JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 2, April 2021, Page 415-421
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i2.2835

Analisis Sentimen Terhadap Review Film Menggunakan Metode *Modified Balanced Random Forest* dan *Mutual Information*

Firdausi Nuzula Zamzami*, Adiwijaya, Mahendra Dwifebri P

Fakultas Informatika, Universitas Telkom, Bandung, Indonesia
Email: [1,*]nuzulaz@students.telkomuniversity.ac.id, [2]adiwijaya@telkomuniversity.ac.id,
[3]mahendradp@telkomuniversity.ac.id
Email Penulis Korespondensi: nuzulaz@students.telkomuniversity.ac.id

Abstrak−Pertukaran informasi saat ini paling banyak terjadi pada internet. Pertukaran informasi dapat dilakukan dengan banyak cara, seperti mengungkapkan ekspresi pada sosial media. Salah satunya adalah me-review sebuah film. Pada saat seseorang melakukan review terhadap sebuah film dia akan menggunakan emosinya untuk mengekspresikan perasaan mereka, itu bisa berupa hal yang positif atau negatif. Pertumbuhan internet yang semakin cepat membuat informasi semakin beragam, banyak, dan tidak terstruktur. Analisis sentiment bisa menangani hal tersebut, karena analisis sentiment merupakan sebuah proses klasifikasi untuk memahami opini, interaksi, dan emosi dari sebuah dokumen atau teks yang dilakukan secara otomatis oleh sistem komputer. Salah satu metode pembelajaran mesin yang cocok adalah Modified Balanced Random Forest. Untuk mengatasi data yang beragam, seleksi fitur yang digunakan adalan Mutual Information. Dengan kedua metode tersebut, sistem mampu mampu menghasilkan nilai akurasi 79% dan nilai F1-scores 75%.

Kata Kunci: Modified Balanced Random Forest; Mutual Information; Analisa Sentimen; Review Film; Klasifikasi

Abstract−Information exchange is currently the most happening on the internet. Information exchange can be done in many ways, such as expressing expressions on social media. One of them is reviewing a film. When someone reviews a film he will use his emotions to express their feelings, it can be positive or negative. The fast growth of the internet has made information more diverse, plentiful and unstructured. Sentiment analysis can handle this, because sentiment analysis is a classification process to understand opinions, interactions, and emotions of a document or text that is carried out automatically by a computer system. One suitable machine learning method is the Modified Balanced Random Forest. To deal with the various data, the feature selection used is Mutual Information. With these two methods, the system is able to produce an accuracy value of 79% and F1-scores value of 75%.

Keywords: Modified Balanced Random Forest; Mutual Information; Sentiment Analysis; Movie Review; Classification

1. PENDAHULUAN

Teknologi terus berkembang internet merupakan sebuah teknologi yang sebagian besar orang memilikinya. Banyak orang menghabiskan waktu untuk melakukan pertukaran informasi pada situs daring. Pertukaran informasi dapat dilakukan dengan banyak cara, seperti mengungkapkan ekspresi pada sosial media[1]. Mengungkapkan ekspresi di internet bisa dilakukan dengan banyak cara, seperti melakukan *review* film.

Pada saat seseorang melakukan *review* terhadap sebuah film dia akan menggunakan emosinya untuk mengekspreksikan perasaan mereka, itu bisa berupa hal yang positif atau negatif[2]. Banyaknya informasi yang berada pada internet menyebabkan data menjadi sangat banyak dan beragam. Hal ini mengakibatkan konsumen sulit memetakan mana informasi yang penting dan tidak[3]. Analisis Sentimen merupakan proses yang paling cocok dalam masalah ini[2]. Analisis Sentimen merupakan sebuah proses klasifikasi untuk memahami opini, interaksi, dan emosi dari sebuah dokumen atau teks[4][1].

Banyak metode penilitan berbeda yang digunakan untuk melakukan pendekatan pada proses klasifikasi sentiment terhadap data document *review* Film, seperti *Support Vector Machine*[4], *Naïve Bayes*[2], *Artifical Neural Network*[5] dan *Random Forest*[6]. Pada penilitan kali ini menggunakan metode *Modified Balanced Random Forest* dalam melakukan klasifikasi sentiment untuk menentukan nilai positif atau negatif dari sebuah dokumen *review* film. Penelitian menggunakan *Modified Balanced Random Forest* karena data dokumen yang dihadapi pada penelitian ini merupakan data yang *imbalanced.* Data yang *imbalanced* akan mempengaruhi performansi dari sistem yang dibangun, oleh karena itu perlu dilakukan sebuah sampling terhadap data yang *imbalanced*[16]. *Modified Balanced Random Forest* melakukan sampling data setiap kali pembangunan *tree* untuk mengatasi data yang *imbalanced*[14]. Pada penelitian sebelumnya *Modified Balanced Random Forest* mampu mempercepat proses klasifikasi dari metode *Random Forest*[14]. Model klasifikasi seperti *Random Forest* sangat bergantung dengan pemilihan parameter yang akan mempengaruhi proses dari klasifikasi sentiment[6]. Pemilihan parameter yang baik akan sulit dilakukan jika data mempunyai dimensi yang tinggi, tidak terstruktur, dan memiliki fitur yang sangat banyak[6]. Seleksi fitur merupakan metode yang sangat cocok untuk menangani masalah tersebut[7]. Penilitian ini menggunakan metode *Mutual Information* untuk melakukan seleksi fitur. Penelitian ini menggunakan *Mutual Information* karena data yang digunakan merupakan teks dengan tipe *categorical data*. Penelitian sebelumnya juga menggunakan fitur seleksi *Mutual Information* terhadap *categorical data* dalam pembangunan sistemnya[15][17]. Pada penelitian sebelumnya *Mutual Information* juga dapat meningkatkan akurasi klasifikasi sekitar 1,7% pada analisis sentiment[15].

Pada paper[1] ini menjelaskan bahwa data terkait review bertambah jutaan setiap harinya. Analisis sentiment mampu menanggulangi masalah tersebut. Pada penilitian ini juga menjelaskan bahwa terdapat tiga level

Submitted: **17/02/2021**; Accepted: **11/03/2021**; Published: **25/04/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 2, April 2021, Page 415-421
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i2.2835

dalam analisis sentiment, yaitu level dokumen, level kalimat, dan level aspek dan entitas. Secara umum proses analisis sentiment terbagi menjadi beberapa tahapan, yaitu (1) data preparation, (2) review analysis, (3) sentiment classification. Penilitian ini juga membahas polaritas dari analisis sentiment. Polaritas pada analisis sentiment dapat dibagi menjadi tiga macam, yaitu binaty approach, multi-class approach, dan contextual or fuzzy approach. Analisis sentiment mempunyai dua metode pendekatan, yaitu machine learning dan lexicon based. Penelitian ini menjelaskan bahwa klasifikasi merupakan peran yang sangat penting dalam analisis sentiment.

Penelitian pada paper[4] menjelaskan bahwa analisis sentiment merupakan cara untuk menentukan opini dari sebuah teks. Pada penelitian ini penulis menggunakan metode Support Vector Machine untuk klasifikasi sentiment. Penelitian ini juga membandingkan beberapa feature extraction seperti TF-IDF, BO, dan TO. Penulis juga menggunakan seleksi fitur Chi-square untuk mengurangi jumlah fitur yang tidak penting. Data yang digunakan adalah 1000 data positif dan negatif dari pang corpus dan 200 data positif dan negatif dari Taboada corpus. Hasil dari perbandingan beberapa metode feature extraction yaitu pada data Taboada corpus dengan menggunakan TF-IDF 81%, BO 88.83%, TO 89.17%. Pada data pang corpus dengan menggunakan TF-IDF 75.17%, BO 87.33%, TO 86.17%.

Pada paper[2] menjelaskan bahwa sekarang sangat mudah untuk mendapatkan feedback terhadap suatu prodak atau film. Feedback yang didapat di internet juga berbagai macam, biasanya berupa sentiment berupa positif atau negatif. Penulis menjelaskan bahwa seleksi fitur pada analisis sentiment dapat digunakan untuk membantu menambah ketepatan dari klasifikasi sentiment. Penelitian ini menggunakan data dari imdb film india dengan Penelitian ini membandingkan berbagai macam seleksi fitur seperti Chi-Square, Gain Ratio, Information Gain, One-R, dan Relief Attribute. Dengan metode klasifikasi Bayesian Classifier hasil dari perbandingannya adalah jika menggunakan evaluasi F-Value performansi terbaik dimiliki oleh One-R yaitu 88.8%, dan jika menggunakan evaluasi FP-Rate performansi terbaik dimiliki oleh Relief Attribute.

Pada paper[7] menjelaskan bahwa terdapat permasalahan data yang terdapat pada movie review saat ini yang menyebabkan analisis sentiment menjadi sangat lambat dan kurang sensitif. Pada penelitian ini penulis melakukan feature selection untuk memilih menggunakan Information Gain untuk menghilangkan fitur yang tidak penting. Penilitian ini menggunakan dataset V2.0 dari Universitas Cornell yang berisi 1000 data positif dan 1000 data negatif. Dengan menggunakan Information Gain mempunyai hasil yang lebih baik daripada metode penilitan sebelumnya. Untuk perbandingan klasifikasinya Information Gain Classifier menghasilkan akurasi sekitar 95%, ini lebih baik dari Support Vector Machine dan Neural Network yang hanya mampu menghasilkan akurasi sekitar 70%.

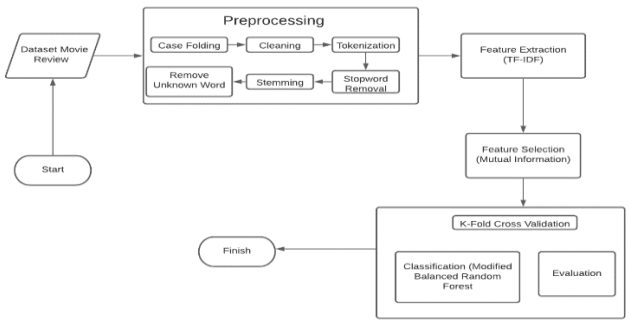
Pada paper[14] menjelaskan bahwa metode klasifikasi Random Forest mempunyai komputasi yang sangat lama untuk data yang besar dan mempunyai data yang imbalanced. Penelitian ini melakukan peningkatan pada metode klasifikasi Random Forest dengan cara melakukan sebagian besar under-sampling data berdasarkan dari clustering, nantinya data under-sampling akan merepresentasikan semua data untuk pembangunan tree dan data. Metode ini disebut Modified Balanced Random Forest. Metode ini mampu mengurangi kelemahan dari under-sampling , dimana sering terjadi pemilihan data penting yang tidak dilakukan input pada saat menjalankan algoritma Random Forest. Dalam penilitian ini pengukuran performansi dilakukan dengan menggunakan AUC, didapat Modified Balanced Random Forest mempunyai nilai AUC 93.51%, nilai ini lebih besar daripada Random Forest karena hanya mampu mempunyai nilai sebesar 78,26%.

Pada paper[15] mengangkat permasalahan komentar terkait pariwisata Lombok di media sosial twitter. Data yang digunakan pada penelitian ini berjumlah 500 data dan menggunakan Bahasa Inggris. Penelitian ini bertujuan untuk membangun sistem analisis sentimen menggunakan metode klasifikasi Naïve Bayes, dan Mutual Information. Pada penelitian ini, metode seleksi fitur Mutual Informaiton mempu mereduksi waktu klasifikasi sebesar 51.52%, dan meningkatkan akurasi sebesar 1.7%.

2. METODOLOGI PENELITIAN

2.1 Rancangan Sistem

Penilitian ini memiliki beberapa tahapan yang perlu dikerjakan seperti pada gambar.



Gambar 1. Gambaran Sistem

Submitted: **17/02/2021**; Accepted: **11/03/2021**; Published: **25/04/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 2, April 2021, Page 415-421
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i2.2835

2.2 Dataset

Pada penelitian ini dataset yang digunakan adalah data *review* film berbahasa Inggris yang diambil dari situs IMDb. Terdapat 5000 jumlah *review* yang terdapat pada dataset, yang terdiri dari 4000 data dengan label negatif dan 1000 data dengan label positif. Tahapan pertama yang dilakukan adalah dengan melakukan pembersihan data atau dengan kata lain membuat data menjadi lebih terstruktur agar lebih mudah diolah oleh sistem.

2.3 *Preprocessing*

Preprocessing merupakan tahapan pertama dalam pembangunan sistem analisis sentimen. Pada tahap ini data terlebih dahulu disiapkan agar proses selanjutnya seperti klasifikasi lebih baik saat digunakan. *Preprocessing* mempunyai beberapa tahapan, pada penilitian ini tahapan yang dilakukan adalah *case folding, cleaning, tokenization, stopword removal, stemming* dan, *remove unknown word*. *Case folding* bertujuan untuk mengubah semua huruf menjadi huruf kecil[9], *cleaning* bertujuan untuk menghilangkan tanda baca atau simbol[12], *tokenization* yaitu bertujuan untuk memotong sebuah kalimat dengan cara memisahkan setiap kata yang terdapat pada kalimat tersebut[9], *stopword removal* bertujuan untuk menghilangkan kata yang tidak penting dan berada pada dokumen[9], *stemming* merupakan proses untuk mengubah kata menjadi sebuah kata dasar[9], dan *remove unknown word* bertujuan untuk menghilangkan kata yang tidak mempunyai makna tersirat pada bahasa inggris seperti br.

2.4 *Feature Extraction* TF-IDF

Feature Extraction merupakan metode untuk mengidentifikasi fitur yang berada pada sebuah dokumen[10]. Model *Feature Extraction* yang digunakan dalam penelitian ini adalah *N-gram*. *N-gram* merupakan teknik untuk memecah kalimat menjadi *N* kata. pada penelitian ini *N* yang digunakan adalah satu atau dalam istilah disebut *unigram.* Setelah dilakukan pemecahan kalimat, selanjutnya adalah dilakukan pembobotan setiap fitur. Pada penilitian ini, metode yang digunakan dalam pembobotan fitur adalah *TF-IDF*. *Term Frequency Inverse Document Frequency(TF-IDF)* merupakan salah satu metode untuk menghitung pembobotan setelah dilakukan proses *feature extraction.* Pembobotan menggunakan *TF-IDF* dilakukan dengan cara mengukur seberapa banyak sebuah kata muncul pada dokumen[4]. *TF-IDF* mempunyai dua tahapan saat pengerjaanya. Pertama harus dilakukan perhitungan pada *TF* yaitu menghitung frekuensi kemunculan *term* pada setiap dokumen. Setelah dilakukan perhitungan pada nilai *TF*, proses selanjutnya adalah melakukan perhitungan *IDF* yaitu proses perhitungan bobot *term* yang terdapat pada seluruh dokumen. Selanjutnya melakukan perhitungan bobot dari *term* dengan menggunakan rumus *TF-IDF* pada persamaan 1

$$TF\ IDF\ (tk, dj) = TF(tk, dj) * log\ (\frac{N}{dft}) \qquad (1)$$

Nilai *TF IDF* pada persamaan 1 menyatakan perhitungan dari frekuensi kemunculan *term* pada setiap dokumen, nilai *N* menyatakan jumlah dari semua dokumen dan *dft* menyatakan banyaknya dokumen yang mengandung *term.*

2.5 *Feature Selection* Mutual Information

Setelah melakukan ekstraksi data, proses selanjutnya adalah seleksi fitur menggunakan *Mutual Information. Mutual Information* melakukan perhitungan dengan mengukur ada atau tidaknya informasi dan kontribusi dari sebuah term[15][17]. Perhitungan dapat dilakukan dengan menggunakan rumus pada persamaan 2.

$$I(U, C) = \sum_{e_c \in \{0,1\}} \sum_{e_i \in \{0,1\}} P(U = e_i, C = e_c)\ \log_2 \frac{P(U = e_i,\ C = e_c)}{P(U = e_i)\ P(\ C = e_c)} \qquad (2)$$

Dimana nilai *P(U,C)* merupakan *join probability*, nilai *P(U)* dan *P(C)* merupakan *marginal probability*[15]. Nilai U pada penelitian ini merujuk ke fitur pada sebuah dokumen dan nilai C merujuk ke arah label kelas.

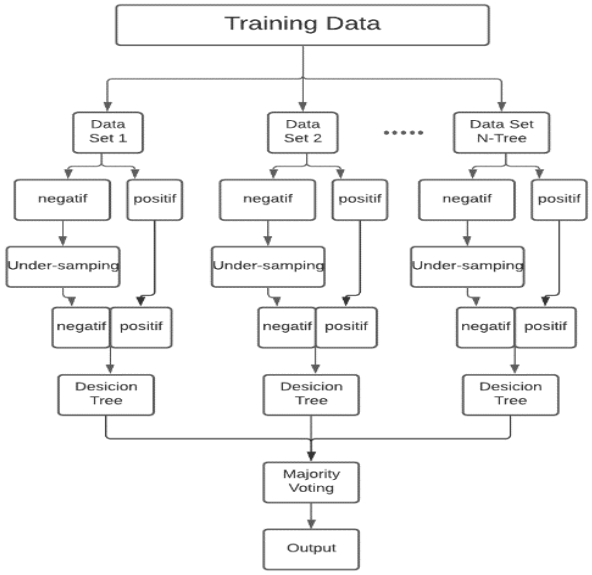
2.5 Klasifikasi Modified Balanced Random Forest

Selanjutnya data akan dipisah secara acak menjadi data *training* 80% dan data *testing* 20% menggunakan *Cross Validation* sebanyak 5 lipatan. Setiap iterasi data akan diklasifikasikan dengan mengunakan *Modified Balanced Random Forest*.

Metode *Modified Balanced Random Forest* merupakan modifikasi dari metode *Random Forest*. *Modified Balanced Random Forest* melakukan *under-sampling* dengan cara menggabungkannya dengan ide dari *ensemble learning*[14]. *Under-sampling* dilakukan dengan cara mengambil sebagian data *majority class* berdasarkan dari dataset *movie review* untuk setiap pembuatan *tree*. Pada penilitian ini metode sampling yang digunakan menggunakan *random under sampling.* Nantinya data tersebut digunakan untuk membangun *tree* yang akan merepresentasikan semua data.

Submitted: **17/02/2021**; Accepted: **11/03/2021**; Published: **25/04/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 2, April 2021, Page 415-421
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i2.2835



Gambar 2. Alur kerja *Modified Balanced Random Forest*\

Untuk cara kerjanya, pertama metode *Modified Balanced Random Forest* akan melakukan sampling sejumlah N data untuk menentukan data mana yang akan digunakan sebagai *tree*. Setelah itu data dengan label dengan kelas *majority* akan dilakukan *under-sampling*, dan data itu akan digabungkan kembali untuk dijalankan menggunakan algoritma *Decision Tree*. Setelah algoritma *Decision Tree* berikut mengeluarkan output maka akan dilakukan *majority voting*, yaitu mengambil kelas yang muncul terbanyak.
Tahap akhir dalam proses sistem ini adalah evaluasi, Pada penilitian ini, evaluasi bertujuan untuk mengukur performansi dari klasifikasi *modified balanced random forest* menggunakan pengukuran *accuracy*, *precission*, *recall*, dan *f1-scores.* Ke-empat pengukuran tersebut dapat dilakukan perhitungan menggunakan *confussion matrix.*

3. HASIL DAN PEMBAHASAN

3.1 Skenario Pengujian

Terdapat tiga fokus utama dalam pengujian ini. Skenario pertama adalah mengukur pengaruh *stemming* pada dataset *movie review* berbahasa Inggris, lalu pada skenario pengujian kedua adalah untuk mengukur mengukur pengaruh dari seleksi fitur *Mutual Information* terhadap klasifikasi *Modified Balanced Random Forest.* Lalu skenario terakhir adalah melakukan perbandingan performansi dari metode *Modified Balanced Random Forest* dan *Random Forest* menggunakan seleksi fitur.

3.2 Hasil Pengujian Skenario 1

Pengujian ini bertujuaan agar dapat mengetauhi pengaruh *stemming* terhadap performansi sistem. Akan dilihat pengaruh dari *stemming* terhadap performansi dari *Modified Balanced Random Forest.* Hasilnya didapat:

Tabel 1. Perbandingan performansi *Modified Balanced Random Forest* menggunakan *stemming* dan tanpa menggunakan *stemming*

MBRF	K	Accuracy	F1-Score
Menggunakan Stemming	1	78%	72%
	2	77%	73%
	3	**79%**	**74%**
	4	75%	71%
	5	77%	71%
Tanpa Menggunakan Stemming	1	76%	70%
	2	77%	73%
	3	**78%**	**73%**
	4	76%	71%
	5	76%	70%

Berdasarkan tabel 1, performansi pada *Modified Balanced Random Forest* mengalami peningkatan 1% dengan menggunakan *stemming* dari pada tidak menggunakan *stemming.* Menggunakan *stemming Modified*

Submitted: **17/02/2021**; Accepted: **11/03/2021**; Published: **25/04/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 2, April 2021, Page 415-421
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i2.2835

Balanced Random Forest memiliki akurasi sebesar 79% dan *f1-score* sebesar 74% sedangkan tanpa menggunakan *stemming* hanya mampu menghasilkan akurasi 78% dan *f1-score* sebesar 73%.

4.3 Hasil Pengujian Skenario 2

Pengujian pada skenario 2 dilakukan untuk mengetahui pengaruh dari seleksi fitur terhadap dataset *movie review*. Penulis terlebih dahulu menentukan *treshold* untuk *mutual information* yaitu 0.01. Setelah itu didapatkan bahwa *Mutual Information* menghilangkan jumlah fitur sebanyak lebih dari 50% pada dataset yang menggunakan *stemming*.

Tabel 2. Perbedaan fitur sebelum dan sesudah seleksi fitur

	Sebelum Mengunakan Seleksi Fitur	Setelah Mengunakan Seleksi Fitur
Jumlah Fitur	35770	17178

Dengan menggunakan seleksi fitur *mutual information* didapatkan lima fitur yang mempunyai nilai *mutual information* terbesar.

Tabel 3. Daftar 5 kata dengan nilai MI tertinggi

Daftar Kata	Nilai MI
bad	0.031
toplevel	0.021
capich	0.021
shop	0.021
perfect	0.021

Kelima kata diatas memiliki nilai MI terbesar, artinya kelima kata tersebut masuk kedalam fitur yang dipilih, sedangkan kata dengan nilai MI kurang dari 0.01 akan dihapus dari dataset. Setelah mendapatkan fitur yang akan dipilih, selanjutnya melihat perbandingan model *Modified Balanced Random Forest* dengan menggunakan seleksi fitur dan tanpa menggunakan seleksi fitur.

Tabel 4. Perbandingan performansi *Modified Balanced Random Forest* menggunakan seleksi fitur dan tanpa seleksi fitur

Classifier	k	Precision	Recall	F1-Score	Accuracy
MBRF Menggunakan Seleksi Fitur	1	70%	80%	72%	78%
	2	72%	81%	73%	77%
	3	**73%**	**82%**	**74%**	**79%**
	4	71%	80%	71%	75%
	5	70%	80%	71%	77%
MBRF Tanpa Menggunakan Seleksi Fitur	1	71%	80%	71%	76%
	2	**73%**	**82%**	**74%**	**78%**
	3	71%	79%	72%	77%
	4	71%	80%	71%	75%
	5	70%	80%	71%	77%

Modified Balanced Random Forest yang menggunakan seleksi fitur mempunyai nilai *f1-scores* 74% sedangkan dengan tanpa seleksi fitur mempunyai nilai *f1-scores* 74%, namun pada akurasi, MBRF tanpa menggunakan seleksi fitur mengalami penurunan sebesar 1%

3.4 Hasil Pengujian Skenario 3

Setelah mendapatkan hasil dari pengaruh seleksi fitur *Mutual Information* selanjutnya adalah melihat bagaimana perbedaan antara *Modified Balanced Random Forest* dengan *Random Forest*. Didapatkan hasilnya :

Tabel 5. Ilustrasi perbandingan performansi Modified Balanced Random Forest dan Random Forest menggunakan seleksi fitur

Classifier	K	Precision	Recall	F1 Score	Max F1 - Score
RF	1	91%	51%	46%	49%
	2	89%	52%	47%	
	3	90%	51%	46%	
	4	**86%**	**52%**	**49%**	
	5	91%	50%	45%	
MBRF	1	70%	80%	72%	74%
	2	72%	81%	73%	
	3	**72%**	**81%**	**74%**	

Submitted: **17/02/2021**; Accepted: **11/03/2021**; Published: **25/04/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 2, April 2021, Page 415-421
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i2.2835

Classifier	K	Precision	Recall	F1 Score	Max F1 - Score
	4	71%	80%	71%	
	5	70%	80%	71%	

Performansi model klasifikasi *Modified Balanced Random Forest* memiliki nilai *f1-scores* terbaik 74% sedangkan *Random Forest* hanya 49%. Hal tersebut dikarenakan jumlah label kelas yang tidak seimbang atau *imbalanced* menyebabkan *Random Forest* memiliki nilai *f1-scores* yang kecil.

3.5 Analisis Hasil Pengujian

Pada skenario pengujian yang pertama penggunaan *stemming* mempunyai pengaruh terhadap performansi *Modified Balanced Random Forest.* Nilai akurasi terbaik pada pengujian ini adalah 79% dengan nilai *f1-score* 74%, yaitu dengan menggunakan *stemming*. Berbeda dengan pengukuran tanpa menggunakan *stemming* nilai akurasi terbesar yang dihasilkan sebesar 78% dan *f1-score* 73%, terjadi peningkatan 1% untuk akurasi dan *f1-score*. Hal tersebut dikarenakan dengan menggunakan *stemming* mengubah kata menjadi ke kata dasarnya, ini memungkinkan ada dua atau lebih kata yang mempunyai makna yang berbeda menjadi kata dasar yang sama. Sehingga menambahkan nilai akurasi dan *f1-score.*
Pada skenario pengujian yang kedua dilakukan agar mengetahui pengaruh dari seleksi fitur *Mutual Informationi* terhadap data yang telah dilakukan *preprocessing* pada skenario sebelumya.

Tabel 6. Perbandingan performansi pengujian terbaik pada MBRF

Classifier	Selection Feature	Precision	Recall	F1 Score	Accuracy
MBRF	Mutual Information	73%	82%	74%	79%
MBRF	-	73%	82%	74%	78%

Tabel 6 diatas diambil berdasarkan nilai akurasi terbesar dari kelima pengujian menggunakan klasifikasi *modified balanced random forest.* Nilai *modified balanced random forest* dengan seleksi fitur memiliki nilai performansi yang lebih tinggi dari pada tanpa menggunakan seleksi fitur. Menggunakan seleksi fitur dan tanpa menggunakan seleksi fitur mempunyai nilai *f1-score* yang sama yaitu 74%. Namun dengan menggunakan seleksi fitur memiliki nilai akurasi yang lebih besar yaitu 79%. Hal tersebut dikarenakan beberapa fitur yang dianggap tidak penting sudah dihapus sebelumnya saat seleksi fitur dilakukan, ini menyebabkan proses *under-sampling* untuk tiap pembangunan *tree* menjadi lebih efisien karena data yang tersisa merupakan data yang penting, sehingga menyebabkan nilai akurasi saat menggunakan seleksi fitur bertambah 1%.Jika dilihat dari waktu *running* program saat pembangunan model, kecepatan waktu *running* saat menggunakan seleksi fitur jauh lebih cepat daripada tanpa menggunakan seleksi fitur. Hal tersebut dikarenakan jumlah data yang diproses pada model klasifikasi berkurang hampir 50% saat menggunakan seleksi fitur.

Hasil dari pengujian skenario ketiga bertujuan untuk melihat perbandingan antara penggunaan model klasifikasi *Modified Balanced Random Forest* dengan *Random Forest.* Berdasarkan pada tabel 5, pengujian ini menunjukan hasil bahwa *Modified Balanced Random Forest* memiliki performansi yang lebih baik. Nilai *recall Modified Balanced Random Forest* memiliki nilai 82% berbeda dengan *Random Forest* yang hanya 52%. Hal tersebut dikarenakan data yang tidak *balanced* menyebabkan nilai *recall* yang kecil pada *Random Forest*. Dengan nilai presisi yang sangat tinggi namun nilai *recall* yang sangat rendah, *Random Forest* hanya memiliki nilai *f1-*score rendah, yaitu hanya sebesar 49% berbeda dengan *Modified Balanced Random Forest* yang mempunyai nilai 75%. Itu memungkinkan model yang dibangun menggunakan *Random Forest* mempunyai nilai yang *overfit.*
Modified Balanced Random Forest memiliki performansi yang lebih baik karena pada setiap saat pembentukan *tree* pada proses klasifikasi *modified balanced random forest*, nilai *majority class* pada kasus ini merupakan kelas dengan label negatif diambil dan dihapus secara acak agar jumlahnya menjadi seimbang dengan *minority class*, berbeda dengan *random forest* yang langsung membangun pohon dengan dataset yang *imbalanced*. Hal tersebut menjelaskan alasan *Random Forest* memiliki nilai *f1-score* yang kecil.

4. KESIMPULAN

Berdasarkan hasil penilitian dan analisa yang telah dilakukan, maka dapat diambil kesimpulan bahwa penggunaan stemming pada tahap preprocessing meningkatkan performansi pada sistem. Lalu seleksi fitur Mutual Information juga dapat mengurangi fitur yang kurang relevan untuk digunakan proses klasifikasi Modified Balanced Random Forest dan meningkatkan performansi proses klasifikasi, dengan nilai akurasi tertinggi 79% dan nilai F1-Scores tertinggi 74%. Kemudian untuk klasifikasi Modified Balanced Random Forest memiliki kinerja yang sangat baik untuk dataset movie review berbahasa inggris. Modified Balanced Random Forest mampu meningkatkan nilai f1-scores dari random forest sebesar 27% untuk dataset moview review berbahasa inggris yang imbalanced. Pada tahap preprocessing penggunaan metode stemming mampu meningkatkan 1% nilai f1-score. Beberapa saran untuk penelitian selanjutnya seperti Mengganti metode under-sampling yang dilakukan secara acak terhadap majority class untuk modified balanced random forest menjadi metode lain dan menambahkan jumlah dataset yang pada dataset movie review berbahasa inggris agar mendapatkan performansi yang lebih baik.

Submitted: **17/02/2021**; Accepted: **11/03/2021**; Published: **25/04/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 2, April 2021, Page 415-421
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i2.2835

REFERENCES

[1] Shivaprasad, T.K., Shetty, J. (2017). " Sentiment Analysis of Product Reviews: A Review ". *International Conference on Inventive Communication and Computational Technologies (ICICCT 2017).* 298-303
[2] Trivedi, S.K., Tripathi, A. (2016). "Sentiment Analyis of Indian Movie Review with Various Feature Selection Techniques". *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*. 181-185.
[3] Amolik, A., Jivane, N., Bandhari, M., Venkatesan. (2016). "Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques". *International Journal of Engineering and Technology (IJET).* 7(6). 2038-2043.
[4] Zainuddin, N., Selamat, A.(2014). "Sentiment Analysis Using Support Vector Machine". *2014 International Conference on Computer, Communications, and Control Technology (I4CT).* 333-337.
[5] Sharma, A., Dey, S. (2012). "An Artificial Neural Network Based Approach for Sentiment Analysis of Opinionated Text". *Proceedings of the 2012 ACM Research in Applied Computation Symposium.* 37-42.
[6] Parmar, H., Bhanderi, S., Shah, G. (2014). " Sentiment Mining of Movie Reviews using Random Forest with Tuned Hyperparameters ". *International Conference on Information Science.* 1-6.
[7] Pratiwi, A.I., Adiwijaya. (2018). "On the Feature Selection and Classification Based on Information Gain for Document Sentiment Analysis". *Hindawi Applied Computational Intelligence and Soft Computing Volume 2018*. 1-5.
[8] Hedge, Y., Padma, S.K. (2017). " Sentiment Analysis using Random Forest Ensemble for Mobile Product Reviews in Kannada". *2017 IEEE 7th International Advance Computing Conference.* 777-782.
[9] Mubarok, M.S., Adiwijaya, Aldhi, M.D. (2017). "Aspect-based sentiment analysis to review products using Naïve Bayes". *International Conference on Mathematics: Pure, Applied and Computation: Empowering Engineering using Mathematics*. 1-8.
[10] Ahmad, I.S., Bakar, A.A., Yaakub, M.R. (2019). " A review of feature selection in sentiment analysis using information gain and domain specific ontology ". *International Journal of Advanced Computer Research*. 9(44). 283-292
[11] Mediamer, G., Adiwijaya, Faraby, S.A. (2019). "Development of Rule-Based Feature Extraction in Multi-label Text Classification". *International Journal on Advanced Science, Engineering and Information Technology*. 9(4). 1460-1465
[12] Bakar, M.Y.A., Adiwijaya, Faraby, S.A. (2018). "Multi-Label Topic Classification of Hadith of Bukhari (Indonesian Language Translation)Using Information Gain and Backpropagation Neural Network". *2018 International Conference on Asian Language Processing (IALP)*. 344-350
[13] Suyanto. 2018. *Machine Learning Tingkat Dasar dan Lanjut.* Bandung: Informatika Bandung.
[14] Agusta, P.A., Adiwijaya. (2019). "Modified balanced random forest for improving imbalanced data prediction". *International Journal of Advances in Intelligent Informatics.* 5(1). 58-65.
[15] Ulfa, M.A., Irmawati, B., Husodo, A.Y. (2018). "Twitter Sentiment Analysis using Naïve Bayes Classifier with Mutual Information Feature Selection". *Journal of Computer Science and Informatics Engineering.* 2(2). 106-111
[16] Kotsiantis, S., Pintelas, P.E. (2005). "Handling imbalanced datasets: A review*". International Transactions on Computer Science and Engineering*. 30
[17] Jones, G., Xu, Y., Li, J., Wang, B., Sun, C. (2007). "A Study on Mutual Information-based Feature Selection for Text Categorization". *Joutnal of Computational Information Systems.* 1007 – 1011.
[18] M. D. Purbolaksono, F.D. Reskyadita, Adiwijaya,A. A. Suryani and A. F. Huda,"Indonesian Text Classification using Back Propagation and Sastrawi Stemming Analysis with Information Gain for Selection Feature," *International Journal on Advanced Science, Engineering and Information Technology, vol. 10, no. 1, pp. 234-238, 2020.*