Jurnal Teknologi Informatika (J-TIFA) V.3.No.2 September 2020
DOI: https://doi.org/10.52046/j-tifa.v3i2.1121
ISSN: 2654-2617 (Cetak)
ISSN: 2654-2633(Online)

Website : jurnal.ummu.ac.id/index.php/J-TIFA

J-TIFA

(Jurnal Teknologi Informatika)

|Teknologi Informasi | Jaringan Komputer | Data Mining |

Klasifikasi Berita Indonesia Menggunakan Naïve Bayes dengan Porter Stemmer

Gamaria Mandar[a], Abdul Haris Muhammad[b], Sakina Sudin[c]
[abc]Informatika, Universitas Muhammadiyah Maluku Utara, Ternate, Indonesia
Email: gamariamandar20@gmail.com[a], agry.arisandi@gmail.com[b], sakinahsudin80@yahoo.co.ic[c]

Abstrak

Pertumbuhan media online yang semakin banyak membuktikankan bahwa pembaca berita lebih tertarik untuk membaca secara online, dikarenakan berita dapat diupdate setiap saat dan kapanpun serta mudah diakses dengan adanya internet. Tercatat ditahun 2019 terdapat 2.700 portal berita yang terverifikasi oleh dewan pers dari total 47.000. hal ini menandahkan bahwa jumlah data berita yang dikelolah setiap hari oleh masing-masing portal cukup sangat banyak. Teknologi website rata-rata digunakan oleh media kabar sudah cukup baik dalam mengelolah informasi berita yang akan ditampilkan, namun banyaknya data berita yang dikelompokan pada jenis-jenis berita saat ini masih dikelompokan secara manual oleh manusia. Oleh karena itu dengan adanya teknik data mining, dapat dimanfatkan dalam pengklasifikasian kategori/jenis/rubik berita yang dilakukan secara otomatis. Salah satunya dengan menggunakan metode *Naive Bayes Classifier*(NBC) namun sebelum diklasifikasi, data berita berupa teks terlebih dulu dilakukan teknik preprosessing untuk menemukan indeks kata dalam berita yang berbobot, diantara teknik *case folding*, *tokenisasi*, *stopword* dan *stemming*, algoritma *stemming* yang digunakan yaitu *porter stemmer*. Dari hasil uji terhadap 15 data berita yang diklasifikasikan oleh NBC pada tiga ketegori berita *sport*, *otomotif* dan *finance* memperoleh hasil lebih banyak relevan dengan data pakar. Sehingga disimpulkan bahwa penelitian ini mampu mengklasifikasi berita sesuai dengan kategori/rubik masing-masing dengan keakuratan sebesar 79%.
.
Kata Kunci : Berita, Data mining, *Naive Bayes Classifier.*

Abstract

The increasing growth of online media proves that news readers are more interested in reading online, because news can be updated at any time and at any time and is easily accessible with the internet. It was recorded that in 2019 there were 2,700 news portals verified by the press council out of a total of 47,000. this indicates that the amount of news data that is managed every day by each portal is quite large. The average website technology used by the news media is good enough in managing the news information that will be displayed, but the amount of news data that is grouped on the types of news is currently still grouped manually by humans. Therefore, with data mining techniques, it can be used in classifying categories / types / news rubik which is done automatically. One of them is by using the Naive Bayes Classifier (NBC) method, but before being classified, news data in the form of text is first carried out with a preprocessing technique to find a word index in a weighty news, including case folding, tokenization, stopword and stemming techniques, the stemming algorithm used is porter. stemmer. From the test results on 15 news data classified by NBC into three categories of sports, automotive and finance news, the results are more relevant to expert data. So it can be concluded that this research is able to classify news according to each category/rubik with an accuracy of 79%.© 2020 J-Tifa. All rights reserved.

Keywords: News, Data mining, Naive Bayes Classifier

Riwayat Artikel: Menerima(8 Agustus 2020), Revisi (20 Agustus 2020), Diterima (22 Agustus 2020), Online (3 September 2020)

Jurnal Teknologi Informatika (J-TIFA) V.3.No.2 September 2020
DOI: https://doi.org/10.52046/j-tifa.v3i2.1121
ISSN: 2654-2617 (Cetak)
ISSN: 2654-2633(Online)

1. Pendahuluan

Media penyebaran informasi berita saat ini mengalami pergeseran melalui teknologi informasi. penyampaian berita yang mulanya dari cetak, TV dan radio kini berintegrasi menjadi media online. Menurut Asosiasi Media Siber Indonesia (AMSI) Tercatat di tahun 2019 terdapat 2.700 media online di Indonesia yang terverifikasi oleh Dewan Press dari 47.000 media online (AMSI, 2019). Hal ini mengambarkan bahwa pertumbuhan media online yang dikenal dengan portal berita lebih banyak daripada media cetak, tv maupun radio.

Portal berita atau media online didefinisikan sebagai jaringan luas komputer, yang dengan perizinan dapat saling berkoneksi antara satu dengan yang lainnya untuk menyebarluaskan dan membagikan digital files serta memperpendek jarak antar negara (Perebinissoff, 2005). Atau yang dikenal saat ini dengan nama *Website*, dengan ada website berita mempermudah orang untuk mengakses berita lebih cepat dan dapat diakases dimana saja. Informasi berita yang disajikanpun beragam tidak hanya teks dan gambar melainkan beberapa website portal berita dilengkapi dengan video pendek. Hal lain yang menjadi perhatian pada portal berita adalah informasi yang tersusun secara sistematis sehingga pembaca berita dengan mudah memilih berita yang ingin dibaca, sebagaimana yang diketahui rubik adalah bagian penting dari sebuah media cetak maupun online.

Rubik adalah suatu ruang khusus pada media surat kabar, majalah atau tabloid yang memuat informasi, berita, opini atau iklan tertentu dimana penanyanganya dilakukan dalam periode bertahap (Prawiro, 2019). Rubik media cetak maupun online selalu berbeda-beda sesuai dengan liputan media tersebut misalkan *detik.com* memiliki beberapa rubik diantaranya *politik, finance, sport, otomotif, food*, daerah dan lain-lain. Rubik secara umum adalah mengelompokan atau pengklasifikasian berita sesuai dengan topiknya. Hal ini dilakukan untuk mempermudah pembaca menemukan berita dengan cepat. Untuk mengelompokkan teks berita pada website umumnya dilakukan pemilihan langsung pada fitur website dengan memilih kategori berita tersebut. Akan tetapi dengan perkembangan yang begitu pesat pada era teknologi pengklasifikasian berita dapat menggunakan teknik data mining. Data mining adalah suatu proses ekstraksi atau penggalian dari data yang belum diketahui sebelumnya (Thomas, 2015). Salah satu teknik data mining yang dikenal adalah menggunakan metode *Naive Bayes Classifier* (NBC). *Naïve bayes* adalah teknik data mining yang digunakan untuk mengklasifikasikan sebuah data atau informasi tertentu.

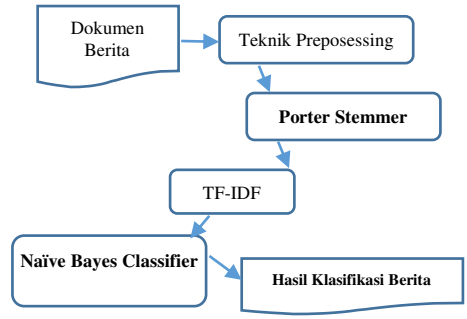
Berdasarkan paparan diatas maka penulis melakukan penelitian dengan klasifikasi berita Indonesia menggunakan metode *Naive Bayes Classifier* dengan *Porter Stemmer*, yang bertujuan untuk mengklasifikasikan teks berita sesuai degan rubik, kategori atau jenis berita secara otomatis dan diharapkan dapat diterapkan pada website portal berita.

2. Penelitian Terdahulu

Penelitian terkait pengklasifikasian berita menggunakan metode *Naive Bayes Classifier* (NBC) telah dilakukan oleh beberapa peniliti sebelumnya diantara klasifikasi berita olahraga menggunakan *Naïve bayes* dengan *Enhanced Confix Stripping Stemmer* (yoga dkk, 2018), dimana pada penelitian ini terlebih dahulu dilakukan teknik prosessing. Dimana pada klasifikasi berita olahraga seperti sepak bola, basket, formula 1, motor GP dan olahraga lain berhasil diklasifikasikan dengan keakuratan sebesar 77%. Dan pada penilitian klasifiksai berita mengunnakan NBC dengan seleksi fitur dan *boosting* oleh boby dengan akurasi 73% (Bobby et al., 2019).

3. Metode Penelitian

3.1 Arsitektur Sistem



Gambar 1. Arsitektur sistem

Riwayat Artikel: Menerima(8 Agustus 2020), Revisi (20 Agustus 2020), Diterima (22 Agustus 2020), Online (3 September 2020)

Jurnal Teknologi Informatika (J-TIFA) V.3.No.2 September 2020
DOI: https://doi.org/10.52046/j-tifa.v3i2.1121
ISSN: 2654-2617 (Cetak)
ISSN: 2654-2633(Online)

Arsitektur sistem yang digunakan pada penelitian ini adalah dimulai dari pengumulan dokumen berita berbahasa Indonesia, teknik prosessing, porter stemmer, algoritma TF-IDF dan *Naïve Bayes Classifier* sehingga memperoleh hasil klasifikasi

3.2 Data

Penelitian ini menggunakan dataset sebanyak 75 artikel berita yang terdiri dari 60 *dataset* dan 15 data traning. Pada dataset masing-masing terdapat 20 artikel berdasarkan kategori *sport, otomotif* dan *finance* yang masing-masing sebanyak 20 dataset dan 5 data traning. Sumber data diambil dari portal berita detik.com dimulai dari tanggal 1 januari sampai dengan 15 Januari 2020.

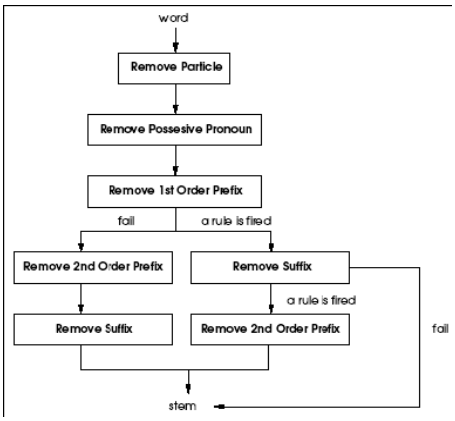
3.3 Preprosessing

Preprosessing bertujuan untuk memilih kata dari sebuah teks berita yang digunakan sebagai *indeks*, dimana *indeks* kata mewakili suatu dokumen, dengan tujuan untuk proses komputasi menjadi lebih efisien (Yoga et al., 2018). Berikut teknik preprosesising yang digunakan pada penelitian ini sebagai berikut (Gamaria, 2017).

a. Melakukan pelabelan pada setiap dataset dokumen berita yang digunakan. Pada penelitian ini menggunakan label sesuai dengan kategori berita yang diteliti yakni *sport, otomotif* dan *finance*.
b. *Case Folding*, Meyeragamkan teks berita dengan cara mengubah semua huruf teks menjadi huruf kecil dan menghilangkan karakter huruf yang dianggap delimeter.
c. *Tokenisasi*, cara memisahkan string kata dari dokumen kalimat.
d. *Stopword* adalah adalah proses untuk mengahapus kata-kata yang dianggap tidak penting, hal ini berfungsi agar dapat memaksimalkan informasi yang penting pada teks berita.

3.4 Porter Stemmer

Stemmer atau dikenal dengan *stemming* adalah proses penyederhaan kata menjadi kata dasar dengan menghilangkan imbuhan kata baik diawal maupun diakhir. *Porter stemmer* adalah algoritma *stemming* yang dikembangkan oleh W.B. Frakes pada tahun 1992 dalam bahasa inggris, dan kemudian kembangkan dalam bahasa Indonesia oleh Fadilla Z Tala pada tahun 2003 (Dian, 2016)



Gambar 2. Desain stemmer untuk bahasa indonesia (Dian, 2016)

3.5 Pembobotan kata menggunakan Algoritma TF-IDF

Pembobotan kata adalah proses untuk menghitung bobot suatu kata pada dokumen yang dilihat dari banyaknya frekuensi kemunculkan kata pada sebuah teks berita atau dokumen kalimat. Algoritma yang digunakan pada penelitian ini adalah menggunakan *Term Frequency-Inverse Document* (TF-IDF) secara matematis dapat ditulis (Winata & Rainarli, 2016):

$$IDF=Log(D/DF) \quad (1)$$
$$W=TF\times IDF \quad (2)$$

Inverse Document Frequency (IDF) adalah hubungan antara banyak dokumen yang memiliki kata (Document Frequency) dengan jumlah dokumen kalimat (D), sedangkan W adalah nilai bobot dari setiap kata pada sebuah dokumen, hasil dari pembobotan kata adalah sebagai dataset yang akan digunakan untuk tahap *Naïve Bayes Classifier* (Gamaria, 2017).

3.6 Klasifikasi Kategori Berita menggunakan Naive Bayes Classifier (NBC)

Naïve Bayes Clasifier merupakan metode pengklasifikasian statistik yang didasarkan pada *teorema bayes* yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu kelas (Muhammad, 2017). *Naïve Bayes Classifier* pada

Riwayat Artikel: Menerima(8 Agustus 2020), Revisi (20 Agustus 2020), Diterima (22 Agustus 2020), Online (3 September 2020)

Jurnal Teknologi Informatika (J-TIFA) V.3.No.2 September 2020
DOI: https://doi.org/10.52046/j-tifa.v3i2.1121
ISSN: 2654-2617 (Cetak)
ISSN: 2654-2633(Online)

penelitian ini digunakan untuk mengklasifikasikan dokumen teks. Pada algoritma *Naïve Bayes* setiap dokumen dipresentasikan dengan masukan atribut "a1, a2, a3,…,an" dimana a1 adalah kata pertama dan berikutnya sampai an (kata ke-n), sedangkan V yaitu label kategori. Selanjutnya yaitu mencari nilai tertinggi dari kategori teks yang diujikan (VMAP) (McCallum, 1998). Persamaan VMAP yaitu sebagai berikut (Bagus, 2016) :

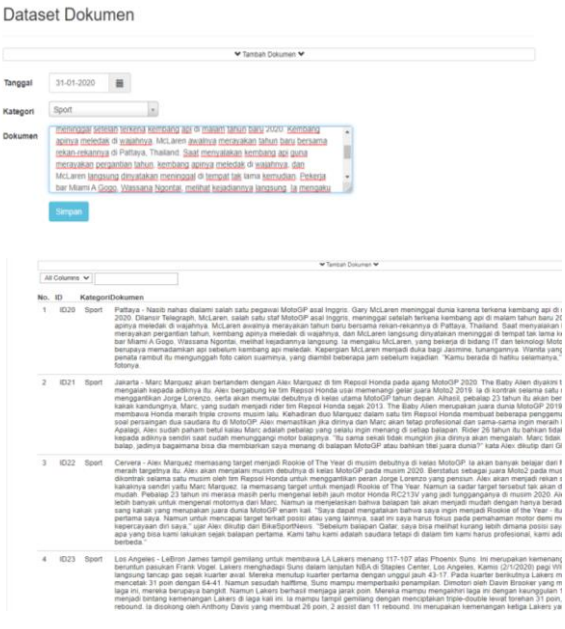
$$V_{MAP} = \underset{vj \in V}{\operatorname{argmax}} P(v_j) \prod_i P(a_i|v_j) \quad (3)$$

$$P(v_j) = \frac{|dokumen\ j|}{|dok.\ training|} \quad (4)$$

$$P(a_i|v_j) = \frac{|n_i+1|}{|n+kosa\ kata|} \quad (5)$$

3.7 Metode Pengujian

Penelitian ini menggunakan metode evaluasi instrik yaitu metode *Precision, Recall* dan *F-Measure* untuk memperoleh hasil akurasi antara hasil klasifikasi pakar dengan sistem.



Gambar 3 Inputan Data Training

Recall ialah kemampuan untuk mengambil peringkat teratas yang sebagian besar relevan (benar), *precision* adalah berapa banyak dokumen yang berhasil diambil oleh sistem, sedangkan untuk mengukur kualitas *recall* dan *precision* menggunakan *F-Measure* (Mustaqhfiri, Abidin, & Kusumawati,2011)

Precison=TP/(TP+FP) (6)
Recall=TP/(TP+FN) (7)
F-Measure=(2*Precison*Recall)/(Recall+Precision) (8)

Dari rumus diatas ada tiga komponen yang penting yaitu True Positive (TP) adalah jumlah dokumen kalimat yang dipilih oleh pakar, False Positive (FP) merupakan jumlah dokumen kalimat yang dipilih oleh sistem benar tetapi menurut pakar salah dan False Negative (FN) adalah jumlah dokumen kalimat yang benar menurut pakar tetapi salah menurut sistem. (Gamaria, 2017).

4. Hasil dan Pembahasan

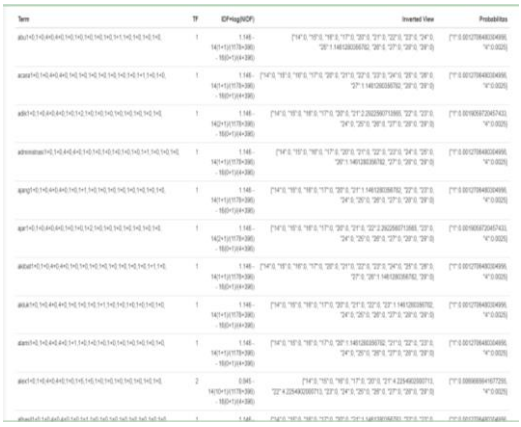
Tahapan penelitian ini diawali dengan proses penginputan dataset yang terdiri 60 artikel berita, dimana masing-masing terdiri dari 20 data kategori *sport, otomotif dan finance*. Pada proses ini disediakan fitur untuk menambahkan dataset baru seperti dipada gambar 3. Setelah dataset terkumpul langkah selanjutnya yang dilakukan adalah tahapan preprosessing dan stemming. teknik prosessing menggunakan beberapa teknik yakni *case folding*, *tokenisasi*, *stopword* dan *stemming* yang menggunakan algoritma *porter* hasilnya dapat dilihat pada gambar 4.

Gambar 4 Tahapan Prosessing dan stemming

Riwayat Artikel: Menerima(8 Agustus 2020), Revisi (20 Agustus 2020), Diterima (22 Agustus 2020), Online (3 September 2020)

Jurnal Teknologi Informatika (J-TIFA) V.3.No.2 September 2020
DOI: https://doi.org/10.52046/j-tifa.v3i2.1121
ISSN: 2654-2617 (Cetak)
ISSN: 2654-2633(Online)

Langkah berikutnya adalah menghitung bobot kata yang dihasilkan oleh tahapan prosessing dan stemming dengan menggunakan algoritma TF-IDF sebagaimana yang dihasilkan pada gambar 5. Setelah memperoleh nilia bobot dari setiap *term* maka dilanjutkan pada proses klasifikasi berita menggunakan *Naive Bayes Classifier* dapat dilihat pada gambar 6.



Gambar 5 Tahapan Pembobotan kata

Gambar 4.3 Tahapan Klasifikasi menggunakan Naïve Bayes

5. Evaluasi

Evaluasi pengujian dilakukan dengan membandingkan hasil klasifikasi berita oleh penelitian ini yakni menggunaakan *Naïve Bayes Clasifier* dan Data pakar yang diambil berdasarkan kategori pada halaman detik.com yang disebut sebagai data testing dapat dilihat pada tabel 1. kemudian menggunakan metode *Precision*, *Recall* dan *F-Measure* untuk memperoleh hasil akurasi antara perdiksi dari sistem atau peneltian ini dengan data pakar. Data pakar yang digunakan pada penelitian ini menggunakan 15 data pakar dimana masing-masing kategori atau jenis berita diambil sebanyak 5 artikel berita. Sehingga memperoleh nilai *precision* yakni 1,000 , *Recall* 0,6666 dan F-Measure 0,7951.

6. Kesimpulan Dan Saran.

Dari penjelasan diatas klasifikasi berita berhahasa Indonesia menggunakan *Naive Bayes Classifier* dan *porter stemmer* mampu mengklasifikasikan berita dengan keakuratan 79%. Selanjutkan untuk penelitian yang serupa penulis menyarankan beberapa hal diantaranya:

1. Pada penelitian selanjutnya diharapkan menambahkan *dataset* lebih banyak dari penelitian ini.
2. Dapat menggunakan metode atau algoritma data mining yang lain.
3. Disarankan untuk menggunakan algoritma stemming selain porter, sehingga dapat melihat apakah perbedaan stemming berpengaruh pada nilai akurasi.

Tabel 1. Hasil Pengujian

No	Judul Detik.Com	Hasil (sistem)	Hasil Pakar	Hasil (Relevan)
1	KOI: Jokowi Minta Jakarta Jadi Tuan Rumah Olimpiade 2032	Sport	Sport	Relevan
2	Soal Virus Corona, Penyelenggara GP Australia dan Bahrain Beda Sikap	Sport	Sport	Relevan
3	Swiss Open Mungkin Batal, All England Masih Sesuai Rencana	Healt	Sport	Tidak Relevan
4	Ingin Dampingi Papua Jadi Tuan Rumah PON 2020, Ini Syaratnya	Sport	Sport	Relevan
5	Kaget Tontowi Ahmad Pensiun, Apriyani Maunya Duet Lebih Lama	Healt	Sport	Tidak Relevan
6	Strategi Baru Aliansi Renault-Nissan-Mitsubishi	Otomotif	Otomotif	Relevan
7	Standar Emisi Hambat Ekspor Kendaraan Komersial	Otomotif	Otomotif	Relevan
8	Yamaha Kenalkan Xmax dan Aerox Versi 'Maxi Signature'	Sport	Otomotif	Tidak Relevan
9	Ekonomi Global Krisis, Ekspor Toyota Rakitan Indonesia Terus Ngegas	Otomotif	Otomotif	Relevan
10	Kia Siapkan World Premier di Indonesia, Bukan Mobil Listrik	Otomotif	Otomotif	Relevan
11	Ahok Buka Suara Revitalisasi Monas, Batik Air Jemput WNI di Wuhan	Healt	finance	Tidak Relevan
12	Proyek 'Kebun Angin' di China Diganggu Virus Corona	Healt	finance	Tidak Relevan
13	Serius Masih Ada yang Terjebak Investasi Bodong di 2020?	finance	finance	Relevan
14	Nasabah Hanson Ramai-ramai Tagih Duit Investasi, Kas Perusahaan Kering	finance	finance	Relevan
15	Saham Gorengan Jadi Buah Bibir, 'Wasit' Pasar Modal Buka Suara	finance	finance	Relevan

Riwayat Artikel: Menerima(8 Agustus 2020), Revisi (20 Agustus 2020), Diterima (22 Agustus 2020), Online (3 September 2020)

Jurnal Teknologi Informatika (J-TIFA) V.3.No.2 September 2020 ISSN: 2654-2617 (Cetak)
DOI: https://doi.org/10.52046/j-tifa.v3i2.1121 ISSN: 2654-2633(Online)

Referensi

Asosiasi Media Siber Indonesia." Dari 47 Ribu, Baru 2.700 Media Online Terverifikasi Dewan Pers"(2019). https://www.amsi.or.id

Dian Novitasari. (2016). Perbandingan Algoritma Stemming Porter Denganarifin Setiono Untuk Menentukan Tingkat Ketepatan Kata Dasar. Jurnal String Vol 1 No 2.

Dwitya Pramudita Yoga, dkk (2018), Klasifikasi Berita Olahraga Menggunakan Metode Naïve Bayes Dengan Enhanced Confix Stripping Stemmer. Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIIK) Vol.5 No 3

Jiawei Han., Micheline Kamber. 2000. Data Mining : Concepts and Techniques.

Mandar Gamaria, Gunawan. (2017). Peringkasan dokumen berita Bahasa Indonesia menggunakan metode Cross Latent Semantic Analysis. Register vol 3 no 2

M. Prawiro. (2019)."Pengertian Rubrik: Arti, Jenis, Syarat, dan Contoh Rubrik". https://www.maxmanroe.com

Mustaqhfiri, M., Abidin, Z., & Kusumawati, R. (2011). Peringkasan teks otomatis berita berbahasa Indonesia menggunakan metode Maximum Marginal Relevance. MATICS, 4(4), 134-147.

Perebinossoff, Philippe. 2005. Programming for TV, Radio and The Internet, Strategi, Development and Evaluation. Second Edition : Focal Press. Elsevier Inc.

Prakoso bobby Suryo, dkk (2019. Klasifikasi Berita Menggunakan Algoritma Naive Bayes Classifer Dengan Seleksi Fitur Dan Boosting. Jurnal Resti. Vol3 No2

Thomas M. Connoly., Carolyn E. Begg. 2015. Database System : A Practical Approach to Design, Implementation, and Management.

Winata, F., & Rainarli, E. (2016). Implementasi Cross method Latent Semantic Analysis untuk meringkas dokumen berita Berbahasa Indonesia. Techno.Com, 15(4), 266-277.

Riwayat Artikel: Menerima(8 Agustus 2020), Revisi (20 Agustus 2020), Diterima (22 Agustus 2020), Online (3 September 2020)