OPEN ACCESS
Indo-JC
Indonesian Journal of Computing
ISSN 2460-9056
socj.telkomuniversity.ac.id/indojc

Ind. Journal on Computing
Vol. 3, Issue. 2, Sept 2018. pp. 83-90
doi:10.21108/indojc.2018.3.2.236

Kategorisasi Berita Menggunakan Metode Pembobotan TF.ABS dan TF.CHI

Muhammad Arif Kurniawan #1, Yuliant Sibaroni #2, Kemas Muslim Lhaksmana #3

Fakultas Informatika, Universitas Telkom, Bandung

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

[2] yuliant@telkomuniversity.ac.id

[3] kemasmuslim@telkomuniversity.ac.id

Abstract

With the current technological advances, news can be found easily and totaling very much in digital form causing the need for a technique to categorize the news into specific topics to make it easier for readers to find the news according to the desired topic. Text categorization is a technique that can categorize news into predefined topics automatically. One important process in categorization is feature extraction where unigram binary is one of the basic feature extraction compared to term weighting which in this study will use the weighting method TF.ABS and TF.CHI to obtain optimal news categorization results. Based on the test results, the average accuracy obtained from three data sources on unigram binary feature extraction of 90.44%. While on the weighting method TF.ABS of 95.74% and TF.CHI of 95.87%. Based on the results of the accuracy, it can be concluded that the term weighting is better than the unigram binary. The weighting methods of TF.ABS and TF.CHI are both good in categorization because they do not differ significantly in performance. Other test results show that the stemming process does not have much effect on the accuracy of news categorization, but this process can make up to 45% time efficiency.

Keywords: Feature extraction, term weighting, text categorization, TF.ABS, TF.CHI, unigram binary.

Abstrak

Dengan kemajuan teknologi saat ini, berita dapat ditemukan dengan mudah dan berjumlah sangat banyak dalam bentuk digital yang menyebabkan diperlukannya suatu teknik untuk mengkategorikan berita-berita tersebut ke dalam topik tertentu agar mempermudah pembaca menemukan berita sesuai dengan topik yang diinginkan. Kategorisasi teks merupakan suatu teknik yang dapat mengkategorikan berita ke dalam topik yang telah ditentukan secara otomatis. Salah satu proses yang penting dalam kategorisasi adalah ekstraksi fitur yang mana *unigram binary* merupakan salah satu ekstraksi fitur yang dasar dibandingkan dengan *term weighting* yang dalam penelitian ini akan menggunakan metode pembobotan TF.ABS dan TF.CHI untuk memperoleh hasil kategorisasi berita yang optimal. Berdasarkan hasil pengujian, rata-rata akurasi yang didapatkan dari tiga sumber data pada ekstraksi fitur *unigram binary* sebesar 90.44%. Sedangkan pada metode pembobotan TF.ABS sebesar 95.74% dan TF.CHI sebesar 95.87%. Berdasarkan hasil akurasi tersebut, dapat disimpulkan bahwa *term weighting* lebih baik dibandingkan dengan *unigram binary*. Metode pembobotan TF.ABS dan TF.CHI sama-sama baik dalam kategorisasi karena tidak berbeda secara signifikan dalam performansinya. Pada hasil pengujian lainnya menunjukkan bahwa proses *stemming* tidak memberikan banyak pengaruh terhadap akurasi kategorisasi berita, namun proses ini dapat mengefisiensikan waktu hingga 45%.

Kata Kunci: Ekstraksi fitur, kategorisasi teks, *term weighting*, TF.ABS, TF.CHI, *unigram binary*.

Received on Jul 2018. Accepted on Sept 2018

I. Pendahuluan

Dengan perkembangan teknologi informasi dan komunikasi saat ini, berita dapat ditemukan dengan mudah dan berjumlah sangat banyak dalam bentuk digital yang berasal dari berbagai sumber. Banyaknya jumlah berita ini menyebabkan diperlukannya suatu teknik pengolahan berita-berita tersebut ke dalam satu atau lebih topik yang telah didefinisikan sebelumnya agar dapat mempermudah pembaca memilih berita sesuai seperti topik yang diinginkan. Berdasarkan permasalahan tersebut, upaya yang dapat dilakukan adalah menerapkan kategorisasi teks untuk mengelompokkan berita berdasarkan topiknya [1].

Kategorisasi pada berita diperlukan suatu data yang terstruktur, namun data yang didapatkan adalah data yang tidak terstruktur atau data yang masih mentah. Permasalahan ini sering dijumpai karena berita yang didapatkan dengan mudah ini berasal dari internet (digital) [2]. Oleh sebab itu, pada kategorisasi teks terdapat suatu proses yang merubah data tersebut menjadi data yang terstruktur yaitu proses *text preprocessing* yang terdiri dari *word tokenization*, *case folding*, *stopwords removing*, dan *word stemming* [3]. Selain itu, *feature extraction* merupakan salah satu proses yang penting dalam kategorisasi yang mana kita dapat menggunakan model *unigram binary* sebagai teknik dasarnya. Model tersebut memberikan nilai *feature* dengan nilai *binary* (Nilai 0 untuk *feature* yang tidak ada dalam dokumen dan nilai 1 untuk *feature* yang terdapat dalam dokumen) [4]. Nilai *binary* ini kurang detail untuk menunjukkan ukuran tingkat kontribusi *feature* terhadap penentuan kategori karena tidak mempertimbangkan frekuensi kemunculan *feature* tersebut pada suatu dokumen, untuk menanggulanginya maka adanya suatu proses yang memberikan bobot pada masing-masing *feature* yang dikenal dengan istilah *term weighting*. *Term weighting* memiliki banyak metode yang dapat digunakan, seperti TF.IDF, TF.RF, TF.IG, TF.OR, TF.CHI, TF.ABS, dan sebagainya. Metode pembobotan yang digunakan akan memberikan pengaruh terhadap performansi kategorisasi suatu dokumen [5].

Penelitian ini menggunakan metode pembobotan TF.ABS dan TF.CHI yang memperhitungkan jumlah kemunculan *feature* pada suatu dokumen dan mempertimbangkan jumlah *feature* yang tidak muncul dalam dokumen. Metode TF.CHI dipilih untuk diuji karena pada penelitian yang menganalisis metode pembobotan TF.CHI dan TF.RF untuk kategorisasi teks berbahasa Indonesia menghasilkan kesimpulan bahwa metode TF.CHI lebih baik dibandingkan dengan metode TF.RF [6]. Sedangkan untuk metode TF.ABS dipilih sebagai perbandingan karena pada penelitian yang meningkatkan metode pembobotan untuk kategorisasi teks menghasilkan kesimpulan bahwa metode TF.ABS lebih baik dibandingkan dengan metode TF.IDF, TF.IG, dan sebagainya [7]. Oleh karena itu, pada penelitian ini melakukan kategorisasi berita menggunakan metode pembobotan TF.ABS dan TF.CHI untuk mengetahui metode pembobotan manakah yang menghasilkan performansi lebih baik dan mengetahui pengaruh dari proses *stemming* terhadap performansi dengan menggunakan metode klasifikasi *Support Vector Machine* (SVM) dan menggunakan *k-Fold Cross Validation*. Metode klasifikasi SVM dipilih karena metode SVM sudah diakui bahwa metode SVM lebih baik atau efektif dibandingkan dengan metode klasifikasi yang lainnya untuk klasifikasi teks [8]. Konsep dari metode SVM ini adalah mencari *hyperplane* terbaik yang berfungsi sebagai pemisah antar kategori atau topik dan didasari oleh prinsip *structural risk minimization* [9].

II. Studi Terkait

Seiring berjalannya waktu, telah banyak penelitian yang mengambil topik kategorisasi teks dengan berbagai metode klasifikasi. Salah satunya yaitu penelitian yang menggunakan metode klasifikasi *Support Vector Machine, k-Nearest Neighbor, Linear Least Square Fit, Neural Network,* dan *Naive Bayes* sebagai pembanding performansi untuk kategorisasi. Dengan menggunakan dataset dari Reuters-21578, penelitian ini menghasilkan kesimpulan bahwa metode klasifikasi SVM lebih baik atau efektif dibandingkan dengan metode klasifikasi yang lainnya untuk kategorisasi teks [8].

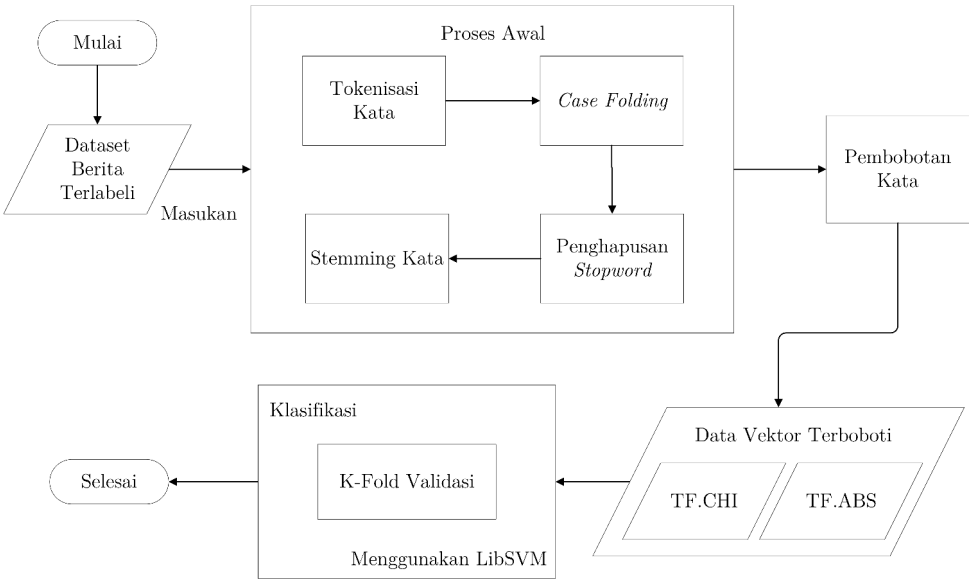
Untuk meningkatkan performansi dari kategorisasi teks, banyak penelitian juga yang berfokus kepada pemberian bobot kata untuk menunjukkan tingkat kepentingannya terhadap kategori yang ada dan mendapatkan hasil performansi yang lebih optimal. Hal ini dikenal dengan istilah proses *term weighting*, Matsunaga dan Ebecken melakukan sebuah penelitian tentang metode pembobotan untuk kategorisasi teks dengan membandingkan metode TF.ABS, TF.IDF, TF.IG, TF.GR, TF.BNS, TF.OR, dan TF.QUI [7].

Penelitian ini menggunakan dataset dari *Federal District Legislative Assembly* di Brazil pada tahun 2003-2004 [10]. Hasil yang diperoleh dari penelitian ini menunjukkan bahwa metode pembobotan TF.ABS lebih baik dibandingkan dengan metode lainnya.

Pada penelitian yang menggunakan 20Newsgroup sebagai dataset untuk kategorisasi teks dengan menggunakan metode pembobotan TF.IDF dan TF.RF menghasilkan kesimpulan bahwa secara umum metode TF.RF lebih unggul dari metode TF.IDF disebagian pengujian yang dilakukan. Kelebihan TF.RF yang memperhitungkan frekuensi kemunculan suatu term pada suatu kategori dan menormalisasikannya ke keseluruhan dokumen, membuat metode ini lebih baik dibanding TF.IDF [11]. Penelitian lainnya menunjukkan bahwa metode pembobotan TF.CHI lebih baik dibandingkan dengan metode TF.RF terhadap kategorisasi teks berbahasa Indonesia dengan menggunakan metode klasifikasi *Support Vector Machine* [6].

III. METODE PENELITIAN

Sistem yang dibangun pada penelitian ini memiliki masukan berupa berita yang berbahasa inggris dari Reuters tahun 1987 (Reuters-21578), AG *News*, dan BBC *News*. Masing-masing sumber berita memiliki sistemnya tersendiri karena adanya perbedaan format dataset yang membuat proses *load* data berbeda. Sistem kemudian dapat melakukan *document preprocessing* yang terdiri dari *word tokenization*, *case folding*, *stopwords removing*, *word stemming*. Proses *stemming* dapat dilakukan atau tidak berdasarkan skenario yang diuji. Setelah proses *preprocessing* selesai, maka dilanjutkan ke tahap ekstraksi fitur dengan menggunakan *unigram binary* dan dua metode pembobotan yaitu metode TF.ABS dan TF.CHI. Proses *term weighting* menghasilkan tiga *file* yang bertipe *.csv (*file* untuk *unigram binary*, TF.ABS, dan TF.CHI). Setelah mendapatkan *file* tersebut, maka dilakukan proses *classification* dengan menggunakan metode klasifikasi *Support Vector Machine* (SVM) dengan *tools* LibSVM yang diintegrasikan pada sistem yang dibangun. Pada proses *classification* ini menggunakan *K-Fold Cross Validation* untuk pembagian data *training* dan *testing* dengan nilai k sebesar lima. Hasil performansi didapatkan setelah proses *classification* selesai yaitu nilai akurasi dari *unigram binary* dan masing-masing metode pembobotan. Berikut ini merupakan gambaran dari sistem yang dibangun.



Gambar 1. Gambaran Sistem

A. Dataset

Dataset yang digunakan merupakan berita yang berbahasa inggris dari tiga sumber yang sudah terlabeli. Dataset pertama berasal dari *ApteMod* Reuters-21578 yang berjumlah 10788 data berkategori 90, dengan pembagian data *training* sebanyak 7769 dan data *test* sebanyak 3019 yang didapatkan dari http://search.cpan.org/~kwilliams/reuters-21578/. Kategori tersebut dirubah menjadi enam kategori yaitu *corporate, commodity, economic, energy, currency*, dan *subject* berdasarkan kategori deksripsi dari Reuters-21578. Pada penelitian ini menggunakan data *test* Reuters-21578 yang dikurangin dengan data berkategori *currency* yang berjumlah 42, sehingga dataset yang digunakan berjumlah 2977 yang memuat lima topik yaitu *corporate, commodity, economic, energy*, dan *subject*. Dataset yang kedua berasal dari BBC *News* yang berjumlah 2225 data dengan lima topik (*business, entertertainment, politic, sport*, dan *tech*) dari tahun 2004-2005 [12]. Jumlah masing-masing dari kategorinya dikurangin sebanyak 100, sehingga dataset BBC yang digunakan berjumlah 1725. Untuk dataset yang terakhir berasal dari AG *News* yang dibuat oleh Xiang Zhang berjumlah 120000 untuk *train* dan 7600 untuk *test* [13]. Data *test* digunakan sebagai dataset penelitian dengan mengurangi masing-masing kategori sebanyak 950, sehingga jumlah dataset AG sebanyak 3800 dengan empat topik yaitu *world, sport, business* , dan *sci/tech*. Pengurangan dari jumlah dataset aslinya dilakukan untuk kelancaran saat menjalankan sistem dengan menyesuaikan hardware sistem penulis. Pembagian jumlah masing-masing topik berdasarkan sumbernya dapat dilihat pada Tabel I.

Tabel I
Pembagian Dataset

	Reuters 21578		**BBC News**		**AG News**	
	Topik	**Jumlah**	**Topik**	**Jumlah**	**Topik**	**Jumlah**
	Corporate	1786	Politics	317	World	950
	Commodity	443	Business	410	Sport	950
	Economic	246	Entertainment	286	Business	950
	Energy	209	Sport	411	Sci/Tech	950
	Subject	293	Tech	301		
Total		2977		1725		3800

B. Preprocessing

Preprocessing dilakukan setelah proses *load* data selesai untuk mereduksi variansi *feature* yang terdiri dari *word tokenization*, *case folding*, *stopwords removing*, *word stemming*. *tokenization* merupakan tahapan pertama pada *preprocessing* yang melakukan pemotongan kalimat yang ada dalam dokumen menjadi tiap kata [14]. Tahapan selanjutnya yaitu *case folding* yang mengubah semua kata menjadi huruf kecil dan menghilangkan karakter selain huruf a sampai dengan z [15]. Setelah itu, *stopword removing* dilakukan dengan menghilangkan kata yang tidak penting seperti kata sandang dan kata hubung yang terdiri dari 571 kata berdasarkan http://search.cpan.org/~kwilliams/reuters-21578/. *Word stemming* merupakan tahapan akhir dari *preprocessing* yang mengubah kata menjadi kata dasarnya menggunakan algoritma Porter, tahapan ini dapat dilakukan atau tidak berdasarkan skenario yang diuji.

C. Term Weighting

Pada dasarnya, kita dapat menggunakan model N-Gram untuk memberikan nilai pada masing-masing *feature* dalam kategorisasi teks. Salah satu model N-Gram yang umum digunakan adalah 1-Gram yang dikenal dengan istilah *unigram binary*. Model tersebut memberikan nilai *feature* dengan nilai *binary*, nilai 1 menunjukkan bahwa kata tersebut terdapat dalam berita. Sedangkan nilai 0 menunjukkan bahwa kata tersebut tidak ada dalam berita itu [4]. Tabel II merupakan distribusi yang diamati untuk *term* t_j dan kategori c_i.

Tabel II
Contingency Table UNTUK *category* DAN *term*

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

Keterangan variabel:
n_{ij}: Merupakan jumlah dokumen dalam kategori c_i yang mengandung *term* t_j
$n_{\bar{i}j}$: Merupakan jumlah dokumen tidak dalam kategori $c_{\bar{i}}$ yang mengandung *term* t_j
$n_{i\bar{j}}$: Merupakan jumlah dokumen dalam kategori c_i yang tidak mengandung *term* $t_{\bar{j}}$
$n_{\bar{i}\bar{j}}$: Merupakan jumlah dokumen tidak dalam kategori $c_{\bar{i}}$ yang tidak mengandung *term* $t_{\bar{j}}$
n_j: Merupakan jumlah dokumen dengan *term* t_j
$n_{\bar{j}}$: Merupakan jumlah dokumen tanpa *term* $t_{\bar{j}}$
n_i: Merupakan jumlah dokumen dengan kategori c_i
$n_{\bar{i}}$: Merupakan jumlah dokumen tanpa kategori $c_{\bar{i}}$
n: Merupakan total atau jumlah dari dokumen
t_j: Merupakan *term* t_j
c_i: Merupakan kategori c_i
$t_{\bar{j}}$: Merupakan tanpa *term* $t_{\bar{j}}$
$c_{\bar{i}}$: Merupakan tanpa kategori $c_{\bar{i}}$.

Pada tahapan ini, *feature* dari hasil *preprocessing* akan diberi bobot dan menyimpannya menjadi representasi vektor agar dapat digunakan sebagai masukan di klasifier. Bobot nilai ini menjadi sebuah ukuran besarnya jumlah dan tingkat kontribusi sebuah kata atau *term* terhadap penentuan suatu kelas atau kategori suatu dokumen [5]. Pembobotan ini menggunakan dua metode yaitu metode TF.ABS dan TF.CHI. Metode TF.ABS merupakan perkalian antar dua metode yaitu metode TF dan metode ABS. Begitu pula metode TF.CHI yaitu perkalian antar metode TF dengan CHI.

Setiap *term* diasumsikan memiliki proporsi kepentingan sesuai dengan jumlah kemunculannya pada suatu dokumen merupakan suatu cara dari pemberian bobot dengan metode *Term Frequency* (TF) [16]. ABS merupakan pengukuran kemungkinan suatu *term* t_j yang ada dalam dokumen dengan kategori c_i dibagi dengan kemungkinan *term* t_j yang tidak ada dalam dokumen dengan kategori tersebut dan menerapkan transformasi logaritmik (basis log e) yang dikenal dengan logit [7]. Perhitungan Abs-logit dengan *term* t_j dan *category* c_i dapat dilihat pada Persamaan 1.

$$ABSL\left(t_j, c_i\right) = \left|\ln\left(\frac{\left(n_{ij}+0.5\right)\left(n_{\bar{i}\bar{j}}+0.5\right)}{\left(n_{\bar{i}j}+0.5\right)\left(n_{i\bar{j}}+0.5\right)}\right)\right| \tag{1}$$

Chi-Square merupakan suatu metode *supervised* yang membutuhkan informasi berasal dari kategori mana suatu dokumen tersebut. Metode ini mempertimbangkan bobot untuk *term* yang tidak muncul dalam dokumen dan *term* yang muncul didalam dokumen [17]. Pembobotan *Chi-Square* dengan *term* t_j dan kategori c_i dapat dilihat pada Persamaan 2.

$$\chi^2\left(t_j, c_i\right) = \frac{n\left(n_{ij}n_{\bar{i}\bar{j}} - n_{\bar{i}j}n_{i\bar{j}}\right)^2}{n_i n_j n_{\bar{i}} n_{\bar{j}}} \tag{2}$$

D. Classification

Classification merupakan tahapan terakhir yang menggunakan metode klasifikasi *Support Vector Machine* (SVM) dengan *tools* LibSVM. *Support Vector Machine* merupakan suatu metode *machine learning* yang mencari *hyperplane* terbaik yang berfungsi sebagai pemisah antar topik. *Hyperplane* terbaik dapat ditemukan dengan mengukur *margin hyperplane* tersebut dengan *pattern* terdekat (*support vector*) dari masing-masing topik [9]. Data masukan akan dibagi menjadi data *training* dan *testing* menggunakan

K-Fold Cross Validation dengan nilai k sebesar lima, yang berarti membagi dataset menjadi lima buah partisi secara acak (data-1, data-2,..., data-5) [18]. Proses *Classification* dari sistem menghasilkan *output* berupa hasil performansi dari *unigram binary* dan masing-masing metode pembobotan.

IV. Hasil dan Analisis

Pengujian dalam penelitian ini dilakukan dengan dua skenario, skenario pertama ditujukan untuk menemukan metode pembobotan manakah yang lebih baik diantara metode TF.ABS dan TF.CHI. Sedangkan skenario kedua ditujukan untuk mengetahui pengaruh proses *stemming* terhadap performansi kategorisasi berita pada penelitian ini.

A. Hasil Pengujian dan Analisis Skenario 1

Skenario 1 bertujuan untuk mengetahui metode pembobotan yang lebih baik diantara TF.ABS dan TF.CHI, skenario ini dilakukan tanpa menerapkan proses *stemming* dengan menggunakan metode pembobotan TF.ABS dan TF.CHI. Penelitian ini menggunakan metode *unigram binary* sebagai pembanding dasar sebelum dibobotin dari kedua metode pembobotan tersebut. Hasil pengujian skenario 1 dapat dilihat pada Tabel III.

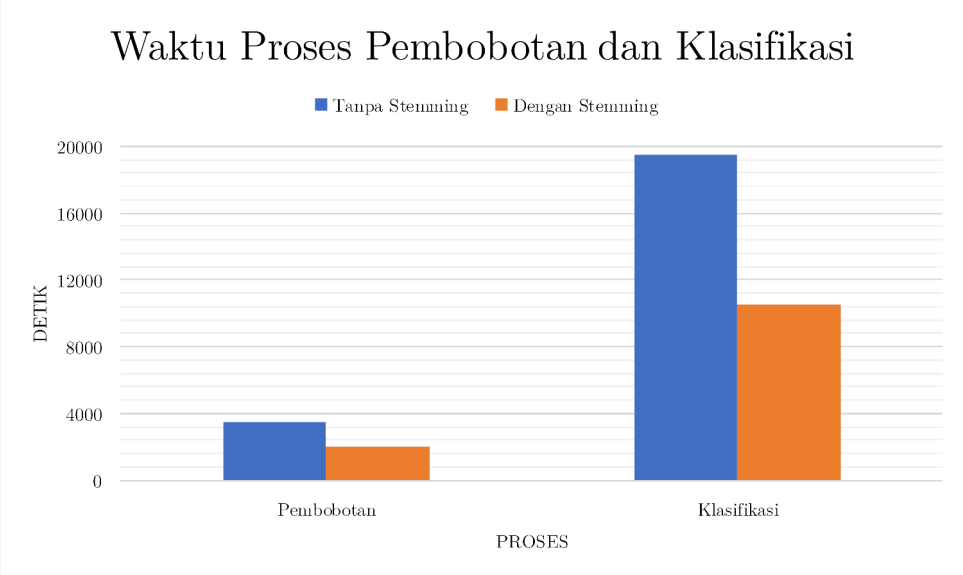
Tabel III
Hasil Akurasi Tanpa *Stemming*

No	Dataset	N Data	N Topik	N kata	Akurasi tanpa *Stemming* (%)		
					Unigram Binary	TF.ABS	TF.CHI
1	Reuters 21578	2977	5	16994	93.08	95.6	95.84
2	BBC News	1725	5	27712	96.75	99.07	99.25
3	AG News	3800	4	14840	81.5	92.55	92.53
				Rata-rata	90.44	95.74	95.87

Berdasarkan rata-rata akurasi pada Tabel III, metode pembobotan TF.ABS dan TF.CHI dapat meningkatkan performa sistem bila dibandingkan dengan *unigram binary* yaitu sebesar 5%. Hal ini dikarenakan *unigram binary* hanya menggunakan nilai biner yang menandakan *feature* tersebut ada pada dokumen, sedangkan metode pembobotan memberikan nilai *feature* yang ada pada dokumen dengan memperhitungkan tingkat kontribusinya. Akurasi metode TF.CHI lebih baik dibandingkan dengan metode TF.ABS untuk dataset yang bersumber dari Reuters-21578 dan BBC *News*. Sedangkan pada dataset AG *News*, metode TF.ABS lebih baik dibandingkan dengan TF.CHI. Perbedaan peningkatan diantara kedua metode tersebut kurang dari 1%, hal ini dapat dilihat pula pada rata-rata akurasi antar metode pembobotan yaitu sebesar 95.74% untuk metode TF.ABS dan 95.87% untuk metode TF.CHI. Kedua metode pembobotan tidak memberikan beda yang signifikan karena sama-sama mempertimbangkan *term* t_j yang muncul maupun tidak muncul dalam dokumen.

B. Hasil Pengujian dan Analisis Skenario 2

Pada pengujian skenario 1, sistem membutuhkan waktu proses yang cukup lama. Oleh sebab itu dilakukannya skenario 2 yang bertujuan untuk mengetahui pengaruh proses *stemming* dengan menggunakan metode pembobotan TF.ABS dan TF.CHI. Penelitian ini menerapkan proses *stemming* sebagai pembanding dengan skenario 1 yang tidak menggunakan proses *stemming*. Perbedaan waktu proses pembobotan dan klasifikasi antara skenario 1 dengan skenario 2 dapat dilihat pada Gambar 2, sedangkan perbandingan hasil akurasi diantara kedua skenario dapat dilihat pada Tabel IV.



Gambar 2. Waktu Proses Pembobotan dan Klasifikasi

Tabel IV
HASIL AKURASI DENGAN *Stemming*

No	Dataset	N Data	N Topik	N kata	Akurasi dengan *Stemming* (%)		
					Unigram Binary	TF.ABS	TF.CHI
1	Reuters 21578	2977	5	12505	93.22	96.17	97.08
2	BBC News	1725	5	18991	97.51	99.13	99.48
3	AG News	3800	4	11072	81.74	92.13	91.71
				Rata-rata	90.82	95.81	96.09

Tabel IV menunjukkan bahwa rata-rata hasil akurasi dengan proses *stemming* untuk *unigram binary* sebesar 90.82%, sedangkan untuk metode pembobotan TF.ABS sebesar 95.81% dan TF.CHI sebesar 96.09%. Bila dibandingkan dengan hasil akurasi tanpa proses *stemming* pada Tabel III, proses *stemming* dapat memberikan peningkatan akurasi kurang dari 1%. Peningkatan yang tidak terlalu signifikan ini dikarenakan perbandingan jumlah data dengan jumlah kata antar dataset, jumlah data yang kecil dan jumlah kata yang banyak akan menghasilkan akurasi yang lebih baik dibandingkan dengan jumlah data yang banyak dan jumlah kata yang kecil. Pada Gambar 2, tanpa stemming merupakan rata-rata waktu proses dari 3 sumber tanpa menerapkan proses *stemming*. Sedangkan dengan stemming merupakan rata-rata waktu yang menerapkan proses *stemming*. Gambar 2 menunjukkan bahwa proses *stemming* dapat mengefisiensikan waktu hingga 45% pada proses pembobotan dan klasifikasi. Hal ini dikarenakan proses *stemming* dapat mengurangi jumlah kata dari dataset hingga 30% yang dapat dilihat perbandingannya pada Tabel III dengan Tabel IV.

V. KESIMPULAN

Berdasarkan dari hasil pengujian dan analisis di atas, maka kesimpulan yang didapatkan pada penelitian ini sebagai berikut:

1) Kategorisasi berita tanpa menerapkan proses *stemming* dengan menggunakan *unigram binary* menghasilkan nilai rata-rata akurasi sebesar 90.44%. Sedangkan bila menggunakan metode pembobotan TF.ABS menghasilkan nilai akurasi sebesar 95.74% dan untuk metode pembobotan TF.CHI sebesar 95.87%. Berdasarkan dari nilai akurasi tersebut dapat disimpulkan bahwa metode pembobotan lebih baik dibandingkan dengan *unigram binary* dengan memberikan peningkatan sebesar 5% , hal ini

dikarenakan *unigram binary* hanya menggunakan nilai biner yang menandakan *feature* tersebut ada pada dokumen, sedangkan metode pembobotan memberikan nilai *feature* yang ada pada dokumen dengan memperhitungkan tingkat kontribusinya. Metode pembobotan TF.CHI dan TF.ABS tidak memberikan beda yang signifikan dalam performansi, hal ini menunjukkan bahwa kedua metode tersebut sama-sama baik dalam kategorisasi berita karena mempertimbangkan *feature* yang muncul maupun tidak dalam dokumen.

2) Proses *stemming* pada tahap *preprocessing* tidak memberikan banyak pengaruh terhadap akurasi dari kategorisai berita dengan menggunakan metode pembobotan TF.ABS dan TF.CHI. Proses *stemming* hanya memberikan peningkatan akurasi kurang dari 1%, namun proses ini mereduksi jumlah kata yang dapat mengefisiensikan waktu proses pembobotan dan klasifikasi hingga 45%. Peningkatan yang tidak terlalu signifikan dalam akurasi ini dikarenakan perbandingan jumlah data dengan jumlah kata antar dataset, jumlah data yang kecil dan jumlah kata yang banyak akan menghasilkan akurasi yang lebih baik dibandingkan dengan jumlah data yang banyak dan jumlah kata yang kecil.

Penelitian ini menunjukkan bahwa *feature* yang didapatkan dari masing-masing sumber dataset berjumlah cukup banyak. Oleh sebab itu, penelitian selanjutnya dapat menggunakan *feature selection* untuk memilih fitur yang berpengaruh dan mengurangi *feature* yang kurang relevan. Selain itu, penelitian kedepannya dapat menggunakan dataset yang berbeda seperti 20 Newsgroup dan *Classifier* yang lainnya sebagai pembanding performansi dari metode TF.ABS dan TF.CHI.

Pustaka

[1] A. Basu, C. Watters, and M. Shepherd. *Support Vector Machines for Text Categorization*. IEEE, 2003.
[2] C. Goutam. *Analysis of Unstructured Data: Applications of Text Analytics and Sentiment Mining*. 2014.
[3] F. Ismail. *Examining Learning Algorithms for Text Classification in Digital Libraries*. Department of Alfa-Informatica, University of Groningen. Netherland, 2004.
[4] T. Christoph and X. Fei. *A phrase-based unigram model for statistical machine translation*. Association for Computational Linguistics Stroudsburg, PA, USA, 2003.
[5] M. Liliam Ayako and E. Nelson Francisco Favilla. *Two Novel Weighting for Text Categorization*. WITPress, 2008.
[6] E. Putri N Kianiseta. *Analisis Perbandingan Metode Pembobotan TF.CHI2 dan TF.RF Terhadap Kategorisasi Teks Berbahasa Indonesia*. Universitas Telkom, 2014.
[7] M. Liliam Ayako and E. Nelson Francisco Favilla. *Term Weighting Approaches for Text Categorization Improving*. IEEE, 2008.
[8] Y. Yiming and L. Xin. *An Re-examination of Text Categorization*. ACM New York, NY, USA, 1999.
[9] W. Ziqiang, S. Xia, and Z. Dexian. *An Optimal Text Categorization Algorithm Based on SVM*. IEEE, 2007.
[10] M. Liliam Ayako. *An Automated Text Categorization Methodology to Distribute the Bills to the Commettees at the Federal Distric Legislative Assembly*. Dept of Civil Engineering, COPPE/Federal University of Rio de Janeiro, 2007.
[11] A. Thopo Martha. *Analisis Perbandingan Metode Pembobotan Kata TF.IDF dan TF.RF Terhadap Performansi Kategorisasi Teks*. Universitas Telkom, 2012.
[12] G. Derek and C. Padraig. *Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering*. ICML, 2006.
[13] Z. Xiang, Z. Junbo, and L. Yann. *Character-level Convolutional Networks for Text Classification*. NIPS, 2015.
[14] M. Christopher D., R. Prabhakar, and S. Hinrich. *Introduction to Information Retrieval*. Cambridge University Press, 2008.
[15] N. Nadia. *Intelligent Text Categorization and Clustering*. Berlin: Springer, 2009.
[16] T. Tokunaga and I. Makoto. *Text Categorization Based On Weighted Inverse Document Frequency*. Tokyo, Japan: Tokyo Institute of Technology, 1994.
[17] S. Catur and Affandy. *Kombinasi Teknik Chi Square dan Singular Value Decomposition Untuk Reduksi Fitur Pada Pengelompokan Dokumen*. Melaka, Malaysia : Universitas Teknikal Malaysia, 2011.
[18] Y. Sanjay and S. Sanyam. *Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification*. IEEE, 2016.