stopword* dan *stemming*		Akurasi
		Terklasifikasi	Tidak terklasifikasi	
1	Artikel	1	0	0.49
2	Fakta	27	7	13.17
3	Frasa	2	0	0.98
4	Gagasan utama	5	0	2.44
5	Kalimat	0	1	0.00
6	Judul	95	9	46.34
7	Karya sastra	17	1	8.29
8	Kutipan	8	3	3.90
9	Opini	5	0	2.44
10	Paragraf	5	0	2.44
11	Puisi	16	0	7.80
12	Tabel	2	1	0.98
	Persentase	183	22	89.27

Hasil dari eksperimen 1 data testing sebanyak 205 soal menghasilkan data yang terklasifikasi yaitu sebanyak 183 soal terklasifkasi dengan baik sesuai kelas kategori masing-masing dan 22 soal tidak terklasifikasi dengan baik.

4.3. Hasil presentase akurasi nilai kebaruan

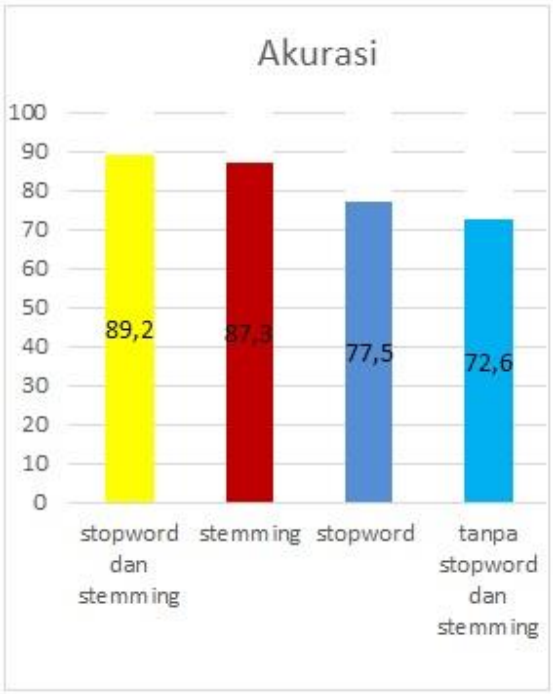
Gambar dibawah ini hasil dari pada nilai akurasi data training yang sudah dilakukan 4 eksperimen. Dengan nilai kebaruan dengan stop word dan stemming porter KBBI menghasilkan akurasi lebih baik.



Gambar 4. hasil presentase akurasi nilai *training*

Dari Gambar 4 menunjukkan bahwa tingkat akurasi data training dari eksperimen 1 lebih baik dibandingakan akurasi eksperimen 2, 3 dan eksperimen 4.

Gambar dibawah ini hasil untuk nilai akurasi data testing yang sudah dilakukan 4 eksperimen.



Gambar 5. Hasil presentase akurasi nilai *testing*

Dari Gambar 5 menunjukkan bahwa tingkat akurasi data testing 1 dari eksperimen 1 lebih baik dibandingakan akurasi eksperimen 2, 3 dan eksperimen 4. Yaitu dengan penggunaan *stopword* dan stemming KBBI hasil akurasi lebih baik.

Dari hasil uji validasi dan evaluasi ada kelas prediksi yang tidak terklasifikasi dengan baik pada kelas label asalnya dikarenakan banyak kata yang muncul sama di kelas prediksi yang berbeda dengan kelas label asalnya, matrik hasil merah menandakan bahwa hasil klasifikasi tidak benar sedangkan hasil matrik hitam merupakan hasil dari klasifikasi yang benar.

Dengan hasil ini analisa kesalahan pada klasifikasi topik soal ujian nasional bisa diketahui dikelas prediksinya, dengan terbentuknya urutan matrik kelas prediksi yang terklasifikasi dengan baik dan tidak terklasifikasi dengan baik.

5. Kesimpulan

Berdasarkan hasil penelitian ini maka didapat beberapa kesimpulan sebagai berikut:

1. Bahwa Dengan menggunakan algoritma stemming porter indonesia Agusta KBBI dalam klasifikasi topik soal UN terbukti memiliki tingkat akurasi yang baik.
2. Dengan menggunakan *Stopword* dalam dlasifikasi topik soal UN terbukti membantu meningkatkan hasil akurasi klasifikasinya.
3. Kesalahan stemming porter agusta pada tahapan proses awal sebelum klasifikasi terjadi karena kata tidak ditemukan di kamus database dan kemudian dianggap kata dasar.
4. Kesalahan klasifikasi terjadi karena kata yang sama muncul pada Beberapa kelas klasifikasi yang berbeda.

Daftar Pustaka

[1] Agusta, Ledy. 2009. Perbandingan algoritma stemming Porter dengan Algoritma Nazief & Adriani untuk stemming dokumen teks bahasa indonesia. konferensi nasional sistem dan informatika. KNS&I09-036.

[2] Ariadi Dio dan Fithriasari Kartika, 2015. Transferring Classification dan Support Vector Machine dengan Confix Stripping Stemmer, JURNAL SAINS ITS.

[3] Chen Kewen, Zhang Zuping , Long Jun, Zhang Hao, 2016.Turning from TF-IDF to TF-IGM for term weighting in text classification, School of Information Science and Engineering, Central South University, China

[4] Cong1, Yao-ban Chan2 & Mark A. Ragan1, 2016. A novel alligment free method for detection of lateral genetic transfer based on TF – IDF, Yingnan, 1Institute for Molecular Bioscience and ARC Centre of Excellence in Bioinformatics, The University of Queensland, St Lucia, Brisbane, QLD 4072, Australia. 2School of Mathematics and Statistics,The University of Melbourne, Parkville, Melbourne, Australia.

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

[6] Jiang Liangxiao, 2016. Deep feature weightin gfornaive Bayes and its application to text classification,, Department of Computer Science,China University of Geosciences,Wuhan 430074,China

[7] Jingnian Chen,2008.Feature selection for text classification with Naïve Bayes, ,Elsevier Ltd. All rights reserved.

[8] Jong-Yeol Yoo, Min-Ho Lee, Grace Aloyce, Dong-Min Yang, 2016. Dept. of Information & Communications Engineering, Daejeon University, Daejeon, Korea,

[9] Kewen Chen, Zuping Zhang , Long, Jun Hao Zhang, 2016.Turning from TF-IDF to TF-IGM for term weighting in text classification, School of Information Science and Engineering, Central South University, China.

[10] Kim S.b, han ks, rim and hc, 2006. Some effective Technicues for naïve bayes text classification., are with department computersscience and engineering, college of information and communication, korea university, Published by the IEEE Computer Society,.

[11] Ong Hong Choon and Low Heng Chin, 2016. Classification Using the General Bayesian m Network, Sau Loong Ang*, Pertanika J. Sci. & Technol. 24 (1): 205 – 211.

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

[13] Silfia Andini, 2013. Klasifikasi dokumen teks menggunakan algoritma naïve bayes dengan bahasa pemrograman java, Jurnal teknologi informasi dan pendidikan.

[14] Sumpeno Surya Destuardi I., 2009. Klasifikasi Emosi Untuk Teks Bahasa Indonesia Menggunakan Metode Naive Bayes, Seminar Nasional Pascasarjana IX – ITS.

[15] Sumpeno surya, hariadi, mochammad 2009. klasifikasi document trxt berbahas Indonesia dengan naïve bayes, joko samodre, seminar teknologi Informatika, Indonesia.

[16] Tala F. Z. 2004. A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia, Institute for Logic, Language and Computation Universiteit van Amsterdam The Netherland.

[17] Tang Bo, He Haibo and Kay Steven, 2015. A Bayesian Classification Approach Using Class-

Specific Features for Text Categorization, are with the Department of Electrical Computer and Biomedical Engineering at the University of Rhode Island.

[18] Ting S.L., W.H. Ip, H.C. Tsang Albert, 2011. Is Naïve Bayes a Good Classifier for Document Classification, International Journal of Software Engineering and Its Applications,hongkong.

[19] Wenyuan Dai† Gui-Rong Xue† Qiang Yang‡ Yong Yu, 2007. Naive Bayes Classifiers for Text Classification,†,Department of Computer Science and Engineering Shanghai Jiao Tong University, Shanghai, China,

[20] Yeol Yoo Jong, Ho Lee Min, Aloyce Grace, 2016. Creating a Naïve Bayes Document Classification Scheme Using an Apriori Algorithm. Dong-Min Yang,Korea, Dept. of Information & Communications Engineering, Daejeon University, Daejeon, Korea,

[21] Zhang A Wen, 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.