Terbit *online* pada laman : http://teknosi.fti.unand.ac.id/

Jurnal Nasional Teknologi dan Sistem Informasi

| ISSN (Print) 2460-3465 | ISSN (Online) 2476-8812 |

Artikel Penelitian

Klasifikasi Dokumen Berita Menggunakan Algoritma *Enhanced Confix Stripping Stemmer* dan *Naïve Bayes Classifier*

Erwin Yudi Hidayat[a,*], *Muhammad Aditya Rizqi*[a]

[a]*Fakultas Ilmu Komputer, Universitas Dian Nuswantoro, Semarang, 50131, Indonesia*

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima Redaksi: 15 November 2020
Revisi Akhir: 14 Juli 2020
Diterbitkan *Online*: 31 Agustus 2020

KATA KUNCI

Klasifikasi
Enhanced Confix Stripping Stemmer
Naïve Bayes Classifier

KORESPONDENSI

E-mail: erwin@dsn.dinus.ac.id*

A B S T R A C T

Berita adalah salah satu sarana informasi bagi masyarakat umum, dengan media *online* sebagai salah satu sarana untuk mengaksesnya. Di Indonesia, media *online* memiliki presentase paling besar dalam penyebarluasan berita. Dibutuhkan mekanisme yang dapat mengklasifikasikan setiap topik berita secara akurat. Klasifikasi adalah proses yang krusial, karena memerlukan tahap *preprocessing* untuk mengubah data tidak terstruktur menjadi informasi yang bermakna. *Preprocessing* diawali dengan *case folding*, *tokenizing*, *stemming*, dan *filtering*, diahiri dengan klasifikasi. Penelitian ini menggunakan *Enhanced Confix Stripping Stemmer* untuk memisahkan kata dasar dari partikel awalan dan imbuhan, yang sebelumnya sulit dilakukan. Algoritma *Naïve Bayes Classifier* kemudian diterapkan untuk proses klasifikasi. Dataset dari portal www.jawapos.com yang digunakan berjumlah 600 dokumen berita. Data tersebut digunakan sebagai data *training*, terbagi merata ke dalam kategori Olahraga, Teknologi, Ekonomi, dan Lain-lain. Dari 40 data *testing* yang diuji, akurasi tiap kategori diperoleh 90%, 90%, 100%, dan 100%, yang menghasilkan rata-rata akurasi keseluruhan kategori sebesar 95%.

1. PENDAHULUAN

Teknologi Informasi saat ini telah berkembang pesat dari waktu ke waktu seiring berjalannya zaman. Di era globalisasi ini manusia memang tidak terlepas dalam membutuhkan segala informasi dan komunikasi yang berasal dari berbagai media untuk dapat mengembangkan kemampuan di bidang teknologi, yang mana saat ini sudah semakin cepat perkembangannya. Tak khayal teknologi informasi juga saat ini telah memberikan banyak sekali manfaat bagi masyarakat umum.

Teknologi informasi sangat penting bagi sektor bisnis, sebagai alat manajerial untuk mengoptimalkan pengolahan informasi, menghasilkan barang dan jasa demi laba perusahaan [1]. Selain berperan dalam memangkas waktu produksi barang, teknologi informasi dan komunikasi juga ternyata dapat memberikan layanan yang cepat dan efektif [2]. Bahkan, terbukti mampu meningkatkan kualitas layanan, seperti yang terdapat pada restoran-restoran cepat saji [3]. Tak hanya perusahaan dan industri besar, Usaha Mikro Kecil Menengah (UMKM) pun sebagai pelaku bisnis ikut, terdampak oleh teknologi informasi ini. Bagi UMKM, teknologi informasi termasuk berperan dominan dalam pengembangan inovasi [4].

Secara umum teknologi informasi dapat dimanfaatkan sebagai media untuk mendapatkan serta menyebarluaskan informasi. Ini berbanding lurus dengan banyaknya jenis media massa yang dapat dimanfaatkan untuk mendapatkan maupun menyebarluaskan sebuah informasi. Contohnya media cetak seperti buku, koran dan majalah. Media elektronik seperti televisi, telepon genggam dan juga komputer, serta media *online* seperti *website* dan media sosial. Media *online* merupakan media pers yang berbasis komputer dan internet, seperti contoh *weblog*, portal berita, dan radio *online*.

Karena kemudahan dan kenyamanan yang ditawarkan, media sosial menjadi salah satu pilihan utama dalam penyebaran berita [5] dan publikasi. Misalnya saja, dalam satu menit ada sebanyak 3,600 foto baru yang diunggah ke Instagram, 48 jam video diunggah ke YouTube, 684,478 konten baru di Facebook, dan

https://doi.org/10.25077/ TEKNOSI.v6i2.2020.90-99

sebanyak 2 juta pencarian dilakukan di mesin pencari Google [6]. Sebagian besar konsumsi berita video terjadi di Facebook (32%) dan YouTube (26%), di mana penyedia berita tersebut dapat menambah keuntungan melalui iklan yang ditayangkan di dalamnya [7].

Berita telah menjadi agen komunikasi sejak lama. Berita mampu menyatukan banyak orang, sekalipun berada pada wilayah dan zona waktu yang berbeda. Dalam dunia penyebaran berita, tahun 1605 adalah sebuah titik bersejarah dalam bidang komunikasi berita, di mana pada tahun ini, surat kabar atau koran cetak diperkenalkan kepada dunia [8]. Sejak saat itu, surat kabar mengalami perubahan dan perkembangan yang berkelanjutan.

Sejak pertama kali internet berkembang, keberadaan bisnis berita cetak kerap menjadi sorotan, terkait kemampuannya dalam mendatangkan keuntungan perusahaan. Teknologi yang kian maju menjadi magnet tersendiri bagi industri berita. Hal ini banyak menyebabkan beberapa bisnis surat kabar berbasis media cetak, melakukan transformasi ke dalam ranah sumber berita berbasis *online* [9]. Sebagai contoh adalah Le Monde, sebuah surat kabar di Perancis yang mulai melakukan digitalisasi basis data dokumen pertama kali pada tahun 1987. Dua tahun kemudian, orang-orang di kantor berita ini mulai berkirim pesan melalui media elektronik. Saat ini, Le Monde tidak hanya menyebarkan berita menggunakan media cetak saja, tapi merambah ke media online [10].

Dalam 20 tahun terakhir, hubungan antara perusahaan penyedia berita dan teknologi internet berada pada titik yang paling kritis. Sirkulasi surat kabar cetak tradisional tercatat semakin menurun sejak awal 1990-an [11]. Perkembangan yang terjadi dalam teknologi media, khususnya penemuan internet, memberikan kebebasan berkomunikasi melalui ruang Waring Wera Wanua (WWW) [12].

Di Indonesia sendiri media *online* menjadi salah satu media yang paling banyak digunakan masyarakat untuk mendapatkan informasi berita terkini yang dibutuhkan karena dapat memberikan segala informasi dari belahan bumi lain dengan kecepatannya yang tinggi. Ditinjau dari segi waktu, media *online* mampu menyajikan informasi lebih cepat dengan harga relatif sangat murah [13]. Sebuah penelitian menunjukkan, Indonesia berada pada urutan ke-15 dunia dengan penetrasi internet sebesar 8% (1,6% dari total pengguna internet dunia, dengan 18.000.000 pengguna dari populasi 224.481.720 jiwa) [14]. Selain itu, Indonesia juga menempati posisi di peringkat nomor 6 terbesar, di antara sekitar 3,6 miliar jumlah pengakses internet dunia [15].

Angka di atas berimbas pada pesatnya perkembangan media baru panyampaian berita seperti portal berita *online*. Saat ini banyak media cetak yang sukses melebarkan sayapnya dengan beralih ke media *online* seperti kompas.com yang juga sukses seperti versi cetaknya. Dan banyak media online lainnya seperti Tempointeraktif.com, Vivanews.com, Metrotv.com, Liputan6.com, Okezone.com dan Detik.com [16].

Pada umumnya di dalam sebuah portal berita terdapat berbagai pilhan kategori berita untuk memudahkan masyarakat dalam mengkases topik berita sesuai apa yang diinginkan, seperti berita politik, olahraga, ekonomi, hiburan sampai berita kesehatan. Dengan banyaknya kategori berita yang ada, serta banyaknya jumlah aliran berita yang harus diunggah ke dalam portal berita, menyebabkan kinerja editor menjadi lebih banyak. Sebab, seorang editor harus mengetahui isi berita secara keseluruhan terlebih dahulu kemudian dikelompokkan secara manual ke dalam jenis kategori yang sesuai [17].

Salah satu solusi agar masalah tersebut dapat ditanggulangi adalah dengan mengubah proses klasifikasi yang sebelumnya dilakukan secara konvensional menjadi proses yang dilakukan secara otomatis dengan komputer dan menggunakan metode tertentu. Konsep ini dinamakan klasifikasi, yaitu memasukkan teks baru yang kategorinya belum diketahui dengan melakukan pelatihan terhadap kumpulan teks yang kategorinya telah diketahui [18]. Klasifikasi teks termasuk proses yang penting dalam klasifikasi dokumen dan *text mining*, yang biasanya dilakukan oleh pakar [19]. Sebelum berita diklasifikasikan, terlebih diolah dahulu karena data dari teks yang ada di dalam berita merupakan data yang tidak terstruktur.

Text mining merupakan salah satu variasi yang terdapat dalam *data mining* yang bertujuan untuk menemukan pola yang menarik menggunakan tahapan dan analisis tertentu di dalam sekumpulan teks dokumen [20]. *Text preprocessing* merupakan tahapan yang ada dalam *text mining* yang bertujuan untuk mengubah teks menjadi data yang akan diolah pada tahap berikutnya. Di dalam *text preprocessing* terdapat beberapa tahapan, yaitu *case folding* yang bertujuan untuk mengubah semua huruf yang ada di dalam dokumen berita menjadi huruf kecil. Selanjutnya *tokenizing* yaitu proses memecah kalimat menjadi sebuah kata. Setelah itu ada proses *filtering* yang merupakan tahap mengambil kata kata penting yang dihasilkan dari proses *tokenizing* dan *stemming* yang merupakan proses penting di dalam *text mining* di mana pada tahap *stemming* terjadi proses penghilangan kata imbuhan sehingga menjadi kata dasar [21].

Pada penelitian sebelumnya telah dilakukan analisa dan percobaan terhadap algoritma *Confix Stripping Stemmer* di mana masih terdapat beberapa kata yang tidak dapat di-*stemming*. Dari analisa tersebut muncul algoritma baru untuk mengatasi kekurangan yang ada, yaitu algoritma *Enhanced Confix Stripping Stemmer.* Algoritma ini mampu mengurangi jumlah *term* hingga sebesar 32,67%, dari nilai 30,94% yang dilakukan menggunakan *Confix Stripping Stemmer* [22]. Maka dari itu penelitian ini menggunakan algoritma perbaikan dari *Confix Stripping Stemmer* pada tahapan *stemming* yaitu algoritma *Enhanced Confix Stripping Stemmer.*

Klasifikasi konten berita secara otomatis dapat dilakukan menggunakan metode *Naïve Baiyes Classifier* [23] yang terbukti bekerja dengan baik dibandingkan metode lainnya [24], serta memiliki kelebihan dalam hal presisi [25]. Proses klasifikasi ini bertujuan agar mampu mengimplementasikan metode *Naïve Baiyes Classifier* serta mengukur tingkat akurasinya dalam mengklasifikasikan kategori berita secara.

2. METODE

2.1. Data Penelitian

Penelitian ini menggunakan sumber data berupa data dokumen berita yang diambil langsung dari kantor Jawa Pos Radar Semarang, yang beralamat Jl. Veteran No.55, Lempongsari, Gajahmungkur, Kota Semarang Jawa Tengah.

Data berita yang digunakan terdiri dari empat kategori berita, yaitu: Olahraga, Teknologi, Ekonomi, dan Lain-lain. Jumlah data *training* yang digunakan sebanyak 600 data, dengan 150 data *training* untuk setiap kategori. Sedangkan untuk data *testing* sebanyak 40 data yang diambil secara acak pada portal berita www.jawapos.com.

Adapun untuk keperluan eksperimen, percobaan klasifikasi dilakukan dengan menggunakan judul berita sebanyak 5 buah. Dokumen tersebut dimanfaatkan sebagai dokumen latih (*training*). Satu artikel yang berbeda lainnya digunakan sebagai data uji (*testing*).

Tabel 1. Data latih dan data uji untuk percobaan klasifikasi

Dokumen Judul	Judul Berita	Kategori
Doc Latih 1	Drama 6 Gol: Barcelona Petik Kemenangan, Messi Bikin Rekor Lagi	Olahraga
Doc Latih 2	YLKI Sebut Tarif MRT Rp 8.500 Cukup Fair dan Diterima Masyarakat	Ekonomi
Doc Latih 3	Zinedine Zidane Resmi Melatih Real Madrid Lagi	Olahraga
Doc Latih 4	Anggaran Naik, Kuota Mudik Gratis Bertambah Jadi 54 Ribu Orang	Ekonomi
Doc Latih 5	Bocoran Jersey Baru Barcelona Dikritik Netizen	Olahraga
Doc Uji 1	Prediksi Real Madrid vs Barcelona: Tuan Rumah Sedang Ditertawakan	Belum Diketahui

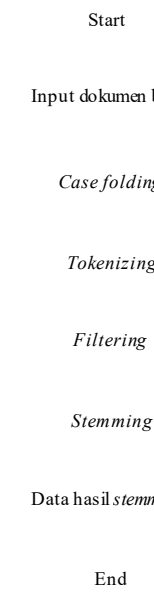
2.2. Preprocessing

Preprocessing merupakan tahap awal penelitian yang bertujuan untuk menyamakan bentuk kata, mengurangi frekuensi kosa kata yang ada dalam dokumen agar menjadi data yang dapat diolah pada proses selanjutnya. Proses ini memiliki beberapa tahapan yaitu *case folding*, *tokenizing*, *filtering*, dan *stemming.* Pada tahap ini akan dilakukan proses pelatihan kemudian dilanjutkan dengan proses pengujian klasifikasi. Diagram alir dari preprocessing klasifikasi disajikan pada Gambar 1.

2.2.1. Case Folding

Case Folding merupakan tahap merubah semua huruf kapital yang ada dalam dokumen menjadi huruf kecil. Proses *case folding* dapat dilihat pada Gambar 2, dan hasil dari proses ini dapat dilihat pada Tabel 2.

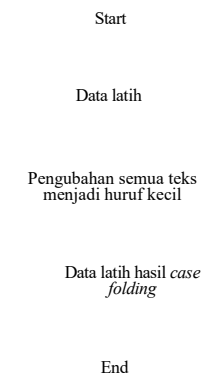
Kolom input merupakan data dokumen yang akan digunakan sebagai data *training.* Sedangkan kolom output berisi data *training* yang sudah melewati tahap c*ase folding* dengan perubahan semua huruf kapital menjadi huruf kecil.



Gambar 1. Diagram alir *preprocessing*

Tabel 2. Contoh dokumen hasil proses *case folding*

Dokumen Judul	Input	Output
Doc Latih 1	Drama 6 Gol: Barcelona Petik Kemenangan, Messi Bikin Rekor Lagi	drama 6 gol: barcelona petik kemenangan, messi bikin rekor lagi
Doc Latih 2	YLKI Sebut Tarif MRT Rp 8.500 Cukup Fair dan Diterima Masyarakat	ylki sebut tarif mrt rp 8.500 cukup fair dan diterima masyarakat
Doc Latih 3	Zinedine Zidane Resmi Melatih Real Madrid Lagi	zinedine zidane resmi melatih real madrid lagi
Doc latih 4	Anggaran Naik, Kuota Mudik Gratis Bertambah Jadi 54 Ribu Orang	anggaran naik, kuota mudik gratis bertambah jadi 54 ribu orang
Doc Latih 5	Bocoran Jersey Baru Barcelona Dikritik Netizen	bocoran jersey baru barcelona dikritik netizen
Doc Uji	Prediksi Real Madrid vs Barcelona: Tuan Rumah Sedang Ditertawakan	prediksi real madrid vs barcelona: tuan rumah sedang ditertawakan

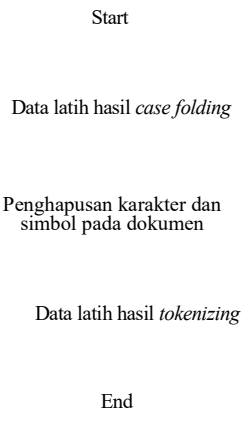


Gambar 2. Diagram alir proses *case folding*

 https://doi.org/10.25077/ TEKNOSI.v6i2.2020.90-99

2.2.2. *Tokenizing*

Pada tahap tokenizing, dilakukan pemotongan kata dalam kalimat, paragraf, maupun halaman menjadi potongan kata dasar atau kata tunggal dan menghilangkan *delimiter* seperti tanda titik (.), koma (,), dan karakter angka yang ada pada dokumen tersebut [26]. Contoh lain karakter yang dihapus adalah '@', '&', '!', '?', dan ':', serta tab dan *enter*. Gambar 3 menunjukkan bagaimana *tokenizing* dilakukan, yang hasilnya diperlihatkan pada Tabel 3.



Gambar 3. Diagram alir proses *tokenizing*

Tabel 3. Contoh dokumen hasil proses *tokenizing*

Dokumen Judul	Input	Output
Doc Latih 1	drama 6 gol: barcelona petik kemenangan, messi bikin rekor lagi	drama gol barcelona petik kemenangan messi bikin rekor lagi
Doc Latih 2	ylki sebut tarif mrt rp 8.500 cukup fair dan diterima masyarakat	ylki sebut tarif mrt rp cukup fair dan diterima masyarakat
Doc Latih 3	zinedine zidane resmi melatih real madrid lagi	zinedine zidane resmi melatih real madrid lagi
Doc Latih 4	anggaran naik, kuota mudik gratis bertambah jadi 54 ribu orang	anggaran naik kuota mudik gratis bertambah jadi ribu orang
Doc Latih 5	bocoran jersey baru barcelona dikritik netizen	bocoran jersey baru barcelona dikritik netizen
Doc Uji	prediksi real madrid vs barcelona: tuan rumah sedang ditertawakan	prediksi real madrid vs barcelona tuan rumah sedang ditertawakan

Kolom input berisi hasil data latih yang telah melewati tahap *case folding.* Pada tahap *tokenizing* terjadi penghilangan tanda baca, karakter angka, serta karakter-karakter selain huruf 'a' sampai 'z' yang hasilnya ditampilkan pada kolom output.

2.2.3. *Filtering*

Proses *filtering* merupakan proses pemotongan untai (*string*) input berdasarkan setiap kata yang menyusunnya serta mengambil kata penting yang didapat dari hasil *tokenizing.* Kata hasil *tokenizing* akan dibandingkan dengan kata yang ada didalam kamus *stopword.* Jika kata tersebut ada dalam *stopword* maka kata tersebut akan dihapus.

Gambar 4. Diagram alir *filtering*

Tabel 4. Contoh dokumen hasil proses *filtering*

Dokumen Judul	Input	Output
Doc Latih 1	drama gol barcelona petik kemenangan messi bikin rekor lagi	'drama', 'gol', 'barcelona', 'petik', 'kemenangan', 'messi', 'bikin', 'rekor'
Doc Latih 2	ylki sebut tarif mrt rp cukup fair dan diterima masyarakat	'ylki', 'tarif', 'mrt', 'rp', 'fair', 'diterima', 'masyarakat'
Doc Latih 3	zinedine zidane resmi melatih real madrid lagi	'zinedine', 'zidane', 'resmi', 'melatih', 'real', 'madrid'
Doc Latih 4	anggaran naik kuota mudik gratis bertambah jadi ribu orang	'anggaran', 'kuota', 'mudik', 'gratis', 'bertambah', 'ribu', 'orang'
Doc Latih 5	bocoran jersey baru barcelona dikritik netizen	'bocoran', 'jersey', 'barcelona', 'dikritik', 'netizen'
Doc Uji	prediksi real madrid vs barcelona tuan rumah sedang ditertawakan	'prediksi', 'real', 'madrid', 'vs', 'barcelona', 'tuan', 'rumah', 'ditertawakan'

Kata hasil dari proses *tokenizing* dilakukan perbandingan dengan daftar *stopword* yang berasal dari [27]. Beberapa contoh *stopword* antara lain "dan", "antara", "bahwa", "berarti", "demikian", "masih", "oleh", dan "tidak". Apabila data hasil dari proses *tokenizing* sama dengan yang ada di daftar *stopword* maka kata dihapus. Jika tidak maka disimpan. Pada kolom output merupakan kata yang telah melalui tahap *filtering* dan kata yang sama dengan daftar *stopword* telah dihapus.

2.2.4. *Stemming dengan Enhanced Confix Stripping Stemmer*

Setelah dilakukan proses *case folding, tokenezing,* dan *filtering,* proses selanjutnya yaitu *stemming. Stemming* yang digunakan

pada penelitian ini menggunakan algoritma *Enhanced Confix Stipping Stemmer,* terdiri dari beberapa langkah:

1) Data *training* hasil dari *filtering* akan dilakukan pengecekan atau pencarian kata-kata yang sesuai dengan kamus umum. Apabila data *training* hasil *filtering* sesuai dengan kamus umum maka kata akan dikeluarkan sementara, karena sudah dianggap sebagai kata dasar.
2) Apabila masih terdapat kata yang tidak termasuk dalam kata dasar maka tahap selanjutnya adalah menghapus *inflection suffixes* yang merupakan akhiran pertama. Kata yang memiliki akhiran *partticles* seperti "-pun", "-kah", "-tah", "-lah" dan akhiran *possessive pronoun* seperti "-mu", "-ku" dan "-nya" dihilangkan.
3) Apabila hasilnya terdapat kata dasar maka kata tersebut dikeluarkan sementara dari algoritma, namun jika belum valid dengan kamus langkah selanjutnya menghapus awalan *derivaion suffixes* yaitu "-kan", "-i" dan "-an".
4) Apabila hasilnya terdapat kata dasar maka kata tersebut dikeluarkan sementara dari algoritma, namun jika belum valid dengan kamus langkah selanjutnya menghapus akhiran *derivation prefixes* yaitu "be-", "pe-", "te-", "-me", "ke-", "di-" dan "se-".
5) Lakukan pengecekan terhadap kata dengan kamus apabila valid maka proses dihentikan. Apabila semua tahapan sudah dilakukan namun masih terdapat kata yang tidak ditemukan setelah pengecekan maka kata tersebut termasuk kata tidak terdeteksi dan dianggap sebagai kata dasar.

Setelah semua langkah dilakukan maka kata dasar yang dikeluarkan sementara pada langkah 1, 2, 3 dan 4 serta kata yang tidak terdeteksi menjadi hasil *stemming* dari algoritma *Enhanced Confix Stripping Stemmer* dan menghasilkan data *training* seperti Tabel 5 dibawah ini.

Tabel 5. Hasil implementasi *Enhanced Confix Stripping Stemmer*

Dokumen Judul	Input	Output
Doc Latih 1	'drama', 'gol', 'barcelona', 'petik', 'kemenangan', 'messi', 'bikin', 'rekor'	'drama', 'gol', 'barcelona', 'petik', 'menang', 'messi', 'bikin', 'rekor'
Doc Latih 2	'ylki', 'tarif', 'mrt', 'rp', 'fair', 'diterima', 'masyarakat'	'ylki', 'tarif', 'mrt', 'rp', 'fair', 'terima', 'masyarakat'
Doc Latih 3	'zinedine', 'zidane', 'resmi', 'melatih', 'real', 'madrid'	'zinedine', 'zidane', 'resmi', 'latih', 'real', 'madrid'
Doc Latih 4	'anggaran', 'kuota', 'mudik', 'gratis', 'bertambah', 'ribu', 'orang'	'anggar', 'kuota', 'mudik', 'gratis', 'tambah', 'ribu', 'orang'
Doc Latih 5	'bocoran', 'jersey', 'barcelona', 'dikritik', 'netizen'	'bocor', 'jersey', 'barcelona', 'kritik', 'netizen'
Doc Uji	'prediksi', 'real', 'madrid', 'vs', 'barcelona', 'tuan', 'rumah', 'ditertawakan'	'prediksi', 'real', 'madrid', 'vs', 'barcelona', 'tuan', 'rumah', 'tawa'

2.3. Klasifikasi dengan Naïve Bayes Classifier

Klasifikasi terhadap sampel dokumen ini menggunakan *Naive Baiyes Classifier.* Berdasarkan teorema Bayes, perhitungan nilai probabilitas dapat dihitung menggunakan persamaan berikut ini:

$$P(H|E) = \frac{P(E|H).P(H)}{P(E)} \quad (1)$$

Di bawah ini merupakan persamaan dari rumus P(H|E) yang sudah mengalami penambahan proses *Laplacian smoothing* yang digunakan untuk proses klasifikasi dokumen:

$$P(W_k \mid C_i) = \frac{n_k+1}{n+|vocabulary|} \quad (2)$$

Tabel 6 memuat sampel judul dokumen yang telah melewati tahap *preprocessing,* yaitu Doc Latih 1 kategori olahraga, Doc Latih 2 kategori ekonomi, Doc Latih 3 kategori olahraga, Doc Latih 4 kategori ekonomi, dan Doc Latih 5 kategori olahraga. Kelima sampel tersebut akan dijadikan sebagai data latih dalam tahap klasifikasi. Sedangkan Doc Uji dijadikan sebagai data uji. Klasifikasi menggunakan *Naive Baiyes Classifier* ini terbagi menjadi 2 tahap yaitu tahap *training* dan tahap *testing.*

Tabel 6. Hasil sampel dokumen setelah *preprocessing*

Dokumen Judul	Kata Dasar	Kategori
Doc Latih 1	drama gol barcelona petik menang messi bikin rekor	Olahraga
Doc Latih 2	ylki tarif mrt rp fair terima masyarakat	Ekonomi
Doc Latih 3	zinedine zidane resmi latih real madrid	Olahraga
Doc Latih 4	anggar kuota mudik gratis tambah ribu orang	Ekonomi
Doc Latih 5	bocor jersey barcelona kritik netizen	Olahraga
Doc Uji	prediksi real madrid vs barcelona tuan rumah tawa	Belum Diketahui

2.3.1. Tahap latih (training)

Tahap *training* menggunakan 5 sampel dokumen latih seperti pada Tabel 6. Persamaan yang digunakan sebagai berikut:

$$P(C_i) = \frac{f_d(C_i)}{|D|} \quad (3)$$

Tabel 7. Kosa kata berbeda dari 5 dokumen *training*

Kosa kata hasil *training*			
drama	ylki	zidane	gratis
gol	tarif	resmi	tambah
barcelona	mrt	latih	ribu
petik	rp	real	orang
menang	fair	madrid	bocor
messi	terima	anggar	jersey
bikin	masyarakat	kuota	kritik
rekor	zinedine	mudik	netizen

Terdapat 32 kosa kata berbeda yang terdapat pada kelima dokumen *training*. Nantinya setiap kosa kata tersebut dihitung probabilitasnya terhadap dokumen yang sudah diketahui kategorinya. Tabel 7 merupakan 32 kosa kata berbeda yang dihitung probabilitasnya, berdasarkan kategori olahraga dan ekonomi menggunakan persamaan (2) dan (3).

 https://doi.org/10.25077/ TEKNOSI.v6i2.2020.90-99

$P(\text{Olahraga}) = \frac{3}{|5|} = 0{,}6$

$P(\text{drama} \mid \text{Olahraga}) = \frac{1+1}{19+|32|} = 0{,}03921$

$P(\text{gol} \mid \text{Olahraga}) = \frac{1+1}{19+|32|} = 0{,}03921$

$P(\text{barcelona} \mid \text{Olahraga}) = \frac{2+1}{19+|32|} = 0{,}0588$

$P(\text{petik} \mid \text{Olahraga}) = \frac{1+1}{19+|32|} = 0{,}03921$

$P(\text{menang} \mid \text{Olahraga}) = \frac{1+1}{19+|32|} = 0{,}03921$

$P(\text{messi} \mid \text{Olahraga} = \frac{1+1}{19+|32|} = 0{,}03921$

$P(\text{bikin} \mid \text{Olahraga}) = \frac{1+1}{19+|32|} = 0{,}03921$

$P(\text{rekor} \mid \text{Olahraga}) = \frac{1+1}{19+|32|} = 0{,}03921$

$P(\text{zinedine} \mid \text{Olahraga}) = \frac{1+1}{19+|32|} = 0{,}03921$

$P(\text{zidane} \mid \text{Olahraga}) = \frac{1+1}{19+|32|} = 0{,}03921$

$P(\text{resmi} \mid \text{Olahraga}) = \frac{1+1}{19+|32|} = 0{,}03921$

$P(\text{latih} \mid \text{Olahraga}) = \frac{1+1}{19+|32|} = 0{,}03921$

$P(\text{real} \mid \text{Olahraga}) = \frac{1+1}{19+|32|} = 0{,}03921$

$P(\text{madrid} \mid \text{Olahraga}) = \frac{1+1}{19+|32|} = 0{,}03921$

$P(\text{bocor} \mid \text{Olahraga}) = \frac{1+1}{19+|32|} = 0{,}03921$

$P(\text{jersey} \mid \text{Olahraga}) = \frac{1+1}{19+|32|} = 0{,}03921$

$P(\text{kritik} \mid \text{Olahraga}) = \frac{1+1}{19+|32|} = 0{,}03921$

$P(\text{netizen} \mid \text{Olahraga}) = \frac{1+1}{19+|32|} = 0{,}03921$

Dengan mengacu pada persamaan yang sama dengan kategori olahraga, perhitungan dokumen berdasarkan kategori ekonomi menghasilkan angka:

$P(\text{Ekonomi}) = \frac{2}{|5|} = 0{,}4$

$P(\text{ylki} \mid \text{Ekonomi}) = \frac{1+1}{14+|32|} = 0{,}04347$

$P(\text{tarif} \mid \text{Ekonomi}) = \frac{1+1}{14+|32|} = 0{,}04347$

$P(\text{mrt} \mid \text{Ekonomi}) = \frac{1+1}{14+|32|} = 0{,}04347$

$P(\text{rp} \mid \text{Ekonomi}) = \frac{1+1}{14+|32|} = 0{,}04347$

$P(\text{fair} \mid \text{Ekonomi}) = \frac{1+1}{14+|32|} = 0{,}04347$

$P(\text{terima} \mid \text{Ekonomi}) = \frac{1+1}{14+|32|} = 0{,}04347$

$P(\text{masyarakat} \mid \text{Ekonomi}) = \frac{1+1}{14+|32|} = 0{,}04347$

$P(\text{anggar} \mid \text{Ekonomi}) = \frac{1+1}{14+|32|} = 0{,}04347$

$P(\text{kuota} \mid \text{Ekonomi}) = \frac{1+1}{14+|32|} = 0{,}04347$

$P(\text{mudik} \mid \text{Ekonomi}) = \frac{1+1}{14+|32|} = 0{,}04347$

$P(\text{gratis} \mid \text{Ekonomi}) = \frac{1+1}{14+|32|} = 0{,}04347$

$P(\text{tambah} \mid \text{Ekonomi}) = \frac{1+1}{14+|32|} = 0{,}04347$

$P(\text{ribu} \mid \text{Ekonomi}) = \frac{1+1}{14+|32|} = 0{,}04347$

$P(\text{orang} \mid \text{Ekonomi}) = \frac{1+1}{14+|32|} = 0{,}04347$

2.3.2. Tahap Uji (testing)

Pada tahap *testing* ini menggunakan 1 dokumen uji dan setiap kata yang ada pada dokumen uji dihitung probabilitasnya berdasarkan kategori yang sudah diketahui yaitu kategori olahraga dan kategori ekonomi. Dokumen *testing* yang akan digunakan seperti pada Tabel 8.

Tabel 8. Dokumen *testing*

Dokumen Judul	Input	Output
Doc Uji	prediksi real madrid vs barcelona tuan rumah tawa	Belum Diketahui

Selanjutnya dihitung probabilitias dokumen uji berdasarkan kategori olahraga dengan mengacu pada persamaan 2 sebagai berikut:

$P(\text{prediksi} \mid \text{Olahraga}) = \frac{0+1}{19+|32|} = 0{,}0196$

$P(\text{real} \mid \text{Olahraga}) = \frac{1+1}{19+|32|} = 0{,}0392$

$P(\text{madrid} \mid \text{Olahraga}) = \frac{1+1}{19+|32|} = 0{,}0392$

$P(\text{vs} \mid \text{Olahraga}) = \frac{0+1}{19+|32|} = 0{,}0196$

$P(\text{barcelona} \mid \text{Olahraga}) = \frac{2+1}{19+|32|} = 0{,}0588$

$P(\text{tuan} \mid \text{Olahraga}) = \frac{0+1}{19+|32|} = 0{,}0196$

$P(\text{rumah} \mid \text{Olahraga}) = \frac{0+1}{19+|32|} = 0{,}0196$

$P(\text{tawa} \mid \text{Olahraga}) = \frac{0+1}{19+|32|} = 0{,}0196$

Adapun perhitungan yang dilakukan terhadap dokumen uji berdasarkan kategori ekonomi menghasilkan angka yang semua sama, yaitu 0,0219.

Hasil probabilitas dokumen *testing* pada kategori olahraga adalah:

$$\begin{aligned} P(\text{Olahraga}) &= P(\text{Olahraga}) + P(\text{prediksi} \mid \text{Olahraga}) + P(\text{real} \mid \text{Olahraga}) + P(\text{madrid} \mid \text{Olahraga}) + P(\text{vs} \mid \text{Olahraga}) + P(\text{barcelona} \mid \text{Olahraga}) + P(\text{tuan} \mid \text{Olahraga}) + P(\text{rumah} \mid \text{Olahraga}) + P(\text{tawa} \mid \text{Olahraga}) \\ &= 0{,}6 + 0{,}0196 + 0{,}0392 + 0{,}0392 + 0{,}0196 + 0{,}0588 \; 0{,}0196 + 0{,}0196 + 0{,}0196 \\ &= 0{,}8352 \end{aligned}$$

Hasil probabilitas dokumen *testing* pada kategori ekonomi adalah:

$$\begin{aligned} P(\text{Ekonomi}) &= P(\text{Ekonomi}) + P(\text{prediksi} \mid \text{Ekonomi}) + P(\text{real} \mid \text{Ekonomi}) + P(\text{madrid} \mid \text{Ekonomi}) + P(\text{vs} \mid \text{Ekonomi}) + P(\text{barcelona} \mid \text{Ekonomi}) + P(\text{tuan} \mid \text{Ekonomi}) + P(\text{rumah} \mid \text{Ekonomi}) + P(\text{tawa} \mid \text{Ekonomi}) \\ &= 0{,}4 + 0{,}0219 + 0{,}0219 + 0{,}0219 + 0{,}0219 + 0{,}0219 \\ &\quad + 0{,}0219 + 0{,}0219 + 0{,}0219 \\ &= 0{,}5752 \end{aligned}$$

Dari hasil perhitungan probabilitas diatas yaitu P(Olahraga) > P(Ekonomi) maka dapat disimpulkan bahwa dokumen uji tersebut dikategorikan sebagai kategori Olahraga.

Perhitungan akurasi dilakukan dengan membandingkan jumlah klasifikasi yang bernilai benar dan jumlah dokumen uji.

$$\text{Akurasi} = \frac{\text{Jumlah klasifikasi benar}}{\text{Jumlah dokumen uji}} \times 100 \quad (4)$$

Maka tingkat akurasi yang didapatan dari percobaan proses klasifikasi menggunakan *Naive Baiyes* yang telah dilakukan sebelumnnya mendapatkan hasil akurasi sebesar 100%

3. HASIL

3.1. Hasil Preprocessing

Tahapan yang dilakukan sekarang adalah implementasi *Enhanced Confix Stipping Stemmer* dan klasifikasi *Naïve Bayes Classifier* menggunakan data set yang terdiri dari 600 dokumen berita. Terbagi menjadi 4 kategori yaitu Olahraga, Teknologi, Ekonomi, dan Lain-lain. Tiap-tiap kategori terdiri atas 150 dokumen berita sebagai data latih. Sedangkan dokumen berita sebanyak 40, digunakan sebagai data uji yang diambil langsung dari portal www.jawapos.com, dengan subdomain 4 kategori yang sama. Tabel 9 merupakan contoh proses dan juga hasil yang telah didapatkan melalui *preprocessing* dengan menggunakan sampel data *training* yang dipilih secara acak dan siap untuk digunakan untuk proses klasifikasi.

Tabel 9. Dokumen latih dataset

Case folding	manchester united makin terpuruk pada pekan terakhir premier league 2018-2019. dari big six, hanya manchester united yang menelan kekalahan pada laga pamungkas. manchester city dan liverpool meraih kemenangan, chelsea imbang di kandang leicester city, tottenham imbang saat menjamu everton, dan arsenal menang saat tandang ke markas burnley.	D-1
Tokenizing	manchester united makin terpuruk pada pekan terakhir premier league dari big six hanya manchester united yang menelan kekalahan pada laga pamungkas manchester city dan liverpool meraih kemenangan chelsea imbang di kandang leicester city tottenham imbang saat menjamu everton dan arsenal menang saat tandang ke markas burnley	
Filtering	manchester united terpuruk pekan premier league big six manchester united menelan kekalahan laga pamungkas manchester city liverpool meraih kemenangan chelsea imbang kandang leicester city tottenham imbang menjamu everton arsenal menang tandang markas burnley	
Stemming	manchester united puruk pekan premier league big six manchester unite telan kalah laga pamungkas manchester city liverpool raih menang chelsea imbang kandang leicester city tottenham imbang jamu everton arsenal menang tandang markas burnley	
Case folding	baru-baru ini xiaomi dan line friends merilis beberapa produk edisi khusus. produk tersebut adalah handset mi 9. *smartphone* xiaomi mi 9 se brown bear edition tampak mentereng bersematkan logo brown bear. logo disesuaikan di bagian belakang perangkat dan di boks penjualan *flagship* xiaomi itu.	
Tokenizing	baru baru ini xiaomi dan line friends merilis beberapa produk edisi khusus produk tersebut adalah handset mi smartphone xiaomi mi se brown bear edition tampak mentereng bersematkan logo brown bear logo disesuaikan di bagian belakang perangkat dan di boks penjualan flagship xiaomi itu	D-2
Filtering	xiaomi line friends merilis produk edisi khusus produk handset mi smartphone xiaomi mi brown bear edition mentereng bersematkan logo brown bear logo disesuaikan perangkat boks penjualan flagship xiaomi	
Stemming	xiaomi line friends rilis produk edisi khusus produk handset mi smartphone xiaomi mi brown bear edition mentereng semat logo brown bear logo sesuai perangkat boks jual flagship xiaomi	

Data uji yang digunakan untuk pengujian adalah sebagaimana terdapat dalam Tabel 10.

Tabel 10. Dokumen uji dataset

Case folding	manchester united berada dalam posisi yang serba salah. sebab apapun hasil yang diraih di derby manchester, rabu (25/4) dini hari wib, mereka akan membuat dua rivalnya, manchester city dan liverpool jadi mendekati gelar juara premier league musim 2018-2019.	DX -1
Tokenizing	manchester united berada dalam posisi yang serba salah sebab apapun hasil yang diraih di derby manchester rabu dini hari wib mereka akan membuat dua rivalnya manchester city dan liverpool jadi mendekati gelar juara premier league musim	
Filtering	manchester united posisi serba salah apapun hasil diraih derby manchester rabu wib rivalnya manchester city liverpool mendekati gelar juara premier league musim	
Stemming	manchester united posisi serba salah apa hasil raih derby manchester rabu wib rival manchester city liverpool dekat gelar juara premier league musim	
Case folding	xiaomi mengggelar peresmian untuk merilis produk edisi mi smartphone xiaomi mi special edition dalam rangka acara ulang tahun xiaomi yang diadakan di beijing cina, 27 Januari 2019.	DX -2
Tokenizing	xiaomi mengggelar peresmian untuk merilis produk edisi mi smartphone xiaomi mi special edition dalam rangka acara ulang tahun xiaomi yang diadakan di beijing januari	
Filtering	xiaomi gelar peresmian merilis produk edisi mi smartphone xiaomi mi special edition rangka acara ulang tahun xiaomi beijing januari	
Stemming	xiaomi gelar resmi rilis produk edisi mi smartphone xiaomi mi special edition rangka acara ulang tahun xiaomi beijing januari	

3.2. Klasifikasi dengan Naïve Bayes

Setelah semua langkah *preprocessing* dilakukan maka didapatkan data *training* yang siap diproses untuk klasifikasi menggunakan algoritma *Naive Baiyes*.

 https://doi.org/10.25077/ TEKNOSI.v6i2.2020.90-99

3.2.1. Tahap training

Pada tahap *training* ini selanjutnya dataset akan dihitung jumlah kosa kata berbeda beserta jumlah kemunculannya untuk setiap dokumen. Sebagai simulasi perhitungan, *training* yang dilakukan melibatkan 10 data latih. Berikut ditampilkan dua di antaranya sebagai contoh.

Tabel 11. Contoh frekuensi kemunculan kosa kata

No	*Term*	D-1	D-2	No	*Term*	D-1	D-2
1	manchester	3	0	11	laga	1	0
2	united	2	0	12	pamungkas	1	0
3	puruk	1	0	13	city	2	0
4	pekan	1	0	14	liverpool	1	0
5	premier	1	0	15	raih	1	0
6	league	1	0	16	menang	2	0
7	big	1	0	17	chelsea	1	0
8	six	1	0	…	…	…	…
9	telan	1	0	166	suka	0	0
10	kalah	1	0	167	kekang	0	0

Probabilitas untuk setiap kategori yang diperoleh adalah:

$$P(\text{Olahraga}) = \frac{3}{|10|} = 0{,}3$$

$$P(\text{Teknologi}) = \frac{3}{|10|} = 0{,}3$$

$$P(\text{Ekonomi}) = \frac{2}{|10|} = 0{,}2$$

$$P(\text{Lain-lain}) = \frac{2}{|10|} = 0{,}2$$

3.2.2. Tahap testing

Pada tahap *testing* dilakukan analisis setiap *term* yang ada pada dokumen uji dan dihitung nilai probabilitasnya, berdasarkan 4 kategori yang sudah diketahui sebelumnya. Berikut adalah data uji yang akan digunakan

Tabel 12. Contoh data uji dari dataset

ID	Konten
DX-1	manchester united posisi serba salah apa hasil raih derby manchester rabu wib rival manchester city liverpool dekat gelar juara premier league musim
DX-2	xiaomi gelar peresmian untuk rilis produk edisi mi smartphone xiaomi mi special edition rangka acara ulang tahun xiaomi beijing januari

Langkah selanjutnya adalah menghitung probabilitas tiap kata terhadap tiap-tiap kategori menggunakan persamaan 2. Hasil perhitungan disajikan pada Tabel 13.

Tabel 13. Probabilitas dokumen uji DX-1

Term	Probabilitas Kategori			
	Olahraga	Teknologi	Ekonomi	Lain-lain
machester	0,0269	0,0042	0,0046	0,0048
united	0,0115	0,0042	0,0046	0,0048
posisi	0,0038	0,0042	0,0046	0,0048
serba	0,0038	0,0042	0,0046	0,0048
salah	0,0038	0,0042	0,0046	0,0048
apa	0,0038	0,0042	0,0046	0,0048
hasil	0,0038	0,0042	0,0046	0,0048
raih	0,0153	0,0042	0,0046	0,0048
derby	0,0038	0,0042	0,0046	0,0048
rabu	0,0038	0,0042	0,0046	0,0048
wib	0,0076	0,0042	0,0046	0,0048
rival	0,0038	0,0042	0,0046	0,0048
city	0,0231	0,0042	0,0046	0,0048
liverpool	0,0192	0,0042	0,0046	0,0048
dekat	0,0038	0,0042	0,0093	0,0048
gelar	0,0076	0,0085	0,0046	0,0048
juara	0,0115	0,0042	0,0046	0,0048
premier	0,0192	0,0042	0,0046	0,0048
lague	0,0192	0,0042	0,0046	0,0048
musim	0,0153	0,0042	0,0046	0,0048
Total	**0,2106**	**0,0883**	**0,0967**	**0,096**

Tabel 14. Probabilitas dokumen uji DX-2

Term	Probabilitas Kategori			
	Olahraga	Teknologi	Ekonomi	Lain-lain
xiaomi	0,0038	0,0341	0,0046	0,0048
gelar	0,0076	0,0086	0,0046	0,0048
resmi	0,0038	0,0042	0,0046	0,0048
rilis	0,0038	0,0086	0,0046	0,0048
produk	0,0038	0,0129	0,0046	0,0048
edisi	0,0038	0,0086	0,0046	0,0048
mi	0,0038	0,0257	0,0046	0,0048
smartphone	0,0038	0,0086	0,0046	0,0048
special	0,0038	0,0042	0,0046	0,0048
edition	0,0038	0,0086	0,0046	0,0048
rangka	0,0038	0,0042	0,0046	0,0048
acara	0,0038	0,0042	0,0046	0,0048
ulang	0,0038	0,0129	0,0046	0,0048
tahun	0,0038	0,0086	0,0046	0,0048
beijing	0,0038	0,0042	0,0046	0,0048
januari	0,0038	0,0042	0,0046	0,0048
Total	**0,0608**	**0,1624**	**0,0736**	**0,0768**

Probabilitas dokumen uji DX-1 terhadap masing-masing kategori:

P(DX-1 | Olahraga|)
= P(Olahraga) + P(manchester | Olahraga) + P(united | Olahraga) + P(posisi | Olahraga) + P(serba | Olahraga) + P(salah | Olahraga) + P(apa | Olahraga) + P(hasil | Olahraga) + P(raih | Olahraga) + P(derby | Olahraga) + P(rabu | Olahraga) + P(wib | Olahraga) + P(rival | Olahraga) + P(city | Olahraga) + P(liverpool | Olahraga) + P(dekat | Olahraga) + P(gelar | Olahraga) + P(juara | Olahraga) + P(premier | Olahraga) + P(league | Olahraga) + P(musim | Olahraga)
= 0,3 + 0,0269 + 0,0115 + 0,0038 + 0,0038 + 0,0038 + 0,0038 + 0,0038 + 0,0153 + 0,0038 + 0,0038 + 0,0076 + 0,0038 + 0,0231 +0,0192 + 0,0038 + 0,0076 + 0,0115 + 0,0192 + 0,0192 + 0,0153
= 0,3 + 0,2106
= **0,5106**

Dengan cara yang sama, hasil perhitungan probabilitas dokumen uji DX-1 dan DX-2 terhadap semua kategori dirangkum pada Tabel 15.

Tabel 15. Hasil perhitungan klasifikasi dokumen uji

ID	Olahraga	Teknologi	Ekonomi	Lain-lain
DX-1	0,5106	0,3883	0,2967	0,2969
DX-2	0,3608	0,4624	0,2736	0,2768

Dengan demikian, maka diperoleh:
P(DX-1 | Olahraga) > P(DX-1 | Teknologi) > P(DX-1 | Lain-lain) > P(DX-1 | Ekonomi)

P(DX-2 | Teknologi) > P(DX-2 | Olahraga) > P(DX-2 | Lain-lain) > P(DX-2 | Ekonomi)

Dari hasil perhitungan di atas maka dapat disimpulkan perhitungan klasifikasi yang dilakukan menghasilkan dokumen DX-1 termasuk kategori Olahraga, dan dokumen DX-2 berkategori Teknologi.

3.3. Pengujian Akurasi Menggunakan Confusion Matrix

Confusion Matrix adalah metode pengujian yang digunakan untuk melakukan perhitungan akurasi pada konsep *data mining*. Menggunakan 40 dokumen uji, pengukuran akurasi dilakukan dengan persaman (5), yaitu menghitung rata-rata nilai akurasi setiap kelas.

$$\text{Akurasi} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \times 100\% \quad (5)$$

Berdasarkan formula tersebut, didapatkan akurasi masing-masing kategori:

Tabel 16. Hasil pengukuran akurasi dengan *Confusion Matrix*

Kategori	Akurasi (%)
Olahraga	$\frac{9+27}{40} \times 100\% = 90$
Teknologi	$\frac{7+29}{40} \times 100\% = 90$
Ekonomi	$\frac{10+30}{40} \times 100\% = 100$
Lain-lain	$\frac{10+30}{40} \times 100\% = 100$

Dengan demikian, rata-rata akurasi seluruh kelas adalah:

$$\frac{90\% + 90\% + 100\% + 100\%}{4} = 95\ \%$$

4. PEMBAHASAN

4.1. Performa Enhanced Confix Stripping Stemmer

Algoritma *Enhanced Confix Stripping Stemmer* yang dikembangkan oleh Putu Adhi Kerta Mahendra pada tahun 2008 ini, merupakan perbaikan dari algoritma sebelumnya yaitu *Confix Stripping Stemmer*. Penambahan istilah *enhanced* ini dilakukan dengan menambah pengambalian akhiran, agar pemenggalan yang tidak harus dilakukan dapat diantisipasi.

Meskipun *Confix Stripping Stemmer* dapat mengupas dan mengambil salah satu awalan ataupun akhiran agar proses pengambilan dan pemisahan imbuhan dokumen menjadi lebih cepat dan akurat, namun algoritma ini nyatanya perlu penyempurnaan. Beberapa kata yang mengandung partikel "di-kan", "ke-nya", "di-nya", "memper- kan", "memper-i","berpeng-an", "memper-kan", "ke-annya", "ke-nya", "peng- an", "pen-an", "peny-kan", "peny-an", "per-kan", "perkan", dan kata imbuhan sisipan, semula tidak dapat dicari kata dasarnya. Dengan kelebihannya, *Enhanced Confix Stripping Stemmer* mampu menentukan kata dasar dengan partikel tersebut secara lebih baik.

Sejumlah 600 data *training* yang digunakan pada penelitian ini, terdiri atas banyak kata dengan partikel imbuhan dan awalan yang kompleks. Karena kemampuannya inilah, proses stemming dengan *Enhanced Confix Stripping Stemmer* turut berkontribusi terhadap tingginya akurasi pada klasifikasi dokumen berita.

4.2. Peran Naïve Bayes Classifier

Naïve Baiyes merupakan pengklasifikasian probabilistik sederhana yang digunakan untuk menghitung atau mencari probabilitas tertinggi dalam pengklasifikasian data uji pada kategori yang sesuai. Dikemukakan pertama kali oleh Thomas Baiyes, seorang ilmuan Inggris, algoritma ini mampu memprediksi peluang di masa depan berdasarkan pengalaman sebelumnya. Klasifikasi yang dilakukan menggunakan *Naïve Baiyes* bekerja berdasarkan teori probabilitas, yang memandang semua fitur data sebagai bukti yang ada di dalam probabilitas.

Karakteristik yang dimiliki *Naïve Bayes Classifier* di antaranya adalah dapat menangani nilai atribut yang salah, dan dapat bekerja optimal terhadap data yang terisolasi dengan karakteristik berbeda.

Pada proses klasifiasi data latih dan data uji dokumen berita yang dilakukan, *Naïve Bayes Classifier* dilakukan dengan pengkodean sederhana, dan mudah dipahami. Dengan data latih masing-masing kategori (kelas) sebesar 150, menunjukkan bahwa algoritma ini hanya membutuhkan sejumlah kecil data *training* untuk memperkirakan parameter yang dibutuhkan untuk klasifikasi dengan cakap.

5. KESIMPULAN

Penelitian terhadap klasifikasi dokumen berita menggunakan algoritma *Enhanced Confix Stripping Stemmer* dan Naïve Bayes berhasil dilakukan. Sebanyak 600 data dilibatkan sebagai data *training*, dengan pembagian setiap 150 data untuk kategori Olahraga, Teknologi, Ekonomi, dan Lain-lain. Sebagai data *testing,* digunakan sejumlah 40 data yang diambil secara acak pada portal berita www.jawapos.com. Tahap awal penelitian dilakukan proses *preprocessing* pada *data training* dan data *testing* agar dataset dapat diklasifikasikan. Proses *preprocessing* terdiri dari beberapa tahapan yaitu *Case Folding, Tokenizing*, *Filtering*, dan *Stemming.*

Klasifikasi dokumen berita yang dilakukan menghasilkan akurasi untuk kategori Olahraga, Teknologi, Ekonomi, dan Lain-lain berturut-turut sebesar 90%, 90%, 100%, dan 100%. Rata-rata akurasi 4 kategori tersebut berada pada angka 95%. Nilai ini diperoleh berkat peran algoritma *Enhanced Confix Stripping Stemmer* yang mangkus, dan metode klasifikasi *Naïve Bayes Classifier* yang terbukti handal.

DAFTAR PUSTAKA

[1] K. Nikoloski, "The Role of Information Technology in the Business Sector," *Int. J. Sci. Res.*, vol. 3, no. 12, pp. 303–309, 2014.

[2] A. Berisha-Shaqiri, "Impact of Information Technology and Internet in Businesses," *Acad. J. Business, Adm. Law Soc. Sci.*, vol. 1, no. 1, pp. 73–79, 2015.

[3] A. F. Malkawi, "The Impact of the Use of Information Technology in Improving the Quality of Services: A Field Study of Fast-Food Restaurants in Jordan," *Eur. Sci. J. August*, vol. 13, no. 23, pp. 359–376, 2017.

[4] J. E. M. Peñalba, G. M. Guzmán, and E. G. de Mojica,

https://doi.org/10.25077/ TEKNOSI.v6i2.2020.90-99

"The Effect of Information and Communication Technology in Innovation Level: The Panama SMEs Case," *J. Bus. Econ. Policy*, vol. 2, no. 2, pp. 124–131, 2015.

[5] A. S. Kümpel, V. Karnowski, and T. Keyling, "News Sharing in Social Media: A Review of Current Research on News Sharing Users, Content, and Networks," *Soc. Media Soc.*, vol. 1, no. 2, pp. 1–14, 2015.

[6] S. Cortesi and U. Gasser, "Youth Online and News: A Phenomenological View on Diversity," *Int. J. Commun.*, vol. 9, no. 1, pp. 1425–1448, 2015.

[7] N. Newman, R. Fletcher, A. Kalogeropoulos, and R. K. Nielsen, "Reuters Institute Digital News Report 2019," 2019.

[8] J. Hillgaertner, "Current Trends in the History of News," *Reformation*, vol. 20, no. 1, pp. 68–76, 2015.

[9] C. E. Everett, "Transformation of Newspapers in the Technology Era," *Elon J. Undergrad. Res. Commun.*, vol. 2, no. 2, pp. 102–115, 2011.

[10] R. Mesquita, "The Transition of a Traditional Newspaper to the Internet Age: An Historical Account of Le Monde's Case," *Observatorio*, vol. 11, no. 1, pp. 54–60, 2017.

[11] M. S. Weber, "The Tumultuous History of News on the Web," in *The Web as History. Using Web Archives to Understand the Past and the Present*, N. Brügger and R. Schroeder, Eds. London: UCL Press, 2017, pp. 83–100.

[12] N. Ahmad, "The Decline of Conventional News Media and Challenges of Immersing in New Technology," no. 25, pp. 71–82, 2016.

[13] Nurkinan, "Dampak Media Online Terhadap Perkembangan Media Konvensional," *J. Polit. Indones.*, vol. 2, no. 2, pp. 28–42, 2017.

[14] M. Rustam, "Internet dan Penggunaannya (Survei di Kalangan Masyarakat Kabupaten Takalar Provinsi Sulawesi Selatan)," *J. Stud. Komun. Dan Media*, vol. 21, no. 1, pp. 13–24, 2017.

[15] L. P. Supratman, "Penggunaan Media Sosial oleh Digital Native," *J. Ilmu Komun.*, vol. 15, no. 1, pp. 47–60, 2018.

[16] C. Juditha, "Akurasi Berita dalam Jurnalisme Online (Kasus Dugaan Korupsi Mahkamah Konstitusi di Portal Berita Detiknews)," *Pekommas*, vol. 16, no. 3, pp. 145–154, 2013.

[17] I. A. Setiawan, T. H. Pudjiantoro, and D. Nursantika, "Klasifikasi Artikel Berita Menggunakan Metode Text Mining dan Naive Bayes Classifier," in *Seminar Nasional Inovasi Dan Aplikasi Teknologi Di Industri*, 2017, pp. 1–6.

[18] P. Widodo, J. A. Putra, S. Afiadi, A. Z. Arifin, and D. Herumurti, "Klasifikasi Kategori Dokumen Berita Berbahasa Indonesia dengan Metode Kategorisasi Multilabel Berbasis Domain Specific Ontology," *J. Ilm. Teknol. Inf. Terap.*, vol. 2, no. 2, pp. 126–137, 2016.

[19] A. Singh and S. K. Chhillar, "A Survey on Machine Learning Techniques for Text Classification," *Int. J. Comput. Sci. Technol.*, vol. 8, no. 2, pp. 205–209, 2017.

[20] B. Kurniawan, S. Effendi, and O. S. Sitompul, "Klasifikasi Konten Berita Dengan Metode Text Mining," *J. Dunia Teknol. Inf.*, vol. 1, no. 1, pp. 14–19, 2012.

[21] S. Andini, "Klasifikasi Dokumen Teks Menggunakan Algoritma Naive Bayes dengan Bahasa Pemprograman Java," *J. Teknol. Inf. Pendidik.*, vol. 6, no. 2, pp. 140–147, 2013.

[22] M. N. Khidfi, Isnawaty, and Jayanti Yusma Sari, "Rancang Bangun Aplikasi Pendeteksian Kesamaan pada Dokumen Teks Menggunakan Algoritma Enhaced Confix Stripping dan Algoritma Winnowing," *Semant. Tek. Inf.*, vol. 4, no. 2, pp. 1–10, 2018.

[23] H. Shimodaira, "Text Classification using Naive Bayes," *Learn. Data Note*, vol. 7, pp. 1–9, 2015.

[24] D. Xhemali, C. J. Hinde, and R. G. Stone, "Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages," *Int. J. Comput. Sci.*, vol. 4, no. 1, pp. 16–23, 2009.

[25] W. Jang, J. K. Lee, J. Lee, and S. H. Han, "Naive Bayesian Classifier for Selecting Good/Bad Projects during the Early Stage of International Construction Bidding Decisions," *Math. Probl. Eng.*, vol. 2015, pp. 1–12, 2015.

[26] D. N. Chandra, G. Indrawan, and I. N. Sukajaya, "Klasifikasi Berita Lokal Radar Malang Menggunakan Metode Naïve Bayes dengan Fitur N-Gram," *J. Ilm. Teknol. Inf. Asia*, vol. 10, no. 1, pp. 11–19, 2016.

[27] F. Z. Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia," University of Amsterdam, 2003.

NOMENKLATUR

P(H|E)	probabilitas akhir bersyarat hipotesis H terjadi
P(E|H)	probabilitas bukti E terjadi akan mempengaruhi
P(H)	probabilitas awal hipotesi H terjadi tanpa melihat bukti apapun
P(E)	probabilitas awal bukti E terjadi tanpa melihat hipotesis apapun atau bukti lain
$P(W_k \mid C_i)$	probabilitas kata k terhadap kategori Ci
nk	nilai kemunculan kata k terhadap kategori Ci
n	jumlah seluruh kata pada kategorii C_i
|vocabulary|	jumlah seluruh kosa kata|
$P(C_i)$	probabilitas dari kategori C_i
$f_d(C_i)$	jumlah dokumen yang memiliki kategori C_i
|D|	jumlah seluruh dokumen *training*
TN	*True Negative*, jumlah data negatif yang terdeteksi dengan benar
TP	*True Positive*, jumlah data positif yang terdeteksi dengan benar
FN	*False Negative*, jumlah data positif terdeteksi dengan salah
FP	False Positive, jumlah data negatif terdeteksi dengan benar

BIODATA PENULIS

Erwin Yudi Hidayat

Menamatkan pendidikan sarjana di Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro, sebagai Sarjana Komputer (S.Kom). Studi jenjang S2 diselesaikan di Universiti Teknikal Malaysia Melaka (UTeM), dengan konsentrasi pada bidang kecerdasan buatan. Penulis tertarik dalam penelitian *machine learning*, *deep learning*, dan pengolahan citra digital. Saat ini menjadi staf pengajar pada almamater di tempat kuliah S1 dulu ditempuh.

Muhammad Aditya Rizqi

Penulis adalah alumni Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro, Semarang. Ketertarikannya akan *data mining*, mengantarkan penulis terhadap penelitian klasifikasi dokumen. Gelar S.Kom yang diperoleh, diselesaikan dengan baik dan lulus dengan kepujian.