Kumpulan jurnaL Ilmu Komputer (KLIK)
Volume 08, No 2 Juni 2021
ISSN: 2406-7857

# TEXT MINING UNTUK MENGKLASIFIKASI JUDUL BERITA ONLINE STUDI KASUS RADAR BANJARMASIN MENGGUNAKAN METODE NAÏVE BAYES

**Muhammad Sholih 'Afif[1*], Muhammad Muzakir[2], Moh. Iqbal Al Ghifari Al Awalaien[3]**
[123]Magister Teknik Informatika
Universitas Amikom Yogyakarta
Jl. Ring Road Utara, Condong Catur, Sleman, Yogyakarta
*afifsholihm@gmail.com

***Abstract***

*As technology advances, it causes a shift from news media which is usually published in newspapers, now changes to follow the times and develops into online news. In online news, news is usually grouped into several categories. The categorization is still manual. So in need of automatic categorization of online news titles. The method used in classifying online news titles is Naïve Bayes. The results of this study obtained an accuracy of 78.75%. Meanwhile, the recall results are 80.56% and the precision is 78.75%.*

***Keywords****: Classification, Text Mining, Naïve Bayes and Online News*

***Abstrak***

*Seiring bertambah majunya teknologi menyebabkan peralihan dari media berita yang biasanya dimuat dalam koran sekarang berubah mengikuti perkembangan zaman dan berkembang menjadi berita online. Pada berita online biasanya dalam berita dikelompokan menjadi beberapa kategori. Pengkategorian tersebut masih manual. Sehingga diperlukan pengkategorian judul berita online yang secara otomatis. Metode yang digunakan dalam pengklasifikasian judul berita online ini adalah Naïve Bayes. Hasil dari penelitian ini mendapatkan hasil akurasi sebesar 78.75%. Sedangkan untuk hasil Recall adalah 80.56% dan Precision adalah78.75%.*

***Kata kunci****: Klasifikasi,Text Mining, Naïve Bayes dan Berita Online*

## 1. PENDAHULUAN

### 1.1 Latar Belakang

Penelitian penambangan tentang bidang teks saat ini telah dipelajari secara ekstensif untuk berbagai tujuan, termasuk penambangan data, pembelajaran mesin, database, serta pemasaran, pengobatan diagnostik, dan tujuan konten. Studi berita dan informasi diterapkan untuk menentukan klasifikasi dokumen [1]. Metode penambangan teks atau penambangan teks yang paling umum digunakan adalah Nave Bayes Classifier (NBC) dan Support Vector Machine (SVM). Hubungan antar dokumen sering diukur dengan menggunakan probabilitas yang dapat dibuktikan dengan menggunakan algoritma klasifikasi lainnya.

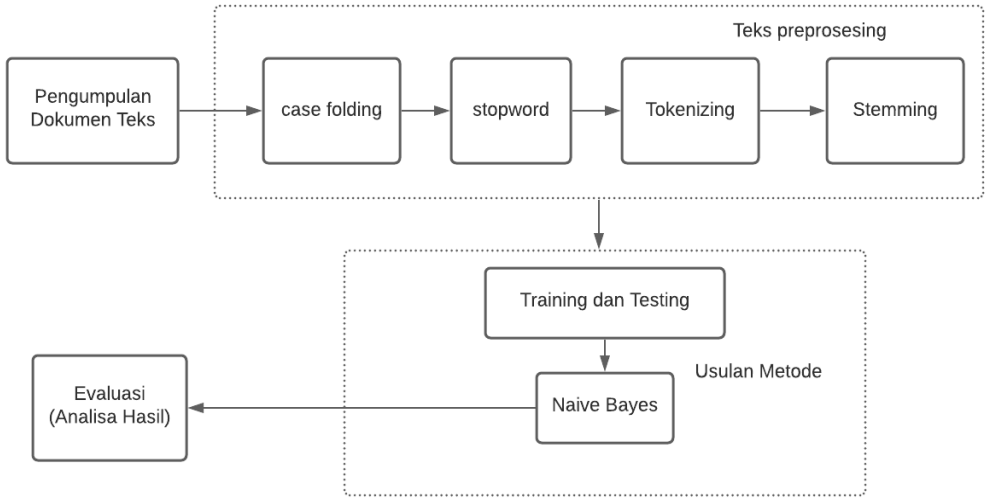
Penggunaan media berita *online* sebagai data untuk dilakukan *text mining* pada dasarnya bisa dilakukan. Dengan menggunakan metode *Naïve Bayes Clasifer* (NBC). Pengklasifikasian secara otomatis bisa di lakukan dengan menggunakan metode ini dengan tingkat akurasi yang tinggi [2]. Informasi adalah sebuah hal yang sangat penting didalam kehidupan dewasa ini. Contoh sumber infromasi adalah *web portal* atau *website* berita. Banyaknya berita yang di keluarkan dalam satu bulan bisa mencapai 300 sampai 400 artikel. Hal ini menyebabkan kinerja seorang editor menjadi lebih banyak di karenakan harus mengkategorikan berita tersebut secara manual dengan kategori yang sudah di tentukan [3].

Penelitian menggunakan metode Naïve Bayes Classifier untuk pengklasifikasian berita online. Berita yang dikumpulkan dalam data penelitian ini adalah sebanyak 18.794. Skema pengujian dilakukan dengan cara pembagian data menjadi 70:30. Hasil dari penelitian dengan metode *Naïve Bayes* mendapatkan hasil akurasi 98,90% dari jumlah total sembilan (9) kelas yang telah teruji. Penelitian ini menyarankan beberapa hal seperti proses *training* yang lebih, menambah jumlah class, menambah fitur klasifikasi seperti sentimen berita dan menambahkan fitur teks [4].

## 2. METODOLOGI PENELITIAN

### 2.1 Text Mining

Tujuan *text mining* hampir mirip untuk sama dengan tujuan *data mining* yaitu menemukan pola pada data agar dapat dimanfaatkan manusia untuk membantu pekerjaannya. Yang berbeda adalah sumber data yang akan digunakan, pada proses *text mining* sumber datanya adalah teks atau dokumen. Perbedaan yang lain adalah ada proses ekstraksi fitur untuk mengubah data teks menjadi data terstruktur [5].




Gambar 1. Diagram proses *text mining*

### 2.2 TF-IDF

Setelah dilakukan *preprocessing* pada *dataset*, hal selanjutnya yang dilakukan adalah memberikan bobot pada setiap judul berita yang *dataset*-nya sudah mempunyai standar kesamaan. Pembobotan kata dengan menggunakan *Term*

*Frequency.* Metode ini adalah untuk pembobotan yang menghitung seberapa sering sebuah kata muncul pada suatu dokumen. Setiap kata di dalam dokumen akan diberikan pembobotan dengan menggunakan persamaan:

$$W_{TF}(t_i, d_j) = f(t_i, d_j) \quad (1)$$

*Inverse Document Frequency* merupakan pendekatan yang menggunakan kata yang frekuensi munculnya lebih dominan di dalam suatu dokumen. Namun jika sebuah kata mempunyai frekuensi munculnya lebih jarang maka pembobotannya menjadi lebih tinggi. Persamaannya adalah:

$$W_{IDF}(t_i) = 1 + LOG\left(\frac{D}{d_{(ti)}}\right) \quad (2)$$

Sehingga untuk pembobotan TF-IDF menggunakan perkalian dari rumus 1 dan 2 dan menghasilkan persamaan:

$$W_{TF.IDF}(t_i, d_j) = f(t_i, d_j)x\left(1 + LOG\left(\frac{D}{d_{(ti)}}\right)\right) \quad (3)$$

Dimana $W_{TF.IDF}(t_i, d_j)$adalah pembobotan kata i pada dokumen j. Sedangkan $f(t_i, d_j)$ untuk banyak kata atau term i pada dokumen j dan $D$ total dokumen di dataset kemudian $t_i$ total dokumen yang memuncul-kan kata i

## 2.3 Naïve Bayes Classifer

Algoritma NBC (Naive Bayes Classifier) adalah pengklasifikasi yang menggunakan metode probabilistik dan statistik yang diusulkan oleh ilmuwan Inggris Thomas Bayes untuk memprediksi potensi masa depan berdasarkan pengalaman masa lalu. Metode NBC melakukan langkah-langkah pelatihan dan klasifikasi dalam proses klasifikasi teks. Selama fase pembelajaran, analisis dilakukan dalam bentuk kata-kata atau kumpulan kosakata dalam dokumen sampel, yaitu kata-kata yang dapat ditampilkan dalam dokumen. Tampilan dokumen. Kemudian tentukan probabilitas sebelumnya untuk setiap kategori berdasarkan dokumen sampel. Nilai kategori dokumen pada tahap klasifikasi ditentukan [4]. Dalam penentuan probabilitas dari tiap-tiap term menggunakan persamaan seperti di bawah ini:

$$P(W_i|C) = \frac{count\ (w_i.C)+1}{count\ (C)+|v|}$$

Dimana $P(W_i|C)$ adalah probabilitas jumlah dari kata $W_i$ didalam kelas dalam C. Count (C) merupakan total kata di dalam kelas C, |v| merupakan jumlah kata- kata.

## 2.4 Modeling

Modeling adalah sebuah tahap pada pemilihan teknik yang akan dipilih pada penambangan dengan menentukan dahulu algoritma mana yang akan dipilih. Penelitian ini akan menggunakan *tools* yang biasa dipakai dalam melakukan

pemodelan dengan teknik yang telah ditentukan sebelumnya, *tools* yang digunakan RapidMiner versi 9.9. Pada penelitian ini digunakan algortima klasifikasi untuk metodenya. Algortima yang digunakan adalah *Naïve Bayes* (NB). Nantinya hasil model *Naive Bayes* (NB) ini digunakan untuk mengklasifikasikan judul berita ke dalam empat buah kategori yakni 'Banua', 'Bisnis', 'Hukum dan Kriminal', dan 'Sport'[6] .

## 2.5 Evaluation

*Evaluation* adalah tahap yang memiliki tujuan untuk menentukan apakah kegunaan model yang telah di buat. Proses validasinya sendiri berasal dari dua buah subproses yang mana, data pelatihan (*training set*) dan data pengujian (*testing set*).

Sistem kalsifikasi kinerja untuk mengetahui seberapa baik sebuah metode dalam menangani *dataset*. *Confuison matrix* adalah salah satu metode yang membantu dalam pengukuran kinerja sebuah model klasifikasi. Aslinya *Confusion matrix* merupakan informasi agar dapat membandingkan hasil dari metode yang digunakan. *Confusion matrix* terdapat empat proses inti pengklasifikasian yaitu :

1. *True Positive* (TP) ialah total dari data positif yang diklasifikasikan data bagian positif.
2. *True Negative* (TN) ialah total dari data negatif yang diklasifikasikan data bagian negatif.
3. *False Positive* (FP) ialah total dari data negatif yang diklasifikasikan data bagian positif.
4. *False Negative* (FN) ialah total dari data positif yang diklasifikasikan data bagian negative .

### *2.5.1. Accuracy*

Nilai akurasi menggambarkan seberapa sistem untuk mengklasifikasi sebuah dataset secara baik. Nilai akurasi diperoleh dengan rumus :

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\%$$

### *2.5.2. Recall*

Nilai *recall* menggambarkan total data positif yang dianggap benar oleh computer untuk mencari nilai *recall* menggunakan rumus :

$$Recall = \frac{TP}{FN+TP} \times 100\%$$

### *2.5.3. Precision*

Nilai *precision* dalam menggambarkan total judul data positif yang sudah terklasifikasi bagian dari positif. Nilai *precision* ini dapat diperoleh menggunakan rumus :

$$Precision = \frac{TP}{FP+TP} \times 100\%$$

### 2.6 Deployment

Pada tahapan *deployment* merupakan tahapan untuk membuat sebuah model implementasi yang dibuat kedalam *tools*. Proses ini juga memasukan model ke dalam RapidMiner untuk nantinya bisa diproses dan didapatkan hasil pengujian dari metode *Naïve Bayes* (NB) dalam *text mining* dalam mengkalasifikasikan judul berita online.

## 3. HASIL DAN PEMBAHASAN

### 3.1 Bahan Penelitian

Penggunaan *Naïve Bayes* untuk pengklasifikasian judul berita online ini menggunakan data yang di ambil dengan menggunakan teknik *scrapping*. Data yang di ambil berupa judul dan mengambil empat buah kategori yaitu 'Banua', 'Bisnis', 'Hukum dan Kriminal', dan 'Sport'. Jumlah data yang di ambil sebanyak 400 data judul berita yang di ambil dari website berita Radar Banjarmasin. Teknik pengumpulan data menggunakan bahasa pemrograman R :

```
for(i in 1:nrow(data_url)) {
  url = data_url[i,1]

  web_page = read_html(url)

  news_content = web_page %>% html_nodes(".content-artikel-judul h3") %>%
html_text(trim = TRUE)

  news_content.all = "";

  for(el in news_content) {
    news_content.all = paste(news_content.all, el)
  }
  if(!exists("main_data")) {
    assign("main_data", c(news_content.all))
  } else {
    main_data = rbind(main_data, c(news_content.all))
  }
}
```

Data yang sudah dikumpulkan akan disimpan pada sebuah file excel. Berikut adalah contoh data yang telah terkumpul :

Tabel 1. Contoh judul berita pada situs Radar Banjarmasin.

| Judul | Kategori |
|---|---|
| Terlalu Lama Tanpa Gubernur, Tim Sahbirin-Muhidin Minta MK Tolak Permohonan Sengketa | Banua |
| Terdakwa Korupsi Tetap Kembalikan Uang Negara | Sport |
| Pentingnya Belajar dari Pengalaman; Janji-Janji Harus Diingat, Kalau Perlu Tagih..!! | Hukum dan Peristiwa |
| Cegah Banjir di Cempaka, Embung Bagian Hulu Dikeruk | Bisnis |
| Penyandang Disabilitas Divaksinasi, Banjarbaru Klaim yang Pertama di Kalsel | Sport |

| | |
|---|---|
| Jalan Kunyit - Atilam Naik Kelas | Bisnis |
| Wabub Ingin Penyaluran BLT-DD Lancar | Banua |
| DWP Gelar Lomba Kreasi Bucket | Hukum dan Peristiwa |
| Gula Semut Diperkenalkan | Sport |
| Menyongsong Generasi Sehat | Hukum dan Peristiwa |
| Intensitas Karhutla Mulai Naik | Banua |
| 8 Kali Sukses Gasak Motor, Residivis ini Kembali Berulah | Sport |
| Niat Mengadvokasi, Denny Indrayana Disuruh Pulang | Hukum dan Peristiwa |
| JAHAT..!! Karena Cemburu, Ibu Ini Aniaya Anak Tiri Selama Dua Bulan | Sport |
| Berbahaya, Perekam Video Viral Aksi Free Style Ikut Diamankan | Bisnis |

### *3.2 Text Preprocesing*

Setelah dilakukan pengambilan data dari *website*, masing-masing data dibagi sesuai dengan kategorinya. Tahapan selanjutnya adalah mengubah data menjadi data yang siap dipakai. Kemudian data yang telah siap akan diolah pada proses *case folding*, *tokenizing, stemming,* dan *stopword*. Berikut adalah contoh proses yang dilakukan :

#### *3.2.1. Case Folding*

*Case folding* merupakan proses dalam mengkonversi sebuah teks menjadi huruf yang di inginkan kecil ataupun besar. Sehingga yang pada awalnya judul berita yang tiap katanya memiliki huruf kapital. Berikut contoh penggunaan *case folding* :

Tabel 2. Contoh konversi *case folding* dalam judul berita.

| **Sebelum** |
|---|
| Banyak Tempat Usaha Tak Berizin, Setiap Bulan Akan Dirazia |
| **Sesudah** |
| banyak tempat usaha tak berizin, setiap bulan akan dirazia |

#### *3.2.2. Tokenizing*

*Tokenizing* adalah proses dimana tahap pemotongan dari kalimat sehingga terpisah tiap katanya. Proses ini juga menyaring tanda baca yang ada dalam sebuah kalimat. Berikut contoh *tokenizing* :

Tabel3. Contoh konversi *tokenizing* pada judul berita.

| **Sebelum** |
|---|
| banyak tempat usaha tak berizin, setiap bulan akan dirazia |
| **Sesudah** |

banyak tempat usaha tak berizin setiap bulan akan dirazia

### *3.2.3. Stopword*

*Stopwod* adalah proses untuk menyaring kata yang ingin digunakan. Dengan melakukan pengecekan tiap kata yang sudah di proses sebelumnya di *tokenizing*. Proses ini menyaring berdasarkan kata yang sesuai Bahasa Indonesia dan memasukan list kata yang ingin dimasukan ataupun dihilangkan. Berikut contoh *stopword* :

Tabel 4. Contoh konversi *stopword* pada judul berita.

| **Sebelum** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| banyak | tempat | usaha | tak | berizin | setiap | bulan | akan | dirazia |
| **Sesudah** | | | | | | | | |
| | tempat | usaha | | berizin | | bulan | | dirazia |

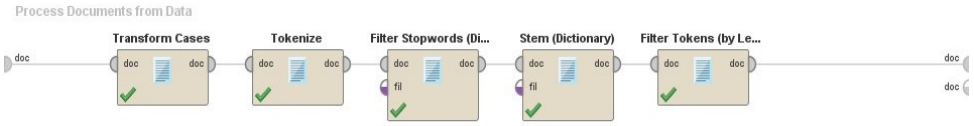
### *3.2.4. Stemming*

*Stemming* ialah proses yang berguna untuk mengubah sebuah kata menjadikan kata dasar dari proses sebelumnya *tokenizing*. Pengubahan kata menjadi kata dasar ini berguna ntuk memperkecil indeks hasil. Berikut contoh *stemming* :

Tabel 5. Contoh konversi *stemming* pada judul berita.

| **Sesudah** | | | | |
|---|---|---|---|---|
| tempat | usaha | berizin | bulan | dirazia |
| **Sesudah** | | | | |
| tempat | usaha | izin | bulan | razia |

Proses pengerjaan *text processing* dilakukan juga didalam *tools* RapidMiner. Pengumpulan data awalnya dengan menyalin 400 *link* artikel yang ada di *website* Radar Banjarmasin lalu di masukan kedalam file Excel. Selanjutnya proses pengambilan data judul dilakukan dengan bantuan pemrograman bahasa R. Setelah data mentah terkumpul lalu di masukan kedalam *tools* RapidMiner untuk *preprocessing text* dengan alur seperti gambar di bawah ini :



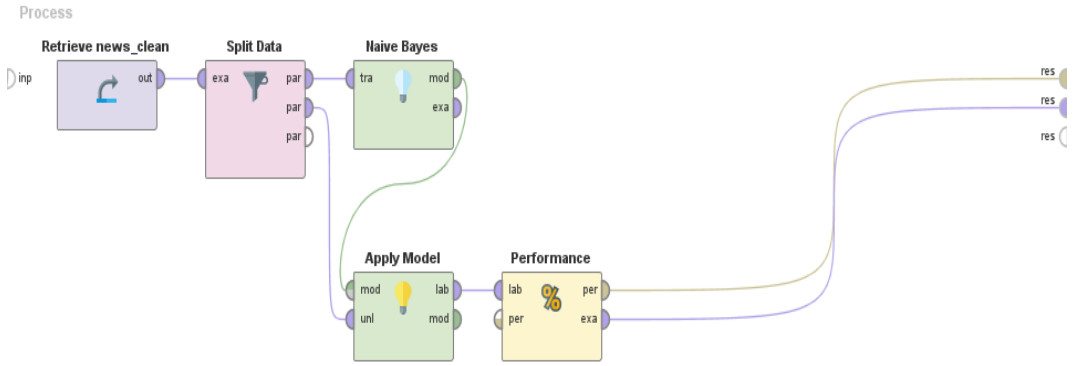

Gambar 1. Proses *preprocessing text*.

Data yang sudah melalui proses *case folding*, *tokenizing*, *stopword* dan *stemming* selanjutnya data ini menjadi data yang terstruktur sehingga data sudah bisa di lakukan proses selanjutnya.

### 3.3 Model Klasifikasi

Pada penelitian ini menggunakan tools atau alat dalam pemrosesan dan pengujian datanya. Perancangan model di dalam RapidMiner ini menggunakan metode *Naïve Bayes*. Penggunaan *tools* RapidMiner versi 9.9 untuk *prepocessing* data teks.

Tahapan awal yang dilakukan adalah memasukan data yang sudah terolah kedalam RapidMiner. *File* yang di unggah adalah berkas Excel yang memiliki data sudah terstruktur. Selanjutnya data dimasukan kedalam algortima *Naïve Bayes*. Setelah data dimasukan kedalam algoritma langkah selanjutnya adalah melihat performanya berikut gambar alur perancangan didalam program RapidMiner :




Gambar . Perancangan di dalam RapidMiner.

### 3.4 Hasil Pengujian

Hasil pengujian dari model yang sudah di rancang dan di jalankan mendapatkan hasil akurasi sebesar 78.75%.

accuracy: 78.75%

| | true banua | true hukum dan peristiwa | true bisnis | true sport | class precision |
|---|---|---|---|---|---|
| pred. banua | 8 | 1 | 2 | 0 | 72.73% |
| pred. hukum dan peristiwa | 0 | 18 | 0 | 0 | 100.00% |
| pred. bisnis | 3 | 0 | 17 | 0 | 85.00% |
| pred. sport | 9 | 1 | 1 | 20 | 64.52% |
| class recall | 40.00% | 90.00% | 85.00% | 100.00% | |

Gambar 3. Hasil akurasi model.

Untuk hasill *precision* mendapatkan hasil sebesar 80.56%.

weighted_mean_precision: 80.56%, weights: 1, 1, 1, 1

| | true banua | true hukum dan peristiwa | true bisnis | true sport | class precision |
|---|---|---|---|---|---|
| pred. banua | 8 | 1 | 2 | 0 | 72.73% |
| pred. hukum dan peristiwa | 0 | 18 | 0 | 0 | 100.00% |
| pred. bisnis | 3 | 0 | 17 | 0 | 85.00% |
| pred. sport | 9 | 1 | 1 | 20 | 64.52% |
| class recall | 40.00% | 90.00% | 85.00% | 100.00% | |

Gambar 4. Hasil Presisi model.

Sedangkan untuk hasil pengujian *recallnya* mendapatkan hasil sebesar 78.75%.

weighted_mean_recall: 78.75%, weights: 1, 1, 1, 1

| | true banua | true hukum dan peristiwa | true bisnis | true sport | class precision |
|---|---|---|---|---|---|
| pred. banua | 8 | 1 | 2 | 0 | 72.73% |
| pred. hukum dan peristiwa | 0 | 18 | 0 | 0 | 100.00% |
| pred. bisnis | 3 | 0 | 17 | 0 | 85.00% |
| pred. sport | 9 | 1 | 1 | 20 | 64.52% |
| class recall | 40.00% | 90.00% | 85.00% | 100.00% | |

Gambar 5. Hasil *Recall* model.

Dari satu kali menjalankan proses pemodelan *Naïve Bayes* dengan menggunakan pemodelan yang telah terimplementasi mendapatkan hasil akurasi sebesar 78.75% . Sedangkan untuk hasil Recall adalah 80.56% dan Precission adalah 78.75%.

## 4. SIMPULAN

Pengujian yang telah di lakukan dengan menggunakan metode *Naïve Bayes* untuk data judul berita dengan studi kasus Radar Banjarmasin. Metode ini dapat di gunakan untuk pengklasifikasian judul berita *online*. Dengan menggunakan 400 data dan 4 kategori di dapatkan bahwa metode *Naïve Bayes* dari model yang dibuat di RapidMiner mendapatkan hasil akurasi akhir sebanyak 78.75%. Sedangkan untuk hasil *recall* adalah 80.56% dan *precision* adalah 78.75%.

## DAFTAR PUSTAKA

[1] W. S. Ari Putra Wibowo, “Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Konten Berita Olahraga,” vol. XV, no. 1, pp. 15–20, 2020.

[2] S. Kurniawan, W. Gata, D. A. Puspitawati, N. -, M. Tabrani, and K. Novel, “Perbandingan Metode Klasifikasi Analisis Sentimen Tokoh Politik Pada Komentar Media Berita Online,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 3, no. 2, pp. 176–183, 2019, doi: 10.29207/resti.v3i2.935.

[3] I. Setiawan and D. Nursantika, “Klasifikasi Artikel Berita Menggunakan Metode Text Mining Dan Naive Bayes Classifier,” *Pros. SENIATI*, pp. 1–6, 2017, [Online]. Available: http://ejournal.itn.ac.id/index.php/seniati/article/view/790.

[4] I. Rasila, U. Ristian, J. Rekayasa Sistem Komputer, and F. H. MIPA Universitas Tanjungpura Jl Hadari Nawawi, “Implementasi Metode Naive Bayes Classifier Pada Sistem Pengklasifikasi Berita Otomatis Berbasis Website (Studi Kasus: Berita Lokal Dari Mediamassa Online Kalimantan Barat),” *Coding J. Komput. dan Apl.*, vol. 07, no. 2, pp. 49–60, 2019.

[5] A. Fattah and R. Setyadi, “Teknologi informasi dan pendidikan,” *J. Teknol. Inf.*

*dan Pendidik.*, vol. 12, no. 2, pp. 1–7, 2019.

[6] A. Y. Muniar, P. Pasnur, and K. R. Lestari, "Penerapan Algoritma K-Nearest Neighbor pada Pengklasifikasian Dokumen Berita Online," *Inspir. J. Teknol. Inf. dan Komun.*, vol. 10, no. 2, p. 137, 2020, doi: 10.35585/inspir.v10i2.2570.