Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 5 Nomor 2, September 2021, pp. 727-735
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

Extractive Text Summerization Pada Berita Berbahasa Indonesia Menggunakan Algoritma *Support Vector Machine*

Thalita Meisya Permata Aulia[1], Asep Jamaludin[2], Tesa Nur Padilah[3]
[123]Universitas Singaperbangsa Karawang
Jl. HS.Ronggo Waluyo, Puseurjaya, Kec. Karawang, Indonesia
thalita.meisya17204@student.unsika.ac.id

Abstract

According to the Program for International Student Assessment (PISA) for the 2018 survey of 61 countries that participated in PISA, the reading interest of the Indonesian people still received a low score of 358 out of an overall average score of 472. One of the consequences of low reading is the difficulty of understanding the content of reading, especially for long and many texts, so it will be easier to read the summary. With advances in text summarization technology can be done using text mining methods. text mining will retrieve information on big data from text-based documents, the summary process will take the main points of news or important sentences without changing the content of the reading or also called extraction techniques. To get maximum results, the weighting is done by extracting sentence features based on numerical data, quotations, sentence length, sentence position in paragraphs, and overall sentence position. The research methodology uses knowledge discovery in database (KDD) and modeling using support vector machine algorithms. Testing or evaluation using recall, precision and F-measure. The best research result is the scenario of comparison of test data and training data 7:3, using the Linear kernel, with accuracy 72,4%, precision 63,4%, recall 51,9%, and F-measure 57,1%.

Keywords*: Summerization text, Extraction Technique, Support Vector Machine, Knowledge Discovery in Database*

Abstrak

Menurut Programme for Internasional Student Assessment (PISA) untuk survei tahun 2018 dari 61 negara yang mengikuti PISA, minat baca masyarakat indonesia masih mendapatkan score rendah yaitu 358 dari nilai rata-rata keseluruhan 472. Salah satu akibat dari rendahnya membaca adalah sulitnya memahami isi bacaan terutama untuk teks yang panjang dan banyak, sehingga akan lebih mudah jika membaca ringkasannya. Dengan kemajuan teknologi peringkasan teks dapat dilakukan dengan menggunakan metode text mining. text mining akan mengambil informasi pada data besar dari dokumen berbasis teks, proses peringkasan akan mengambil poin utama berita atau kalimat penting tanpa mengubah isi bacaan atau disebut juga dengan teknik ekstraksi. Untuk mendapatkan hasil yang maksimal pembobotan dilakukan dengan ekstraksi fitur kalimat berdasarkan data numerik, kutipan, panjang kalimat, posisi kalimat dalam paragraf, posisi keseluruhan kalimat. Metodologi penelitian menggunakan knowledge discovery in database (KDD) dan pemodelan menggunakan algoiritma support vector machine. Pengujian atau evaluasi menggunakan recall, precision dan F-measure. Hasil penelitan yang terbaik adalah dengan skenario perbandingan data uji dan data latih 7:3, menggunakan kernel Linear, dengan hasil accuracy 72,4%, precision 63,49%, recall 51,9%, dan F-measure 57,1%.

Kata kunci*: Teks Summerisasi, Teknik Ekstraksi, Mesin Vektor Dukungan, Penemuan Pengetahuan dalam Basis Data*

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 5 Nomor 2, September 2021, pp. 727-735
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

1. PENDAHULUAN

Indonesia masih memiliki minat baca rendah berdasarkan data dari *Programme for International Student Assessment* (PISA) untuk tahun 2000 - 2018. Hasil survei terbaru dari 61 negara yang mengikuti PISA pada tahun 2018, minat baca Indonesia mendapatkan score sebesar 358 dari nilai rata-rata keseluruhan 472[1]. Akibat dari rendahnya minat membaca adalah sering sekali ditemukan masyarakat kesulitan untuk memahami isi teks bacaan terutama jika teks panjang dan banyak, maka cara untuk dapat memahami isi dokumen dengan cepat dan tepat adalah dengan membaca ringkasannya[2]. Dengan kemajuan teknologi, peringkasan teks dapat dilakukan secara otomatis dengan metode penambangan teks. Peringkasan dilakukan untuk mengambil poin utama pada berita sehingga dapat dengan mudah untuk dipahami.

Text mining atau penambangan teks sangat populer akhir-akhir ini, karena begitu banyak data text yang bisa didapatkan dan diolah, misalnya adalah data pada jejaring sosial berupa *tweet* pada Twitter, komentar di sosial media Youtube, Instagram, berita online dan lain-lain. *Text mining* sendiri merupakan proses penggalian informasi bermakna dari data atau dokumen teks[3]. Salah satu yang menarik adalah peringkasan teks otomatis atau *summarization* teks. *Summarization* teks menjadi hal yang menarik untuk dipelajari lebih lanjut karena bertambahnya jumlah data tekstual di internet meningkat dari berbagai macam artikel berita, karya ilmiah, dokumen hukum, dan lain-lain [4]. *Summarization* teks dibedakan menjadi ekstraktif, abstraktif dan hibrida. Pendekatan abstraktif akan mengambil kalimat inti pada dokumen teks dari sumber, kemudian dibuat menjadi kalimat baru yang berbeda namun memiliki inti yang sama. Pendekatan hibrid menggabungkan teknik abstraksi dan ekstraksi, dimana teknik ekstraksi akan memilih kalimat penting dari dokumen sumber kemudian digabungkan menjadi paragraf baru yang lebih pendek [5].

SVM diperkenalkan oleh Vapnik sebagai model Machine learning untuk regresi dan klasifikasi dengan berbasiskan kernel. SVM memiliki kemampuan generalisasi yang cukup tinggi, bahkan jika menggunakan data latih yang sedikit, SVM juga memiliki ruang dimensi yang tinggi dari ruang input, namun jika data bersifat non linear, maka SVM harus menggunakan ruang kernel [6]. *Support Vector Machine* (SVM) memiliki dua buah kelas pemisah, atau disebut *hyperplane. Hyperplane* dibagi lagi menjadi dua yaitu *hyperlane linear* dan *non linear*. *Hyperplane non linear* akan ditransformasi ke ruang fitur yang memiliki dimensi lebih tinggi, kemudian data akan terpisah seperti *hyperplane linear.* Beberapa *kernel* yang umum digunakan adalah *linear kernel, polynomial kernel, sigmoid kernel* dan RBF *kernel.*

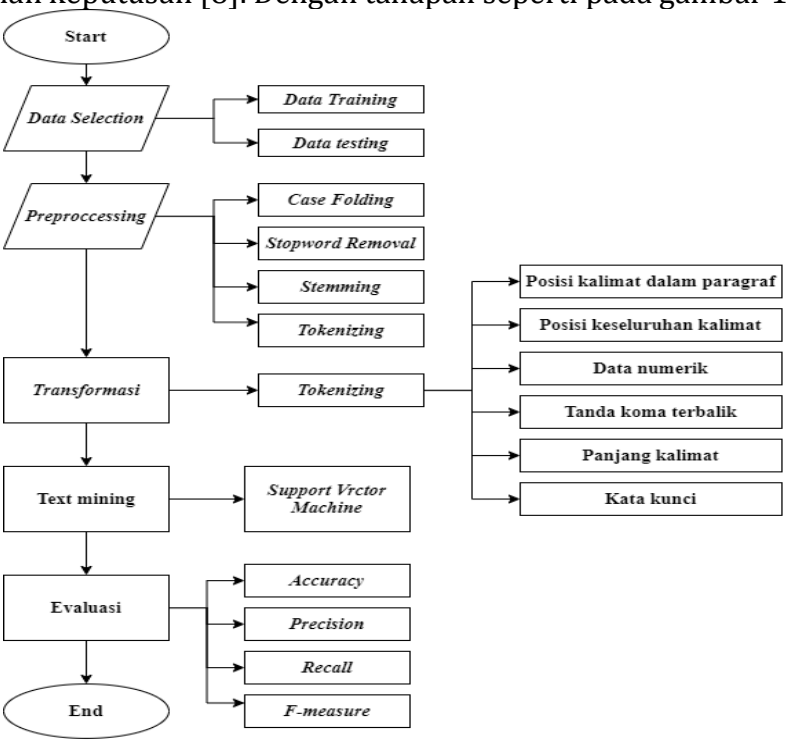
Tidak hanya algoritma namun dalam penelitian juga memerlukan tahapan *preprocessing* yang baik. Pada kasus *text mining* proses *preprocessing* sangat penting, tahapan yang umumnya dilakukan adalah *case folding, stopword removal, tokenization, filtering, stemming.* Kemudian pengukuran

evaluasi yang umum digunakan dan dianggap mengukur hasil klasifikasi dengan baik, selain *F-measure* dan akurasi ada cara untuk mengukur hasil klasifikasi yang lain, diantanya *accuracy, recall, precision,* dan *rand accuracy*[7]. *Precision* akan menghitung hasil tingkat ketepatan, sedangkan *recall* akan menghitung nilai keberhasilan ringkasan pada sistem peringkasan teks. Maka dari itu nilai *F-measure* akan bergantung pada nilai *recall* dan *precision.*

Data yang digunakan merupakan dokumen berita yang diambil dari website Radar Karawang. Harian Umum Radar Karawang adalah surat kabar lokal terbesar yang berada di Kabupaten Karawang, Provinsi Jawa Barat yang dikelola oleh tenaga muda *profesional* sehingga menempatkan koran ini sebagai *market leader* di wilayah Kabupaten Karawang. Untuk mempertahankan eksistensinya di tengah banyaknya media digital, Radar Karawang mengunggah berita secara *online* yang dapat diakses gratis dan berbasis digital pada halaman *website* radarkarawang.id.

2. METODOLOGI PENELITIAN

Menjelaskan Metodologi penelitian yang digunakan adalah *Knowledge Discovery in Database* (KDD). KDD atau *Knowledge Discovery in Database* merupakan salah satu metode yang digunakan untuk proses *data mining,* dalam sebuah *database* biasanya ditemukan tabel-tabel yang saling terhubung, pengetahuan yang didapat dalam proses tersebut merupakan *knowledge base* atau berbasis pengetahuan, yang kemudian digunakan dalam pengambilan keputusan [8]. Dengan tahapan seperti pada gambar 1.



Gambar 1. Alur penelitian

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 5 Nomor 2, September 2021, pp. 727-735
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

2.1. *Data selection*

Data Penelitian ini memerlukan data masukan berupa data teks berita berbahasa Indonesia dengan ekstensi .txt sebanyak 50 data berita nasional. Pada tahap ini data sebanyak 50 dkokumen akan dibagi menjadi dua yaitu data training dan data testing dengan perbandingan 7:3, 8:2, dan 9:1 atau disebut juga dengan *traintestsplit*.

2.2. *Preprocesing*

Tahap *preprocessing* dilakukan untuk memperbaiki data sebelum masuk ke tahap *text mining. Data preprocessing* adalah salah satu fase utama dalam proses *data mining,* tahapan *preprocessing* diperlukan untuk mengatasi kesalahan pada data seperti, inkonsisten data, *missing value, outlier* dan redudansi data [9]. Tahap *preprocessing* yang dilakukan adalah:

1. *Case folding*

Pada penelitian *text mining, preprocessing case folding* diperlukan karena proses bersifat *case sensitive* dan tidak semua teks menggunakan huruf kapital secara konsisten. *Case folding* adalah proses mengubah seluruh isi dokumen menjadi *lowcase.*

2. *Tokenizing*

Dokumen dari hasil *case folding* selanjutnya akan dipecah menjadi kumpulan beberapa kata berdasarkan spasi atau disebut dengan *tokenizing*.

3. *Stopword removal*

Stopword removal merupakan proses untuk menghilangkan kata yang tidak penting, berfungsi untuk mengurangi dimensi teks masukan sehingga proses lebih ringan. Contohnya adalah kata yang, atau, dan.

4. *Stemming*

Stemming adalah proses untuk mengubah kata di dalam dokumen menjadi bentuk kata dasarnya, salah satunya adalah menghapus imbuhan. Contohnya adalah kata 'berbagi' menjadi 'bagi'.

2.3. *Transformation*

Transformation adalah mengubah skala data menjadi bentuk tertentu sesuai dengan kebutuhaan sebelum data diproses. Proses transformasi adalah ekstraksi fitur berdasarkan:

a) Posisi Kalimat

Kalimat awal dalam sebuah paragraf umumnya merupakan bagian penting atau topik berita dan berpotensi besar menjadi poin penting. Berikut merupakan persamaan untuk menghitung posisi kalimat dalam sebuah paragraf.

$$\text{Posisi kalimat} = \frac{n - i}{n} \quad (1)$$

Dengan n merupakan jumlah kalimat dalam paragraf dan i merupakan posisi kalimat ke-i.

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 5 Nomor 2, September 2021, pp. 727-735
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

b) Posisi keseluruhan kalimat

Perhitungan posisi keseluruhan kalimat dilakukan dengan cara memberi nilai terbesar untuk awal kalimat dan terkecil untuk akhir kalimat. Persamaan untuk menghitung posisi keseluruhan kalimat :

$$\text{Posisi keseluruhan} = \frac{n - i}{n} \quad (2)$$

Dengan n merupakan jumlah kalimat dalam dokumen dan i posisi kalimat ke-i.

c) Data numerik

Data numerik merupakan representasi untuk informasi penting yang dianggap seperti hasil survei, umur, tanggal, keuangan, dan sebagainya. Persamaan untuk data numerik adalah sebagai berikut :

$$\text{Data numerik} = \frac{\text{Total data numerik dalam kalimat}}{\text{Total kata dalam kalimat}} \quad (3)$$

d) Kutipan

Tanda koma terbalik biasanya digunakan untuk kutipan judul, percakapan langsung atau informasi penting lainnya. Persamaan untuk menghitung tanda koma terbalik adalah sebagai berikut :

$$Kutipan = \frac{Total\ banyak\ \ kata\ dalam\ tanda\ koma\ terbalik}{Total\ kata\ dalam\ kalimat} \quad (4)$$

e) Panjang kalimat

Kalimat yang panjang belum tentu dapat merepresentasikan topik berita, juga dengan kalimat pendek. Maka dari itu panjang kalimat dirumuskan sebagai berikut :

$$Panjang\ Kalimat = \frac{Total\ kata\ dalam\ kalimat}{Total\ kata\ dalam\ kalimat\ terpanjang\ pada\ sebuah\ paragraf} \quad (5)$$

f) Kata kunci

Kata yang memiliki frekuensi tinggi pada dokumen berita, disebut juga kata kunci. Kata kunci bisa jadi digunakan untuk menentukan kata penting pada dokumen. Persamaan untuk menghitung kata kunci dalam kalimat adalah sebagai berikut :

$$Kata\ kunci = \frac{Total\ banyaknya\ kata\ kunci\ dalam\ kalimat}{Total\ kata\ dalam\ kalimat} \quad (6)$$

2.4. *Text mining*

Sama halnya dengan *data mining* yang digunakan untuk mengekstrak data menjadi sebuah informasi kemudian menjadi pengetahuan yang dapat digunakan untuk mengambil pengetahuan, *text mining* juga berfokus mengambil informasi pada data besar dari dokumen berbasis teks [10]. Data yang sebelumnya telah dilakukan *preprocessing*, transformasi kemudian selanjutnya akan diproses dengan menerapkan *text mining* menggunakan algoritma *Support Vector Machine* dengan empat *kernel SVM* yaitu *linear kernel, polynomial kernel, sigmoid kernel* dan RBF *kernel*, bahasa pemrograman python.

2.5. Evaluasi

Pada penelitian ini akan menggunakan *recall, precision, F-measure* dan *accuracy*.

Precision

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

Recall

$$\text{Recall} = \frac{TP}{TP + FN} \quad (8)$$

F-*measure*

$$\text{F} - \text{measure} = 2\,X\,\frac{Recall\ X\ Precision}{Precision + recall} \quad (9)$$

Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad (10)$$

Dimana TP adalah *true positive*, TN adalah *true negative*, FP adalah *False Positive*, dan FN adalah *False Negative*.

3. HASIL DAN PEMBAHASAN

Hasil Penelitian ini adalah meringkas berita dengan teknik ekstraksi, yaitu mengambil poin utama pada berita tanpa mengubah isi bacaan. Algoritma yang digunakan adalah algoritma Support Vector Machine, dengan ekstraksi fitur berdasarkan fitur kalimat.

3.1. *Data Selection*

Data diambil dari webiste resmi Radar Karawang yaitu radarkarawang.id sebanyak 50 data berita, berisi judul berita, tanggal terbit dan juga isi berita.

3.2. *Preprocessing*

Preprocessing yang dilakuan pertama adalah *case folding* untuk mengubah seluruh kalimat menjadi kalimat *lowcase.* Kemudian data berita akan dipecah menjadi kalimat-kalimat dengan menggunakan *text parsing.* Setelah itu dilakukan *tokenizing* untuk mengubah data menjadi kumpulan kata. *Tokenizing* diperlukan untuk tahapan *preprocessing selanjutnya* yaitu *stopword removal,* dan *stemming.* Hasil *preprocessing* dapat dilihat pada Tabel 1

Tabel 1. Hasil preprocessing

Kalimat berita	Hasil *preprocessing*
Keributan, kriminal, pencabulan biasanya berawal dari pesta minum-minuman keras.	'ribut' 'kriminal' 'cabul' 'pesta' 'minum' 'minum' 'keras'
Untuk mengantisipasi hal tersebut, anggota polsek cikampek mencari para penjual minuman keras.	'antisipasi' 'anggota' 'polsek' 'cikampek' 'cari' 'jual' 'minum' 'keras'

Kalimat berita	Hasil *preprocessing*
Alhasil petugas menemukan minuman keras di toko pendi desa dawuan tengah kecamatan cikampek.	'alhasil' 'tugas' 'temu' 'minum' 'keras' 'toko' 'pendi' 'desa' 'dawuan' 'tengah' 'camat' 'cikampek'

3.3. *Tranformation*

Tranformation dilakukan dengan cara ekstraksi fitur berdasarkan fitur kalimat diantaranya adalah berdasarkan posisi kalimat dalam paragraf (F1), posisi keseluruhan kalimat (F2), data numerik (F3), tanda koma terbalik (F4), panjang kalimat (F5), dan kata kunci (F6).

Tabel 2. Hasil *Transformation*

Kalimat berita	F1	F2	F3	F4	F5	F6
Keributan, kriminal, pencabulan biasanya berawal dari pesta minum-minuman keras.	0,5	0,88	0	0	0,3	0
Untuk mengantisipasi hal tersebut, anggota polsek cikampek mencari para penjual minuman keras.	0	0,77	0	0	0,4	0
Alhasil petugas menemukan minuman keras di toko pendi desa dawuan tengah kecamatan cikampek.	0,5	0,66	0	0	0,43	0
”Dari hasil kegiatan kita berhasil amankan lima botol arak kecil, dan lima botol arak besar,” kata anggota Reskrim Polsek Cikampek Ipda Kadek.	0	0,55	0	0,68	0,73	0,28
Dia menyampaikan himbauan agar warga ikut mencegah penyebaran Covid-19 dan selalu terapkan protokol kesehatan dengan tepat.	0,33	0,11	0,11	0	0,53	0

3.4. *Text mining*

Tahap *text mining* adalah tahap implementasi dari algoritma *Support Vector Machine* (SVM) dengan menggunakan empat *kernel* yaitu *linear kernel, polynomial kernel, sigmoid kernel* dan RBF *kernel.* Setelah mendapatkan Fitur-fitur berdasarkan posisi kalimat dalam paragraf, posisi keseluruhan kalimat, data numerik, tanda koma terbalik, panjang kalimat dan kata kunci selanjutnya adalah melakukan pemodelan dengan menggunakan SVM. SVM merupakan algoritma klasifikasi dan merupakan *supervised learning* sehingga komputer perlu diberi pembelajaran terlebih dahulu. Pada penelitian ini proses pembelajaran akan menggunakan *data training,* selanjutnya jika model berhasil dibangun, model klasifikasi akan diuji menggunakan *data testing.* Klasifikasi dibedakan menjadi dua kelas, yaitu kelas Ringkasan dan Bukan Ringkasan. Skenario penelitian dibagi menjadi tiga skenario pembagian data. Skenario pertama adalah 70% *data training* dan 30% *data testing,* kemudian skenario kedua adalah 80% *data training*

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 5 Nomor 2, September 2021, pp. 727-735
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

dan 20% *data testing*, dan terakhir skenario pertama 90% *data training* dan 10% *data testing.*

3.5. Evaluasi

Evaluasi penelitian dilakukan dengan melakukan pengujian terhadap model yang telah dilakukan diantaranya adalah menghitung nilai *accuracy, precision, recall* dan F-*measure.* Berikut hasil perhitungan *accuracy, precision, recall* dan F-*measure* dapat dilihat pada Tabel 3.

Tabel 3. Hasil evaluasi

Skenario	Kernel	Hasil			
		Accuracy	***Precision***	***Recall***	***F-Measure***
Skenario 1 (7:3)	*Linear*	0,72	0,63	0,51	0,57
	Polynomial	0,67	0,57	0,33	0,42
	RBF	0,69	0,58	0,42	0,49
	Sigmoid	0,64	0,50	0,46	0,49
Skenario 2 (8:2)	*Linear*	0,70	0,54	0,51	0,52
	Polynomial	0,69	0,56	0,29	0,38
	RBF	0,69	0,53	0,48	0,51
	Sigmoid	0,63	0,43	0,44	0,44
Skenario 3 (9:1)	*Linear*	0,64	0,52	0,44	0,48
	Polynomial	0,64	0,52	0,33	0,40
	RBF	0,64	0,52	0,44	0,48
	Sigmoid	0,65	0,53	0,51	0,52

4. SIMPULAN

Berdasarkan penelitian yang telah dilakukan dapat disimpulkan beberapa hal berikut:

a) Berdasarkan Tabel 3 Hasil akurasi tertinggi diperoleh pada skenario 1, dengan *data training* 70% dan *data testing* 30%, untuk *linear kernel* yaitu *Accuracy* 72%, *precision* 63%, *recall* 51% dan *F-measure* 57%.
b) Model yang dibangun menggunakan empat kernel SVM, ternyata perbedaan setiap kernel memengaruhi hasil evaluasi. Untuk skenario pertama hasil linear kernel adalah Accuracy 72%, precision 63%, recall 51% dan F-measure 57%. Kemudian hasil polynomial kernel adalah Accuracy 67%, precision 57%, recall 33% dan F-measure 42%. Selanjutnya RBF kernel hasilnya adalah Accuracy 69%, precision 58%, recall 42% dan F-measure 49%. Dan keempat adalah sigmoid kernel hasilnya Accuracy 64%, precision 50%, recall 46% dan F-measure 49%. Jika dilihat hasil menggunakan skenario pembagian data set yang sama yaitu 70% data train dan 30% data testing perbedaan kernel cukup signifikan memberikan hasil yang berbeda.
c) Pembuatan model menggunakan tiga skenario pembagian data set. Yang pertama adalah 70% *data train* dan 30%, selanjutnya skenario kedua adalah 80% *data train* dan 20%, dan skenario ketiga adalah

90% *data train* dan 10%. Pembagian skenario dataset juga dapat mempengaruhi hasil evaluasi. Salah satu contohnya adalah hasil *kernel linear* menghasilkan nilai evaluasi paling tinggi pada penelitian ini yaitu *Accuracy* 72%, *precision* 63%, *recall* 51% dan *F-measure* 57% dengan menggunakan skenario pertama. Namun skenario kedua memiliki hasil yang berbeda yaitu *Accuracy* 70, *precision* 54%, *recall* 51% dan *F-measure* 52%. Begitupun dengan skenario ketiga, masih menggunakan kernel linear namun menghasilkan nilai yang berbeda yaitu *Accuracy* 64%, *precision* 52%, *recall* 44% dan *F-measure* 48%.

DAFTAR PUSTAKA

[1] PISA, "Reading performance (Programme for international Student Assesment)," 2020. [Online]. Available: https://data.oecd.org.

[2] N. S. W. Gotami, Indriati dan K. R. Dewi, "Peringkasan teks otomatis secara ekstraktif pada artikel berita kesehatan berbahasa indonesia dengan menggunakan metode latent semantic analysis," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer,* pp. 2821-2828, 2018.

[3] M. Allahyari, S. Puriyeh, M. Assefi, S. Safaei, E. Trippe, J. B. Gutierrez dan K. Kochut, "Brief survey of text mining : classification, clustering and extraction techniques," *ArXIv,* 2017.

[4] W. S. El-Kassas, C. R. Salma, A. A. Rafea dan H. K. Mohamed, "Automatic text summerization : a comprehensive survey," *Elsevier, 165,* pp. 2-26, 2020.

[5] N. Moratanch dan S. Citrakala, "A survey on extractive text summerization," *IEEE International Conference on Computer and Signal Processing,* pp. 1-7, 2017.

[6] j. Carventes, F. G. Lamont, L. R. Mazahua dan A. Lopez, "A comphresive survey on support vector machine classication: application challenges and trends," *Neurocomputing, 408,* pp. 189-215, 2019.

[7] P. D. M. W, "Evaluation : From Precision, Recall and F-measure to ROC, Informedness, Markedness, & Correlation," *International Journal of Machine Learning Technology,* pp. 37-63, 2011.

[8] S. R. Gallego, B. Krawczyk, S. Garcia, M. Wozniak dan F. Hererra, "A survey on data preprocessing for data stream mining : current status and feature direction," *Neurocomputing, 239,* pp. 39-57, 2017.

[9] M. Yuli, "Data mining : Klasifikasi Menggunakan Algoritma C4.5," *Jurnal Edik Informatika, 2(2),* pp. 213-219, 2017.

[10] R. Suresh dan R. S. Harshni, "Data mining and text mining - a survey," *International conference on computation of power, energy, information and comunacation, 21,* pp. 412-420, 2017.