Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.201853810
Vol. 5, No. 3, Agustus 2018, hlm. 269-276 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 51/E/KPT/2017 e-ISSN: 2528-6579

KLASIFIKASI BERITA OLAHRAGA MENGGUNAKAN METODE *NAÏVE BAYES* DENGAN *ENHANCED CONFIX STRIPPING STEMMER*

Yoga Dwitya Pramudita[1], Sigit Susanto Putro[2], Nurul Makhmud[3]

[123]Prodi Teknik Informatika Universitas trunojoyo
Email: [1]yoga@trunojoyo.ac.id, [2]sigit.putro@trunojoyo.ac.id

(Naskah masuk: 6 Mei 2018, diterima untuk diterbitkan: 7 Agustus 2018)

Abstrak

Dokumen berita olahraga dalam bentuk web kini memiliki jumlah yang besar dalam kurun waktu singkat. Untuk kemudahan akses dokumen perlu melakukan pengelompokan dokumen berita kedalam beberapa kategori. Hal tersebut bertujuan agar berita olahraga tersusun sesuai dengan kategori yang ditentukan. Berita dapat dikelompokkan secara manual oleh manusia, akan tetapi hal tersebut membutuhkan waktu yang lama untuk melakukan kategorisasi. Metode klasifikasi diusulkan dalam penelitian ini untuk melakukan pengkategorian secara otomatis dokumen berita. Tujuan dilakukannya klasifikasi adalah untuk mempercepat dan mempermudah dalam pemberian kategori, sehingga dapat meningkatkan efisiensi waktu. Pada penelitian ini menggunakan metode klasifikasi Naïve Bayes Classifier. Sebelum dilakukan klasifikasi ada proses preprocessing dengan menggunakan Enhanced Confix Striping Stemmer. Hal ini bertujuan untuk mengembalikan ke bentuk kata dasar, sehingga data berkurang dan proses komputasi menjadi lebih efisien. Pengujian dilakukan menggunakan 18 berita olahraga yang dipilih secara acak oleh user atau tester, dari 18 berita yang diujikan terdapat 14 berita yang bernilai benar atau relevan dengan analisis yang dilakukan use atau tester pada berita uji. Dari penelitian ini dapat disimpulkan bahwa Aplikasi Klasifikasi Berita Olahraga menggunakan Metode Naïve Bayes dengan Enhanced Confix Striping Stemmer mampu mengklasifikasi berita olahraga sesuai dengan kategori masing-masing, seperti Sepak Bola, Basket, Raket, Formula 1, Moto GP dan olahraga lainnya dengan keakuratan sebesar 77%.

Kata kunci: *Klasifikasi, Berita Olahraga, Naïve Bayes Classifier, Enhance Confix Stripping Stemmer*

SPORTS NEWS CLASSIFICATION USING NAÏVE BAYES WITH ENHANCED CONFIX STRIPPING STEMMER

Abstract

Web-based sports news currently has a considerable amount of documents. News documents need to be grouped into multiple categories for easy access. The goal is that sports news is structured according to the specified category. News can be grouped manually by humans, but it takes a long time to categorize if it involves large documents. Classification method is proposed in this research to categorize automatically news document. The purpose of doing the classification is to accelerate and simplify the granting of categories, thereby increasing the efficiency of time. In this research using the Naïve Bayes Classifier classification method. Prior to classification there is a preprocessing process using Enhanced Confix Striping Stemmer. It aims to return to the basic word form, so the data is reduced and the computing process becomes more efficient. From the test using 18 sports news randomly selected by the user or tester, there are 14 news stories that are true or relevant to the analysis by the user or the tester on the test news. This study concludes that the Sports News Classification Application using the Naïve Bayes Method with Enhanced Confix Striping Stemmer is able to classify sports news according to their respective categories, such as Football, Basket, Racquet, Formula 1, Moto GP and other sports with accuracy of 77%.

Keywords: *Classification, Sports News, Naïve Bayes Classifier, Enhance Confix Stripping Stemmer*

1. PENDAHULUAN

Berita adalah informasi berdasarkan fakta atau laporan mengenai suatu kejadian yang sedang atau telah terjadi dan dipublikasikan melalui media cetak, siaran, internet maupun dari mulut ke mulut. Dengan adanya berita, masyarakat menjadi lebih tahu mengenai kejadian terkini. Berita olahraga merupakan salah satu berita yang paling banyak diakses oleh masyarakat. Pada alexa.com situs berita olahraga seperti sport.detik.com, bola.net masuk pada

25 *top site* di Indonesia. Hal tersebut menunjukkan bahwa masyarakat lebih memilih mengakses berita olahraga melalui internet. Berbeda dengan media cetak ataupun siaran, berita melalui internet dapat kapan saja diakses tanpa adanya batas waktu. Akan tetapi seiring dengan berjalannya waktu, jumlah berita olahraga pada internet akan semakin besar. Besarnya jumlah berita membuat berita tersebut perlu diorganisasi kedalam kelompok atau kategorisasi sesuai dengan isi dari berita. Dengan tujuan agar berita tersebut terorganisir sehingga memudahkan dalam hal akses. Kategorisasi dapat dilakukan dengan menerapkan teknik *data mining* yaitu klasifikasi.

Data mining adalah suatu proses ekstraksi atau penggalian dari data yang belum diketahui sebelumnya (Thomas, 2015). Data tersebut digali dari database besar dan digunakan untuk membuat suatu keputusan bisnis yang sangat penting. Sedangkan klasifikasi adalah pengolahan untuk menemukan kumpulan model (atau fungsi) yang menggambarkan dan membedakan kelas data atau konsep, dengan tujuan mampu menggunakan model untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui. Model didasarkan pada analisis kumpulan data latih (jiawei, 2000).

Sebelum berita dapat diklasifikasi, perlu melakukan tahap *preprocessing*. *Preprocessing* itu sendiri terdiri dari beberapa tahapan, yaitu : *case folding, tokenizing/parsing, filtering* dan *stemming* (Triawati, 2009). Pada tahapan *preprocessing*, proses *stemming* merupakan proses yang terpenting karena pada *stemming* terjadi penghilangan kata imbuhan sehingga menghasilkan kata dasar. Algoritma yang digunakan pada proses *stemming* adalah *Algoritma Enhanced Confix Stripping Stemmer*. Algoritma tersebut merupakan pengembangan dari *Algoritma Confix Stripping* dengan melakukan beberapa perbaikan aturan sehingga memiliki tingkat kesalahan paling sedikit dibandingkan algoritma sebelumnya (Putu, 2008).

Klasifikasi yang digunakan adalah *Naïve Bayes Classifier*. Klasifikasi ini adalah model probalistik yang simpel untuk klasifikasi data kedalam kelas yang spesifik berdasarkan fitur data yang berbeda (Friedman, 1997). *Naïve Bayes* telah menjadi metode utama yang digunakan pada klasifikasi untuk variasi pada data mulai dari medis, jaringan komputer dan teks karena kesederhanaan, keefektifan, dan kapabilitas menangkap penalaran data dalam model grafis (Abraham, 2009; Zgan & Gao, 2011; Mukherjee & Sharmaa, 2012). Kelebihan *Naïve Bayes Classifier* yaitu konsep yang mudah dimengerti, tidak sensitif terhadap fitur yang relevan, dapat menangani data *real* maupun diskrit (Eamonn). Meskipun asumsi bahwa fitur-fitur dataset bersifat *independent* adalah sebuah asumsi yang kurang baik tetapi kenyataannya hasil klasifikasi berbasis *Naïve Bayes* memiliki kinerja yang mampu berkompetisi dengan metode-metode lain yang lebih kompleks dalam aspek komputasinya (Bagus, 2016). *Naïve Bayes Classifier* memiliki keunggulan dalam hal waktu komputasi dibandingkan dengan algoritma klasifikasi lainnya (Dwi, 2012).

Penelitian tentang klasifikasi berita sudah pernah dilakukan sebelumnya, diantaranya yaitu penelitian menggunakan metode *Naïve Bayes* dengan fitur *N-Gram* untuk mengklasifikasi berita lokal radar Malang. Dokumen latih diambil dari web portal www.kompas.com dimana terdapat beberapa kategori dengan dokumen politik, ekonomi, *news*, edukasi, kesehatan, *travel*, dan olahraga. Sedangkan dokumen uji diambil pada portal www.radarmalang.co.id. Pada penelitian ini dilakukan 5 ujicoba. Pada kelima ujicoba yang dilakukan didapatkan ketepatan prediksi atau akurasi sebesar 78.66%, 68.20%, 59.24%, 65.93%, 74.39% (Denny, 2016).

Penelitian selanjutnya menggunakan metode *Naïve Bayes* dengan *Natural Langage Processing* untuk mengklasifikasi jenis berita pada arsip pemberitaan. Pada penelitian ini, awalnya setiap artikel akan dijabarkan kata per kata dan dipisah atau dibersihkan dari penggunaan tanda baca, kalimat positif dan negatif dengan metoda *natural language processing*. Kemudian setiap paragraf akan diambil kata kunci. Kata kunci akan di input sebagai data latih ke dalam *Database*. Setelah pemecahan kata, langkah selanjutnya adalah setiap data akan diklasifikasikan dengan metode *Naïve Bayes*. Dari percobaan yang dilakukan dengan 4 kategori artikel masing-masing memberikan nilai akurasi hingga lebih dari 82%., dimana pada percobaan 1 didapatkan akurasi sebesar 82.5%, pada percobaan 2 didapatkan akurasi sebesar 89%, pada percobaan 3 didapatkan akurasi sebesar 89.375%, pada percobaan 4 didapatkan akurasi sebesar 90%, dan pada percobaan didapatkan 5 akurasi sebesar 94.16667% (Novia, 2016).

Penelitian berikutnya menggunakan metode *Naïve Bayesian Classification* dan *Support Vector Machine* dengan *Confix Stripping Stemmer* untuk mengklasifikasi berita Indonesia. Dalam penelitian ini data telah dibagi menjadi dua yaitu data latih dan uji dengan proporsi 70:30. Jumlah *word vector* yang akan diuji coba pada data latih adalah 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, dan 10000. Sedangkan untuk data uji karena jumlah artikel berita lebih sedikit maka *word vector* yang akan digunakan adalah 1000, 1500, 2000, 2500, dan 3000. Pada penelitian ini dihasilkan bahwa semakin banyak jumlah *word vector* pada data latih maka semakin tepat hasil akurasi yang dihasilkan yaitu dengan akurasi sebesar 82,22% untuk klasifikasi menggunakan metode *Naïve Bayes Classifier*. Sedangkan klasifikasi dengan menggunakan metode *Support Vector Machine*, hasil ketepatan klasifikasi pada data latih dipengaruhi oleh parameter C dan *Gamma*. Semakin mengecil nilai *gamma* semakin mengurangi ketepatan klasifikasi sehingga perlu ditambahkan parameter C yang lebih besar. Nilai parameter C yang digunakan 10-2 hingga 104. Hasil

akurasi yang dihasilkan pada data uji menggunakan kernel linear yaitu sebesar 88,1% (Dio, 2015).

Implementasi Algoritma *Naïve Bayes Classifier* Berbasis *Particle Swarm Optimization* (PSO) Untuk Klasifikasi Konten Berita Digital Bahasa Indonesia. Pada penelitian ini, peneliti melakukan pengujian model dengan menggunakan teknik 10 *cross validation*, di mana proses tersebut membagi data secara acak ke dalam 10 bagian. Proses pengujian dimulai dengan pembentukan model dengan data pada bagian pertama. Model yang terbentuk akan diujikan pada 9 bagian data sisanya. Setelah itu proses akurasi dihitung dengan melihat seberapa banyak data yang sudah terklasifikasi dengan benar. Dari hasil pengujian yang dilakukan membuktikan bahwa algoritma *Naïve Bayes Classifier* berbasis *Particle Swarm Optimization* memiliki nilai akurasi sebesar 94.17% yang didapat saat proses ke-5 *fold cross validation* (Nurhadi, 2016).

Implementasi dari Metode *K-Nearest Neighbor* dengan *Decision Rule* dilakukan untuk Klasifikasi Subtopik Berita. Pada penelitian ini dilakukan 3 kali percobaan dengan nilai k yang berbeda yaitu k=3, k=5, k=7, didapatkan hasil akurasi 88,29% untuk k=3, 88,29% untuk k=5, dan 87,23% untuk k=7, dari percobaan yang dilakukan disimpulkan bahwa nilai k tidak banyak berpengaruh pada hasil akhir *K-Nearest* karena persentase keakuratan rata-rata masih diatas 80% (tergolong baik). Sedangkan keakuratan klasifikasi dengan menggunakan *K-Nearest* dengan *Decision Rule*, dengan nilai k yang sama yaitu k=3, k=5, k=7 didapatkan hasil yang tidak jauh berbeda dengan klasifikasi menggunakan *K-Nearest* yaitu dengan akurasi sebesar 89,36% untuk k=3, 86,17% untuk k=5, 87,23% untuk k=7. Hanya terjadi peningkatan akurasi sebesar 2% pada k=3. Jadi dapat disimpulkan penggunaan k=3 merupakan hasil klasifikasi yang menunjukkan presentase terbaik dalam *K-Nearest Neighbor* maupun *K-Nearest Neighbor with Decision Rule* (Yoseph dkk, 2015).

Dari paparan diatas, pada penelitian ini melakukan klasifikasi berita olahraga menggunakan metode *Naïve Bayes* dengan *Enhanced Confix Stripping Stemmer* sebagai algoritma *stemming*. Tujuannya adalah mengetahui seberapa akurat klasifikasi yang dihasilkan jika menggunakan *preprocessing* tersebut.

2. METODE

Klasifikasi adalah proses pencarian se-kumpulan model atau fungsi yang menggambarkan dan membedakan kelas data dengan tujuan agar model tersebut dapat digunakan untuk memprediksi kelas dari suatu objek yang belum diketahui kelasnya. Klasifikasi memiliki dua proses yaitu membangun model klasifikasi dari sekumpulan kelas data yang sudah didefinisikan sebelumnya (*training data set*) dan menggunakan model tersebut untuk klasifikasi data tes serta mengukur akurasi dari model (Rachli, 2007).

2.1 TEXT PREPROCESSING

Text Preprocessing adalah suatu proses pengubahan bentuk data yang belum terstruktur atau tidak terstruktur menjadi data yang terstruktur (mengubah teks menjadi *term index*). Tujuannya adalah untuk memperkecil dimensi data sehingga proses komputasi lebih menjadi efisien dan diharapkan lebih presisi. *Preprocessing* terdiri dari beberapa tahapan. Adapun tahapan preprocessing berdasarkan, yaitu : *case folding*, *tokenizing/ parsing*, *filtering* dan *stemming* (Triawati, 2009).

1. *Case Folding* adalah mengubah semua huruf dalam teks menjadi huruf kecil.
2. *Tokenizing* adalah sebuah proses untuk memilah isi teks sehingga menjadi satuan kata-kata.
3. *Filtering* adalah tahap mengambil kata-kata penting dari hasil token. Bisa menggunakan algoritma *stoplist* (membuang kata yang kurang penting) atau *wordlist* (menyimpan kata penting).
4. *Stemming* adalah suatu proses untuk mereduksi kata ke bentuk dasarnya. Tahap *stemming* merupakan tahap mencari akar (*root*) kata dari tiap kata hasil *filtering* (Agus dkk, 2015).

2.2 ALGORITMA ENHANCED CONFIX STRIPPING STEMMER

Algoritma *Enhanced Confix Stripping Stemmer* merupakan perbaikan dari algoritma sebelumnya yaitu, algoritma *Confix Stripping Stemmer*. Setelah dilakukan beberapa percobaan dan analisis, ditemukan beberapa kata yang tidak dapat di-*stemming* menggunakan *Confix Stripping Stemmer* yaitu sebagai berikut (Andita dkk, 2010) :

1. Kurangnya aturan pemenggalan awalan untuk kata-kata dengan format “mem+p...”, Contohnya yaitu terjadi pada kata “mempromosikan”, “memproteksi”
2. Kurangnya aturan pemenggalan awalan untuk kata-kata dengan format “men+s...”, Contohnya yaitu terjadi pada kata “mensyaratkan”, “mensyukuri”.
3. Kurangnya aturan pemenggalan awalan untuk kata-kata dengan format “menge+...”. Contohnya yaitu terjadi pada kata “mengerem”.
4. Kurangnya aturan pemenggalan awalan untuk kata-kata dengan format “penge+...”. Contohnya yaitu terjadi pada kata “pengeboman”.
5. Kurangnya aturan pemenggalan awalan untuk kata-kata dengan format “peng+k...”. Contohnya yaitu terjadi pada kata “pengkajian”.

6. Adanya elemen pada beberapa kata dasar yang menyerupai suatu imbuhan. Kata-kata seperti "pelanggan", "perpolitikan", dan "pelaku" gagal distemming karena akhiran "-an", "-kan" dan "-ku" seharusnya tidak dihilangkan.

Pada Algoritma *Enhanced Confix Stripping Stemmer* terdapat beberapa modifikasi aturan pemenggalan yang yang dapat dilihat pada tabel 1.

Tabel 1. Modifikasi dan Tambahan Aturan Oleh *Enhanced Confix Stripping Stemmer*

Aturan	Format Kata	Pemenggalan
14	men{e|c|d|j|z}...	men-{e|c|d|j|z}...
17	mengV...	meng-V... | meng-kV... | (mengV-... Jika V='e')
19	mempA...	mem-pA...dimana A!='e'
28	pengC...	peng-C...
29	pengV...	peng-V... | peng-kV... | (pengV-... Jika V='e')

Untuk memperbaiki kesalahan yang telah disebutkan diatas, algoritma *Enhanced Confix Stripping Stemmer* melakukan beberapa perbaikan sebagai berikut (Andita dkk, 2010):

1. Melakukan modifikasi beberapa aturan.
2. Menambahkan suatu algoritma tambahan untuk mengatasi kesalahan pemenggalan akhiran. Algoritma tambahan kemudian disebut dengan loopPengembalianAkhiran. Langkah ini dilakukan apabila proses *recoding* gagal.

Algoritma *loopPengembalianAkhiran* di-definisikan sebagai berikut :

1. Mengembalikan seluruh awalan yang telah dihilangkan sebelumnya, sehingga menghasilkan model kata seperti berikut : [DP+[DP+[DP]]] + Kata Dasar.
   Pemenggalan awalan dilanjutkan dengan proses pencarian di kamus, kemudian dilakukan pada kata yang telah dikembalikan menjadi model tersebut. Jika proses tersebut sukses maka akan dihentikan, apabila tidak sukses maka langkah selanjutnya akan dilakukan.
2. Mengembalian akhiran yang telah dihilangkan sebelumnya. Hal ini artinya bahwa pengembalian dimulai dari DS ("-i", "-kan", "-an"), lalu PP ("-ku", "-mu", "-nya"), dan terakhir adalah P ("-lah", "-kah", "-tah", "-pun"). Untuk setiap pengembalian, lakukan langkah 3) hingga 5) berikut. Khusus untuk akhiran "-kan", pengembalian pertama dimulai dengan "k", baru kemudian dilanjutkan dengan "an".
3. Lakukan pengecekan di kamus kata dasar. Apabila ditemukan, proses dihentikan. Apabila gagal, maka lakukan proses pemenggalan awalan berdasarkan aturan.
4. Lakukan *recoding* apabila diperlukan.

Apabila pengecekan di kamus kata dasar tetap gagal setelah *recoding*, maka awalan-awalan yang telah dihilangkan dikembalikan lagi.

2.3 TERM FREQUENCY

Term Frequency merupakan salah satu metode untuk menghitung bobot tiap *term* dalam *text*. Pada metode ini, tiap term diasumsikan memiliki nilai kepentingan yang sebanding dengan jumlah kemunculan term tersebut pada *text* (Diah, 2010).

2.4 NAÏVE BAYES CLASSIFIER

Naïve Bayes Clasifier atau disebut juga dengan Bayesian Classification merupakan metode pengklasifikasian statistik yang didasarkan pada teorema bayes yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu kelas (Muhammad, 2017).

Naïve Bayes Classifier pada penelitian ini digunakan untuk mengklasifikasikan dokumen teks. Pada algoritma *Naïve Bayes* setiap dokumen dipresentasikan dengan masukan atribut "a_1, a_2, a_3,…,a_n" dimana a_1 adalah kata pertama dan berikutnya sampai a_n (kata ke-n), sedangkan V yaitu label kategori. Selanjutnya yaitu mencari nilai tertinggi dari kategori teks yang diujikan (V_{MAP}) (McCallum, 1998). Persamaan V_{MAP} yaitu sebagai berikut (Bagus, 2016) :

$$V_{MAP} = \underset{vj \in V}{\operatorname{argmax}} P(v_j) \prod_i P(a_i|v_j) \quad (1)$$

Nilai $P(v_j)$ dihitung pada saat data latih, dengan rumus sebagai berikut :

$$P(v_j) = \frac{|dokumen\ j|}{|dok.\ training|} \quad (2)$$

Dimana :
|dokumen j| adalah jumlah dokumen yang memiliki kategori j pada dokumen latih.
|dok. training| adalah jumlah dokumen latih.

$$P(a_i|v_j) = \frac{|n_i+1|}{|n+kosa\ kata|} \quad (3)$$

Dimana :

1. n_i adalah jumlah kemunculan kata a_i pada dokumen yang berkategori v_j
2. n adalah jumlah seluruh kata pada dokumen yang berkategori v_j

3. kosakata adalah jumlah kata pada seluruh dokumen latih.

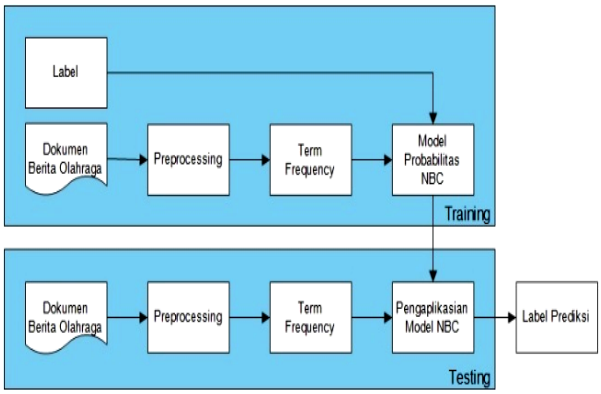
Untuk mengetahui tingkat akurasi dari luaran klasifikasi maka bisa dihitung dengan menggunakan rumus akurasi berikut.

Perhitungan Akurasi :

$$\text{Akurasi} = \frac{\text{Jumlah Dokumen Diuji Benar}}{\text{Jumlah Dokumen diujikan}} x\ 100\% \qquad (4)$$

3. ARSITEKTUR SISTEM

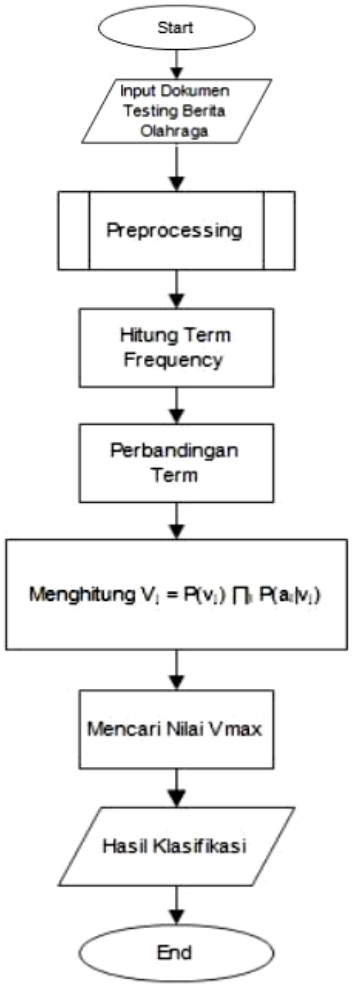
Pada gambar 1 menunjukkan arsitektur klasifikasi dokumen berita olahraga. Terdapat dua pembagian dokumen, yaitu dokumen latih dan dokumen uji. Pada dokumen latih dilakukan proses pembelajaran (*learning*) pada setiap kategori untuk menghasilkan model probabilitas.



Gambar 1. Gambaran umum sistem

Sebelum dilakukan proses pembelajaran (*learning*), dokumen terlebih dahulu melalui tahap *preprocessing* dan perhitungan *term frequency*. Sama halnya dengan dokumen latih, dokumen uji juga melalui tahap *preprocessing* dan perhitungan *term frequency* setelah selanjutnya dilakukan proses pengklasifikasian yang mengacu pada model probabilitas yang dihasilkan pada proses pembelajaran (*learning*), dan label kategori pada dokumen uji dapat ditentukan dengan mencari nilai V (nilai tertinggi dari kategori teks yang diujikan.

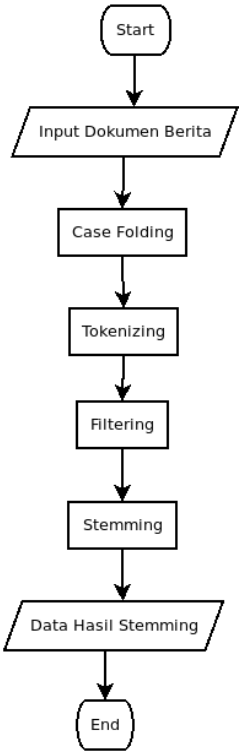
Arsitektur dari sistem yang akan dibuat, akan dideskripsikan pada beberapa *flowchart* pada gambar 2. Flowchart tersebut menunjukan alur dari proses pengklasifikasian dengan metode *Naïve Bayes*. Adapun tahapannya yaitu input-kan dokumen uji berita olahraga. Lalu dokumen tersebut dilakukan *preprocessing*. Setelah dokumen dilakukan *preprocessing*, selanjutnya menghitung *term frequency*. Langkah selanjutnya yaitu menghitung $V_j = P(v_j) \prod_i P(a_i|v_j)$ untuk setiap kategori dengan mengacu pada model probabilitas yang telah diketahui pada proses pelatihan sebelumnya. Kategori dapat ditentukan dengan mencari nilai V_j maksimal.



Gambar 2. Flowchart Naïve Bayes Classifier

Pada gambar 3 menunjukan *flowchart* dari proses *preprocessing*. Adapun tahapan *preprocessing* berdasarkan, yaitu : *case folding*, *tokenizing/parsing*, *filtering* dan *stemming*. Tahapan yang pertama setelah dokumen berita dimasukkankan yaitu tahapan *case folding*, *case folding* adalah mengubah huruf kapital (besar) menjadi huruf kecil. Tahap selanjutnya yaitu tahap *tokenizing*, yaitu proses untuk memilah isi teks sehingga menjadi satuan kata-kata. Setelah melalui tahap *tokenizing*, selanjutnya yaitu tahap *filtering* yaitu mengambil kata-kata penting dari hasil token. Pada tahap *filtering* akan menghilangkan kata *stopword* atau kata yang dianggap tidak penting. Dan tahapan terakhir pada *preprocessing* yaitu *stemming*, *stemming* adalah mengubah kata ke dalam bentuk aslinya.

Algoritma *stemming* yang digunakan dalam penelitian ini adalah Algoritma *Enhanced Confix Stripping Stemmer*. Pada gambar 4 merupakan *flowchart Enhanced Confix Stripping Stemmer*. Algoritma tersebut merupakan algoritma perbaikan dari algoritma sebelumnya yaitu algoritma *Confix Stripping Stemmer*.



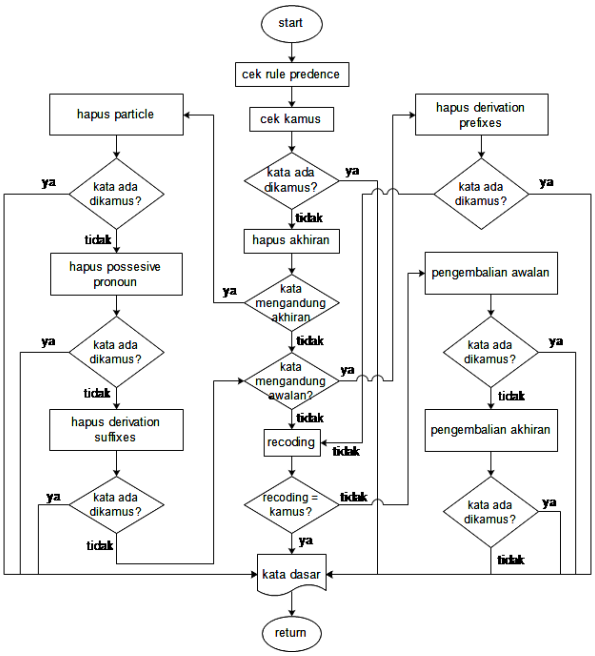
Gambar 3. Flowchart Preprocessing

Langkah pertama pada algoritma *stemmer* adalah melakukan pengecekan *rule Precedence*, yaitu kombinasi awalan dan akhiran yang tidak diperbolehkan, kemudian mencocokkan kata yang diinputkan dengan kata dasar pada *database*. Apabila kata yang dimasukkan dengan kata dasar sama, maka kata tersebut merupakan kata dasar. Jika tidak sama, maka dilanjutkan ke proses hapus akhiran.

Pada hapus akhiran terdapat 3 *rule* yaitu hapus *particle*, hapus *possesive pronoun*, dan hapus *derivation suffixes*, apabila ketiga *rule* tersebut telah diproses dan kata yang diproses tadi terdapat kesamaan dengan kata dasar, maka kata tersebut merupakan kata dasar. Jika ketiga *rule* tersebut telah diproses akan tetapi kata yang diproses tidak terdapat kesamaan dengan kata dasar, maka selanjutnya yaitu jalankan proses hapus awalan.

Pada hapus awalan terdapat proses hapus *derivation prefixes*, dimana apabila kata yang diproses tadi terdapat kesamaan dengan kata dasar, maka kata tersebut merupakan kata dasar. Apabila telah dilakukan proses hapus awalan akan tetapi kata dasar belum ditemukan, maka proses *recoding* dilakukan.

Proses *recoding* merupakan penyusunan kembali kata-kata yang mengalami proses *stemming* yang berlebih. Pada proses *recoding* akan dilakukan proses pemenggalan kata. Apabila kata yang mengalami proses *recoding* sama dengan kata dasar yang ada pada *database*, maka kata tersebut merupakan kata dasar. Apabila tidak ditemukan kesamaan dengan kata dasar yang ada pada *database*, maka proses *recoding* dikatakan gagal, dan akan dilakukan *loopPengembalianAkhiran.*



Gambar 4. Flowchart Enhanced Confix Stripping Stemmer

Proses berikutnya adalah mengembalikan seluruh awalan yang telah dihilangkan sebelumnya, sehingga menghasilkan model kata seperti berikut : [DP+[DP+[DP]]] + Kata Dasar. Pemenggalan awalan dilanjutkan dengan proses pencarian di kamus kemudian dilakukan pada kata yang telah dikembalikan menjadi model tersebut. Jika proses tersebut sukses maka akan dihentikan, apabila tidak sukses maka langkah selanjutnya akan dilakukan.

Selanjutnya yaitu mengembalikan akhiran yang telah dihilangkan sebelumnya. Lalu lakukan pengecekan di kamus kata dasar. Apabila ditemukan, proses dihentikan. Apabila tidak ditemukan maka kata tersebut dianggap sebagai kata dasar.

4. ANALISIS HASIL UJI COBA

Pada klasifikasi berita olahraga menggunakan metode *Naïve Bayes* terdapat 2 jenis dokumen berita yaitu berita latih dan berita uji. Berita latih didapat dari situs berita olahraga yaitu *sport.detik.com*. Peneliti mengambil 151 berita latih dari 6 kategori yaitu sepakbola, basket, raket, motoGP, formula 1 dan berita olahraga lainnya. Berita olahraga lainnya berisi berita olahraga dengan kategori selain dari kelima kategori tersebut, seperti taekwondo, tinju, lari, voli dan lainnya. Pengambilan berita latih dilakukan secara acak oleh peneliti tanpa melakukan seleksi sebelumnya.

Tabel 2. Hasil pengujian

Pengguna : (Wartawan Madura TV)

No.	Judul Berita	Kategori (Sistem)	Kategori (*User*)	Hasil (Relevan)
1	Indonesia Jadi Runner Up Kejuaraan Asia Bulutangkis Junior	Lainnya	Raket	X
2	Chris Froome Juara Tour de France untuk Keempat Kalinya	Lainnya	Lainnya	V
3	FIA Konfirmasi Halo Akan Digunakan di F1 Musim Depan	Lainnya	Formula 1	X
4	Antusiasme Gerry Salim cs Hadapi Seri Balapan ATC di Sepang	Formula 1	Moto GP	X
5	Jakarta 10K Jadi Pemanasan Agus Prayogo Jelang SEA Games	Lainnya	Lainnya	V
6	Jelang Wimbledon, Djokovic Rebut Gelar Juara	Raket	Raket	V
7	Kala Dhean Fazrin Tiru Tendangan Maut Hwoarang	Lainnya	Lainnya	V
8	Maldini: Bonucci Pembelian Bagus untuk Milan, tapi…	Sepakbola	Sepakbola	V
9	McGregor Santai Tanggapi Prediksi Bakal Mati Lawan Mayweather	Lainnya	Lainnya	V
10	MU Harus Bahagia Dulu Sebelum Kejar Target Musim Depan	Sepakbola	Sepakbola	V
11	Pelajaran di Assen Jadi Modal Vinales Hadapi Paruh Kedua Musim	Moto GP	Moto GP	V
12	Praveen Debby Gagal Persembahkan Gelar Juara	Raket	Raket	V
13	Russell Westbrook Sabet Gelar MVP	Basket	Basket	V
14	Terkait Latihan di Lithuania, Timnas Basket Tunggu Surat Setneg	Lainnya	Basket	X
15	Warriors Juara, Messi dan Barca Beri Ucapan Selamat	Basket	Basket	V
16	Yamaha Yakin Vinales Akan Raih Sejumlah Titel MotoGP	Moto GP	Moto GP	V
17	Tiga Pebalap ABM Motorsport Naik Podium di ISSOM Seri Ketiga	Formula 1	Formula 1	V
18	Balasan Serena Setelah Disebut 'Hanya Peringkat 700 di Tenis Putra'	Raket	Raket	V

Dari ke-151 berita latih tersebut, akan dilakukan tahap *preprocessing*, dan perhitungan model probabilitas tiap kata x terhadap kategori y. Model probabilitas akan menjadi acuan berita uji untuk menemukan label kategori.

Selanjutnya melakukan pengujian terhadap sistem, pengujian dilakukan dengan menggunakan dokumen uji berita olahraga yang diambil dari situs *sport.detik.com* secara acak dengan jumlah 30 dokumen berita. 30 dokumen berita disimpan dengan file berformat txt. Dari 30 berita uji yang ada, *user* atau *tester* hanya memilih 18 berita secara acak.

Dari hasil uji coba yang dilakukan oleh *user*, didapatkan hasil bahwa terdapat 14 berita yang kategorinya dinyatakan relevan antara pendapat *user* atau *tester* dengan hasil dari sistem klasifikasi yang dapat dilihat pada tabel 2. Pada pengujian yang dilakukan dengan menggunakan 18 berita uji, dihasilkan keakuratan sebesar 77%, dengan *error rate* sebesar 23%.

$$\text{Akurasi} = \frac{14}{18}\ x\ 100\% = 77\%$$

5. KESIMPULAN DAN SARAN

Dari hasil penelitian yang telah dilakukan, dapat ditarik kesimpulan sebagai berikut :

Aplikasi klasifikasi berita olahraga menggunakan metode *Naïve Bayes* dengan *Enhanced Confix Stripping Stemmer* mampu mengklasifikasi berita olahraga dengan keakuratan sebesar 77% dengan tingkat kesalahan mencapai 23%.

Saran yang dapat peneliti sampaikan :

1. Diharapkan menggunakan metode *feature selection* agar token yang dihasilkan lebih teringkas.
2. Diharapkan menggunakan metode klasifikasi lainnya seperti *Support Vector Machine*, untuk mengetahui perbandingan keakuratannya.
3. Jumlah dokumen latih dan uji ditambah dan skenario pemilihan dokumen diperbanyak untuk melihat sejauh mana tingkat akurasi dari motode yang digunakan.

6. DAFTAR PUSTAKA

THOMAS M. CONNOLY., CAROLYN E. BEGG. 2015. *Database System : A Practical Approach to Design, Implementation, and Management*.

JIAWEI HAN., MICHELINE KAMBER. 2000. *Data Mining : Concepts and Techniques.*

TRIAWATI., CHANDRA. 2009. *Metode Pembobotan Statistical Concept Based untuk Klastering dan Kategorisasi Dokumen*

Berbahasa Indonesia. Institut Teknologi Telkom, Bandung.

I PUTU ADI DKK. 2008. *Penggunaan Algoritma Semut Dan Confix Stripping Stemmer Untuk Klasifikasi Dokumen Berita Berbahasa Indonesia.*

FRIEDMAN ET AL. 1997. *Bayesian Network Classifiers.*, Volume 29, Issue 2–3, pp 131–163.

ABRAHAM, R.., SIMHA, J.B. & IYENGAR, S. 2009. *Effective Discretization and Hybrid Feature Selection Using Naïve Bayesian Classifier For Medical Data Mining*. International Journal of Computational Intelligence Research 4

ZGAN & GAO. 2011. *An Improvement to Naive Bayes for Text Classification.*

MUKHERJEE & SHARMAA. 2012. *Intrusion Detection using Naive Bayes Classifier with Feature Reduction*. A Department of Computer Science, Banasthali University, Jaipur, Rajasthan, 304022, India.

EAMONN KEOGH, *Naïve Bayes Classifier.*

BAGUS SETYA RINTYARNA. 2016. *Pengaruh Seleksi Fitur Pada Skema Klasifikasi Naive Bayes Berbasis Gaussian dan Kernel Density*. Volume 01, Nomor 01.

DWI WIDIASTUTI. 2012. *Analisa Perbandingan Algoritma Svm, Naive Bayes, Dan Decision Tree Dalam Mengklasifikasikan Serangan (Attacks) Pada Sistem Pendeteksi Intrusi.* Universitas Gunadarma.

DENNY NATHANIEL CHANDRA., GEDE INDRAWAN., I NYOMAN SUKAJAYA. 2016. *Klasifikasi Berita Lokal Radar Malang Menggunakan Metode Naïve Bayes Dengan Fitur N-Gram*. Vol. 10, No. 1.

NOVIA BUSIARLI., LIAN AGA ADITYA., ALBERTUS YOKI ANDIKA. 2016. *Penerapan Algoritma Naïve Bayes & Natural Language Processing Untuk Mengklasifikasi Jenis Berita Pada Arsip Pemberitaan.*

DIO ARIADI., KARTIKA FITHRIASARI. 2015. *Klasifikasi Berita Indonesia Menggunakan Metode Naïve Bayesian Classification dan Support Vector Machine dengan Confix Stripping Stemmer*. Vol. 4. No.2.

ACMAD NURHADI. 2016. *Implementasi Algoritma Naïve Bayes Classifier Berbasis Particle Swarm Optimization (PSO) Untuk Klasifikasi Konten Berita Digital Bahasa Indonesia*. Volume 8 No 3.

YOSEPH SAMUEL., ROSA DELIMA., ANTONIUS RACHMAT. 2015. *Implementasi Metode K-Nearest Neighbor dengan Decision Rule untuk Klasifikasi Subtopik Berita.*

RONEN FELDMAN., JAMES SANGER. 2007. *The Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data.*

ELLY MUNINGSIH. 2016. *Klasifikasi Konten Berita Digital Bahasa Indonesia Menggunakan Support Vector Machines (SVM) Berbasis Particle Swarm Optimization (PSO).* AMIK BSI Yogyakarta.

MUHAMAD RACHLI. 2007. *Email Filtering Menggunakan Naïve Bayesian.* Institut Teknologi Bandung.

AGUS SETIAWAN., INDAH FITRI ASTUTI., AWANG HARSA KRIDALAKSANA. 2015. Klasifikasi dan Pencarian Buku Referensi Akademik Menggunakan Metode Naïve Bayes Classifier (NBC). Vol. 10 No. 1.

ANDITA DWIYOGA TAHITOE., DIANA PURWITASARI. 2010. *Implementasi Modifikasi Enhanced Confix Stripping Stemmer Untuk Bahasa Indonesia Dengan Metode Corpus Based Stemming.*

DIAH PUDI LANGGENI1., ZK. ABDURAHMAN BAIZAL., YANUAR FIRDAUS A.W. 2010. *Clustering Artikel Berita Berbahasa Indonesia Menggunakan Unsupervised Feature Selection.*

MCCALLUM., NIGAM. 1998. *A Comparison of Event Models for Naive Bayes Text Classification.*

MUHAMAD, HUSIN et al. Optimasi Naïve Bayes Classifier Dengan Menggunakan Particle Swarm Optimization Pada Data Iris. Jurnal Teknologi Informasi dan Ilmu Komputer, [S.l.], v. 4, n. 3, p. 180-184, sep. 2017. ISSN 2528-6579.