Klasifikasi Konten Berita Dengan Metode *Text Mining*

[1]Bambang Kurniawan, [1]Syahril Effendi, [1]Opim Salim Sitompul

[1]Program Studi S1 Teknologi Informasi
Fakultas Ilmu Komputer dan Teknologi Informasi
Universitas Sumatera Utara
E-mail: bambangkur_niawan@yahoo.co.id, syahril1@usu.ac.id, opim@usu.ac.id

Abstrak -- Banyak instansi yang bergerak dalam penyaluran informasi atau berita sudah mulai menggunakan sisitem berbasis *web* untuk menyampaikan berita secara *up to date*. Pada umumnya berita yang disampaikan dalam portal tersebut terdiri dari beberapa kategori seperti berita politik, olahraga, ekonomi dan lain sebagainya. Namun, dalam membagi berita ke dalam kategori-kategori tersebut untuk saat ini masih dilakukan secara manual. Hal ini sangat merepotkan apabila berita yang ingin diunggah berjumlah banyak. Oleh karena itu perlu adanya sistem yang bisa mengklasifikasikan berita secara otomatis. *Text mining* merupakan metode klasifikasi yang merupakan variasi dari data mining berusaha menemukan pola yang menarik dari sekumpulan data tekstual yang berjumlah besar. Sedangkan algoritma *naïve bayes classifier* merupakan lagoritmape ndukung untuk melakukan klasifikasi. Dalam penelitian ini data yang digunakan berupa berita yang berasal dari beberapa media online. Berita terdiri dari 4 kategori yaitu politik, ekonomi, olahraga, entertainment. Setiap kategori tediri dari 100 berita; 90 berita digunakan untuk proses *training* dan 10 berita digunakan untuk proses *testing*. Hasil dari penelitian ini menghasilkan sistem klasifikasi berita berbasis *web* dengan menggunakan bahasa pemrograman PHP dan database MySQL menunjukkan bahwa berita *testing* bisa terklasifikasi secara otomatis seluruhnya.

Kata kunci : Sistem Klasifikasi, Berita, *Text Mining*, *Naïve Bayes Classifier*.

I. PENDAHULUAN

Banyak instansi yang bergerak dalam penyaluran informasi masyarakat atau berita yang pada awalnya menyampaikan berita melalui media Televisi, Surat Kabar, Majalah atau Radio sudah mulai menggunakan sistem berbasis *web* untuk menyampaikan beritanya secara *up to date* [1].

Pada umumnya berita yang disampaikan dalam portal tersebut terdiri dari beberapa kategori seperti berita politik, olahraga, ekonomi, kesehatan, dan lain - lain (sebagai contoh pada website *www.kompas.com*, *www.waspada.com, dan www.vivanews.com*). Namun, dalam membagi berita ke dalam kategori-kategori tersebut untuk saat ini masih dilakukan secara manual, artinya dalam mengunggah berita pengunggah harus terlebih dahulu mengetahui isi dari berita yang akan diunggah secara kesuluruhan untuk selanjutnya dimasukkan ke dalam kategori yang tepat [2]. Hal ini sangat merepotkan bagi para pengunggah berita apabila jumlah berita yang ingin diunggah berjumlah banyak. Oleh karena itu, perlu adanya sistem bebasis *web* dimana sistem tersebut dapat mengklasifikasikan berita secara otomatis sesuai dengan kategori-kategori berita yang ada sehingga bisa membantu para pengunggah berita dalam mengunggah beritanya.

Text mining adalah salah satu teknik yang dapat digunakan untuk melakukan klasifikasi dimana, t*ext mining* merupakan variasi dari data mining yang berusaha menemukan pola yang menarik dari sekumpulan data tekstual yang berjumlah besar [3]. Selain klasifikasi, *text mining* juga digunakan untuk menangani masalah *clustering*, *information extraction*, dan *information retrival* [4].

II. INDENTIFIKASI MASALAH

Dalam penelitian ini akan dibangun sebuah sistem berbasis *web* dimana sistem tersebut dapat mengklasifikasikan berita secara otomatis. sehingga dapat dibuat rumusan masalahnya yaitu : bagaimana mengklasifikasikan berita secar otomatis.

Agar tulisan ini tidak keluar dari pokok permasalahan yang dirumuskan maka ruang lingkup pembahasan dibatasi pada algoritma yang digunakan dalam pengklasifikasian adalah *naïve bayes classifier*, perancangan program aplikasi sistem pengklasifikasian ini menggunakan bahasa pemrograman PHP dan *database server* Mysql, kategori berita yang digunakan hanya 4 kategori yaitu berita politik, ekonomi, olahraga dan *entertainment* dimana data berita tersebut diambil dari media berita *online*, berita yang digunakan dalam penelitian ini hanya berita berbahasa Indonesia, pada tahap *text mining* tidak dilakukan tahap *tagging* karena tidak menangani teks yang berbahasa inggris, penelitian ini tidak melakukan perbandiangan algoritma, sistem yang dibangun tidak disatukan dengan media berita yang sudah ada tetapi dengan membuat *homepage* sendiri dan menggunakan jaringan *offline*.

Dari permasalahan diatas, maka tujuan yang harus dicapai dan dilakukan dalam penelitian ini adalah membangun aplikasi pengklasifikasian berita dengan *text mining* menggunakan NBC (*Naïve Bayes Classifier*) sehingga bisa mempercepat proses klasifikasi dan menghasilkan kategori berita yang sesuai.

Manfaat yang diharapkan dari penelitian ini adalah memberikan efisiensi waktu dan efisiensi kerja bagi para penyedia berita dalam mengklasifikasikan berita dan membantu para pencari berita untuk mendapatkan berita yang mereka inginkan.

III. PENELITIAN TERDAHULU

Berdasarkan penelitian sebelumnya, ada beberapa algoritma stemming yang bisa digunakan untuk *stemming* bahasa Indonesia diantaranya algoritma *confix-stripping*, algoritma Porter *stemmer* bahasa Indonesia, algoritma Arifin dan Sutiono, dan Algorima Idris [8]. Dimana, Algoritma *confix-stripping stemmer* adalah algoritma yang akurat dalam *stemming* bahasa Indonesia [6].

IV. METODE PENELITIAN

A. Pengumpulan Data

Data yang digunakan pada penelitian ini berupa data berita yang didapat dari beberapa media online. Data berita tersebut berjumlah 400 data berita dan dibagi menjadi 4 kategori berita yaitu berita politik, berita ekonomi, berita olahraga, dan berita *entertaintment* dimana masing-masing kategori berjumlah 100 data berita. Dari 100 data berita tersebut 90 data berita dijadikan sebagai data *training* dan 10 data berita dijadikan sebagai data *testing*.

B. Text Mining

Text mining merupakan variasi dari data *mining* yang berusaha menemukan pola yang menarik dari sekumpulan data tekstual yang berjumlah besar [3]. Langkah-langkah yang dilakukan dalam *text mining* adalah sebagai berikut :

1. Text Preprocessing

Tindakan yang dilakukan pada tahap ini adalah *toLowerCase*, yaitu mengubah semua karakter huruf menjadi huruf kecil, dan *Tokenizing* yaitu proses penguraian deskripsi yang semula berupa kalimat – kalimat menjadi kata-kata dan menghilangkan delimiter-delimiter seperti tanda titik(.), koma(,), spasi dan karakter angka yang ada pada kata tersebut [7].

2. Feature Selection

Pada tahap ini tindakan yang dilakukan adalah menghilangkan *stopword* (*stopword removal*) dan *stemming* terhadap kata yang berimbuhan [3][4].

Stopword adalah kosakata yang bukan merupakan ciri (kata unik) dari suatu dokumen [5]. Misalnya “di”, “oleh”, “pada”, “sebuah”, “karena” dan lain sebagainya.

Stemming adalah proses pemetaan dan penguraian berbagai bentuk (*variants*) dari suatu kata menjadi bentuk kata dasarnya *(stem)* [6]. Tujuan dari proses *stemming* adalah menghilangkan imbuhan-imbuhan baik itu berupa prefiks, sufiks, maupun konfiks yang ada pada setiap kata.

C. Algoritma Confix-stripping Stemmer

Algoritma *confix-stripping stemmer* adalah algoritma yang digunakan untuk melakukan proses stemming terhadap kata-kata berimbuhan [8]. Algoritma *Confix-stripping stemmer* mempunyai aturan imbuhan sendiri dengan model sebagai berikut :

[[[AW +]AW +]AW +] Kata-Dasar [[+AK][+KK][+P]

AW : Awalan	AK : Akhiran
KK : Kata ganti kepunyaan	P : Partikel

Langkah-langkah algoritma *confix-stripping stemmer* adalah sebagai berikut :

1. Kata yang belum di-*stemming* dibandingkan ke dalam database kamus kata dasar. Jika ketemu, maka kata tersebut diasumsikan sebagai kata dasar dan algoritma berhenti. Jika kata tidak sesuai dengan kata dalam kamus, lanjut ke langkah 2.

2. Jika kata di-*input* memiliki pasangan awalan-akhiran “be-lah”,”be-an”,”me-i”,”di-i”,”pe-i”, atau “te-i” maka langkah stemming selanjutnya adalah 5, 3, 4, 5, 6, tetapi jika kata yang di-*input* tidak memiliki pasangan awalan-akhiran tersebut, langkah *stemming* berjalan normal yaitu 3, 4, 5, 6, 7.

3. Hilangkan partikel dan kata ganti kepunyaan. Pertama hilangkan partikel (“-lah”, “-kah”, “-tah”, “-pun”). Setelah itu hilangkan juga kata ganti kepunyaan (“-ku”, “-mu”, atau “-nya”). Contoh : kata “bajumulah”, proses *stemming* pertama menjadi “bajumu” dan proses stemming kedua menjadi “baju”. Jika kata “baju” ada di dalam kamus maka algoritma berhenti. Sesuai dengan model imbuhan, menjadi :

[[[AW+]AW+]AW+] Kata Dasar [+AK]

4. Hilangkan juga Akhiran (“-i”,”-an”, dan “-kan”), sesuai dengan model imbuhan, maka menjadi :

[[[AW+]AW+]AW+] Kata Dasar

Contoh : kata “membelikan” di-*stemming* menjadi ”membeli”, jika tidak ada dalam *database* kata dasar maka dilakukan proses penghilangan awalan.

5. Penghilangan awalan (“be-“, ”di-“, ”ke-“, ”me-”, ”pe-”, ”se-”, dan “te-“) mengikuti langkah-langkah berikut :

a. Algoritma akan berhenti jika :
   i. Awalan diidentifikasi bentuk sepasang imbuhan yang tidak diperbolehkan dengan akhiran (berdasarkan tabel 1) yang dihapus pada langkah 3.
   ii. Diidentifikasi awalan yang sekarang identik dengan awalan yang telah dihapus sebelumnya atau,
   iii. Kata tersebut sudah tidak memiliki awalan.

b. Identifikasi jenis awalan dan peluruhannya bila diperlukan.jenis awalan ditentukan dengan aturan dibawah ini.
   i. Jika awalan dari kata adalah “di-“, “ke-“, atau “se-“ maka awalan dapat langsung dihilangkan.
   ii. Hapus awalan “te-“, “be-“, “me-“, atau “pe-“ yang menggunakan aturan peluruhan yang dijelaskan pada tabel 2.
   Sebagai contoh kata “menangkap”, setelah menghilangkan awalan “me-“ maka kata yang didapat adalah “nangkap”. Karena kata “nangkap” tidak ditemukan dalam database kata dasar maka karakter “n” diganti dengan karakter “t” sehingga dihasilkan kata “tangkap” dan kata “tangkap” merupakan kata yang sesuai dengan kata yang ada di database kata dasar, maka algoritma berhenti.

6. Jika semua langkah gagal, maka kata yang diuji pada algoritma ini dianggap sebagai kata dasar.

Tabel 1. Kombinasi Prefix dan Sufiks yang tidak diperbolehkan [8].

Awalan (Prefiks)	Akhiran (Suffiks)
be-	-i
di-	-an
ke-	-i –kan
me-	-an
se-	-i –kan
te-	-an

Tabel 2. Aturan peluruhan kata dasar [8]

Aturan	Bentuk Awalan	Peluruhan
1	berV…	ber-V… | be-rV…
2	belajar…	bel-ajar
3	beC_1erC_2…	be-C_1erC_2…dimana C_1!={'r'|'l'}
4	terV…	ter-V…|te-rV…
5	terCer…	ter-Cer…dimana C!='r'
6	teC_1erC_2	te-C_1erC_2…dimana C_1!='r'
7	me{l|r|w|y}V…	me-{l|r|w|y}V…
8	mem{b|f|v}…	mem-{b|f|v}…
9	mempe…	mem-pe…
10	mem{rV|V}…	me-m{rV|V}…|me-p{rV|V}…
11	men{c|d|j|z}…	men-{c|d|j|z}…
12	menV…	me-nV…|me-tV…
13	meng{g|h|q|k}…	meng-{g|h|q|k}…
14	mengV…	meng-V…|meng-kV…
15	mengeC	menge-C
16	menyV…	me-ny…| meny-sV…
17	mempV…	mem-pV…
18	pe{w|y}V…	pe-{w|y}V…
19	perV…	per-V…|pe-rV…
20	pem{b|f|v}…	pem-{b|f|v}…
21	pem{rV|V}…	pe-m{rV|V}…|pe-p{rV|V}
22	pen{c|d|j|z}…	pen-{c|d|j|z}…
23	penV…	pe-nV…|pe-tV…
24	peng{g|h|q}	peng-{g|h|q}
25	pengV	peng-V|peng-kV
26	penyV…	pe-nya|peny-sV
27	pelV..	pe-lV…; kecuali untuk kata "pelajar" menjadi "ajar"
28	peCP	pe-CP…dimana C!={r|w|y|l|m|n} dan P!='er'

29	perCerV	Per-CerV… dimana C!={r|w|y|l|m|n}	

D. Algoritma Naïve Bayes Classifier

Algoritma *naive bayes classifier* merupakan algoritma yang digunakan untuk mencari nilai probabilitas tertinggi untuk mengklasifikasi data uji pada kategori yang paling tepat [3]. Dalam penelitian ini yang menjadi data uji adalah dokumen berita. Ada dua tahap pada klasifikasi dokumen. Tahap pertama adalah pelatihan terhadap dokumen yang sudah diketahui kategorinya. Sedangkan tahap kedua adalah proses klasifikasi dokumen yang belum diketahui kategorinya.

Dalam algoritma *naïve bayes classifier* setiap dokumen direpresentasikan dengan pasangan atribut "$x_1, x_2, x_3,...x_n$" dimana x_1 adalah kata pertama, x_2 adalah kata kedua dan seterusnya. Sedangkan V adalah himpunan kategori berita. Pada saat klasifikasi algoritma akan mencari probabilitas tertinggi dari semua kategori dokumen yang diujikan (V_{MAP}), dimana persamaannya adalah sebagai berikut :

$$V_{MAP} = \underset{V_j \, e \, V}{\arg\max} \left(\frac{P(x_1,x_2,x_3,...x_n|V_j)P(V_j)}{P(x_1,x_2,x_3,...x_n)} \right) \quad [1]$$

Untuk $P(x_1, x_2, x_3,...x_n)$ nilainya konstan untuk semua kategori (V_j) sehingga persamaan dapat ditulis sebagai berikut :

$$V_{MAP} = \underset{V_j \, e \, V}{argmax} \left(P(x_1, x_2, x_3, \dots x_n|V_j)P(V_j) \right) \quad [2]$$

Persamaan diatas dapat disederhanakan menjadi sebagai berikut :

$$V_{MAP} = \underset{V_j \, e \, V}{\arg max} \prod_{i=1}^{n} \left(P(x_i|V_j)P(V_j) \right) \quad [3]$$

Keterangan :

V_j = Kategori berita j =1, 2, 3,…n. Dimana dalam penelitian ini j_1 = kategori berita politik, j_2 = kategori berita ekonomi, j_3 = kategori berita olahraga, dan j_4 = kategori berita *entertainment*

$P(x_i|V_j)$ = Probabilitas x_i pada kategori V_j

$P(V_j)$ = Probabilitas dari V_j

Untuk $P(V_j)$ dan $P(x_i|V_j)$ dihitung pada saat pelatihan dimana persamaannya adalah sebagai berikut :

$$P(V_j) = \frac{|docs\ j|}{|contoh|} \quad [4]$$

$$P(x_i|V_j) = \frac{n_k + 1}{n + |kosakata|} \quad [5]$$

Keterangan :

|docs j| = jumlah dokumen setiap kategori j

|contoh| = jumlah dokumen dari semua kategori

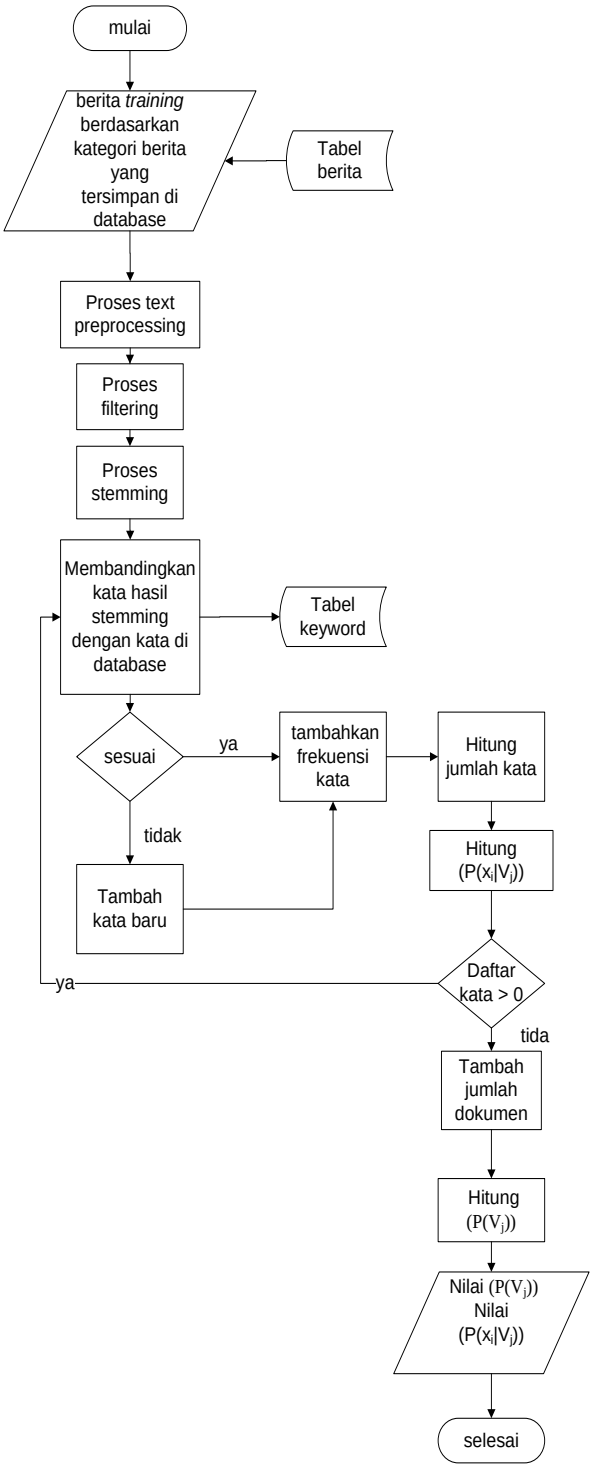
n_k = jumlah frekuensi kemunculan setiap kata

n = jumlah frekuensi kemunculan kata dari setiap kategori

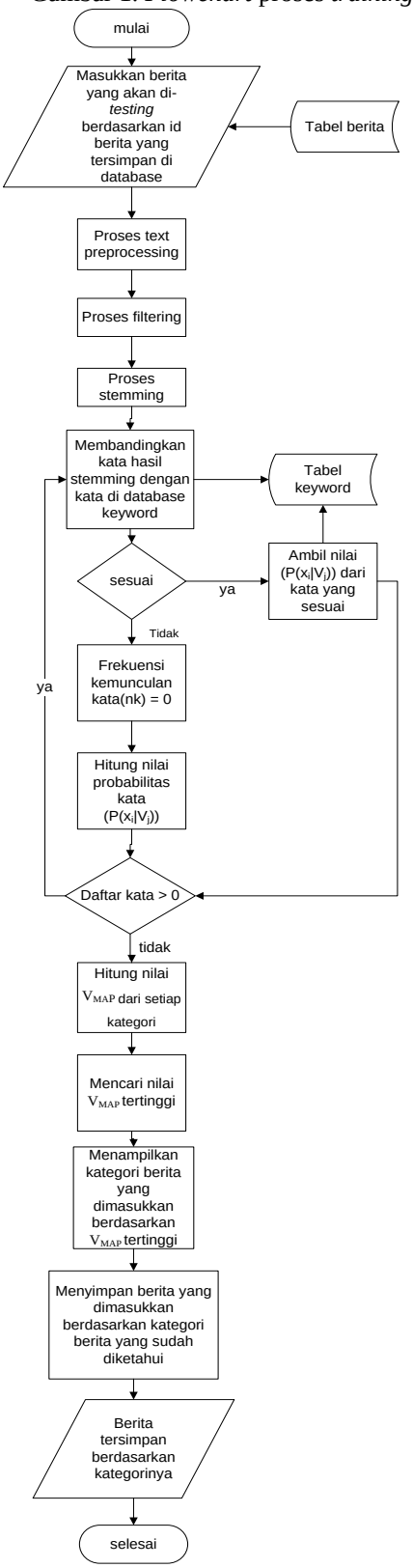
|kosakata| = jumlah semua kata dari semua kategori

V. HASIL DAN PEMBAHASAN

A. Pengujian Sistem



Gambar 1. *Flowchart* proses *training*



Gambar 2. *Flowchart* proses *testing*

Dalam penelitian ini sistem mempunyai 2 tahapan proses yaitu tahapan pertama adalah tahap pembelajaran atau *training* yaitu tahap pengklasifikasian terhadap berita yang sudah diketahui kategorinya. Gambar 1 dan gambar 2 merupakan *flowchart* dari proses *training* dan proses *testing*

Pada pengujian tahap *testing* hal-hal yang dilakukan adalah dengan melakukan klasifikasi berita yang belum diketahui kategorinya. Berita yang dijadikan pengujian tahap *testing* berjumlah 10 berita untuk masing-masing kategori. Berita-berita yang digunakan untuk proses *testing* dapat dilihat pada gambar 3.

Id	Kategori	Judul
1		KPK Minta Imigrasi Cegah Emir Moeis
2		PPP: Penyebutan JK Bukan untuk Pencitraan
3		Kemenpora Digeledah, Ini Komentar Andi Mallarangeng
4		PKS Anggap Hasil Survei 'Kompas' Peringatan Dini
5		Anis: JK Pantas Jadi Capres!
6		Pencapresan JK Tidak Akan Berdampak Buruk buat Ical
7		Hasil Survei, JK Unggul dari Ical
8		Prabowo-JK Pasangan Terpopuler Pemilu 2014
9		PKS-PPP Dukung Foke? Ini Komentar Kubu Jokowi-Ahok....
10		Kurang Sosialisasi, Penyebab Kekalahan Hidayat-Didik
11		Rossi Kembali ke Yamaha? Lorenzo Tetap Santai
12		10 Atlet Terseksi di Olimpiade London
13		Arsenal Resmi Gaet Cazorla
14		Spies Tinggalkan Yamaha
15		Honda Sediakan 4 Motor bagi Pedrosa
16		Atlet Spanyol Cemooh Seragam Olimpiade
17		Rossi Tak Pernah Jadi Pilihan untuk Gantikan Stoner!
18		Dovizioso Ukir Sejarah Baru bagi Tech 3 di MotoGP
19		Adam Tak Punya Maksud Jahat terhadap Bale
20		Tak Ada Bikini di Voli Pantai Olimpiade

Gambar 3. Berita *testing*

Setelah melakukan pengujian proses *testing* maka berita-berita yang belum berkategori akan mendapatkan kategori yang sesuai. Berita-berita hasil pengujian proses *testing* dapat dilihat pada gambar 4.

21	olahraga	10 Atlet Terseksi di Olimpiade London
22	olahraga	Arsenal Resmi Gaet Cazorla
23	olahraga	Rossi Kembali ke Yamaha? Lorenzo Tetap Santai
24	olahraga	Tak Ada Bikini di Voli Pantai Olimpiade
25	olahraga	Adam Tak Punya Maksud Jahat terhadap Bale
26	olahraga	Dovizioso Ukir Sejarah Baru bagi Tech 3 di MotoGP
27	olahraga	Rossi Tak Pernah Jadi Pilihan untuk Gantikan Stoner!
28	olahraga	Atlet Spanyol Cemooh Seragam Olimpiade
29	olahraga	Honda Sediakan 4 Motor bagi Pedrosa
30	olahraga	Spies Tinggalkan Yamaha
31	politik	Anis: JK Pantas Jadi Capres!
32	politik	Pencapresan JK Tidak Akan Berdampak Buruk buat Ical
33	politik	Kurang Sosialisasi, Penyebab Kekalahan Hidayat-Didik
34	politik	PKS-PPP Dukung Foke? Ini Komentar Kubu Jokowi-Ahok....
35	politik	Prabowo-JK Pasangan Terpopuler Pemilu 2014
36	politik	Hasil Survei, JK Unggul dari Ical
37	politik	KPK Minta Imigrasi Cegah Emir Moeis
38	politik	PPP: Penyebutan JK Bukan untuk Pencitraan
39	politik	PKS Anggap Hasil Survei \
40	politik	Kemenpora Digeledah, Ini Komentar Andi Mallarangeng

Gambar 4. Hasil pengujian tahap *testing*

VI. KESIMPULAN

Setelah melakukan studi literatur, perancangan, analisis, implementasi dan pengujian aplikasi pengklasifikasian berita secara otomatis maka dapat disimpulkan Aplikasi ini sudah mampu melakukan proses klasifikasi data berita secara otomatis dan proses klasifikasi semakin akurat jika data latih yang digunakan dalam pembelajaran berjumlah banyak. Untuk penelitian berikutnya diharapkan sistem ini tidak hanya untuk mengklasifikasi berita melainkan bisa juga digunakan untuk mengklasifikasikan dokumen lain seperti kesenian, olahraga, dan jurnal.

DAFTAR PUSTAKA

[1] Fajar, Muhammad. 2008. Media cetak era digital. www.emfajar.net/internet/media-cetak-di-era-digital/. Diakses tanggal 5 Juli 2011.

[2] Lin, S. 2008. A document classification and retrival system for R&D in semiconductor industry-A hybrid approach.*Expert System* 18, 2:4753-4764.

[3] Feldman, R & Sanger, J. 2007. The Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press : New York.

[4] Berry, M.W. & Kogan, J. 2010. Text Mining Aplication and theory. WILEY : United Kingdom.

[5] Dragut, E., Fang, F., Sistla, P., Yu, S. & Meng, W. 2009. Stop Word and Related Problems in Web Interface Integration. http://www.vldb.org/pvldb/2/vldb09-384.pdf. Diakses tanggal 8 Desember 2011.

[6] Tala, Fadillah Z. 2003. A Study of Stemming Efects on Information Retrieval in Bahasa Indonesia. Institute for Logic, Language and ComputationUniversiteit van Amsterdam The Netherlands.http://www.illc.uva.nl/Research/Reports/MoL-2003-02.text.pdf. Diakses tanggal 29 September 2011.

[7] Weiss, S.M., Indurkhya, N., Zhang, T., Damerau, F.J. 2005. Text Mining : Predictive Methods fo Analyzing Unstructered Information. Springer : New York.

[8] Adriani, M., Asian, J., Nazief, B., Tahaghoghi, S.M.M., Williams, H.E. 2007. Stemming Indonesian : A Confix-Stripping Approach. Transaction on Asian Langeage Information Processing. Vol. 6, No. 4, Artikel 13. Association for Computing Machinery : New York .