ISBN: 978-602-60280-1-3

Uji Akurasi Pengategorian Otomatis Berita Berbahasa Indonesia Dengan Algoritma Multinomial Naive Bayes Classifier (NBC)

1st Dwi Yulian Rachmanto Lingke
Program Studi Teknik Informatika
Sekolah Tinggi Ilmu Komputer PGRI Banyuwangi
Email : lingkeku@gmail.com

2nd Ahmad Su'udi
Program Studi Teknik Informatika
Sekolah Tinggi Ilmu Komputer PGRI Banyuwangi
Email : osing.banyuwangi@gmail.com

***Abstrak*— Program pengategorian otomatis menjadi kebutuhan menyusul banyaknya berita online yang diproduksi media-media *mainstream* atau media arus utama, maupun media alternatif di Indonesia. *Multinomial Naive Bayes Classifier* (NBC) merupakan sebuah metode klasifikasi teks yang berakar pada teorema Bayes. Metode pengklasifikasian dengan menggunakan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai Teorema Bayes. Ciri utama dari Multinomial NBC ini adalah asumsi yang sangat kuat (naif) akan independensi dari masing-masing kondisi atau kejadian. Penelitian ini bertujuan mengetahui akurasi hasil program berdasarkan algoritma Multinomial NBC dengan hasil kerja pakar konten media mainstream. Hasil dari penelitian menunjukkan bahwa proses pengambilan keputusan pada kategori politik secara tematik menghasilkan akurasi 40% dengan 100 data training dan 60% dengan 200 data training. Proses pengambilan keputusan pada kategori ekonomi secara tematik menghasilkan akurasi 90%, dengan 100 data training, maupuan 200 data training.**

Kata Kunci*— *Pengategorian, Berita Online, Multinomial Naive Bayes Classifier, NBC

I. PENDAHULUAN

A. Latar Belakang

Pertumbuhan jumlah website di international network (internet) kian pesat. Sampai September 2017, di dunia ada 331.9 juta *top level domain* (tld) yang telah terdaftar [1]. Setiap website memiliki konten-konten yang diupload untuk dibagikan kepada mereka yang membutuhkan informasi.

Sementara itu Dewan Pers RI merilis data bahwa Indonesia memiliki 44.300 media online [2]. Salah satu di antaranya, adalah Okezone.com yang memroduksi 1.000 berita per hari hingga tercatat di Museum Rekor Indonesia (MURI) [3].

Produksi konten media online yang tinggi memerlukan pengategorian teks otomatis untuk memudahkan editor menangani berita yang masuk. Tanpa harus menentukan kategori berita, editor bisa lebih fokus pada editing naskah, baik dari isi berita maupun tata Bahasa Indonesia.

Penelitian kali ini menggunakan algoritma Multinomial Naive Bayes Classifier (NBC) untuk membangun aplikasi pengategorian berita berbahasa Indonesia secara otomatis. Mempermudah pekerjaan wartawan dan editor, dalam hal menentukan kategori berita sehingga bisa fokus pada tugas lain, mengingat yang dilakukan editor berita masih secara manual.

1.1 Tujuan

Tujuan dari penelitian ini adalah untuk menerapkan algoritma Multinomial NBC dengan framework PHP Codeigniter (CI) dalam pengategorian berita berbahasa Indonesia secara otomatis dan menghitung akurasinya.

II. TINJAUAN PUSTAKA

2.1 Penelitian Terkait

Inspirasi yang diperoleh penulis untuk penelitian kali ini adalah dari karya-karya yang banyak menyinggung tema dari pembahasan yang menyangkut pengategorian teks otomatis dan algoritma NBC.

Tinjauan pustaka pertama berjudul Kategorisasi Dokumen Web Secara Otomatis Berdasarkan Folksonomy Menggunakan Multinomial Naive Bayes Classifier, yang ditulis Hendry Irawan mahasiswa Sekolah Tinggi Teknologi (STT) Telkom Bandung, tahun 2005. Penelitian itu menggali metode folksonomy yang juga digunakan di http://del.icio.us, sebuah web social bookmark yang menyuguhkan link konten web-web lain dalam beberapa kategori sesuai dengan minat masing-masing pembaca. Penelitian itu menggabungkan metode folksonomy dengan multinomial NBC. [4]

Tinjauan pustaka berikutnya berjudul Sistem Klasifikasi Surat Masuk Menggunakan Multinomial Naive Bayes, karya Fran Sandy Herlin Hanopo, Universitas Sanata Dharma Yogyakarta, tahun 2013. Penelitian itu berisi uji akurasi algoritma Multinomial NBC ketika diterapkan pada surat-surat masuk di kampus penulis itu sendiri. Salah satu saran yang diberikan adalah, penyiapan data training melibatkan pakar agar menghasilkan model-model teks yang layak dijadikan data training. [5]

Tinjauan pustaka yang lain berjudul Analisis Sentimen Pada Lirik Lagu Menggunakan Metode Multinomial Naive Bayes, yang ditulis Elsa Rika Octaviana tahun 2015, di Universitas Sanata Dharma Yogyakarta. Penelitian ini bertujuan mengetahui lagu mana saja yang baik dan tidak baik untuk anak berdasarkan lirik masing-masing lagu. Salah satu kesimpulannya mengatakan, bahwa dalam algoritma Multinomial NBC, frekuensi kemunculan kata sangat berpengaruh terhadap hasil yang didapat. [6]

2.2 Algoritma Multinomial Naive Bayes Classifier (NBC)

Naive Bayes Classifier (NBC) adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class. Algoritma itu didasarkan pada teorema Naive Bayes yang memiliki kemampuan klasifikasi serupa dengan decision tree dan neural network [7]. NBC juga terbukti mampu memberikan hasil dengan akurasi dan kecepatan yang tinggi

ISBN: 978-602-60280-1-3

saat diaplikasikan ke dalam database dengan data yang besar. Multinomial Naive Bayes merupakan proses pengambilan jumlah kata yang muncul dalam setiap dokumen, dengan mengasumsikan dokumen memiliki beberapa kejadian dalam kata dengan panjang yang tidak tergantung dari kelasnya dalam dokumen [8]. Hendry Irawan menulis Multinomial NBC adalah algoritma yang menggabungkan Teori Probabilitas dan Satistika, Teorema Bayes, Bayes Classifier, Binary Independence Model dan Multinomial Model untuk diterapkan pada permasalahan klasifikasi teks secara otomatis [4].

- Menghitung Multinomial NBC

Menurut Elsa Rifka Octaviana dalam Manning, Raghavan, Schutze, probabilitas sebuah dokumen d berada di kelas c, kondisi berikut dapat dinyatakan dengan rumus : [9]

$$P(c|d) \propto P(c) \prod_{1 \leqslant k \leqslant n_d} P(t_k|c)$$

Keterangan :

$P(t_k|c)$ = Merupakan *conditional probabilities* dari kata tk yang terdapat dalam sebuah dokumen dari kelas c.

P(c) = Merupakan *prior probabilities* dari sebuah dokumen yang terdapat dalam kelas c.

K = Merupakan urutan angka dari 1 sampai pada nd.

n_d = Jumlah token atau atribut dalam dokumen d.

- Menghitung P(c)

Untuk memperkirakan prior probability P(C) dinyatakan dengan rumus :

$$P(c) = \frac{Nc}{N}$$

Keterangan :

Nc = Jumlah dokumen training dalam kategori c.

N = Jumlah seluruh dokumen training dari seluruh kelas.

- Menghitung P(t|c)

Untuk perkiraan conditional probability P(t|c) dinyatakan dengan rumus :

$$P(t|c) = \frac{Tct}{\sum t' \in v\, Tct'}$$

Keterangan :

Tct = Jumlah kemunculan kata t dalam sebuah dokumen training di kelas c.

$\sum t' \in v\, Tct'$ = Jumlah keseluruhan term dalam dokumen training di kelas c.

Untuk menghilangkan nilai nol pada sebuah dokumen, digunakan *laplace smoothing* sebagai proses penambahan nilai 1 pada setiap nilai Tct pada perhitungan conditional probabilities dan dinyatakan dengan rumus :

$$P(tPtk|c) = \frac{Tct + 1}{\sum t' \in v\, Tct + B'}$$

B' = Total kata unik pada seluruh kelas dalam dokumen training (2015).

2.3 Algoritma Nazief & Adriani

Stemming untuk Bahasa Indonesia dan Bahasa Inggris saling berbeda karena berbedanya tata bahasa yang dimiliki. Bahasa Indonesia memiliki imbuhan depan (prefiks), akhiran (sufiks), sisipan (infiks) dan konfiks yang berupa awalan dan akhiran sekaligus [10]. Algoritma Nazief & Adriani merupakan algoritma untuk mengubah kata yang memiliki *sufiks*, *prefix*, *infiks* dan *konfiks* menjadi bentuk kata dasar [11].

Setidaknya ada 2 algoritma stemming yang bisa diterapkan untuk bahasa Indonesia, yakni algoritma Porter untuk Bahasa Indonesia dan algoritma Nazief & Adriani. Algoritma Nazief & Adriani membutuhkan database kosakata Bahasa Indonesia yang lengkap, seperti halnya dalam Kamus Besar Bahasa Indonesia (KBBI) [12].

Dalam penelitian yang dilakukan Keke, Dyan, dkk, dijelaskan langkah-langkah yang ada dalam algoritma Nazief dan Adriani sebagai berikut : [13]

1. Pertama cari kata yang akan diistem dalam kamus kata dasar. Jika ditemukan, maka diasumsikan kata adalah root word. Maka algoritma berhenti.
2. *Inflection Suffixes* ("-lah", "-kah", "-ku", "-mu", atau "-nya") dibuang. Jika berupa particles ("-lah", "-kah", "-tah" atau "-pun") maka langkah ini diulangi lagi untuk menghapus *Possesive Pronouns* ("-ku", "-mu", atau "-nya"), jika ada.
3. Hapus *Derivation Suffixes* ("-i", "-an" atau "-kan"). Jika kata ditemukan di kamus, maka algoritma berhenti. Jika tidak maka ke langkah 3a.
   a. Jika "-an" telah dihapus dan huruf terakhir dari kata tersebut adalah "-k", maka "-k" juga ikut dihapus. Jika kata tersebut ditemukan dalam kamus maka algoritma berhenti. Jika tidak ditemukan maka lakukan langkah 3b.
   b. Akhiran yang dihapus ("-i", "-an" atau "-kan") dikembalikan, lanjut ke langkah 4.
4. Hilangkan derivation prefixes ("di-", "ke-", "se-", "me-", "be-", "pe", "te-") dengan iterasi maksimum adalah 3 kali
   a. Langkah 4 berhenti jika :
   - Ada kombinasi awalan dan akhiran pada Tabel 1.

TABEL 1. PASANGAN AWALAN DAN AKHIRAN TERLARANG

Awalan	Akhiran Ditolak
be	i
di	an
ke	i, kan
me	an
se	i, kan
te	an

   - Awalan yang dideteksi saat ini sama dengan awalan yang dihilangkan.
   - Tiga awalan telah dihilangkan.

ISBN: 978-602-60280-1-3

b. Identifikasikan tipe awalan dan hilangkan. Awalan ada tipe:
   - Standar : “di-”, “ke-”, “se-” yang dapat langsung dihilangkan dari kata.
   - Kompleks : “me-”, “be-”, “pe”, “te-” adalah tipe-tipe awalan yang dapat bermorfologi sesuai kata dasar yang mengikutinya. Oleh karena itu, gunakan aturan pada Tabel 2 untuk mendapatkan pemenggalan yang tepat.

TABEL 2. VARIASI AWALAN DAN CARA MEMOTONGNYA

NO	Sifat Imbuhan	Cara Pemotongan
1	berV...	ber-V... | be-rV...
2	belajar...	bel-ajar
3	beC$_1$erC$_2$...	be-C$_1$erC$_2$... dimana C$_1$!={'r'|'l'}
4	terV...	ter-V... | te-rV...
5	terCer...	ter-Cer... dimana C!='r'
6	teC$_1$erC$_2$...	te-C$_1$erC$_2$... dimana C!='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-nyV... | meny-sV...
17	mempV...	mem-pV...
18	pe{w|y}V...	pe-{w|y}V...
19	perV...	per-V...
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-nyV... | peny-sV...
27	pelV...	pe-lV... kecuali untuk kata 'pelajar' yang seharusnya berubah 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}

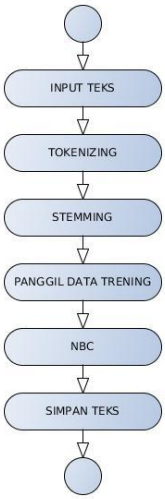
c. Cari kata yang telah dihilangkan awalannya ini di dalam kamus. Apabila tidak ditemukan, maka langkah 4 diulangi kembali. Apabila ditemukan, maka keseluruhan proses dihentikan

5. Apabila setelah langkah 4 kata dasar masih belum ditemukan, maka proses recoding dilakukan dengan mengacu pada aturan pada Tabel 2. Recoding dilakukan dengan menambahkan karakter recoding di awal kata yang dipenggal. Pada Tabel 2, karakter recoding adalah huruf kecil setelah tanda hubung (‘-’) dan terkadang berada sebelum tanda kurung. Sebagai contoh, kata “menangkap” (aturan 15), setelah dipenggal menjadi “nangkap”. Karena tidak valid, maka recoding dilakukan dan menghasilkan kata “tangkap”.
6. Jika semua langkah telah selesai tetapi tidak juga berhasil maka kata awal diasumsikan sebagai root word atau kata dasar.
7. Proses selesai.

III. METODE PENELITIAN

3.1 Activity Diagram Utama

Activity diagram yang dibuat adalah aktifitas program sesuai dengan langkah-langkah yang tertera pada algoritma



Multinomial NBC. Beberapa aktivitas program desainnya dijelaskan dalam diagram tersendiri untuk menyederhanakan gambaran program.

Gambar 1. Activity diagram umum algoritma NBC

Gambar 1 adalah activity diagram umum yang menjelaskan tahapan kerja dari pra proses sampai pada langkah perhitungan algoritma Multinomial NBC. Diawali penginputan teks, tokenizing yang memecah teks menjadi kata-kata dasar yang berdiri sendiri menjadi token, stemming untuk mengubah kata-kata itu menjadi kata dasar, pemanggilan data training, proses NBC, hingga penyimpanan teks baru.

3.2 Activity Diagram Tokenizing

Gambar 2. Activity diagram proses tokenizing

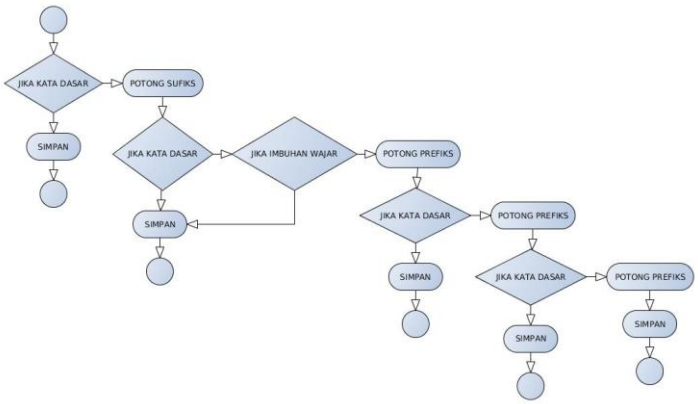
Gambar 2 adalah activity diagram proses tokizing sebelum masuk dalam proses inti algoritma NBC. Prosesnya meliputi penghapusan tanda baca, memperkecil semua huruf, mengubah teks menjadi kata-kata yang saling terpisah, yang

ISBN: 978-602-60280-1-3

juga biasa disebut sebagai token, eliminasi spasi dan eliminasi stopword.

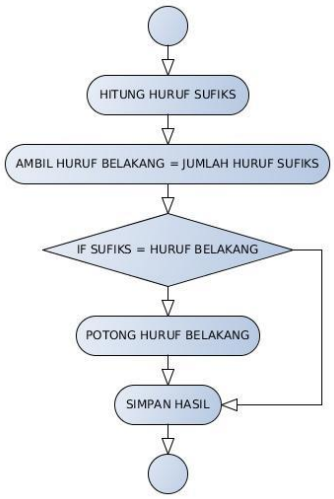
3.3 Activity Diagram Stemming

Gambar 3 adalah activity diagram untuk proses stemming, yakni mengubah kata-kata yang dihasilkan dari proses sebelumnya menjadi kata dasar. Proses ini menggunakan algoritma Nazief & Adriani yang pronsip kerjanya menemukan dan memotong imbuhan yang menempel di kata tersebut, baik sufiks, prefiks, infiks maupun konfiks.



Gambar 3. Activity diagram proses stemming

3.4 Activity Diagram Pemotongan Sufiks (Imbuhan Akhiran)



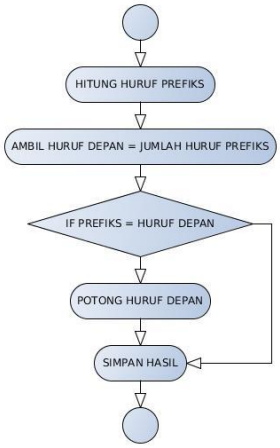
Gambar 4. Activity diagram pemotongan sufiks

Gambar 4 adalah gambar activity diagram pemotongan sufiks atau imbuhan akhiran dari kata yang sedang diproses. Setidaknya ada 5 aktifitas dalam tahapan ini, dari penghitungan jumlah huruf imbuhan hingga penyimpanan hasil pemotongan untuk diikutkan dalam tahapan selanjutnya. Keseluruhan proses ini dilakukan secara looping sesuai dengan jumlah imbuhan yang diproses atau terjadi pemotongan imbuhan akhiran.

1. Langkah pertama adalah menghitung huruf sufiks. Ada sufiks yang terdiri dari 1 huruf, misalnya ‘i’, ada pula yang terdiri dari 2 huruf, misal ‘mu’, atau yang terdiri dari 3 huruf seperti ‘kan’. Maka diperlukan penghitungan jumlah huruf imbuhan yang sedang diproses, agar tahu juga, berapa huruf di akhir kata yang sedang diproses yang harus dipotong. Misalkan imbuhan yang diproses ‘kan’, yang artinya terdiri dari 3 huruf, maka 3 huruf akhir dari kata yang sedang diproses diambil untuk diperbandingkan. Keputusan pengambilan huruf akhir sejumlah huruf imbuhan ini dilakukan di langkah ke-2.
2. Langkah selanjutnya merupakan decision atas kesamaan huruf-huruf yang dipotong dengan imbuhan yang sedang diproses. Jika tidak sama, berarti imbuhan yang tengah diproses tidak terkandung dalam kata yang sedang diproses, dan tidak akan dilakukan pemotongan imbuhan dari kata yang sedang diproses tersebut. Jika sama, berarti kata itu jelas mengandung imbuhan yang tengah diproses, dan akan masuk ke proses selanjutnya, yakni proses pembuangan imbuhan itu.
3. Bila dalam langkah sebelumnya dipastikan kata yang tengah diproses mengandung imbuhan, akan masuk kemari dan dibuang imbuhan belakangnya.
4. Pemotongan imbuhan belakang terjadi atau tidak, akan masuk ke proses akhir ini, dimana hasil disimpan dalam variabel hasil.

Sering kali didapati sebuah kata memiliki 1, 2 atau 3 sufiks yang mengikutinya, sehingga diperlukan langkah khusus urutan khusus untuk memotongnya. Pemotongan dengan algoritma Nazief dan Adriani, mula-mula mendeteksi dan memotong sufiks lah, kah, ku, mu, nya. Namun bila ditemukan partikel lah, kah, tah, pun, maka proses dilanjutkan dengan deteksi dan pemotongan prefiks ku, mu dan nya.

3.5 Activity Diagram Pemotongan Prefiks (Imbuhan Awalan)



Gambar 5. Activity diagram pemotongan prefiks

Gambar 5 adalah gambar activity diagram pemotongan prefiks atau imbuhan awalan dari kata yang sedang diproses. Setidaknya ada 5 aktifitas dalam tahapan ini, dari penghitungan jumlah huruf imbuhan hingga penyimpanan hasil pemotongan untuk diikutkan dalam tahapan selanjutnya. Keseluruhan proses ini dilakukan secara looping sesuai dengan jumlah imbuhan yang diproses atau terjadi pemotongan imbuhan awalan.

1. Langkah pertama adalah menghitung jumlah huruf prefiks yang tengah diproses. Ada prefiks yang terdiri

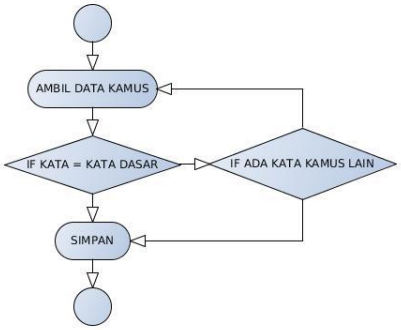
ISBN: 978-602-60280-1-3

dari 2 huruf, misalnya ‘me’, ada pula yang terdiri dari 3 huruf, misal ‘men’, atau yang terdiri dari 4 huruf seperti ‘meng’. Maka diperlukan penghitungan jumlah huruf imbuhan yang sedang diproses, agar tahu juga, berapa huruf di awal kata yang sedang diproses yang harus dipotong.

2. Misalkan imbuhan yang diproses ‘men’, yang artinya terdiri dari 3 huruf, maka 3 huruf awal dari kata yang sedang diproses diambil untuk diperbandingkan. Pengambilan huruf akhir sejumlah huruf imbuhan ini dilakukan di langkah ke-2.
3. Langkah selanjutnya merupakan decision atas kesamaan huruf-huruf yang dipotong dengan imbuhan yang sedang diproses. Jika tidak sama, berarti imbuhan yang tengah diproses tidak terkandung dalam kata yang yang sedang diproses, dan tidak akan dilakukan pembuangan huruf dari kata yang sedang diproses tersebut. Jika sama, berarti kata itu jelas mengandung imbuhan yang tengah diproses, dan akan masuk ke proses selanjutnya, yakni proses pembuangan imbuhan tersebut.
4. Bila dalam langkah sebelumnya dipastikan kata yang tengah diproses mengandung imbuhan, akan masuk kemari dan dibuang imbuhan awalannya.
5. Pemotongan imbuhan awalan terjadi atau tidak, akan masuk ke proses akhir ini yang keluarannya disimpan sebagai hasil.

Dalam tahapan pemotongan prefiks ini, perlu diperhatikan bahwa ada 5 kali pemotongan, pertama untuk awalan di, ke dan se yang tidak memiliki perubahan bunyi atau fonem. 4 bagian pemotongan prefiks lain adalah, untuk awalan yang mengalami morfologi, yakni be, ter, me, dan pe, masing memiliki banyak variasi imbuhan sehingga memerlukan proses sendiri-sendiri. Maka setiap kata yang diproses akan mengalami pemrosesan pemotongan pada 5 kelompok prefiks tersebut.

3.6 Activity Diagram Pengecekan Kesesuaian Kata Dasar

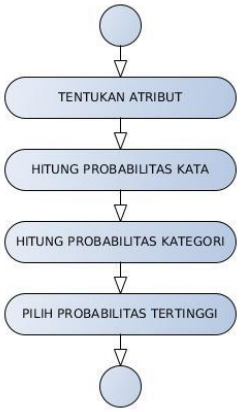


Gambar 6. Activity diagram pengecekan kesesuaian kata

Gambar 6 adalah gambar activity diagram untuk proses pengecekan kesesuaian kata yang tengah diproses dengan kata dasar dari database. Bila sama, berarti kata dasar telah ditemukan, dan program akan lanjut memproses kata selanjutnya dari teks berbahasa Indonesia yang diinputkan. Mula-mula program mengambil daftar kata-kata dasar Kamus Bahasa Indonesia dari database. Lalu dicek kesesuaiannya satu-satu dengan kata yang sedang diproses.

Jika ditemukan kata yang sama, maka kata yang sedang diproses telah menjadi kata dasar yang bisa disimpan sebagai hasil. Bila sampai akhir pengulangan tidak ada kata dasar yang sesuai dengan kata yang sedang diproses, maka kata yang sedang diproses disimpan sebagai hasil.

3.7 Multinomial NBC



Gambar 7. Activity diagram utama algoritma NBC

Gambar 7 adalah activity diagram utama dalam pembahasan algoritma Multinomial NBC. Langkah pertama menentukan atribut mana saja yang digunakan dalam perhitungan probabilitas, yakni atribut yang ada di kelas training dan dalam teks baru. Dalam penelitian ini, atribut berupa kata-kata yang telah lolos dari proses-proses sebelumnya. Bila sebuah atribut hanya dimiliki dokumen kelas training, atau hanya dimiliki dokumen baru, maka tidak dipakai [6]. Langkah kedua adalah menghitung probabilitas masing-masing atribut di masing-masing klas training. Dari probabilitas masing-masing atribut, kemudian dihitung peluang dokumen baru masuk kategori mana.

IV. HASIL DAN PEMBAHASAN

NO	Judul	Tema	Menggunakan 100 Training		Menggunakan 200 Training	
			True	False	True	False
1	Genjot Bisnis Properti, Ini Strategi Pembangunan Jaya Ancol	bisnis properti	1		1	
2	Sukses Bisnis Properti dengan Modal Rp 20 Juta			1		1
3	Satgas Pangan Cek Penyebab Lonjakan Harga Barang Kebutuhan Pokok	harga kebutuhan pokok	1		1	
4	Menjelang Ramadan, Harga Kebutuhan Pokok Naik		1		1	
5	Inflasi Rendah, BI Tak Kendurkan Kebijakan Moneter	kebijakan bi	1		1	
6	BI: Kebijakan Moneter Perlu Instrumen Kebijakan Fiskal		1		1	
7	Kurs Rupiah Melemah Jadi 13.778 Akibat Pidato Pejabat The Fed	kurs rupiah	1		1	
8	Kurs Rupiah Rp 14.000 per Dolar AS, Ini Rencana Sri Mulyani		1		1	
9	Suku Bunga Bank Indonesia Naik, Berapa Kenaikan Bunga Deposito?	suku bunga	1		1	
10	BI Naikkan Suku Bunga Acuan, Ini Dampak Positif Negatifnya		1		1	
Jumlah Hasil Tepat			9	1	9	1
Prosentase (Jumlah hasil tepat / 10 X 100)			90 %	10 %	90 %	10 %

A. Skenario Uji Coba

Percobaan sudah dilakukan dengan melibatkan total 240 teks berita. Sebanyak 100 teks merupakan berita untuk data training berkategori politik dan 100 teks lagi untuk data training kategori ekonomi. Sebanyak 10 teks menjadi bahan uji kategori politik secara tematik, 10 untuk uji kategori politik secara acak, 10 untuk uji kategori ekonomi secara tematik, dan 10 teks terakhir untuk uji kategori secara acak. Untuk memastikan teks yang dimasukkan tidak ada yang

ISBN: 978-602-60280-1-3

sama, penulis memilih berita dari Tempo.co dalam rentang tahun terbit yang berbeda, sebagaimana dijelaskan berikut :

1. 50 data training kategori politik dalam 5 tema terbitan tahun 2018.
2. 50 data training kategori ekonomi dalam 5 tema terbitan tahun 2018.
3. 50 data training kategori politik dalam 5 tema terbitan tahun 2013-2016.
4. 50 data training kategori ekonomi dalam 5 tema terbitan tahun 2013-2016.
5. 10 teks uji kategori politik dalam 5 tema terbitan tahun 2017.
6. 10 teks uji kategori ekonomi dalam 5 tema terbitan tahun 2017.
7. 10 teks uji kategori politik acak terbitan tahun 2017.
8. 10 teks uji kategori ekonomi acak terbitan tahun 2017.

B. Hasil Uji Coba

TABEL 3. PERBANDINGAN HASIL UJI POLITIK 100 TRAINING DAN 200 TRAINING

NO	Judul	Tema	Menggunakan 100 Training		Menggunakan 200 Training	
			True	Flase	True	False
1	Sebelum Umumkan Cawapres, Begini Strategi Koalisi Jokowi	koalisi jokowi	1			1
2	Oesman: Koalisi Jokowi Belum Bahas Nama Cawapres 2019		1			1
3	Partai Demokrat Sesalkan Pernyataan Hasto tentang E-KTP	korupsi e-ktp		1	1	
4	Wasekjen Demokrat Tanggapi Hasto Kristiyanto Soal Korupsi E-KTP			1	1	
5	Semua Menteri Kabinet Kerja dari PKB Mendaftar sebagai Caleg	pendaftaran caleg	1			1
6	Sudirman Said Diminta Langsung Prabowo Jadi Caleg			1	1	
7	Barikade Gus Dur Deklarasi Dukung Khoifah-Emil di Pilgub Jatim	pilgub jatim		1	1	
8	Pilgub Jatim, Khoifah - Emil Rebut Basis Suara PKB - PDIP			1	1	
9	Wiranto: RUU Terorisme Tidak untuk Kepentingan Politik	ruu terorisme		1	1	
10	Buntut Bom Surabaya, Kapolri Minta DPR Sahkan RUU Terorisme		1			1
Jumlah Hasil Tepat			4	6	6	4
Prosentase (Jumlah hasil tepat / 10 X 100)			40 %	60 %	60 %	40 %

Tabel 3 memperlihatkan hasil uji akurasi pada pengategorian otomatis pada kategori politik secara tematik ketika menggunakan data training 100 teks atau 50 teks per kategori, dan menggunakan 200 teks training atau 100 teks tarining per kategori. Untuk kategori politik, pengujian menggunakan 200 teks training mendapatkan nilai akurasi yang lebih tinggi, yakni 60 persen, dibandingkan dengan 100 teks training yang menghasilkan nilai akurasi 40 persen.

Tabel 4 memperlihatkan hasil uji akurasi pada pengategorian otomatis pada kategori ekonomi secara tematik ketika menggunakan data training 100 teks atau 50 teks per kategori, dan menggunakan 200 teks training atau 100 teks tarining per kategori. Untuk kategori ekonomi, pengujian menggunakan 200 teks training mendapatkan nilai akurasi sama tinggi dengan ketika menggunakan 100 teks training yang menghasilkan nilai akurasi 90 persen.

TABEL 4. PERBANDINGAN HASIL UJI EKONOMI 100 TRAINING DAN 200 TRAINING

Data di atas merupakan hasil uji yang telah dilakukan dengan 100 data training kategori politik dan 100 data training ekonomi. Memperlihatkan bahwa proses pada kategori politik dengan cara tematik memiliki akurasi 60 persen, lebih baik daripada dengan cara acak yang sebesar 50 persen.

Pengujian secara tematik dilakukan dengan menyesuaikan tema data training dan teks baru yang dimasukkan. Dalam penelitian ini ada 5 tema yang di ambil pada kategori politik, yakni koalisi jokowi, korupsi e-ktp, pendaftaran caleg, pilgub jatim, dan ruu terorisme. Masing-masing tema mendapatkan 20 data training dan 2 teks baru untuk bahan uji. Sedangkan cara uji acak dilakukan dengan mencari teks baru tanpa memperhatikan kesesuaian pada 5 tema politik yang sudah disebutkan di atas.

TABEL 5. HASIL UJI KATEGORI POLITIK DENGAN PENENTUAN TEMA

NO	Politik Tematik 200 Training			
	Judul	Tema	True	False
1	Sebelum Umumkan Cawapres, Begini Strategi Koalisi Jokowi	koalisi jokowi		1
2	Oesman: Koalisi Jokowi Belum Bahas Nama Cawapres 2019			1
3	Partai Demokrat Sesalkan Pernyataan Hasto tentang E-KTP	korupsi e-ktp	1	
4	Wasekjen Demokrat Tanggapi Hasto Kristiyanto Soal Korupsi E-KTP		1	
5	Semua Menteri Kabinet Kerja dari PKB Mendaftar sebagai Caleg	pendaftaran caleg		1
6	Sudirman Said Diminta Langsung Prabowo Jadi Caleg		1	
7	Barikade Gus Dur Deklarasi Dukung Khoifah-Emil di Pilgub Jatim	pilgub jatim	1	
8	Pilgub Jatim, Khoifah - Emil Rebut Basis Suara PKB - PDIP		1	
9	Wiranto: RUU Terorisme Tidak untuk Kepentingan Politik	ruu terorisme	1	
10	Buntut Bom Surabaya, Kapolri Minta DPR Sahkan RUU Terorisme			1
Jumlah Hasil Tepat			6	4
Prosentase (Jumlah hasil tepat / 10 X 100)			60 %	40 %

TABEL 6. HASIL UJI KATEGORI POLITIK TANPA PENENTUAN TEMA

NO	Politik Acak 200 Training			
	Judul	Tema	True	False
1	PDIP Bisa Cetak Sejarah Politik Indonesia			1
2	Megawati Dilaporkan, Sekjen PDIP: Ada Kepentingan Politik			1
3	Pengamat Politik: Isu Umat Bakal Menyerang Jokowi di Pilpres 2019		1	
4	Cerita Busyro Soal Pramnih Cukong Politik di Tiap Pilkada		1	
5	Partai Politik Ingin Ketemu Jokowi, Wiranto: Ketemu Menteri Dulu			1
6	Jokowi: Generasi Y Bakal Pengaruhi Politik dan Ekonomi 5-10 Tahun			1
7	SMRC: Isu Kebangkitan PKI Berkaitan dengan Politik		1	
8	Jokowi: Kalau Nyari Dukungan Politik, Saya Bangun Jawa Saja			1
9	Bawaslu Sebut Parpol Keluhkan Sistem Informasi Pencalonan KPU		1	
10	Fahri Hamzah Sarankan Jokowi Bersiap Hadapi Tahun Ribut Politik		1	
Jumlah Hasil Tepat			5	5
Prosentase (Jumlah hasil tepat / 10 X 100)			50 %	50 %

Tabel 7 merupakan perbandingan antara uji akurasi tematik ekonomi dengan uji akurasi secara acak. Hasilnya lagi-lagi uji akurasi secara tematik memberikan hasil lebih baik, yakni 90 persen, daripada dengan cara acak yang menghasilkan nilai akurasi 60 persen.

ISBN: 978-602-60280-1-3

TABEL 7. PERBANDINGAN HASIL UJI POLITIK 100 TRAINING DAN 200 TRAINING

NO	Ekonomi Tematik 200 Training			
	Judul	Tema	True	False
1	Genjot Bisnis Properti, Ini Strategi Pembangunan Jaya Ancol	bisnis properti	1	
2	Sukses Bisnis Properti dengan Modal Rp 20 Juta			1
3	Satgas Pangan Cek Penyebab Lonjakan Harga Barang Kebutuhan Pokok	harga kebutuhan pokok	1	
4	Menjelang Ramadan, Harga Kebutuhan Pokok Naik		1	
5	Inflasi Rendah, BI Tak Kendurkan Kebijakan Moneter	kebijakan bi	1	
6	BI: Kebijakan Moneter Perlu Instrumen Kebijakan Fiskal		1	
7	Kurs Rupiah Melemah Jadi 13.778 Akibat Pidato Pejabat The Fed	kurs rupiah	1	
8	Kurs Rupiah Rp 14.000 per Dolar AS, Ini Rencana Sri Mulyani		1	
9	Suku Bunga Bank Indonesia Naik, Berapa Kenaikan Bunga Deposito?	suku bunga	1	
10	BI Naikkan Suku Bunga Acuan, Ini Dampak Positif Negatifnya		1	
Jumlah Hasil Tepat			9	1
Prosentase (Jumlah hasil tepat / 10 X 100)			90 %	10 %

V. KESIMPULAN

Berdasarkan hasil uji coba menyebutkan bahwa algoritma bekerja lebih baik dengan cara tematik daripada dengan cara acak. Lebih banyak data training per kategori dan per tema, menghasilkan keputusan yang lebih mendekati kebenaran. Proses pengambilan keputusan pada kategori politik secara tematik menghasilkan akurasi 40% dengan 100 data training dan 60% dengan 200 data training. Proses pengambilan keputusan pada kategori ekonomi secara tematik menghasilkan akurasi 90%, dengan 100 data training, maupuan 200 data training.

DAFTAR PUSTAKA

[1] Verisign.com, Verisign Q4 2017, "Domain Name Industry Brief : Internet Grows to 332,4 Million Domain Name Registrations the Fourth Quarter of 2017", 2017, Internet.

[2] Merdeka.com, "Dewan Pers : Media Online Ada 43.000, tapi cuma 0,04 Persen yang Profesional", 2018, Internet.

[3] Okezone.com, "Pencapaian Luar Biasa Okezone, Produksi 1.000 Berita per Hari", 2017, Internet.

[4] Irawan, Hendry, "Kategorisasi Dokumen Web Secara Otomatis Berdasarkan Folksonomy Menggunakan Multinomial Naive Bayes Classifier", Skripsi, 2005, Sekolah Tinggi Teknologi (STT) Telkom, Bandung.

[5] Hanopo, Fran Sandy Herlin, "Sistem Klasifikasi Surat Masuk Menggunakan Multinomial Naive Bayes", Skripsi, 2013, Universitas Sanata Dharma, Yogyakarta.

[6] Octaviana, Elsa Rika, "Analisis Sentimen Pada Lirik Lagu Menggunakan Metode Multinomial Naive Bayes", Skripsi, 2015, Universitas Sanata Dharma, Yogyakarta.

[7] Kusrini dan Luthfi, Emha Taufiq , "Algoritma Data Mining", 2009, Andi, Yogyakarta.

[8] Octaviana, Elsa Rika, "Analisis Sentimen Pada Lirik Lagu Menggunakan Metode Multinomial Naive Bayes", 2015, Skripsi, Universitas Sanata Dharma. Yogyakarta.

[9] Manning, Raghavan, Schutze, "Introduction to Information Retrieval", 2008, Cambridge University Press, New York, NY, USA

[10] Pramudita, Hafiz Ridha, "Penerapan Algoritma Nazief & Adriani dan Similarity pada Penerimaan Judul Thesis", 2001, (online) Dasi, 15(4)

[11] Abdurrasyid, Shandy Ardiansyah, dan Suyanto, "Implementasi dan Optimalisasi Algoritma Nazief dan Adriani untuk Dokumen Bahasa Indonesia", 2012, Fakultas Teknik Informatika, Universitas Telkom.

[12] Agusta, Ledy, "Perbandingan Algoritma Stemming Porter dengan Algoritma Nazief & Adriani untuk Stemming Teks Dokumen Bahasa Indonesia", 2009, (online) Konferensi Nasional Sistem Informasi, Bali.

[13] Keke, Dyan, dan Chikita, Rian, dan Prayogo, Agus Dwi, "Sistem Temu Balik Informasi Algoritma Nazief dan Adriani", 2012, Fakultas Ilmu Komputer, Universitas Gadjah Mada.

[14] Herdianto, R. Anggit, "Analisis dan Pengembangan Sistem Informasi Penerimaan Siswa Baru Berbasis Codeigniter PHP Framework di SMK Ma'arif 1 Mungkid", 2014, Skripsi, Universitas Negeri Yogyakarta. Yogyakarta.

[15] Kurniawan, Imam, dan Sutanta, Edhy, dan Tritono, Joko, "Raspberry Pi High Availability Cluster Server dengan Metode Synchronous Replication, Redundancy, dan Failover", 2017, (online) Jarkom, 6(1).

[16] Tala, Fadillah Z., "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia", 2003, Unpublised Mater's Thesis. Institute for Logic, Language and Computation, University of Amsterdam.