JURNAL MEDIA INFORMATIKA BUDIDARMA, Vol 3, No 4, Oktober 2019
ISSN 2614-5278 (media cetak)
ISSN 2548-8368 (media online)
Hal 284-292 | DOI: 10.30865/mib.v3i4.1410

Klasifikasi Berita Bahasa Indonesia Menggunakan *Mutual Information* dan *Support Vector Machine*

Lalu Gias Irham, Adiwijaya, Untari Novia Wisesty

School of Computing, Telkom University, Bandung, Indonesia
Email: giasirham@students.telkomuniversity.ac.id, [1]adiwijaya@telkomuniversity.ac.id , [1]untarinw@telkomuniversity.ac.id

Abstrak

Berita merupakan sumber informasi yang disebarkan dalam berbagai macam media. Agar memudahkan pembaca berita dalam memperoleh berita yang diinginkan, maka berita perlu diklasifikasikan. Banyaknya jumlah berita yang tersebar menimbulkan kesulitan dalam mengelompokkan berita tersebut berdasarkan topiknya. Oleh sebab itu penulis melakukan penelitian untuk mengklasifikasikan berita ke dalam 12 kelas (budaya, ekonomi, hiburan, hukum, kesehatan, gaya hidup, otomotif, pendidikan, politik, olahraga, teknologi, dan wisata) secara otomatis terhadap 360 data berita Bahasa Indonesia. Pada penelitian ini dilakukan beberapa scenario pengujian untuk melihat pengaruh dari metode *stopword removal* dan *stemming* pada *preprocessing* data, pengaruh *mutual information* dalam menyeleksi fitur, dan performansi *Support Vector Machine* dalam mengklasifikasikan data berita. Hasil pengujian menunjukkan data yang hanya menggunakan *stemming* tanpa *stopword removal*, menggunakan fitur seleksi MI dan metode klasifikasi SVM menghasilkan hasil terbaik yaitu 94.24%, dibandingkan dengan metode yang lainnya.

Kata Kunci: Berita, Klasifikasi Teks, *Support Vector Machine*, Seleksi Fitur, *Mutual Information*

Abstract

News is a source of information disseminated in various types of media. In order to make it easier for news readers to obtain the desired news, the news needs to be classified. The large number of scattered news creates difficulties in classifying the news based on the topic. Therefore the author conducted a study to classify news into 12 classes (culture, economy, entertainment, law, health, life, automotive, education, politics, sports, technology, and tourism) automatically against 360 Indonesian news data. In this study several test scenarios were conducted to see the effect of stopword removal and stemming methods on data preprocessing, the effect of mutual information in selecting features, and performance of Support Vector Machine in classifying news data. The test results showed that the data using only stemming without stopword removal, using the MI selection feature and SVM classification method produced the best results of 94.24%, compared to the other methods.

Keywords: *News, Text Classification, Support Vector Machine, Feature Selection, Mutual Information*

1. PENDAHULUAN

Berita merupakan sebutan yang mengacu kepada informasi yang disebarkan oleh surat kabar, radio, televisi, internet, dan media lainnya [1]. Ratusan berita dituliskan setiap harinya di berbagai portal berita Indonesia berbasis *online,* dikarenakan banyaknya portal berita yang beralih dari media cetak menjadi media elektronik yang dapat diakses secara *online* menggunakan internet [2]. Untuk mempermudah pembaca memperoleh berita yang diinginkan dari ribuan berita yang ada, maka berita butuh diklasifikasikan. Hasil dari penelitian terakhir telah membuktikan jika klasifikasi berita menggunakan komputer lebih efisien dibandingkan klasifikasi yang dilakukan oleh manusia [3]. Maka dari itu, untuk mengatasi masalah kekeliruan klasifikasi [2] dan meningkatkan efisiensi proses klasifikasi berita, dibutuhkan pembuatan sistem klasifikasi otomatis yang dapat dilakukan dengan menggunakan klasifikasi teks.

Banyak penelitian yang telah dilakukan dalam mengklasifikasikan berita menggunakan klasifikasi teks, dalam penelitian yang dilakukan Septian dan tim [4] serta Asy'arie dan partnernya [5] dalam mengklasifikasikan berita menggunakan metode klasifikasi *Naïve Bayes* menunjukkan bahwa walaupun *Naïve Bayes* merupakan metode yang sederhana dengan akurasi yang tinggi, namun menimbulkan masalah performansi jika memproses data yang besar. Sedangkan penelitian yang dilakukan Tej dan partnernya [6] serta penelitian yang dilakukan Dyah dan partnernya [7] menunjukkan bahwa metode *Support Vector Machine* dapat mengatasi masalah dimensi dan memberikan hasil lebih baik dibandingkan metode lainnya dalam mengklasifikasikan berita.

Sehingga pada penelitian ini, untuk membuktikan bahwa SVM merupakan metode terbaik untuk mengklasifikasikan data teks maka penulis membandingkan penggunaan metode Support Vector Machine dengan metode klasifikasi lainnya seperti K-Nearest Neighbor (KNN), Neural Network (NN), dan Naïve Bayes (NB) untuk mengklasifikasikan berita Bahasa Indonesia. Untuk menangani masalah fitur yang sangat besar pada klasifikasi teks [5][6][7] dan meningkatkan kinerja sistem klasifikasi, maka pada penelitian ini akan dilakukan pengurangan fitur dengan metode seleksi fitur yaitu Mutual Information. Mutual Information dipilih sebagai metode seleksi fitur pada penelitian ini dikarenakan memiliki titik fokus terhadap hubungan term kata dengan suatu kelas, sehingga fitur yang dihasilkan dari proses ini mampu meningkatkan akurasi dari proses klasifikasi yang digunakan [9]. Selain itu, pada penelitian ini juga akan dilihat pengaruh dari proses dalam preprocessing data dalam performansi maupun akurasi dari model yang dibangun.

Pada penelitian ini penulis mengangkat topik permasalahan mengenai cara membangun klasifikasi data berita Bahasa Indonesia yang cukup banyak secara tepat. Fokus penelitian ini mencari tahu pengaruh dari penerapan metode-metode pada preprocessing, feature selection, dan klasifikasi yang diterapkan. Adapun pada

JURNAL MEDIA INFORMATIKA BUDIDARMA, Vol 3, No 4, Oktober 2019
ISSN 2614-5278 (media cetak)
ISSN 2548-8368 (media online)
Hal 284-292 | DOI: 10.30865/mib.v3i4.1410

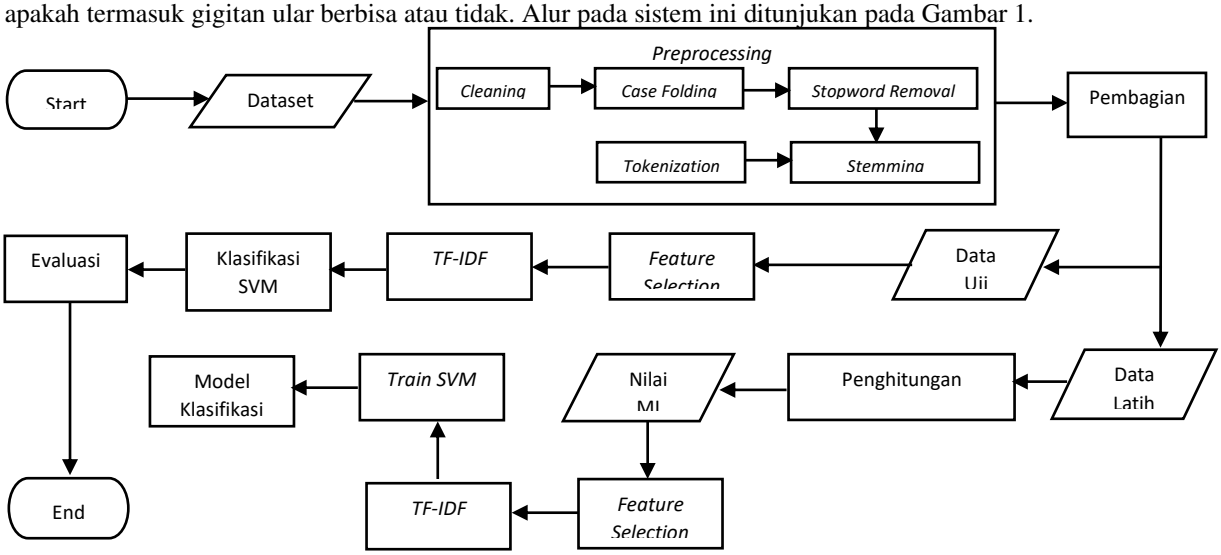
penelitian ini akan menggunakan stopword removal dan stemming pada tahap preprocessing, penggunaan Mutual Information dalam menyeleksi fitur yang akan digunakan, dan menggunakan metode klasifikasi Support Vector Machine. Terdapat beberapa hal yang dijadikan sebagai batasan masalah pada penelitian ini. Pertama, banyak kelas yang digunakan untuk klasifikasi terdiri dari 12 kelas yaitu budaya, ekonomi, hiburan, kriminal, kesehatan, gayahidup, otomotif, pendidikan, politik, olahraga, teknologi, dan wisata. Jumlah berita yang digunakan sebanyak 360 berita, dimana masing-masing topik memiliki 30 berita. Data didapat dari penelitian sebelumnya yang dilakukan oleh Fahmi [9].

Penelitian ini bertujuan untuk mempermudah pembaca berita dalam memilih berita yang akan dibaca dengan cara mengelompokkan berita yang ada kedalam topik berita seperti budaya, ekonomi, hiburan, hukum, kesehatan, gaya hidup, otomotif, pendidikan, politik, sport, tekno, dan wisata. Penelitian ini juga bertujuan untuk menganalisis pengaruh penggunaan *stopword removal* dan *stemming* dalam *preprocessing* data terhadap akurasi klasifikasi berita Bahasa Indonesia. Serta penelitian ini bertujuan untuk menganalisis keefektifan penggunaan *Mutual Information* dalam menyeleksi fitur-fitur yang ada, dan performansi metode klasifikasi *SVM* dibandingkan dengan metode klasifikasi lainnya dalam mengklasifikasikan berita Bahasa Indonesia.

2. METODE PENELITIAN

2.1 Perancangan Sistem

Penelitian ini membangun sistem untuk mengenali gigitan ular pada citra dan selanjutnya citra tersebut dikategorikan gigitan ular berbisa atau tidak berbisa. *Input* dari sistem berupa sebuah citra yang menggambarkan pola gigitan yang terdapat pada suatu media, dan hasil *output* berupa kategori hasil identifikasi bekas gigitan yang didapat. Secara garis besar, sistem yang akan digunakan terdiri dari empat tahap, yaitu: tahap *preprocessing*, tahap ekstraksi fitur pada citra, tahap pelatihan klasifikasi dan terakhir tahap pengujian dimana data akan diidentifikasi apakah termasuk gigitan ular berbisa atau tidak. Alur pada sistem ini ditunjukan pada Gambar 1.



Gambar 1. Alur Sistem

2.2 Representasi Data

Pada penelitian ini, dibangun sistem yang mampu mengklasifikasikan banyak data berita Bahasa Indonesia kedalam topik-topik yang tersedia. Data yang digunakan merupakan 360 data teks yang telah diberi label berdasarkan topiknya, dimana setiap topik mempunyai 30 data berita. Kelas yang digunakan pada dataset terdiri dari 12 kelas yaitu budaya, ekonomi, hiburan, kriminal, kesehatan, gayahidup, otomotif, pendidikan, politik, olahraga, teknologi, dan wisata**.** Data didapat dari penelitian sebelumnya yang dilakukan oleh Fahmi [9]. Representasi dari data berita yang digunakan dalam penelitian ini ditunjukkan pada Tabel 1.

Tabel 1. Representasi Data Berita

Berita Bahasa Indonesia	Kelas
Menpora Imam Nahrawi hari Kamis (16/2) siang kemarin menerima pianis cilik Jefri Setiawan yang akan melakukan pemecahan rekor dunia bermain piano dengan mata tertutup di Inggris…	1

Pada Tabel 1, dapat dilihat dataset berupa teks berita Bahasa Indonesia dimana nilai kelasnya adalah satu. Ini menunjukkan bahwa teks berita tersebut termasuk kedalam topik budaya. Penomeran kelas dilakukan berturut-turut dari 1-12 dengan urutan kelas yaitu budaya, ekonomi, hiburan, kriminal, kesehatan, gayahidup, otomotif, pendidikan, politik, olahraga, teknologi, dan wisata.

JURNAL MEDIA INFORMATIKA BUDIDARMA, Vol 3, No 4, Oktober 2019
ISSN 2614-5278 (media cetak)
ISSN 2548-8368 (media online)
Hal 284-292 | DOI: 10.30865/mib.v3i4.1410

2.3 *Preprocessing*

Sistem yang dibentuk dimulai dari memasukkan dataset berupa data teks ke dalam proses *preprocessing*. Proses ini sendiri memiliki beberapa tahap yaitu *cleaning*, *case folding*, *stopword removal, stemming,* dan *tokenization*. *Cleaning* adalah proses untuk menghilangkan angka, simbol, tanda baca pada kalimat. Lalu *case folding* akan mengubah semua huruf yang ada menjadi huruf kecil. *Stopword removal* menghilangkan kata-kata yang dianggap tidak penting dalam kalimat seperti kata hubung yang, dan, dll. Kemudian *stemming* akan mengubah semua kata ke dalam bentuk kata dasarnya. Pada penelitian ini algoritma stemming yang akan digunakan adalah Algoritma Nazief dan Adriani, karena pada penelitian sebelumnya [4], terbukti bahwa algoritma ini efektif untuk digunakan pada data teks berbahasa Indonesia. Dan yang terakhir, kata-kata dalam kalimat tersebut akan dipotong-potong sehingga berdiri sendiri dalam bentuk token pada proses *tokenization*. Berikut contoh proses *preprocessing* dari data teks berita dimasukkan hingga hasil dari setiap prosesnya dapat dilihat pada Tabel 2.

Tabel 2. *Preprocessing* Dataset Berita

Nama Proses	Output Kalimat
DataSet	Presiden Direktur FWD Life Rudi Kamdani menuturkan, Batam merupakan kawasan strategis untuk mendukung FWD Life dengan pemasaran.
Cleaning	Presiden Direktur FWD Life Rudi Kamdani menuturkan Batam merupakan kawasan strategis untuk mendukung FWD Life dengan pemasaran
Case Folding	presiden direktur fwd life rudi kamdani menuturkan batam merupakan kawasan strategis untuk mendukung fwd life dengan pemasaran
Stopword Removal	presiden direktur fwd life rudi kamdani menuturkan batam merupakan kawasan strategis mendukung fwd life pemasaran
Stemming	presiden direktur fwd life rudi kamdani tutur batam rupa kawasan strategis dukung fwd life pasar
Tokenization	“presiden”, “direktur”, “fwd”, “life”, “rudi”, “kamdani”, “tutur”, “batam”, “rupa”, “kawasan”, “strategis”, “dukung”, “fwd”, “life”, “pasar”.

Hasil token dari proses *preprocessing* disimpan kedalam *array* berdasarkan dokumen-dokumennya. Kemudian dengan menggunakan *k-fold* data berita tersebut akan dibagi menjadi dua bagian, yaitu data latih dan data uji.

2.4 *Mutual Information*

Setelah data dipecah, data uji akan digunakan untuk proses selanjutnya yaitu *feature selection*. Pada tahap ini kata-kata pada *array* akan diseleksi dan dipilih untuk diambil fitur kata yang paling relevan terhadap masing-masing kelas, untuk digunakan pada proses selanjutnya, sedangkan fitur kata yang tidak relevan akan dibuang [8]. Proses ini dilakukan dengan tujuan untuk mengurangi fitur yang sangat banyak menjadi lebih pendek, informative, dan efektif [11] Metode yang digunakan untuk proses seleksi fitur pada penelitian ini adalah *Mutual Information.* Metode *Mutual Information* atau MI memiliki konsep untung menghitung seberapa banyak informasi yang terkandung dalam *term*, dan kontribusinya untuk membuat keputusan klasifikasi yang tepat pada suatu kelas [9][16]. Misalkan *term* “sidang” memiliki nilai MI tertinggi untuk kelas politik. Maka jika dalam sebuah berita terdapat kata sidang, besar kemungkinan berita tersebut termasuk ke dalam berita politik. Berdasarkan penelitian sebelumnya [10][12], untuk memperoleh nilai MI dibutuhkan beberapa nilai pendukung yaitu frekuensi *term* x pada kelas A, frekuensi *term* lain pada kelas A, frekuensi *term* x di kelas selain A, frekuensi *term* selain x di kelas selain A, dan jumlah semua *term* yang ada. Setelah frekuensi dari *term* diperoleh, tahap selanjutnya adalah menghitung nilai MI dari masing-masing *term* yang ada. Formula perhitungan nilai MI dapat dilihat pada persamaan 1.

$$I(U,C) = \sum_{et\in\{1,0\}} \sum_{ec\in\{1.0\}} P(U=et, C=ec) \log_2 \frac{P\,(U=et, C=ec)}{P(U=et)P(C=ec)} \qquad (1)$$

Berdasarkan persamaan 1, variabel U merupakan variabel acak dengan nilai $et = 1$ (mengandung *term* t) dan $et = 0$ (tidak mengandung term t). Sementara untuk C merupakan variabel acak dengan nilai $ec = 1$ (kata berada di kelas c) dan $ec = 0$ (kata tidak berada di kelas c). Sehingga U merupakan nilai dari *term* sedangkan C merupakan kelasnya. Jika diuraikan lebih detail maka perhitungan nilai MI dapat dilihat pada persamaan 2 berikut.

$$I(U,C) = \frac{N_{11}}{N}\log_2\frac{N.N_{11}}{N_1.N_1} + \frac{N_{01}}{N}\log_2\frac{N.N_{01}}{N_0.N_1} + \frac{N_{10}}{N}\log_2\frac{N.N_{10}}{N_1.N_0} + \frac{N_{00}}{N}\log_2\frac{N.N_{00}}{N_0.N_0} \qquad (2)$$

Berdasarkan persamaan 2, variabel N merupakan total banyak dari semua *term* yang ada. Sedangkan variabel N_{uc} merupakan jumlah kalimat yang bernilai et dan ec. Sebagai contoh N_{10} merupakan kalimat yang mengandung *term* t ($et = 1$) namun tidak di kelas c ($ec = 0$). Sedangkan $N_1 = N_{11}$ dan N_{10} adalah jumlah dari kalimat yang mengandung *term* t ($et = 1$) dan jumlah kalimat independen anggota kelas tersebut ($ec \in \{1,0\}$). Dengan menggunakan persamaan diatas, akan diperoleh nilai MI untung setiap *feature*. Semakin tinggi nilainya

JURNAL MEDIA INFORMATIKA BUDIDARMA, Vol 3, No 4, Oktober 2019
ISSN 2614-5278 (media cetak)
ISSN 2548-8368 (media online)
Hal 284-292 | DOI: 10.30865/mib.v3i4.1410

maka semakin besar pengaruh dari *feature* tersebut terhadap suatu kelas. Kemudian dari fitur-fitur dengan nilai tertinggi yang ada, akan diambil sebagian untuk digunakan kedalam proses *feature extraction*. Contoh perhitungan MI pada penelitian ini dapat dilihat beruturut-turut pada Tabel 3, 4, dan 5.

Tabel 3. Contoh Data Perhitungan *Mutual Information*

Berita	Kelas	Panjang
Es krim menu dessert sangat digemari orang	Wisata	7
Kepentingan nasional dan keberlanjutan investasi. Semua pihak sudah tahu kalau Freeport	Ekonomi	11
Pimpinan Pusat Partai Amanat Nasional menunggu momentum untuk mendeklarasikan dukungan terhadap calon Gubernur	Politik	13
Organisasi massa pendiri Partai Golkar melakukan musyawarah nasional di Cilegon	Politik	10

Tabel 4. Contoh Truth Table Fitur Nasional Terhadap Kelas Politik

Nasional \ Politik	ec = e politik = 1	ec = e larangan = 0
et = e nasional = 1	**2**	**1**
et = e nasional = 0	**21**	**17**

Tabel 5. Contoh Perhitungan Menggunakan Rumus *Mutual Information* dan Hasil Perhitungan Semua Fitur Terhadap Kelas Politik Setelah Dilakukan Perhitungan Nilai *Mutual Information.*

Perhitungan	Hasil Perhitungan	
$(nasional, politik) = \frac{2}{41}\log_2\frac{41x2}{(2+1)x(2+21)} + \frac{21}{41}\log_2\frac{41x21}{(21+17)x(21+2)} + \frac{1}{41}\log_2\frac{41x1}{(1+2)x(1+17)} + \frac{17}{41}\log_2\frac{41x17}{(17+1)x(17+21)} \approx 0{,}002645$	**Fitur**	**Nilai**
	Nasional	0.002645
	Agenda	0.003876
	Jajah	0.001332
	Zaman	0.002874
	…	…
	Pusat	0.003554

Setelah semua fitur dihitung nilainya untuk 1 kelas, maka akan dilakukan perhitungan nilai MI fitur tersebut untuk kelas lainnya. Kemudian semua hasil nilai MI fitur tersebut terhadap semua kelas akan dibandingkan, dan akan disimpan nilai MI terbesar dari fitur tersebut kedalam sebuah tabel. Misalnya nilai MI fitur nasional terhadap kelas ekonomi adalah 0.001329, sedangkan nilai fitur nasional terhadap kelas politik adalah 0.002645, maka pada tabel gabungan nilai MI nilai fitur nasional yang disimpan adalah 0.002645. Kemudian fitur akan diurutkan berdasarkan nilai MI yang paling tinggi ke yang paling rendah. Berikut dapat dilihat hasil akhir dari proses *feature selection* penelitian ini pada Tabel 6.

Tabel 6. Tabel Semua *Feature* dengan Nilai *Mutual Information*

No. Fitur	Fitur	Nilai MI
1	Nasution	0.43657
2	Kasih	0.43541
3	Musibah	0.43297
4	Posting	0.43203
5	Tegur	0.43112
…	…	…
9043	Zamora	0.00329

2.5 *Term Frequency – Inverse Document Frequency* (TF-IDF)

Pada proses selanjutnya yaitu proses pengubahan kata kedalam bentuk angka agar dapat diproses *classifier,* dengan cara sistem akan memberikan nilai bobot dari setiap fitur yang ada. Pada penelitian ini metode *Term Frequency – Inverse Document Frequency (TF-IDF),* dimana pembobotan dalam metode ini berdasarkan gabungan antara *document-based* (*TF*) dan *collection based (IDF)* [13]. Singkatnya *Term Frequency (TF)* adalah berapa banyak sebuah *term* muncul dalam sebuah dokumen, sedangkan *Inverse Document Frequency (IDF)* adalah perhitungan bagaimana sebuah *term* didistribusikan pada koleksi dokumen [4][15]. Hasil dari proses ini berupa matriks yang terdiri dari data sebagai baris, fitur sebagai kolomnya, dan isi dari matriks tersebut merupakan bobot nilai dari setiap fitur terhadap dokumen setelah dilakukan perhitungan TF-IDF. Formula perhitungan bobot TF-IDF dapat dilihat pada persamaan 3.

$$W_{i,j} = tf_{i,j} \times idf = tf_{ij} \times log\left(\frac{N}{df_i}\right) \qquad (3)$$

JURNAL MEDIA INFORMATIKA BUDIDARMA, Vol 3, No 4, Oktober 2019
ISSN 2614-5278 (media cetak)
ISSN 2548-8368 (media online)
Hal 284-292 | DOI: 10.30865/mib.v3i4.1410

Berdasarkan persamaan 3, $W_{i,j}$ merupakan bobot kata t_j terhadap dokumen d_i, tf_{ij} mmerupakan jumlah kemunculan kata kata t_j terhadap dokumen d_i, N adalah jumlah dokumen keseluruhan dan df_i adalah jumlah dokumen keseluruhan yang mengandung kata j. Berikut adalah contoh input dan output pada proses TF-IDF penelitian ini dapat dilihat pada Tabel 7.

Tabel 7. Matriks TF-IDF

No. Fitur / No. Doc.	1	2	3	4	5	6	7	8	9	10	…	5000
1	0	0	90.792	0	0	0	0	0	0	0	…	0
2	0	0	5.502	0	0	0	0	0	0	0	…	0
3	0	0	2.751	0	0	2.503	0	0	0	0	…	0
4	0	0	22.010	0	0	2.503	0	0	0	4.965	…	0
5	0	0	13.756	0	0	0	0	0	0	0	…	0
…	…	…	…	…	…	…	…	…	…	…	…	…
71	0	0	30.2641	0	0	0	0	0	0	0	…	0

2.5 *Support Vector Machine* untuk Klasifikasi

Berikutnya, matriks yang telah didapat pada proses *TF-IDF* akan digunakan sebagai inputan proses klasifikasi, dimana metode yang digunakan pada proses klasifikasi penelitian ini adalah metode *Support Vector Machine*. Berdasarkan penelitian sebelumnya [6][7], SVM dikatakan sebagai metode klasifikasi yang memiliki akurasi lebih tinggi dibandingkan metode klasifikasi yang lain dan dapat mengatasi masalah *nonlinear* yang kompleks. Ini disebabkan karena SVM menerapkan prinsip *Structural Risk Minimization (SRM)*, dimana prinsip ini memerlukan pencarian *hyperlane* pemisah yang optimal [14]. *Hyperlane* merupakan pemisah antara sampel positif dan negatif, dan dikatakan optimal jika margin yang terbentuk memiliki jarak paling dekat dengan sampel positif dan negatif. Sehingga ketika ada sampel baru yang masuk, jika sampel tersebut di sisi positif maka dia termasuk kedalam kelas positif dan begitu juga sebaliknya. Diberikan data uji, T= $\{(x_1, y_1),.. (x_n, y_n)\}$, dimana $x_i \in R^n$ adalah pola inputan, dan $y_i \in \{-1,1\}$ adalah label kelas positif dan negatif. Pada penelitian ini akan digunakan SVM Linier untuk mengklasifikasikan data, dimana pada metode ini masalah utamanya adalah menemukan *hyperlane* terbaik dengan meminimalkan nilai *margin* (jarak antara *hyperlane* dengan *pattern* terdekat) dengan menggunakan persamaan 4 berikut.

$$\frac{1}{2}\|w\|^2 = \frac{1}{2}\left(w_1{}^2 + w_2{}^2\right) \qquad (4)$$

Dengan syarat persamaan 5 sebagai berikut.

$$y_i(w.x_i + b) \geq 1, \qquad i = 1,2,3, \ldots \ldots, N \qquad (5)$$

Berdasarkan persamaan 4 dan 5, y_i merupakan kelas dari data ke-i, w adalah normal bidang, dan b adalah posisi alternatif terhadap pusat koordinat. *Hyperplane* yang optimal tersebut akan digunakan untuk memisahkan data yang ada, sehingga data yang masih belum diprediksi dapat diketahui kelasnya dengan melihat disisi mana data tersebut berada. Pada penelitian ini SVM akan menerima input data berupa matriks TF-IDF dari data train yang sudah melalui proses *feature selection* sebelumnya, beserta kelas-kelas sebenarnya dari setiap dokumen untuk dipelajari polanya oleh mesin menggunakan perhitungan metode SVM diatas. Dikarenakan SVM hanya dapat memproses dua buah nilai kelas (1 dan -1), maka pada penelitian ini inputan kelas dirubah menjadi bentuk nilai 0 dan 1 seperti yang direpresentasikan pada Tabel 8.

Tabel 8. Representasi Dokumen Beserta Kelasnya Pada Klasifikasi SVM

No. Doc / No. Kelas	1	2	3	4	5	6	7	8	9	10	…	71
1	1	1	1	1	1	1	1	1	1	1	…	0
2	0	0	0	0	0	0	0	0	0	0	…	0
3	0	0	0	0	0	0	0	0	0	0	…	0
…	…	…	…	…	…	…	…	…	…	…	…	…
12	0	0	0	0	0	0	0	0	0	0	…	1

Dapat dilihat pada tabel 8, dokumen pertama memiliki nilai 1 pada kelas 1 dan nilai 0 pada kelas lainnya, ini artinya dokumen 1 termasuk kedalam kelas Budaya. Sedangkan pada kelas dokumen ke-71, nilai 1 ditemukan pada kelas 12, sehingga dokumen ke-71 termasuk kedalam kelas Wisata. Setelah dipelajari, pola ini akan dijadikan acuan untuk mengklasifikasikan data test yang belum memiliki kelas. Data test yang ingin diklasifikasikan kelasnya, harus terlebih dahulu diubah kedalam bentuk matriks melalui proses TF-IDF. Contoh output dari proses klasifikasi pada penelitian ini dapat dilihat pada Tabel 9, cara membaca data sama dengan cara membaca data pada Tabel 8 yang telah diterangkan sebelumnya.

JURNAL MEDIA INFORMATIKA BUDIDARMA, Vol 3, No 4, Oktober 2019
ISSN 2614-5278 (media cetak)
ISSN 2548-8368 (media online)
Hal 284-292 | DOI: 10.30865/mib.v3i4.1410

Tabel 9. Contoh Output Hasil Klasifikasi SVM

No. Doc / No. Kelas	1	2	3	4	5	6	7	8	9	10	…	71
1	1	1	0	1	0	0	1	1	1	0	…	0
2	0	0	0	0	0	0	0	0	0	1	…	0
3	0	0	1	0	1	0	0	0	0	0	…	0
…	…	…	…	…	…	…	…	…	…	…	…	…
12	0	0	0	0	0	1	0	0	0	0	…	1

2.6 Evaluasi

Tahap terakhir pada penelitian ini adalah proses evaluasi. Pada tahap ini akan dihitung akurasi dari model yang ada dengan membandingkan hasil prediksi yang ada dengan kelas sebenarnya. Perhitungan dilakukan menggunakan persamaan 6 berikut.

$$Accuracy = \frac{Total\ Benar}{N} x\ 100\% \quad (6)$$

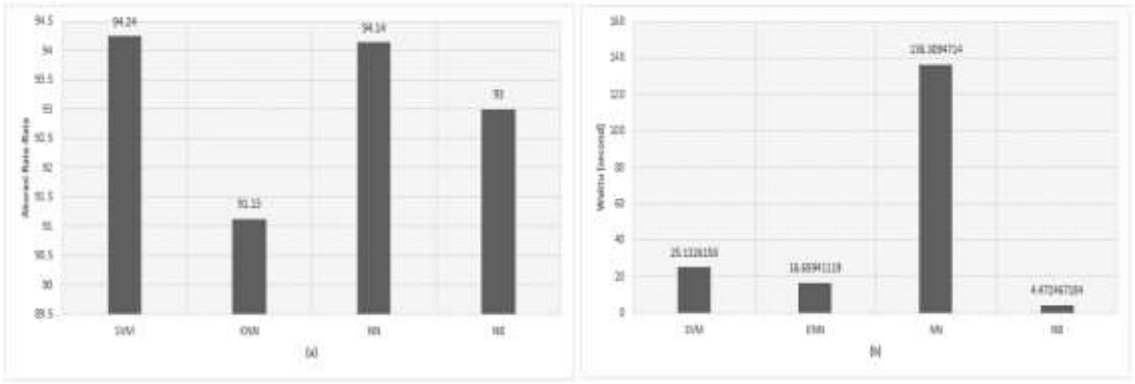
Berdasarkan persamaan 6, total prediksi benar yang dihasilkan oleh sistem akan dibagi dengan jumlah seluruh hasil prediksi dikali dengan 100%. Hasil ini menunjukkan seberapa akurat model yang dibangun dalam memprediksi data berita berbahasa Indonesia.

3. ANALISA DAN PEMBAHASAN

Pada tahap ini dilakukan beberapa skenario pengujian. Penelitian ini memiliki tiga titik fokus yaitu pada tahap *preprocessing*, tahap *feature selection*, dan tahap klasifikasi. Pada penelitian ini juga diterapkan metode *k-fold cross validation* dalam melakukan pengujian, dimana dataset *X* akan dibagi menjadi *k*-bagian dengan jumlah yang sama ($X = x1, x2, x3, \dots, xk$). Untuk setiap fold data akan dibagi menjadi data latih dan data uji, misalnya fold pertama x1 sebagai data uji dan sisanya (x1,x2,x3,…,xk) akan digunakan sebagai data latih. Adapun akurasi yang didapat dengan menghitung rata-rata akurasi yang dihasilkan oleh total semua fold. Penelitian ini menggunakan 5-fold yang diterapkan pada 360 data berita Bahasa Indonesia.

3.1 Pengujian Pertama

Skenario pengujian yang pertama adalah menguji pengaruh proses pada tahap *proceprocessing* terhadap hasil dan kinerja model dengan mengacu pada penggunaan *stopword removal* dan *stemming*. Dalam penelitian ini algoritma *stemming* yang digunakan adalah Nazief-Andriani dengan *library* Sastrawi. Untuk menguji, peneliti menggunakan SVM dengan kernel *linear* sebagai metode klasifikasi dan menggunakan MI sebagai seleksi fiturnya. Hasil akurasi yang ditampilkan merupakan rata-rata dari hasil perhitungan akurasi klasifikasi pada 5-fold. Hasil pengujian dapat dilihat pada Gambar 2.



Gambar 2. (a) Hasil pengujian pengaruh penggunaan *stopword removal* dan *stemming* terhadap hasil akurasi, (b) Hasil pengujian pengaruh penggunaan *stopword removal* dan *stemming* terhadap waktu komputasi sistem

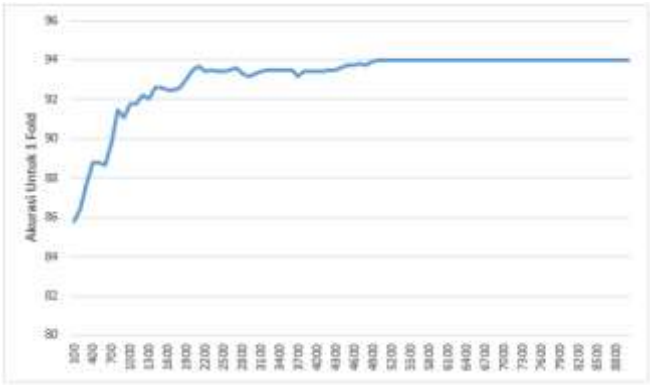
Dapat dilihat dari hasil pengujian pada Gambar 2, nilai akurasi terbaik didapat dari pengujian dengan menggunakan *stemming* namun tidak menggunakan *stopword removal*. Hal ini disebabkan karena, pada proses *stemming* setiap kata yang ada pada data berita diubah ke bentuk kata dasarnya sehingga meminimalisir fitur-fitur yang memiliki makna sama namun karena diberikan akhiran seperti “-lah”,”-kah”,”-nya” dll. awalan, sisipan kata, ataupun imbuhan menjadi memiliki makna yang berbeda. Sedangkan pada proses *stopword* yang dihilangkan hanya kata-kata yang tidak penting atau tidak memiliki makna. Akan tetapi bukan menghapus secara keseluruhan, namun hanya mengurangi jumlah kata yang ada. Sehingga kata yang mengalami proses *stopword removal* seperti kata dan, di, jika, dll. akan banyak digunakan sebagai fitur pada proses MI. Hal ini dapat menyebabkan

JURNAL MEDIA INFORMATIKA BUDIDARMA, Vol 3, No 4, Oktober 2019
ISSN 2614-5278 (media cetak)
ISSN 2548-8368 (media online)
Hal 284-292 | DOI: 10.30865/mib.v3i4.1410

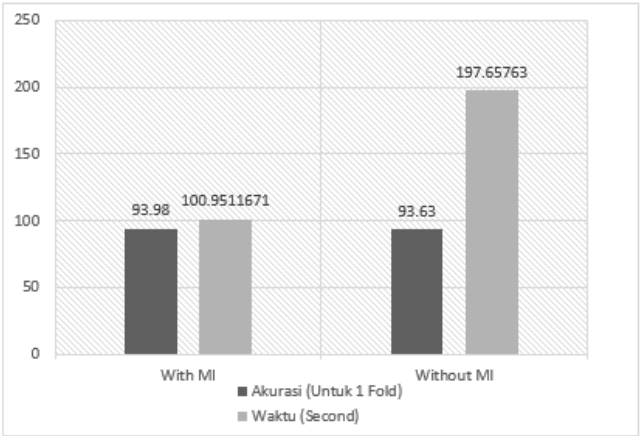
menurunnya hasil akurasi dari sistem, karena saat melakukan klasifikasi fitur yang digunakan tidak menggambarkan suatu kelas. Sementara untuk hasil waktu komputasi tercepat dihasilkan oleh pengujian dengan menggunakan *stopword removal* namun tidak menggunakan *stemming*. Hal ini disebabkan karena semakin sedikit kata yang diproses maka waktu sistem dalam mengolah kata yang ada pun semakin cepat. Serta penghapusan kata pada *stopword removal* dilakukan dalam waktu yang lebih singkat dibandingkan dengan pemotongan kata yang dilakukan pada *stemming*.

3.2 Pengujian Kedua

Skenario pengujian kedua adalah menguji pengaruh dari pengunaan metode *Mutual Information* terhadap hasil dan kinerja model yang dibangun. Dalam pengujian ini, peneliti menggunakan SVM dengan kernel *linear* sebagai metode klasifikasi dan data yang digunakan adalah data dengan *stemming* tanpa *stopword*. Hasil akurasi yang ditampilkan hanya berdasarkan perhitungan akurasi 1-fold yaitu index fold pertama dari 5-fold yang digunakan sebelumnya. Hal ini disebabkan, jika melakukan proses klasifikasi tanpa MI, maka fitur yang digunakan akan sangat banyak, dan waktu komputasi akan sangat tinggi. Jumlah fitur yang digunakan pada pengujian dengan MI adalah sebanyak 5000 sedangkan tanpa MI adalah sebanyak 9043. Hal ini dikarenakan, setelah dilakukan percobaan klasifikasi menggunakan fitur yang telah diurutkan sesuai nilai MI-nya dan jumlah fitur yang digunakan digandakan kelipatan 100 untuk setiap klasifikasi, didapat jumlah fitur yang menghasilkan akurasi terbesar adalah 5000 fitur yaitu sebesar 93.98%. Hasil pengujian dapat dilihat pada Gambar 3 dan 4.



Gambar 3. Hasil pengujian perbandingan jumlah fitur yang digunakan setelah diurutkan berdasarkan nilai *mutual information* terhadap akurasi model klasifikasi.



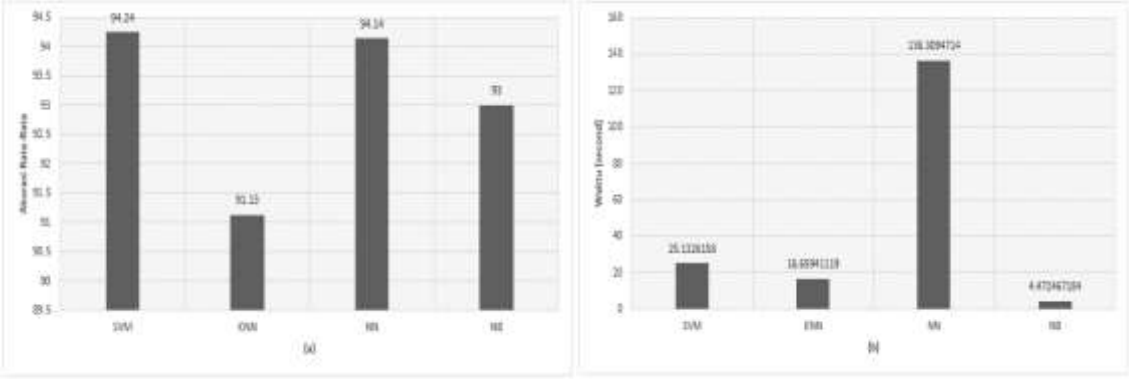
Gambar 4. Hasil pengujian perbandingan penggunaan metode *mutual information* untuk *feature selection* terhadap akurasi dan waktu komputasi system

Dapat dilihat dari hasil pengujian pada Gambar 3, hasil akurasi terbaik diperoleh ketika menggunakan seleksi fitur *Mutual Information* yaitu sebesar 93.81%, dan memiliki waktu komputasi sistem lebih sedikit dibandingkan dengan tidak menggunakan MI. Ini disebabkan karena tanpa adanya seleksi fitur, maka fitur yang digunakan akan sangat banyak dan fitur-fitur tersebut tidak merepresentasikan kelas mereka, ini menyebabkan ketepatan sistem dalam memprediksi jadi berkurang. Sehingga dapat disimpulkan bahwa menggunakan metode ini sebagai seleksi fitur dapat menaikkan efektis dan performansi sistem yang dibangun. Dapat dilihat juga dari hasil pengujian pada Gambar 4 bahwa terdapat titik jenuh, dimana semakin banyak fitur yang digunakan tidak selalu mempengaruhi meningkatnya akurasi klasifikasi. Terdapat titik dimana jumlah fitur yang digunakan sudah cukup mewakili masing-masing kelasnya untuk dilakukan klasifikasi, pada penelitian ini adalah sebanyak 5000 fitur.

JURNAL MEDIA INFORMATIKA BUDIDARMA, Vol 3, No 4, Oktober 2019
ISSN 2614-5278 (media cetak)
ISSN 2548-8368 (media online)
Hal 284-292 | DOI: 10.30865/mib.v3i4.1410

3.3 Pengujian Ketiga

Skenario pengujian ketiga adalah menguji metode klasifikasi yang dibangun untuk membuktikan bahwa metode *Support Vector Machine* adalah metode terbaik untuk mengklasifikasikan berita Bahasa Indonesia. Pengujian ini dilakukan dengan membandingkan hasil akurasi yang dihasilkan oleh SVM (kernel *linear*) dengan metode klasifikasi *Naïve Bayes* (NB), *K-Nearest Neighbor* (KNN) dengan parameter K = 5, dan *Neural Network* (NN) dengan arsitektur *Multi-layer Perceptron* yang menggunakan 100 neuron per-hidden layer. Pengujian dilakukan menggunakan dataset dengan *stemming* tanpa *stopword* serta telah melalui seleksi fitur menggunakan metode MI. Hasil akurasi yang ditampilkan merupakan rata-rata dari hasil perhitungan akurasi klasifikasi pada 5-fold. Hasil pengujian dapat dilihat pada Gambar 5.



Gambar 5. (a) Hasil pengujian perbandingan penggunaan berbagai jenis metode klasifikasi terhadap hasil akurasi, (b) Hasil pengujian pengaruh penggunaan berbagai jenis metode klasifikasi terhadap waktu komputasi

Dapat dilihat dari hasil pengujian pada gambar 4, klasifikasi terbaik dihasilkan oleh metode SVM dengan 94.24% sedangkan metode lain berturut-turut yaitu NN dengan nilai 94.14%, KNN 91.13%, dan NB 93.00%. Selain itu waktu proses komputasi yang dilakukan oleh metode SVM juga jauh lebih kecil dibandingkan dengan metode NN yang memakan waktu cukup banyak. Penelitian ini membuktikan bahwa metode SVM mampu menangani data multi dimensi, ddikarenakan setiap data berita bersifat multi dimensi dimana jumlah dimensi data tergantung dari jumlah fitur yang digunakan. SVM juga memperhatikan fitur, dimana fitur-fitur tersebut memiliki bobot keterhubungan antara kelasnya, sehingga fitur cenderung berada pada salah satu sisi positif atau negatif. Sedangkan metode NN merupakan metode yang mempelajari pola berdasarkan data latih yang digunakan, jadi semakin banyak data latih semakin mudah sistem dalam mengklasifikasikan data. Pada penelitian ini hanya menggunakan 288 data latih disetiap foldnya, sehingga hal ini memungkinkan proses *learning* NN tidak maksimal. Proses learning yang terjadi pada NN juga memakan waktu yang lama karena memerlukan proses *forward pass* dan *backward pass* berulang-ulang untuk mendapatkan nilai weight terbaik dan error sekecil mungkin. Pada metode KNN, hasil akurasi yang dihasilkan relatif kecil dibandingkan dengan metode lainnya dikarenakan KNN tidak mempelajari bobot fitur sehingga tidak mengerti pengaruh fitur dalam penentuan kelas. Metode ini hanya memperhatikan tetangga disekitar k-data latih yang paling dekat jaraknya dengan data uji, sehingga kurang efektif dalam mengklasifikasikan berita yang membutuhkan pengaruh fitur dalam penentuan kelas. Sementara itu metode NB melakukan perhitungan berdasarkan probabilitas kemunculan kata dan tidak memperhatikan pengaruh fitur dalam penentuan kelas. Sehingga jika probabilitas kelas negatif lebih besar dibandingkan kelas positif, maka data tersebut akan dikategorikan sebagai kelas negative. Namun karena perhitungan probabilitas yang tidak memakan waktu yang lama, metode NB menjadi metode dengan sistem komputasi tercepat dibandingkan dengan metode yang lainnya.

4. KESIMPULAN

Berdasarkan hasil dari pengujian-pengujian yang dilakukan dalam penelitian ini, klasifikasi berita Bahasa Indonesia menggunakan *stemming* tanpa *stopword removal*, MI sebagai metode seleksi fitur, dan SVM sebagai metode klasifikasi memberikan hasil prediksi yang paling akurat dibandingkan dengan metode yang lainnya yaitu sebesar 94.24%. *Stemming* dapat meningkatkan efektifitas kinerja sistem dikarenakan dapat meminimalisir perbedaan kata setelah diubah menjadi kata dasar terutama pada dataset berita yang cenderung memiliki sangat banyak kata dan tidak terlalu memperhatikan perbedaan maknanya ketika diberikan imbuhan. *Mutual Information* dapat membantu menyeleksi fitur yang ada sehingga fitur-fitur yang tidak menggambarkan sebuah kelas dapat diabaikan, sehingga performansi sistem akan lebih cepat dan efektif karena tidak perlu menggunakan semua fitur yang ada. SVM merupakan metode yang tepat digunakan dalam memprediksi data teks terutama untuk berita, karena memiliki akurasi yang paling tinggi dibanding metode yang lainnya dengan waktu komputasi yang tidak terlalu lama.

Beberapa saran yang bisa diberikan untuk penelitian selanjutnya adalah penambahan jumlah dataset berita, dimana berita yang ada sudah sangat banyak namun dalam penelitian ini baru menggunakan 360 berita saja, dimana setiap kelas terdiri dari 30 berita. Pengklasifikasian pun sebaiknya dilakukan secara multi-label. Artinya ada kemungkinan berita pada kelas A termasuk juga kedalam kelas B, sehingga pengklasifikasian berita dapat lebih tepat dibandingkan hanya single-label seperti pada penelitian ini. Serta dataset yang digunakan harus diberi label oleh para ahli sehingga label yang diberikan baik single maupun multi-label dapat lebih tepat, untuk dipelajari oleh sistem yang akan dibangun.

REFERENCES

[1] Li-juan Zhu, Feng, Z., Qing-qing, P., Xin, Y., & Zheng-tao, Y. (2015). A classification method of Vietnamese news events based on maximum entropy model. 2015 34th Chinese Control Conference (CCC).

[2] Rizaldy, A., & Santoso, H. A. (2017). Performance improvement of Support Vector Machine (SVM) With information gain on categorization of Indonesian news documents. 2017 International Seminar on Application for Technology of Information and Communication (iSemantic).

[3] Dadgar, S. M. H., Araghi, M. S., & Farahani, M. M. (2016). A novel text mining approach based on TF-IDF and Support Vector Machine for news classification. 2016 IEEE International Conference on Engineering and Technology (ICETECH).

[4] Septian, G., Susanto, A., & Shidik, G. F. (2017). Indonesian news classification based on NaBaNA. 2017 International Seminar on Application for Technology of Information and Communication (iSemantic).

[5] Asy'arie, A. D., & Pribadi, A. W. (2009). Automatic news articles classification in Indonesian language by using Naive Bayes Classifier method. Proceedings of the 11th International Conference on Information Integration and Web-Based Applications & Services - iiWAS '09.

[6] Shahi, T. B., & Pant, A. K. (2018). Nepali news classification using Naïve Bayes, Support Vector Machines and Neural Networks. 2018 International Conference on Communication Information and Computing Technology (ICCICT).

[7] Rahmawati, D., & Khodra, M. L. (2015). Automatic multilabel classification for Indonesian news articles. 2015 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA).

[8] Purbolaksono, M. D., Widiastuti, K. C., Adiwijaya, Mubarok, M. S., & Ma'ruf, F. A. (2018, March). Implementation of mutual information and bayes theorem for classification microarray data. In Journal of Physics: Conference Series (Vol. 971, No. 1, p. 012011). IOP Publishing.

[9] Nurfikri, Fahmi Salman, and Mohamad Syahrul Mubarok. "News Topic Classification Using Mutual Information and Bayesian Network." In 2018 6th International Conference on Information and Communication Technology (ICoICT), pp. 162-166. IEEE, 2018.

[10] Zhili Pei, Yuxin Zhou, Lisha Liu, Lihua Wang, Yinan Lu, & Ying Kong. (2010). A mutual information and information entropy pair based feature selection method in text classification. 2010 International Conference on Computer Application and System Modeling (ICCASM 2010).

[11] Zareapoor, M., and Seeja, K. Feature extraction or feature selection for text classification: A case study on phishing email detection. International Journal of Information Engineering and Electronic Business 7, 2 (2015), 60.Prijono, Agus., Wijaya, Marvin Ch. (2007). Pengolahan Citra Digital Menggunakan MatLAB Image Processing Toolbox. Bandung : Informatika.

[12] Al Faraby, S., Jasin, E.R.R. and Kusumaningrum, A., 2018, March. Classification of hadith into positive suggestion, negative suggestion, and information. In Journal of Physics: Conference Series (Vol. 971, No. 1, p. 012046). IOP Publishing.

[13] Kuang, Q., & Xu, X. (2010). Improvement and Application of TF•IDF Method Based on Text Classification. 2010 International Conference on

[14] Vijayan, V. K., Bindu, K. R., & Parameswaran, L. (2017). A comprehensive study of text classification algorithms. 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI).

[15] Naf'an, M. Z., Bimantara, A. A., Larasati, A., Risondang, E. M., & Nugraha, N. A. S. (2019). Sentiment Analysis of Cyberbullying on Instagram User Comments. Journal of Data Science and Its Applications, 2(1), 88-98.

[16] Asriyanti Indah Pratiwi, Adiwijaya. (2018). On the Feature Selection and Classification Based on Information Gain for Document Sentiment Analysis. Applied Computational Intelligence and Soft Computing, 2018.