

## Klasifikasi Data Paten Berdasarkan Judul Menggunakan Metode *Naïve Bayes Classifier* (Studi Kasus: Direktorat Inovasi dan Kawasan Sains dan Teknologi (DIKST) Universitas Brawijaya)

Syafira Alifah<sup>1</sup>, Dian Eka Ratnawati<sup>2</sup>, Budi Darma Setiawan<sup>3</sup>

Program Studi Teknologi Informasi, Fakultas Ilmu Komputer, Universitas Brawijaya  
Email: <sup>1</sup>syafiralifah@student.ub.ac.id, <sup>2</sup>dian\_ilkom@ub.ac.id, <sup>3</sup>s.budidarma@ub.ac.id

### Abstrak

Paten adalah salah satu perlindungan Hak Kekayaan Intelektual yang dimiliki inventor karena hasil penemuannya (invensi) dalam jangka waktu tertentu. Pendaftaran paten di lingkungan Universitas Brawijaya dapat diajukan ke Direktorat Inovasi dan Kawasan Sains dan Teknologi (DIKST). Klasifikasi secara manual pada jumlah paten yang banyak membutuhkan waktu dan beresiko terjadi kesalahan dalam pengelompokannya, sehingga perlu dilakukan klasifikasi data paten secara otomatis. Penelitian ini menerapkan algoritma klasifikasi *Naïve Bayes* untuk mengategorikan paten berdasarkan judul. Judul paten akan diolah melalui langkah-langkah seperti *text preprocessing*, penerapan TF-IDF, klasifikasi *Naïve Bayes*, serta tahap validasi. Salah satu tahap pada *text preprocessing* akan dilakukan pengurangan untuk mengetahui pengaruh *text preprocessing* terhadap performa algoritma *Naïve Bayes*. Kategori data yang diolah dalam penelitian ini adalah Metode, Proses, Alat, dan Formula dengan jumlah keseluruhan data, yaitu 376. Hasil penelitian klasifikasi paten berdasarkan judul menunjukkan performa lebih unggul menggunakan seluruh tahapan *text preprocessing* dengan *accuracy* sebesar 83.77%, *precision* sebesar 86.52%, *recall* sebesar 83.77%, dan *f-measure* sebesar 83.44%.

**Kata kunci:** klasifikasi, paten, kategori, DIKST, *naïve bayes*

### Abstract

Patent is one of the Intellectual Property Rights protections owned by inventors because of their inventions within a certain period of time. Patent registration within Brawijaya University can be submitted to the Direktorat Inovasi dan Kawasan Sains dan Teknologi (DIKST). Manual classification of a large number of patents takes time and risks errors in grouping, so it is necessary to classify patent data automatically. This research applies the *Naïve Bayes* classification algorithm to categorize patents based on title. Patent titles will be processed through steps such as *text preprocessing*, TF-IDF application, *Naïve Bayes* classification, and validation stages. One of the stages in *text preprocessing* will be reduced to determine the effect of *text preprocessing* on the performance of the *Naïve Bayes* algorithm. The data categories processed in this research are Methods, Processes, Tools, and Formulas with a total of 376 data. The results of patent classification research based on the title show superior performance using all stages of *text preprocessing* with *accuracy* of 83.77%, *precision* of 86.52%, *recall* of 83.77%, and *f-measure* of 83.44%.

**Keywords:** classification, patent, category, DIKST, *naïve bayes*

### 1. PENDAHULUAN

Menurut (Direktorat Jenderal Kekayaan Intelektual, 2020), paten memberikan hak khusus kepada penemu untuk melindungi invensinya selama periode tertentu. Pada konteks paten, penemu dianggap sebagai individu dengan ide terbaru atau dikenal sebagai inventor dan ciptaan yang dihasilkan disebut sebagai invensi. Invensi atau hasil penciptaan dapat berwujud produk, proses, atau

dapat merupakan gabungan keduanya (Direktorat Jenderal Kekayaan Intelektual, 2020). Hak Kekayaan Intelektual akan melindungi invensi jika terbukti unik dan belum pernah ada sebelumnya.

Pengajuan perlindungan Hak Kekayaan Intelektual paten pada lingkungan Universitas Brawijaya dapat didaftarkan pada Direktorat Inovasi dan Kawasan Sains dan Teknologi (DIKST). Kategori paten yang dapat diajukan sebagai hasil invensi adalah Metode, Proses,

Alat, dan Formula. Paten yang diajukan ke Direktorat Inovasi dan Kawasan Sains dan Teknologi (DIKST) telah banyak terdaftar sehingga paten perlu dikelompokkan sesuai kategori. Ketika mengklasifikasikan secara manual, pengelompokan paten membutuhkan waktu dan dapat berpotensi masuk ke dalam kategori yang salah.

Berdasarkan permasalahan di atas, penelitian ini dilakukan untuk menerapkan algoritma *Naïve Bayes* dalam klasifikasi judul paten Direktorat Inovasi dan Kawasan Sains dan Teknologi (DIKST). Hal ini bertujuan untuk menilai keakuratan algoritma tersebut dalam mengklasifikasikan paten ke dalam kategori yang tepat serta untuk mengetahui pengaruh pengurangan salah satu tahapan *text preprocessing* pada performa algoritma tersebut.

## 2. STUDI PUSTAKA

### 2.1. Penelitian Terkait

Nurdin, et al. (2021) menerapkan algoritma klasifikasi *Naïve Bayes* untuk mengelompokkan karya ilmiah dari Jurusan Teknik Informatika Universitas Malikussaleh. Karya ilmiah tersebut dibagi menjadi lima kategori, meliputi Sistem Pakar, Sistem Pengambilan Keputusan, Data Mining, Pengolahan Citra, dan Sistem Informasi Geografis (SIG). Data yang digunakan berjumlah 170, meliputi data latih sejumlah 150 dan data uji sejumlah 20. Penelitian klasifikasi pada karya ilmiah ini menghasilkan rata-rata senilai 86.68% pada *accuracy*.

Penelitian selanjutnya dilakukan oleh (Pujianto, et al., 2019) pada skripsi dan tugas akhir. Data skripsi dan tugas akhir diklasifikasikan menggunakan algoritma *Naïve Bayes* berdasarkan tema penelitian Kelompok Bidang Keahlian (KBK) di Jurusan Teknik Elektro FT UM. Jumlah data yang digunakan adalah 1103 data menggunakan 9 kategori yang berasal dari lingkup teknik informatika, teknik elektro, dan pendidikan vokasi. Klasifikasi pada penelitian oleh (Pujian, et al., 2019) mendapatkan rata-rata nilai 94% pada *accuracy*, 80% nilai *precision*, dan 69% nilai *recall*.

Penelitian lainnya yang menggunakan algoritma *Naïve Bayes* untuk mengklasifikasikan data dilakukan oleh (Putra, et al., 2019). Penelitian tersebut mengklasifikasikan Al-Qur'an terjemahan bahasa Indonesia ke dalam 4 kelas, yaitu Morality, Faith, Knowledge, dan *Muamalah*.

Jumlah data yang digunakan sebanyak 440 data. Klasifikasi dilakukan dengan 2 skenario pengujian, yaitu skenario pertama menggunakan 2 kelas dan skenario kedua menggunakan 4 kelas. Hasil performa terbaik didapatkan pada skenario pengujian pertama dengan nilai 90.95% pada *accuracy*.

### 2.2. Paten

Paten memberikan Hak Kekayaan Intelektual kepada pencipta ide terbaru yang belum pernah ada sebelumnya untuk mendaftarkan hasil pemikirannya. Seorang pencipta ide dalam konteks paten disebut inventor sedangkan hasil ciptaannya disebut invensi. Paten memiliki perlindungan yang terbatas dan berlangsung selama 20 tahun, dimulai dari invensi diajukan untuk permohonan paten. Jenis paten terbagi menjadi dua, paten sederhana dan paten. Paten sederhana tidak membutuhkan penelitian atau pengembangan, sedangkan paten umumnya mencakup beberapa klaim dan memerlukan proses penelitian atau pengembangan (Erlina & Safitri, 2020).

### 2.3. Text Mining

*Text mining* adalah metode menemukan pola dari data yang berbentuk teks untuk tujuan analisis tertentu (Dewi, 2023). Metode ini melakukan transformasi format teks yang tidak terstruktur ke dalam sebuah format yang terstruktur untuk mendapatkan informasi dan pengetahuan yang baru. Beberapa contoh yang memanfaatkan *text mining* dalam bidang industri antara lain, *customer service*, *healthcare*, dan *filter spam*.

### 2.4. Text Preprocessing

Metode *text preprocessing* dilakukan untuk memisahkan teks ke dalam format yang dipahami oleh komputer. Tahapan *text preprocessing* meliputi *Case Folding*, *Tokenization*, *Stopword Removal*, dan *Stemming*. Tujuan dari metode ini adalah memproses data agar lebih terstruktur dan memudahkan jalannya pemodelan (Rahayu, et al., 2023).

### 2.5. Term Weighting

Pada konteks pengolahan teks, *term weighting* adalah pembobotan yang digunakan untuk memperoleh representatif teks berdasarkan bobot pada tiap istilah. Metode

*Term Weighting* yang digunakan dalam penelitian ini adalah *Term Frequency* dan *Inverse Document Frequency* (TF-IDF). Metode TF-IDF bertujuan untuk menilai pentingnya sebuah kata dalam dokumen (Annisa & Kalifia, 2024).

### 2.6. Naïve Bayes Classifier

Algoritma *Naïve Bayes* mengelompokkan data teks ke dalam data latih dan data uji. Data latih berperan dalam melatih dan mengembangkan model, sementara data uji berperan untuk melakukan uji pada model klasifikasi setelah dilatih. Rumus penghitungan *Multinomial Naïve Bayes* ditunjukkan dalam Persamaan (1) berikut.

$$P(y|t) = \frac{P(y) \times P(t|y)}{P(t)}$$

(1)

Keterangan:

$P(y|t)$  = Probabilitas *posterior* dari kategori  $y$  pada saat *term* (kata)  $t$  muncul

$P(y)$  = Probabilitas *prior* dari dokumen yang muncul sebagai kategori  $y$

$P(t|y)$  = Probabilitas *likelihood* dari *term* (kata)  $t$  pada kategori  $y$

$P(t)$  = *Evidence*, probabilitas kemunculan *term* (kata)  $t$

*Multinomial Naïve Bayes* menjadi salah satu jenis algoritma *Naïve Bayes* yang diterapkan pada penelitian ini. Algoritma ini menggunakan konsep frekuensi *term* dalam dokumen yang mengevaluasi setiap kali kata muncul dalam dokumen tertentu (Singh et al., 2019). *Multinomial Naïve Bayes* melakukan penghitungan *likelihood* menggunakan hasil pembobotan TF-IDF (Rahman, Wiranto, & Doewes, 2017). Rumus penghitungan *likelihood Multinomial Naïve Bayes* dapat dilihat pada Persamaan (1) berikut.

$$P(t_n|y) = \frac{W_{yt} + 1}{(\sum W'_{\epsilon V W'_{yt}}) + B'}$$

(2)

Keterangan:

$W_{yt}$  = Nilai TF-IDF atau  $W$  dari *term* (kata)  $t$  dalam kategori  $y$

$\sum W'_{\epsilon V W'_{yt}}$  = Jumlah total  $W$  dari seluruh

*term* (kata)  $t$  dalam kategori  $y$

$B'$  = Jumlah  $W$  *term* (kata) unik (nilai IDF) pada seluruh dokumen

### 3. METODOLOGI PENELITIAN

Tahap awal menjalankan penelitian ini adalah mengidentifikasi masalah lalu dilanjutkan studi literatur terkait penelitian terdahul. Langkah selanjutnya mengumpulkan data yang akan diolah. Data yang diolah adalah data judul paten dan kategori paten dari Direktorat Inovasi dan Kawasan Sains dan Teknologi (DIKST). Data paten yang digunakan adalah 4 kategori, antara lain, Proses, Metode, Alat, dan Formula dengan total data 376. Data kemudian dilakukan pelabelan manual oleh pakar pemberi label. Setelah diberikan label, data diolah melalui tahapan *text preprocessing* untuk dibersihkan sehingga menghasilkan data yang berkualitas untuk memperlancar proses klasifikasi. Tahapan *text preprocessing* yang diterapkan pada penelitian ini dijabarkan sebagai berikut:

#### 1. Case Folding

Tahapan yang melibatkan konversi semua karakter ke bentuk huruf kecil dan menyeleksi karakter yang tidak berhubungan selain alfabet a-z disebut *Case Folding* (Nurdin et al., 2021).

#### 2. Tokenization

*Tokenization* merupakan tahapan yang memecah teks menjadi sebuah kata (Nurdin et al., 2021).

#### 3. Stopword Removal

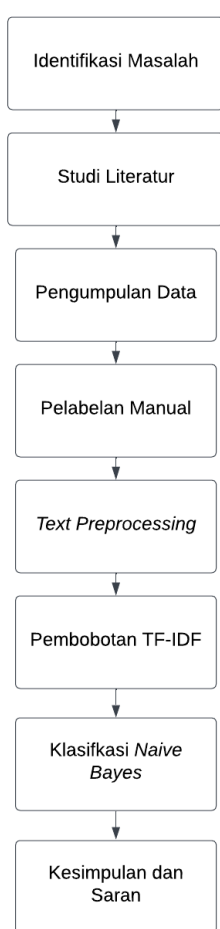
Metode *Stopword Removal* diterapkan untuk mengeliminasi kata-kata yang tidak memiliki relevansi terhadap data teks. Menghapus *stopword* bukan hanya memperkecil dimensi vektor, tetapi juga dapat meningkatkan performa akurasi, kalkulasi, dan eksekusi (Kaur & Buttar, 2018).

#### 4. Stemming

*Stemming* pada bahasa Indonesia menggunakan kata sisipan (infiks) dan kombinasi awal dan akhiran (konfiks) (Adriani et al., 2007). Penggunaan *Stemming* dilakukan dengan tujuan untuk menggantikan suatu kata ke bentuk dasarnya.

Tahapan setelah data teks melalui *preprocessing* adalah *Term Frequency-Inverse Document Frequency* (TF-IDF) yang berperan untuk menghitung bobotnya menggunakan.

Proses ini akan mengukur pentingnya kata yang muncul dalam dokumen (Annisa & Kalifia, 2024). Data yang selesai diproses pada tahap pembobotan selanjutnya dipisahkan menjadi data latih dan data uji. Proses klasifikasi akan dilakukan pada kedua data tersebut menggunakan algoritma *Naïve Bayes*. Hasil klasifikasi dari algoritma tersebut akan dilakukan pengujian untuk memastikan ketepatan hasil klasifikasi. Selanjutnya, penelitian ditutup dengan membuat kesimpulan dan saran berdasarkan penelitian yang telah dilakukan. Detail Metodologi penelitian ditampilkan dalam Gambar 1.



Gambar 1. Metodologi Penelitian

#### 4. HASIL DAN PEMBAHASAN

Pengujian dilakukan untuk mengevaluasi performa algoritma *Naïve Bayes* dalam mengklasifikasikan data paten berdasarkan judul. Data judul paten yang digunakan sebanyak 376 data yang meliputi kategori Metode sebanyak 107, Proses sebanyak 96, Alat sebanyak 90, dan Formula sebanyak 83. Pengujian dilakukan dengan 4 skenario. Skenario pengujian pertama adalah pengujian

*Naïve Bayes* menggunakan seluruh tahapan *text preprocessing* dan skenario pengujian kedua adalah pengujian *Naïve Bayes* tanpa satu tahapan *text preprocessing*. Kedua skenario pengujian tersebut menggunakan empat perbandingan data latih dan data uji yang bervariasi, meliputi 80%:20%, 70%:30%, 60%:40%, dan 50%:50%.

Skenario pengujian selanjutnya melakukan validasi melalui penerapan *K-Fold Cross Validation*. Skenario pengujian ketiga memvalidasi hasil klasifikasi algoritma *Naïve Bayes* menggunakan semua tahapan *text preprocessing*, sementara skenario pengujian keempat melakukan validasi hasil klasifikasi tanpa satu tahapan *text preprocessing*. Kedua skenario tersebut menggunakan jumlah *k* sebanyak 10 lipatan. Hasil setiap lipatan akan dirata-ratakan untuk mendapatkan hasil pengujian validasi *K-Fold Cross Validation*.

#### 4.1. Hasil Pengujian *Naïve Bayes*

Pengujian Pengujian pertama adalah pengujian *Naïve Bayes* menggunakan seluruh tahapan *text preprocessing*. Hasil pengujian algoritma *Naïve Bayes* pada skenario pertama menggunakan seluruh tahapan *text preprocessing* yang tercantum dalam Tabel 1 berikut.

Tabel 1. Hasil Pengujian Variasi Data Latih:Data Uji

Persentase Klasifikasi				
Data Latih & Data Uji	Accuracy	Precision	Recall	F-Measure
80% : 20%	83.77%	86.52%	83.77%	83.44%
70% : 30%	81.40%	84.22%	81.40%	81.15%
60% : 40%	81.12%	84.98%	81.12%	80.65%
50% : 50%	81.10%	84.55%	81.10%	80.70%

Berdasarkan Tabel 1, hasil pengujian tertinggi diperoleh pada persentase data latih dan data uji sebesar 80%:20% dengan nilai 83.77% untuk *accuracy*, nilai 86.52% untuk *precision*, nilai 83.77% untuk *recall*, serta nilai 83.44% untuk *f-measure*. Hasil tersebut membuktikan bahwa semakin banyak data latih yang digunakan maka model klasifikasi semakin paham pada pola data.

### 4.2. Hasil Pengujian Naïve Bayes Tanpa Satu Tahapan Text Preprocessing

Pengujian dengan pengurangan satu tahapan *text preprocessing* menggunakan 376 data yang mencakup kategori Metode, Proses, Alat, dan Formula. Pengujian ini bertujuan untuk mengetahui pengaruh *text preprocessing* pada hasil pengujian algoritma *Naïve Bayes*. Data judul paten dibagi menjadi data latih dan data uji dengan persentase perbandingan yang bervariasi. Hasil pengujian ini tersaji dalam Tabel 2 berikut.

Tabel 2 Hasil Pengujian Tanpa Satu Tahapan Text Preprocessing

Hasil Pengujian		Tanpa Case Folding	Tanpa Tokenization	Tanpa Stopword Removal	Tanpa Stemming
80%: 20%	Accuracy	79.83%	79.54%	79.51%	80.58%
	Precision	84.33%	83.88%	83.74%	84.57%
	Recall	79.83%	79.54%	79.51%	80.58%
	F-measure	79.40%	79.25%	78.79%	79.90%
70%: 30%	Accuracy	79.49%	79.24%	79.50%	80.24%
	Precision	84.03%	83.61%	84.60%	85.14%
	Recall	79.49%	79.24%	79.50%	80.24%
	F-measure	78.78%	78.73%	78.93%	80.05%
60%: 40%	Accuracy	79.02%	79.00%	79.27%	79.82%
	Precision	82.95%	83.69%	83.80%	84.31%
	Recall	79.02%	79.00%	79.27%	79.82%
	F-measure	78.86%	78.38%	78.81%	79.44%
50%: 50%	Accuracy	78.73%	74.86%	78.74%	79.73%
	Precision	83.09%	82.80%	83.50%	84.04%
	Recall	78.73%	74.86%	78.74%	79.73%
	F-measure	78.30%	78.11%	78.33%	79.41%

Pengujian pertama dengan pengurangan satu tahapan *text preprocessing* adalah tanpa *Case Folding*. Tahapan *text preprocessing* dimulai dengan *Tokenization*, *Stopword Removal*, dan *Stemming*. Hasil pengujian tanpa *Case Folding* pada Tabel 2 menunjukkan performa tertinggi pada data latih dan data uji, yaitu 80%:20% yang bernilai 79.83% untuk nilai *accuracy*, 84.33% untuk nilai *precision*, 79.83% untuk nilai *recall*, 79.40% untuk nilai *f-*

*measure*.

Pengujian kedua adalah pengurangan tahap *Tokenization*. Tahapan awal *text preprocessing* adalah *Case Folding*, *Stopword Removal*, dan *Stemming*. Performa terbaik yang ditunjukkan pada Tabel 2 dicapai oleh rasio pembagian 80% data latih dan 20% data uji. Hasil pengujian tersebut mendapatkan nilai 79.54% nilai *accuracy*, 83.88% nilai *precision*, 79.54% nilai *recall*, dan 79.25% nilai *f-measure*.

Pengujian ketiga adalah pengurangan tahap *Stopword Removal* pada *text preprocessing*. Tahapan *text preprocessing* tanpa *Stopword Removal* dimulai dari *Case Folding*, *Tokenization*, dan *Stemming*. Berdasarkan Tabel 2, hasil yang unggul dicapai ketika rasio pembagian data latih dan data uji pada 80%:20% dengan angka yang diperoleh 79.51% nilai *accuracy*, 83.74% untuk nilai *precision*, 79.51% untuk nilai *recall*, dan 78.79% untuk nilai *f-measure*.

Pengujian keempat adalah pengurangan tahap *Stemming* pada *text preprocessing*. Tahapan *text preprocessing* tanpa *Stemming* dimulai dari *Case Folding*, *Tokenization*, dan *Stopword Removal*. Pada Tabel 2 ditunjukkan hasil pengujian tanpa *Stemming* mendapatkan performa terbaik pada persentase 80% data latih dan 20% untuk data uji, dengan hasil 80.58% untuk *accuracy*, 84.57% untuk *precision*, 80.58% untuk *recall*, dan 79.90% untuk *f-measure*.

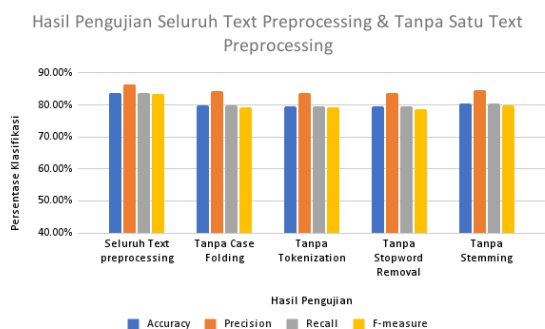
### 4.5. Analisis Hasil Pengujian Naïve Bayes

Pada tahap ini, analisis ditujukan untuk mengetahui hasil pengujian terbaik dari penggunaan semua tahap *text preprocessing* dan tanpa satu tahap *text preprocessing*. Pengujian dilakukan dengan membandingkan hasil pengujian keduanya berdasarkan hasil *K-Fold Cross Validation*. Tabel 3 berikut menampilkan hasil perbandingan kedua pengujian tersebut.

Tabel 3. Hasil Pengujian Seluruh Tahapan *Text Preprocessing* dan Tanpa Satu *Text Preprocessing*

Hasil Pengujian	Seluruh <i>Text Preprocessing</i>	Tanpa <i>Case Folding</i>	Tanpa <i>Tokenization</i>	Tanpa <i>Stopword Removal</i>	Tanpa <i>Stemming</i>
Accuracy	83.77%	79.83%	79.54%	79.51%	80.58%
Precision	86.52%	84.33%	83.88%	83.74%	84.57%
Recall	83.77%	79.83%	79.54%	79.51%	80.58%
F-measure	83.44%	79.40%	79.25%	78.79%	79.90%

Hasil pengujian berdasarkan Tabel 3 menunjukkan bahwa penggunaan seluruh tahapan *text preprocessing* menghasilkan performa yang lebih baik dibandingkan dengan pengurangan satu tahapan *text preprocessing*. Hasil pengujian menggunakan seluruh tahapan *text preprocessing* menunjukkan nilai sebesar 83.77% untuk *accuracy*, 86.52% untuk *precision*, 83.77% untuk *recall*, dan 83.44% untuk *f-measure*. Hal tersebut menunjukkan bahwa metode *preprocessing* berperan sangat penting dalam konteks *text mining* karena merupakan tahapan awal yang dilakukan dalam pengelolaan teks (Vijayarani, Ilamathi, & Nithya, 2015). Hasil pengujian tanpa satu *text preprocessing* memperoleh performa terbaik pada pengujian tanpa *Stemming* dengan nilai *accuracy* 80.58%, *precision* 84.57%, *recall* 80.58%, dan *f-measure* 79.90%. Grafik seluruh tahapan *text preprocessing* dan tanpa satu tahapan *text preprocessing* ditampilkan dalam Gambar 2 berikut.



Gambar 2. Hasil Pengujian Seluruh Tahapan *Text Preprocessing* dan Tanpa Satu *Text Preprocessing*

### 5. PENUTUP

Berdasarkan hasil pengujian dan analisis yang telah dilakukan, dapat ditarik kesimpulan

sebagai berikut:

1. Hasil pengujian algoritma *Naive Bayes* mendapatkan performa terbaik pada rasio data latih dan uji 80%:20% yang menghasilkan 83.77% nilai *accuracy*, 86.52% nilai *precision*, 83.77% nilai *recall*, serta 83.44% nilai *f-measure*
2. Hasil penggunaan seluruh tahapan *text preprocessing* menunjukkan performa yang lebih baik dibandingkan mengurangi satu tahapan *text preprocessing*. Hasil pengujian tanpa satu tahapan *text preprocessing* mendapatkan performa terbaik pada pengujian tanpa *Stemming* yang menghasilkan *accuracy* 80.58%, *precision* 84.57%, *recall* 80.58%, dan *f-measure* 79.90%. Hasil pengujian tersebut menunjukkan bahwa penggunaan *text preprocessing* mempengaruhi hasil akurasi klasifikasi algoritma *Naive Bayes*.

Saran untuk pengembangan penelitian selanjutnya adalah menambahkan jumlah dataset dan abstrak paten untuk memperbanyak variasi data saat proses klasifikasi guna meningkatkan performa klasifikasi.

### 6. DAFTAR PUSTAKA

Direktorat Jenderal Kekayaan Intelektual. 2020. Modul Kekayaan Intelektual Tingkat Dasar Bidang Paten (Edisi 2020).

Nurdin, N., Suhendri, M., Afrilia, Y. dan Rizal, R., 2021. Klasifikasi Karya Ilmiah (Tugas Akhir) Mahasiswa Menggunakan Metode Naive Bayes Classifier (NBC). *Sistemasi: Jurnal Sistem Informasi*, 10(2), pp.268-279.

Pujianto, U., Widiyaningtyas, T., Prasetya, D.D. dan Romadhon, B., 2019. Penerapan algoritma naive bayes classifier untuk klasifikasi judul skripsi dan tugas akhir berdasarkan Kelompok Bidang Keahlian. *TEKNO: Jurnal Teknologi Elektro Dan Kejuruan*, 27(1), pp.79-92.

Putra, S.J., Sugiarti, Y., Dimas, G., Gunawan, M.N., Sutabri, T. dan Suryatno, A., 2019, November. Document Classification using Naïve Bayes for Indonesian Translation of the Quran. In *2019 7th International Conference on Cyber and IT Service Management (CITSM)* (Vol. 7, pp. 1-4). IEEE.

- Erlina, B. dan Safitri, M., 2020. Analisis Komparasi Antara Perlindungan Paten Biasa dengan Paten Sederhana Berdasarkan Undang-Undang Paten. *Pranata Hukum*, 15(1), pp.30-42.
- Dewi, N.K., 2023. Identifikasi Berita Hoax dengan Menerapkan Algoritma Text Mining. *Journal of Informatics, Electrical and Electronics Engineering*, 2(3), pp.65-74.
- Rahayu, K., Fitria, V., Septhya, D., Rahmadden, R. dan Efrizoni, L., 2023. Klasifikasi Teks untuk Mendeteksi Depresi dan Kecemasan pada Pengguna Twitter Berbasis Machine Learning: Text Classification for Detecting Depression and Anxiety among Twitter Users based on Machine Learning. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 3(2), pp.108-114.
- Annisa, L. dan Kalifia, A.D., 2024. Analisis Teknik TF-IDF Dalam Identifikasi Faktor-Faktor Penyebab Depresi Pada Individu. *Gudang Jurnal Multidisiplin Ilmu*, 2(1), pp.302-307.
- Rahman, A., Wiranto, W. dan Doewes, A., 2017. Online news classification using multinomial naive bayes. *ITSMART: Jurnal Teknologi dan Informasi*, 6(1), pp.32-38.
- Kaur, J. dan Buttar, P.K., 2018. A systematic review on stopword removal algorithms. *International Journal on Future Revolution in Computer Science & Communication Engineering*, 4(4), pp.207-210.
- Adriani, M., Asian, J., Nazief, B., Tahaghoghi, S.M. dan Williams, H.E., 2007. Stemming Indonesian: A confix-stripping approach. *ACM Transactions on Asian Language Information Processing (TALIP)*, 6(4), pp.1-33.
- Vijayarani, S., Ilamathi, M.J. dan Nithya, M., 2015. Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, 5(1), pp.7-16.