Jurnal Teknologia
Aliansi Perguruan Tinggi (APERTI) BUMN
Vol. 2, No. 1, Agustus 2019, ISSN 2654-5683

PERBANDINGAN PEMBOBOTAN UNTUK KLASIFIKASI TOPIK BERITA MENGGUNAKAN *DECISION TREE*

Henri Tantyoko[1], Adiwijaya[2], dan Untari Novia Wisesty[3]

[1]henritanyoko@students.telkomuniversity.ac.id

[2]adiwijaya@telkomuniversity.ac.id
[3]untarinw@telkomuniversity.ac.id

ABSTRACT

News is a media to add insight into the outside world, many events that can not be known directly, because it is news that can make it easier to find out more extensive information about the increase. News dissemination consists of online for internet and offline for print media. In the present era, the development of the internet is very fast, making it easier to access information, media delivery of news becomes varied with the internet. Many news available online cause problems because news published by publishers can make mistakes in categorizing news content into the right category. Need technical contributions to categorize news automatically. Categorization of the method used. In this study, the authors used the Decision Tree classification method. A process that is no less important before classification is the word weighting technique. To get optimal accuracy, the authors combine classification techniques using Decision Tree with word weighting techniques TF.ABS, TF.CHI2, TF.RF and TF.IDF. Receive TF.ABS which has the

Keywords: *Decision Tree, Term weighting, TF.ABS, TF.CHI2, TF.RF,TF.IDF*

ABSTRAK

Berita merupakan suatu media untuk menambah wawasan terhadap dunia luar, banyak kejadian yang tidak bisa diketahui secara langsung, oleh karena itu adanya berita dapat memudahkan untuk mengetahui informasi yang lebih luas cakupannya. Penyebaran berita terdiri dari online untuk internet dan offline untuk media cetak. Di zaman sekarang perkembangan internet sangat pesat sehingga memudahkan dalam hal mengakses informasi, media penyampaian berita menjadi variatif dengan adanya internet. Banyak berita yang tersedia secara online sehingga menimbulkan masalah karena berita yang diterbitkan oleh penerbit dapat membuat kesalahan dalam mengkategorikan isi berita kedalam kategori yang tepat. Perlu adanya teknik klasifikasi untuk mengkategorikan berita secara otomatis. Tantangan menggunakan teknik klasifikasi terletak pada akurasi kebenaran pengkategorian dari metode yang digunakan. Dalam penelitian ini penulis menggunakan metode klasifikasi Decision Tree. Proses yang tidak kalah pentingnya sebelum klasifikasi adalah teknik pembobotan kata. Untuk mendapatkan akurasi yang optimal, penulis mengkombinasikan teknik klasifikasi menggunakan Decision Tree dengan teknik pembobotan kata TF.ABS, TF.CHI2, TF.RF dan TF.IDF. Hasilnya TF.ABS memiliki akurasi yang paling tinggi yaitu 82,22% jika Tree tidak dibatasi parameter ketinggiannya.

Kata kunci: *Decision Tree, Pembobotan kata, TF.ABS, TF.CHI2, TF.RF,TF.IDF*

1. PENDAHULUAN

Berita adalah informasi yang dibuat untuk melaporkan kejadian atau peristiwa yang terjadi agar semua orang tahu kondisi selain di sekitarnya [1]. Perkembangan jaman sudah sangat maju, media penyampaian berita tidak hanya media cetak saja tetapi juga ada televisi, radio dan juga internet. Walaupun banyak media penyampaian berita tetapi media cetak tetap ada sampai saat ini karena jika dibandingkan dengan Internet, media cetak jauh lebih terbukti kebenarannya karena sumber beritanya jelas tidak seperti berita di Internet yang semua orang bisa menerbitkan berita kapan pun. Kelemahan media cetak ada di segi lingkungan. Media cetak tidak ramah lingkungan karena memakai kertas yang merupakan hasil pengolahan kayu untuk mencetak berita. Jika dibandingkan dengan media elektronik seperti radio dan televisi mempunyai keunggulan dapat memvisualisasikan berita sehingga penikmat berita lebih memahami isi berita daripada media cetak yang hanya berupa tulisan tapi kekurangannya adalah penayangan beritanya hanya sekali tidak bisa diulang.

Perkembangan internet di Indonesia sangat pesat, berdasarkan survei yang dilakukan oleh Asosiasi Penyedia Jasa Internet Indonesia (APJII) pengguna internet di Indonesia terus meningkat dari tahun 1998 sampai 2017. Di tahun 2017 menurut internetworldstat pengguna internet Indonesia mencapai 143,26 juta pengguna, meningkat 7% dari tahun 2016 yang penggunanya terdapat 132,7 juta pengguna. Indonesia merupakan negara dengan pengguna internet terbesar kelima di dunia setelah China, India, United States, Brazil, oleh karena itu media penyampaian berita melalui internet berpotensi besar menguasai pasar. Banyak berita yang diterbitkan dari internet karena target pembacanya lebih tinggi daripada melalui media cetak. Penerbit berlomba lomba untuk menghasilkan berita yang terbaru, hal tersebut dapat menimbulkan masalah dalam mengkategorikan berita. Contohnya saya ingin membaca berita tentang olahraga, tetapi yang muncul pada kategori olahraga ternyata berita politik. Untuk mengurangi kesalahan pengkategorian perlu adanya teknik klasifikasi yang dapat mengkategorikan berita secara otomatis.

Klasifikasi merupakan bagian terpenting dalam data mining. Klasifikasi merupakan cara teknik untuk mempelajari kumpulan data yang banyak sehingga menghasilkan aturan untuk mengenali data baru yang belum pernah dipelajari [2]. Keluaran dari teknik klasifikasi adalah hasil akurasi yang menunjukkan seberapa benar teknik klasifikasi yang diterapkan dalam memprediksi. Semakin tinggi akurasi maka semakin bagus karena banyak data yang diklasifikasikan benar. Salah satu cara untuk meningkatkan akurasi yaitu dengan cara teknik pembobotan setiap kata. Pembobotan kata merupakan masalah dasar dalam klasifikasi teks dan langsung

berdampak terhadap klasifikasi akurasi, Pembobotan yang paling populer dan menjadi teknik yang banyak digunakan dalam memberi bobot setiap kata saat ini adalah TF.IDF (Term Frequency Inverse Document Frequency), tapi sekarang teknik pembobotan TF.IDF kurang efektif untuk memberikan bobot sehingga muncul penelitian lain untuk menemukan teknik pembobotan yang efektif dikombinasikan dengan teknik klasifikasi sehingga akan menghasilkan akurasi yang optimal [3].

Teknik klasifikasi memiliki beberapa macam metode klasifikasi. Dalam penelitian lainnya dengan studi kasus teks Arab menggunakan *Decision tree* diperoleh akurasi 93% untuk *scentific corpus* [9]. Sedangkan klasifikasi teks berita Indonesia bisa menggunakan beberapa macam metode untuk memprediksi hasil dari kategori yang akan diprediksi. Menurut Rini Wongso dkk dari perbandingan berbagai macam metode klasifikasi *seperti Multinomial Naïve Bayes, Multivariate Bernoulli Naïve Bayes*, dan *Support Vector Machine* dikombinasikan dengan algoritma pembobotan TF.IDF dan SVD diperoleh *Multinomial Naive Bayes* dengan pembobotan TF.IDF yang menghasilkan akurasi yang tinggi sebesar 85% [10]. Penelitian lainnya dilakukan Reynaldi Ananda Pane dkk terhadap topik Quran dengan versi Bahasa Inggris menggunakan klasifikasi *Multinomial Naive Bayes* yang dikombinasikan dengan teknik pembobotan *bag of words* yang hanya memeperhatikan setiap kata yang muncul dalam dokumen, hasilnya *hamming loss* 0.1247 atau akurasi sebesar 0.8753% [11].

Berdasarkan penelitian sebelumnya dengan membandingkan pembobotan *Term Frequency Absolute* (TF.ABS) dengan *Term Frequency Chi Square* (TF.CHI2) menghasilkan akurasi yang tidak jauh berbeda yaitu 95,74% untuk TF.ABS dan akurasi sebesar 95,87% untuk TF.CHI2 [4]. Penelitian lain ada yang membandingkan antara TF.CHI2, *Term Frequency Inverse Document Frequency* (TF.IDF) dan *Term Frequency Relevance Frequency* (TF.RF) diperoleh hasil TF.CHI2 menghasilkan akurasi 89,71%, TF.RF dengan akurasi 88,07% dan TF.IDF hanya mendapat akurasi sebesar 74,93% [5]. Penelitian dari Rifqi Abdul Aziz dkk mengklasifikasikan topik pada lirik lagu menggunakan *Multinomial Naive Bayes* dengan seleksi fitur *chi-square* dan dengan pembobotan *bag of words* memperoleh akurasi 96% [6]. Penelitian yang dilakukan oleh Matsunaga dan Ebecken yang fokusnya untuk menemukan teknik pembobotan mana yang akan menghasilkan akurasi yang optimal. Hasilnya TF.ABS (T*erm Frequency Absolute*) memiliki akurasi 72% yang lebih tinggi daripada TF.IDF,TF.IG,TF.GR, TF.BNS, TF.OR [7]. Dibalik kepopuleran pembobotan TF.IDF ternyata ada yang lebih bagus lagi dari pembobotan TF.IDF yaitu TF.ABS dan TF.CHI, menurut penelitian yang dilakukan Man Lan dkk TF.RF (*Term Frequency Relevance Frequency*) lebih baik daripada TF.IDF dengan

akurasi sebesar 78%. TF.IDF merupakan metode yang *supervised* dimana pembobotannya dilakukan melalui pengawasan oleh sebab itu TF.IDF lebih terfokus ke data latih sedangkan TF.RF merupakan *unspervised*, tidak ada pengawasan terhadap data yang diproses oleh karena itu lebih *powerfull* untuk data uji dan tidak mengalami *over fitting*, TF.RF lebih baik daripada TF.IDF [8].

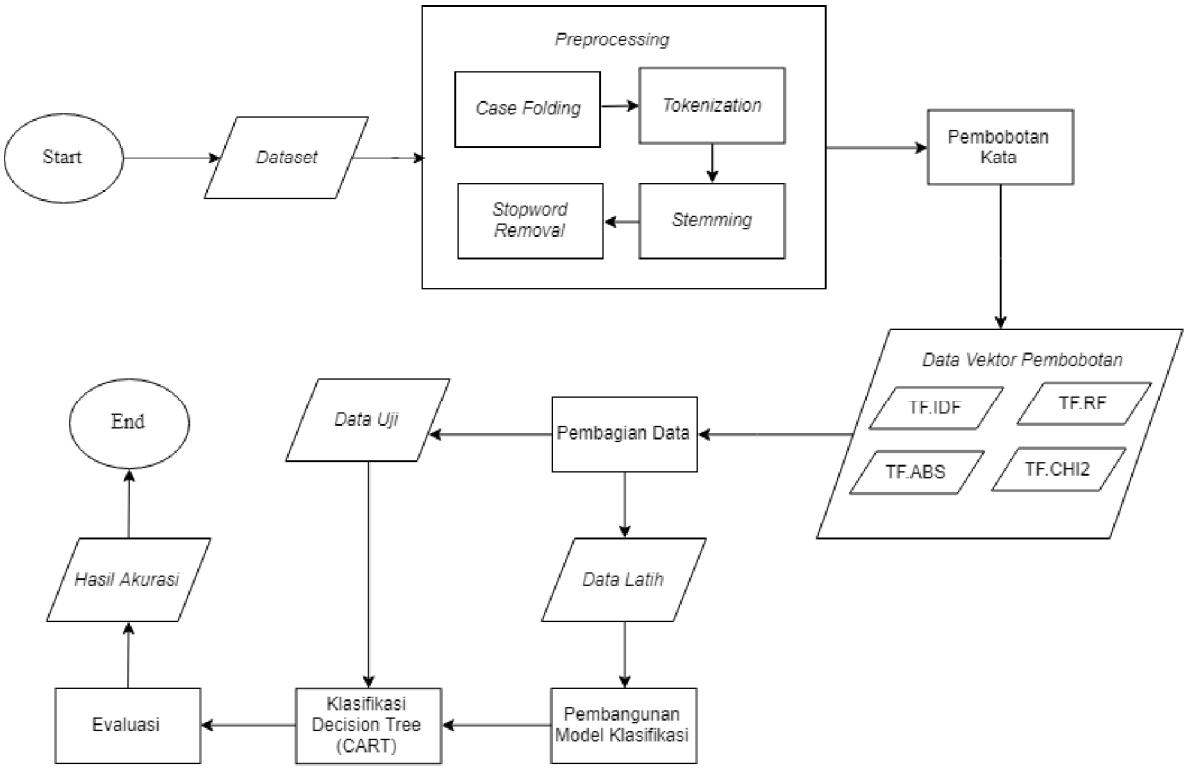
Dari penelitian sebelumnya, penulis mengambil parameter TF.ABS, TF.CHI2, TF.RF, TF.IDF untuk mengetahui akurasi mana yang akan menghasilkan akurasi yang optimal jika dikombinasikan dengan teknik klasifikasi *Decision Tree*. Topik yang diangkat dalam penelitian ini adalah membandingkan beberapa teknik pembobotan dengan mengkombinasikan teknik klasifikasi *Decision Tree* untuk mengklasifikasikan teks berita berdasarkan kategorinya. Berdasakan topik di atas, terdapat beberapa batasan masalah yaitu:

1. Kategori berita terdiri dari 12 kelas yaitu Ekonomi, Hiburan, Hukum, Kesehatan, Gaya Hidup, Otomotif, Pendidikan, Politik, Sosial Budaya, Olahraga, Teknologi Wisata.
2. Jumlah dokumen *dataset* sebanyak 360 dokumen.
3. Fitur menggunakan *unigram*.

Tujuan dari penelitian ini untuk mengetahui teknik pembobotan apa yang menghasilkan akurasi optimal menggunakan klasifikasi *Decision Tree*.

2. DASAR TEORI /MATERIAL DAN METODOLOGI/PERANCANGAN

Proses untuk memprediksi berita yang belum diketahui kategorinya mulai dari pengambilan data, *preprocessing*, proses, dan evaluasi terhadap hasil. Diagram alur sistem dapat dilihat pada gambar 1.



Gambar 6. Gambaran Umum Sistem

2.1 Dataset

Dataset yang diperoleh untuk penelitian penulis adalah artikel berita bahasa Indonesia yang terdiri dari 12 kategori berita seperti politik, budaya, kesehatan, pendidikan, dan lain-lain yang terdiri 360 dokumen. Data ini diambil dari beberapa portal berita seperti *website* kompas.com, tribunnews.com, republika.com, sindonews.com, mediaindonesia.com dari Agustus 2016 – Februari 2017. *Dataset* ini digunakan untuk membangun sistem klasifikasi berita menggunakan *Decision Tree* dengan mengkombinasikan pembobotan menggunakan empat parameter yaitu TF.CHI, TF.IDF, TF.RF dan TF.ABS berdasarkan kategori berita. Tabel kumpulan data set yang diambil untuk diteliti dapat dilihat padat tabel 1.

Tabel 3. Pembagian Dataset

No	Kategori Berita	Jumlah Dokumen	No	Kategori Berita	Jumlah Dokumen
1	Ekonomi	30	7	Otomotif	30
2	Gaya Hidup	30	8	Pendidikan	30
3	Kesehatan	30	9	Politik	30
4	Hiburan	30	10	Budaya	30
5	Hukum	30	11	Teknologi	30
6	Olahraga	30	12	Wisata	30

Data set tersebut diolah sudah diambil dan disimpan dalam bentuk file berekstensi *.txt* sebanyak 360 file dengan 12 kategori.

2.2 Preprocessing

Preprocessing adalah tahap untuk mempersiapkan data sehingga siap untuk diolah, proses ini dilakukan setelah melakukan *load dataset* dengan tujuan untuk mengurangi kata-kata yang dianggap tidak penting dan simbol-simbol yang tidak memiliki makna sehingga saat proses klasifikasi, metode yang digunakan dapat memprediksi teks kedalam kategori yang tepat. *Preprocessing* terdiri dari *case folding* sebagai langkah awal untuk mengubah kata menjadi huruf kecil semua dan menghilangkan huruf selain a sampai huruf z, tahap yang kedua yaitu *tokenization* untuk memotong kalimat pada suatu dokumen menjadi levelnya per kata yang akan digunakan sebagai fitur untuk perhitungan pada saat proses klasifikasi, tahap ketiga adalah *stemming*. Tahap ini bertugas untuk menjadikan suatu kata sebagai kata dasar, dan tahap terakhir adalah *stopword removal* yang berfungsi untuk menghilangkan kata-kata yang tidak penting seperti kata sandang dan kata hubung [12][13], *stopword* yang digunakan adalah stopword bahasa Indonesia yang disusun oleh Tala [14]. Adapun contoh tahapan *preprocessing* yang terdiri dari *input* dan *output* yang dapat dilihat pada tabel 2 agar memudahkan dalam pemahaman.

Tabel 4. *Input* dan *Output Preprocessing*

Preprocessing	Input	Output
Case Folding	Menpora Imam Nahrawi memberikan penghargaan kepada Timnas Indonesia yang telah berjuang di final Sea Games tahun ini	menpora imam nahrawi memberikan penghargaan kepada timnas indonesia yang telah berjuang di final sea games tahun ini
Tokenization	menpora imam nahrawi memberikan penghargaan kepada timnas indonesia yang telah berjuang di final sea games tahun ini	menpora, imam, nahrawi, memberikan, penghargaan, kepada, timnas, indonesia, yang, telah, berjuang, di, final, sea, games, tahun, ini
Stemming	menpora, imam, nahrawi, memberikan, penghargaan, kepada, timnas, Indonesia, yang, telah, berjuang, di, final, sea, games, tahun, ini	menpora, imam, nahrawi, beri, harga, kepada, timnas, indonesia, yang, telah, juang, di, final, sea, games, tahun, ini
Stopword Removal	menpora, imam, nahrawi, beri, harga, kepada, timnas, indonesia, yang, telah, juang, di, final, sea, games, tahun, ini	menpora, imam, nahrawi, beri, harga, timnas, indonesia, juang, final, sea, games, tahun

2.3 Pembobotan

Pembobotan kata adalah teknik untuk memberikan nilai yang bertujuan untuk mengetahui seberapa penting makna suatu kata yang merujuk ke kategori berita yang sudah ditentukan, teknik klasifikasi akan meningkat jika dikombinasikan dengan pembobotan yang tepat. Untuk mengetahui teknik pembobotan yang paling baik adalah dengan membandingkan antara pembobotan TF.IDF, TF.RF,TF.CHI^2, dan TF.ABS. Berikut adalah cara menghitung pembobotan untuk keempat teknik pembobotan untuk mempermudah perhitungan pembobotan, berikut adalah tabel distribusi untuk *term* t_j dan c_i.

Tabel 5. *Contingency* tabel untuk *category* dan *term*

	c_i	$c_{\sim i}$	Total
t_j	n_{ij}	$n_{\sim ij}$	n_j
$t_{\sim j}$	$n_{i\sim j}$	$n_{\sim i\sim j}$	$n_{\sim j}$
Total	n_i	$n_{\sim i}$	n

Keterangan variabel:

n_{ij} : Jumlah dokumen dalam kategori c_j yang mengandung term t_j

$n_{\sim ij}$: Jumlah dokumen tidak dalam kategori c_i yang mengandung term t_j

$n_{i\sim j}$: Jumlah dokumen dalam kategori c_i yang tidak mengandung term t_j

$n_{\sim i\sim j}$: Jumlah dokumen tidak dalam kategori c_i yang tidak mengandung term t_j

n_j : Jumlah dokumen dengan term t_j

$n_{\sim j}$: Jumlah dokumen dengan term selain $t_{\sim j}$
n_i : Jumlah dokumen dengan kategori c_j
$n_{\sim i}$: Jumlah dokumen tanpa kategori c_j
n : Total atau jumlah dari dokumen
t_j : Merupakan term t_j
c_i : Merupakan kategori c_j
$t_{\sim j}$: Merupakan term selain term t_j
$c_{\sim j}$: Merupakan kategori selain c_j.

2.4 *Term Frequency* (TF)

TF adalah salah satu metode pembobotan term yang paling sederhana TF disebut juga dengan *bag of words* karena. Pada metode ini, setiap term diasumsikan memiliki proporsi kepentingan sesuai dengan jumlah terjadinya (munculnya) *term* tersebut dalam dokumen. Dengan metode ini, nilai kontribusi (bobot) suatu *term* pada suatu dokumen adalah sama dengan jumlah munculnya *term* tersebut pada dokumen, pemberian bobot menggunakan metode ini tidak efektif jika mempertimbangkan kata yang banyak muncul yang memiliki bobot paling besar. Rumus TF dapat dilihat pada persamaan 1.

$$W(d,t) = TF(\mathrm{d},\mathrm{t}) \tag{1}$$

Dimana $TF(\mathrm{d},\mathrm{t})$ adalah frekuensi kemunculan *term* t pada dokumen d.

2.5 *Inverse Document Frequency* (IDF)

Pada metode ini, term yang dianggap bernilai adalah term yang jarang muncul pada koleksi/kumpulan dokumen. Selain itu, tingkat kepentingan nilai (bobot) dari suatu term juga diasumsikan berbanding terbalik dengan jumlah dokumen yang mengandung term tersebut . Oleh karena itu, bila suatu term banyak muncul di kumpulan dokumen, maka term tersebut akan dianggap tidak bernilai/berharga. Rumus IDF dapat dilihat pada persamaan 2.

$$IDF(t) = log\frac{\mathrm{n}}{\mathrm{df(t)}} \tag{2}$$

Dimana df(t) adalah banyak dokumen yang mengandung term t.

2.6 *Term Frequency – Inverse Document Frequency* (TF.IDF)

TF.IDF adalah pembobotan yang menggunakan skema dengan cara menghitung jumlah term pada suatu dokumen dan jumlah dokumen yang mengandung kata tersebut. Rumus untuk menghitung bobot menggunakan TF.IDF bisa dilihat di persamaan 3.

$$TF.IDF = \mathrm{TF(d,t)} * IDF(t) \tag{3}$$

2.7 *Term Frequency - Relevance Frequency* (TF.RF)

Relevance frequency merupakan metode yang muncul sebagai upaya perbaikan terhadap metode-metode yang sudah ada. Sebagai contoh, metode IDF hanya akan menilai term berdasarkan kemunculan (ada atau tidaknya saja) term pada suatu dokumen. Berbeda dengan metode RF yang diusulkan oleh Man Lan, metode ini mempertimbangkan relevansi dokumen dilihat dari frekuensi kemunculan term di kategori yang berkaitan.. Persamaan untuk menghitung RF dapat dilihat di persamaan 4.

$$RF(t_j, c_i) = \log\left(2 + \frac{n_{ij}}{\max(1, n_{\sim ij})}\right) \tag{4}$$

Setelah menghitung RF untuk selanjutnya melakukan perhitungan TF.RF dengan cara melakukan perkalian matriks TF dengan RF sesuai persamaan 5.

$$TF.RF = \mathrm{TF}(d,t) * RF(t_j, c_i) \tag{5}$$

2.8 *Term Frequency-Chi-Square* (TF.CHI^2)

Chi-Square merupakan teknik menghitung jumlah term yang muncul pada setiap dokumen TF dan mempertimbangkan bobot untuk term yang tidak muncul didalam dokumen dan term yang muncul didalam dokumen. Sama seperti teknik pembobotan lainnya, hasil pembobotan TF.CHI^2 digunakan untuk membandingkan dengan teknik pembobotan lainnya. Setelah diketahui nilai pembobotan *Chi-Square* langkah selanjutnya akan dikombinasikan dengan metode klasifikasi untuk mendapatkan hasil yang maksimal dengan akurasi yang tinggi. Persamaan rumus menghitung *Chi-Square* dapat dilihat di persamaan 6.

$$CHI^2(t_j, c_i) = \frac{n(n_{ij}n_{\sim i\sim j} - n_{\sim ij}n_{i\sim j})}{(n_i n_j n_{\sim i} n_{\sim j})} \tag{6}$$

Setelah menghitung CHI^2 untuk selanjutnya melakukan perhitungan TF. CHI^2 dengan cara melakukan perkalian matriks TF dengan CHI^2 sesuai persamaan 7.

$$TF.CHI^2 = TF(d,t) * CHI^2\left(t_j, c_i\right) \quad (7)$$

2.9 *Term Frequency Absolute (TF.ABS)*

Pembobotan TF.ABS menggunakan teknik menghitung jumlah term yang muncul pada setiap dokumen dan mengukur kemungkinan suatu term yang ada dalam dokumen dengan kategori dibagi dengan kemungkinan term yang tidak ada dalam dokumen dengan kategori tersebut. Perhitungan ABS dilihat pada persamaan 8.

$$ABS\left(t_j, c_i\right) = |\ln(\frac{(n_{ij}+0.5)(n_{\sim i\sim j}+0.5)}{(n_{\sim ij}+0.5)(n_{i\sim j}+0.5)})| \quad (8)$$

Setelah menghitung ABS untuk selanjutnya melakukan perhitungan TF. ABS dengan cara melakukan perkalian matriks TF dengan ABS sesuai persamaan 9.

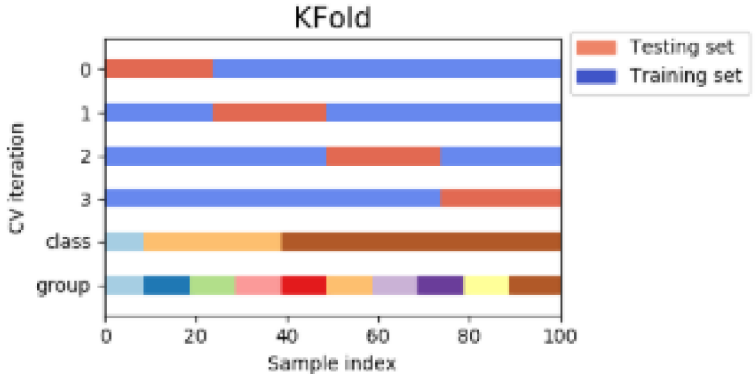
$$TF.ABS\ = TF(d,t) *\ ABS\left(t_j, c_i\right) \quad (9)$$

2.10 *K-fold Cross Validation*

Pada tahap ini penulis membagi *dataset* menjadi data latih dan data uji berdasarkan *k-fold cross validation* berdasarkan nilai k sebesar 10. Artinya *dataset* akan dibagi menjadi 10 partisi secara acak. Data yang sudah menjadi *testing* pada k sebelumnya tidak bisa menjadi *testing* lagi [2], Visualisasi dapat dilihat pada gambar 2.

2.11 Klasifikasi *Decision Tree*

Pada tahap klasifikasi *decision tree*, menggunakan *library python decision tree*



Gambar 2. Visualisasi *k-fold cross validation.*

dari *sklearn* yang menggunakan metode Classification and Regression Tree (CART) dan mendukung inputan berupa numerik maupun kategorik. Langkah pertama adalah menentukan fitur mana yang akan menjadi *root* dengan cara membagi data setiap fitur menjadi dua berdasarkan nilai tengah (median). Contoh kata "belajar" muncul sebanyak 6 kali pada dokumen pertama, 5 kali pada dokumen kedua, dan 3 kali pada dokumen ketiga. Maka dari itu nilai 3 digunakan untuk memisahkan data pada fitur tersebut yaitu =< 3 dan > 3. Setelah itu hitung nilai *Gini Split Index* pada bagian kiri dan kanan pemisah yang dapat dilihat persamaan 10.

$$IGini(t_{ji}) = 1 - \sum_{i=1}^{n} p_{ji}^2 \quad (10)$$

Dimana t_{ji} adalah banyaknya term t_j yang berada pada kelas ke i pada pemisah tertentu, p adalah probabilitas banyaknya term t_j berdasarkan kelas ke i. Jadi masing-masing fitur memiliki nilai dua gini, yang pertama gini sebelah kanan pemisah dan yang kedua adalah sebelah kiri dari pemisah. Kemudian jumlah masing-masing gini dengan menambahkan bobot seperti persamaan 11.

$$Gini(Target, t_{ji}) = \sum_{r=1}^{n} Target_r \,.\, Gini(t_{ji}) \quad (11)$$

Dimana Target adalah peluang pemisah batas kiri dan kanan. Hasilnya akan diperoleh *Gini* untuk setiap fitur. Pemilihan fitur yang digunakan untuk menjadi simpul adalah *Gini* dengan nilai yang rendah karena mengandung *impurity* yang

rendah. *Impurity* rendah artinya keberagamannya rendah sehingga lebih cepat menjadikannya daun yang artinya mengandung satu kelas atau kategori.

2.12 Pengukuran Akurasi

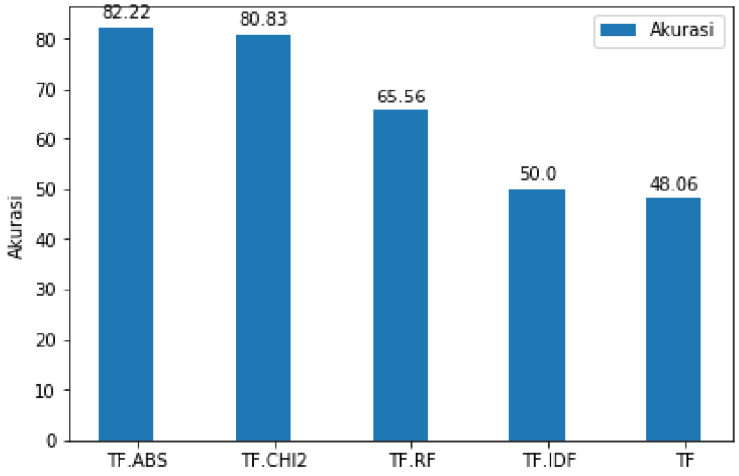
Akurasi merupakan keluaran dari hasil klasifikasi. Metode klasifikasi dikatakan baik jika memiliki akurasi yang tinggi karena banyaknya dokumen yang diprediksi benar ada banyak [15]. Banyak penelitian-penelitian yang berusaha meningkatkan akurasi dengan memilih metode-metode yang tepat digunakan oleh *dataset* acuan. Cara menghitung akurasi dapat dilihat pada persamaan 12.

$$Akurasi = \frac{Banyaknya\ dokumen\ yang\ diprediksi\ benar}{Total\ Dokumen} \quad (12)$$

3. HASIL DAN PEMBAHASAN

3.1 Hasil Pengujian dan Analisis Skenario Pertama

Skenario pengujian ini bertujuan untuk mengetahui teknik pembobotan yang paling baik diantara TF.ABS, TF.CHI2, TF.RF, TF.IDF dan TF. Penelitian ini menggunakan metode *unigram* dengan menggunakan teknik klasifikasi *Decision Tree*. Hasil pengujian skenario dapat dilihat pada gambar 3.



Gambar 3. Hasil perbandingan pembobotan dengan klasifikasi *decision tree*

Berdasarkan pengujian pada skenario pertama diperoleh hasil bahwa TF.ABS menghasilkan akurasi yang paling tinggi sebesar 82,22% berselisih sedikit jika dibandingkan dengan TF.CHI^2 yang akurasinya 80.83, karena jika dilihat dari kedua rumusnya sama-sama mempertimbangkan kemunculan kata yang muncul maupun

tidak muncul pada kategori tersebut dan selain kategori tersebut, sehingga bobot yang dihasilkan oleh kedua teknik ini lebih sesuai, contohnya ketika kata "belajar" sering muncul pada dokumen 1 yang kategorinya A, kalau tidak mempertimbangkan aspek kemunculan kata yang muncul maupun tidak muncul pada kategori c_i dan mempertimbangkan kemunculan kata maupun tidak pada kategori selain kategori c_i maka bobot belajar pada kategori A akan bernilai tinggi padahal kata belajar belum tentu bobotnya tinggi di dokumen lainnya.

TF.RF, TF.IDF, dan TF akurasinya dibawah TF.ABS dan TF.CHI^2 karena teknik-teknik tersebut tidak mempertimbangkan kemunculan kata selain kata t_j dan kemunculan kata selain di kategori c_i di dokumen, TF.RF dan TF.IDF hanya fokus pada kemunculan kata tertentu saja tanpa mempertimbangkan ketidakmunculan kata tersebut, TF.RF mempertimbangkan kemunculan kata berdasarkan kategori tertentu tidak seperti TF.IDF yang tanpa melihat dokumen tersebut kategorinya apa sehingga nilai akurasi untuk TF.RF jauh lebih tinggi yaitu 65,56% daripada TF.IDF yang hanya sebesar 50%.

TF artinya perhitungan klasifikasi hanya bergantung pada banyaknya kata yang muncul pada setiap dokumen saja tanpa dimodifikasi, jadi jika pada dokumen 2 mengandung kata "makan" sebanyak 20 kali maka kata tersebut memiliki nilai bobot yang besar pada kategori B, teknik ini akan membuat akurasi yang rendah karena belum tentu kata makan merupakan kategori B karena bisa saja kata "makan" banyak muncul di banyak dokumen juga. Akurasi dari TF ini paling rendah dari parameter yang diteliti yaitu 48,06% .

Semakin tinggi akurasi dari percobaan semakin sedikit *missclassification* yang dihasilkan. *Missclassification* pada *dataset* yang diperoleh terjadi karena dua faktor, yang pertama adanya keambiguan isi berita yang hampir sama dengan kategori berita yang berbeda contohnya sering terjadinya *missclassification* kategori berita hiburan dengan wisata, politik dengan hukum, budaya dengan wisata, faktor kedua mempunyai kesinambungan dengan faktor pertama yaitu dalam pembuatan pohon keputusan. Pemilihan fitur yang akan dijadikan simpul pohon sangat mempengaruhi kualitas pohon untuk membuat keputusan akhir berupa daun yang merupakan kategori berita.

3.2 Hasil Pengujian dan Analisis Skenario Kedua

Pada *Decision Tree*, pengklasifikasian dilakukan dengan cara membangun pohon keputusan, terdapat *Root* sebagai simpul awal, cabang pohon, dan daun. Ketiga komponen tersebut dibangun berdasarkan aturan dari algoritma CART *Decision Tree*. Semakin tinggi pohonnya semakin kompleks keputusan yang dibuat Skenario

kedua bertujuan untuk mengetahui pengaruh tinggi pohon terhadap akurasi masing-masing teknik pembobotan. Dapat dilihat pada tabel 4 dan 5.

Tabel 4 .Perbandingan teknik pembobotan setelah dibatasi ketinggiannya sebesar 20%

No	Pembobotan	Ketinggian Pohon	*Resize* Ketinggian (%)	Batas Ketinggian Pohon	Akurasi tanpa Batasan ketinggian(%)	Akurasi dengan batasan (%)
1	TF.ABS	22	20	18	82.22	77.5
2	TF.CHI2	19	20	15	80.83	80.55
3	TF.RF	31	20	25	65.56	65.28
4	TF.IDF	38	20	33	50.00	52.5
5	TF	34	20	27	48.06	49.72

Tabel 5. Perbandingan pembobotan setelah dibatasi ketinggiannya sebesar 40%

No	Pembobotan	Ketinggian Pohon	*Resize* Ketinggian (%)	Batas Ketinggian Pohon	Akurasi tanpa Batasan ketinggian(%)	Akurasi dengan batasan (%)
1	TF.ABS	22	40	13	82.22	69.45
2	$TF.CHI^2$	19	40	11	80.83	77.78
3	TF.RF	31	40	19	65.56	58.61
4	TF.IDF	38	40	25	50.00	50.56
5	TF	34	40	20	48.06	48.89

Dilihat dari hasilnya menunjukkan bahwa akurasi setelah adanya pembatasan ketinggian pohon membuat TF.ABS, TF.CHI2, dan TF.RF menjadi turun akurasinya karena ketinggian pohon relatif lebih kecil daripada TF.IDF dan TF, jika ketinggian pohon pendek maka kemampuan untuk membuat simpul menjadi daun akan lebih cepat dibandingkan ketinggian pohon yang tinggi dan jika ketinggian pohon yang pendek dibatasi pertumbuhannya maka akan ada informasi yang gagal di dapatkan untuk memprediksi. Lain halnya dengan ketinggian pohon yang tinggi. Semakin tinggi semakin banyak simpul yang terbentuk maka semakin banyak pohon keputusan memproses simpul menjadi daun untuk menjadikan kelas yang murni. Untuk pembatasan ketinggian pohon dilakukan dengan cara memangkas pohon yang awalnya tanpa batasan kemudian dipangkas sesuai skenario yaitu 20% dan 40%. Pemangkasan 20% dan 40% membuat TF.CHI2 memiliki akurasi tertinggi yaitu 80,55% dan 77,78% dan TF yang terendah dengan 49,72% dan 45,89%. Sesuai dengan persamaan dari CHI2, nilai dari masing-masing parameter diperhitungkan

yaitu muncul dan ketidakmunculan suatu kata pada suatu kategori dan selain kategori tersebut sehingga diperoleh nilai bobot yang tepat, saat terjadi pemangkasan pohon tidak akan terlalu berpengaruh karena pembobotan yang dihasilkan merepresentasikan sesuai kategori yang tepat, berbeda dengan teknik pembobotan yang lainnya yang kurang memperhatikan parameter yang seharusnya diperhitungkan. Dengan adanya pembatasan ketinggian pohon maka pohon yang sudah jadi akan dipangkas sesuai ketinggian yang diinginkan, simpul dipaksa menjadi daun dan jumlah kategori yang paling banyak dijadikan kelas pada daun tersebut. Pembatasan ketinggian yang efektif akan meningkatkan akurasi teknik pembobotan jika ketinggian pohon terlalu tinggi. Contohnya pada teknik pembobotan TF.IDF dan TF yang menghasilkan ketinggian pohon yang tinggi membuat akurasi naik jika dibandingkan tanpa membatasi ketinggian pohon karena akan memangkas pohon yang informasinya sebenarnya sudah didapatkan pada ketinggian pohon sebelumnya.

4. KESIMPULAN DAN SARAN

Dari pengujian beberapa skenario dalam penelitian ini membandingkan pembobotan mana yang menghasilkan akurasi yang tinggi jika dikombinasikan dengan teknik klasifikasi *Decision Tree* hasilnya jika tidak menggunakan teknik pembobotan hanya mendapatkan akurasi sebesar 48,06% jauh berbeda signifikan dengan TF.ABS yang merupakan teknik pembobotan terbesar dari parameter yang diambil, nilai akurasinya sebesar 82,22% . TF.ABS juga mengalahkan teknik pembobotan yang paling populer yaitu TF.IDF yang hanya sebesar 50% akurasinya.

Decision Tree merupakan klasifikasi dengan membuat suatu pohon keputusan, semakin tinggi pohon terebut semakin banyak simpul yang dihasilkan, untuk ketinggian pohon yang dibatasi sebanyak 20% dan 40% dari ketinggian yang dibentuk pohon untuk setiap pembobotan. Dan menempatkan TF.CHI2 sebagai pembobotan dengan akurasi terbaik dengan nilai akurasi 80,55% untuk pembatasan 20% dan 77,78% untuk pembatasan 40%. Untuk pembobotan TF.IDF dan TF mengalami kenaikan akurasi jika ketinggian dibatasi 20% maupun 40% karena ketinggian pohon TF.IDF dan TF jauh lebih tinggi dari ketiga parameter lainnya, artinya jika ketinggian dari pohon itu dikategorikan tinggi maka perlu diterapkan pembatasan ketinggian agar teknik klasifikasi ini tidak terlalu banyak memproses terlalu banyak dan membuat banyak simpul yang sebenarnya pada ketinggian yang tepat dapat dibatasi sehingga simpul terakhir dapat merepresentasikan simpul yang hilang karena pembatasan.

Saran untuk penelitian selanjutnya adalah mencoba membandingkan juga seleksi fitur untuk klasifikasi karena dari fitur yang banyak tentunya ada fitur-fitur

yang dianggap tidak penting dan hanya membuat akurasi menjadi rendah karena pembagi fiturnya tinggi. Bandingkan seleksi fitur mana yang cocok untuk klasifikasi *Decision Tree* agar pemrosesan jauh lebih cepat dan akurasi yang dihasilkan bisa tinggi dan juga membuat mengkategorikan dengan *multilable* untuk mengurangi keambiguan isi berita yang memiliki kategori yang karakteristiknya hampir sama.

UCAPAN TERIMAKASIH

Dalam pengerjaan jurnal ini penulis mendapatkan begitu banyak bantuan.Pada kesempatan kali ini penulis ingin menyampaikan terima kasih yang sebesar-besarnya kepada Telkom University yang senantiasa memberikan dukungan untuk publikasi jurnal penulis.

DAFTAR PUSTAKA

[1] I. M. R. Prawira, Adiwijaya and M. S. Mubarok, "Klasifikasi Multi-Label Pada Topik Berita Berbahasa Indonesia Menggunakan Multinomial Naïve Bayes," vol. 5, no. 3, pp. 7774–7781, 2018.

[2] A. F. Irene, and Adiwijaya "Klasifikasi Sentimen Review Film Menggunakan Algoritma Support Vector Machine Sentiment Classification of Movie Reviews Using Algorithm Support Vector Machine," vol. 4, no. 3, pp. 4740–4750, 2017.

[3] K. Chen, Z. Zhang, J. Long, and H. Zhang, "Turning from TF-IDF to TF-IGM for term weighting in text classification," *Expert Syst. Appl.*, vol. 66, pp. 1339–1351, 2016.

[4] M. A. Kurniawan, Y. Sibaroni, and K. L. Muslim, "Kategorisasi Berita Menggunakan Metode Pembobotan TF.ABS dan TF.CHI," *Indones. J. Comput.*, vol. 3, no. 2, p. 83, 2018.

[5] P. N. Bandung, "ANALISIS AKURASI METODE TERM WEIGHTING INDONESIA DENGAN K-NEAREST NEIGHBOR."

[6] R. Abdul Aziz and M. Syahrul Mubarok, "Klasifikasi Topik pada Lirik Lagu dengan Metode Multinomial Naïve Bayes," *Indosc 2016*, no. 1, pp. 139–148, 2016.

[7] L. A. Matsunaga and N. F. F. Ebecken, "Term weighting approaches for text categorization improving," *Proc. - 8th Int. Conf. Intell. Syst. Des. Appl. ISDA 2008*, vol. 1, pp. 409–414, 2008.

[8] Man Lan, Chew Lim Tan, Jian Su, and Yue Lu, "Supervised and Traditional Term Weighting Methods for Automatic Text Categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 721–735, 2009.

[9] F. Harrag, E. El-Qawasmeh, and P. Pichappan, "Improving Arabic text

categorization using decision trees," *2009 1st Int. Conf. Networked Digit. Technol. NDT 2009*, no. August 2009, pp. 110–115, 2009.

[10] R. Wongso, F. A. Luwinda, B. C. Trisnajaya, O. Rusli, and Rudy, "News Article Text Classification in Indonesian Language," *Procedia Comput. Sci.*, vol. 116, pp. 137–143, 2017.

[11] R. A. Pane, M. S. Mubarok, N. S. Huda, and Adiwijaya, "A multi-lable classification on topics of Quranic verses in English translation using multinomial naive bayes," *2018 6th Int. Conf. Inf. Commun. Technol. ICoICT 2018*, no. May, pp. 481–484, 2018.

[12] G. I. Ulumudin, A. Adiwijaya, and M. S. Mubarok, "A multilabel classification on topics of qur'anic verses in English translation using K-Nearest Neighbor method with Weighted TF-IDF," *J. Phys. Conf. Ser.*, vol. 1192, no. 1, pp. 103–106, 2019.

[13] I. R. Ponilan, Adiwijaya, M. A. Bijaksana, and A. S. Raharusun, "Search relevant retrieval on indonesian translation hadith document using query expansion and smoothing probabilistic model," *J. Phys. Conf. Ser.*, vol. 1192, no. 1, 2019.

[14] F. Z. Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia," *M.Sc. Thesis, Append. D*, vol. pp, pp. 39–46, 2003.