

**MODEL HYBRID PRA-PEMROSESAN DAN SELEKSI FITUR  
UNTUK KLASIFIKASI BERITA AKADEMIK: INTEGRASI  
*ENHANCED CONFIX STRIPPING* DAN *CHI-SQUARED* PADA  
*NAÏVE BAYES CLASSIFIER***

**TUGAS AKHIR**



Oleh:

**M. NAUFAL MUHADZIB AL-FARUQ**

NIM.2108096034

**PROGRAM STUDI TEKNOLOGI INFORMASI  
FAKULTAS SAINS DAN TEKNOLOGI  
UNIVERSITAS ISLAM NEGERI WALISONGO SEMARANG**

**2024**

**MODEL HYBRID PRA-PEMROSESAN DAN SELEKSI FITUR  
UNTUK KLASIFIKASI BERITA AKADEMIK: INTEGRASI  
*ENHANCED CONFIX STRIPPING* DAN *CHI-SQUARED* PADA  
*NAÏVE BAYES CLASSIFIER***

**TUGAS AKHIR**



Oleh:

**M. NAUFAL MUHADZIB AL-FARUQ**

NIM.2108096034

**PROGRAM STUDI TEKNOLOGI INFORMASI  
FAKULTAS SAINS DAN TEKNOLOGI  
UNIVERSITAS ISLAM NEGERI WALISONGO SEMARANG**

**2024**



## PERNYATAAN KEASLIAN NASKAH

Yang bertandatangan di bawah ini

Nama : M. Naufal Muhadzib Al-Faruq

NIM : 2108096034

Jurusan : Teknologi Informasi

Menyatakan bahwa skripsi berjudul:

**MODEL HYBRID PRA-PEMROSESAN DAN SELEKSI FITUR  
UNTUK KLASIFIKASI BERITA AKADEMIK: INTEGRASI  
*ENHANCED CONFIX STRIPPING* DAN *CHI-SQUARED* PADA  
*NAÏVE BAYES CLASSIFIER***

Secara keseluruhan adalah hasil penelitian/karya sendiri,  
kecuali bagian tertentu yang dirujuk sumbernya.

Semarang, 18 Juni 2025

Pembuat Pernyataan



A handwritten signature in black ink, appearing to be 'Fauz' with a stylized flourish at the end.

M. Naufal Muhadzib Al-Faruq  
NIM. 2108096026



# LEMBAR PENGESAHAN



KEMENTERIAN AGAMA  
UNIVERSITAS ISLAM NEGERI WALISONGO  
FAKULTAS SAINS DAN TEKNOLOGI  
Jl. Prof. Dr. Hamka Ngaliyan Semarang  
Telp.024-7601259 Fax. 7615387

## LEMBAR PENGESAHAN

Naskah skripsi berikut ini:

Judul : Model Hybrid Pra-Pemrosesan dan Seleksi Fitur  
untuk Klasifikasi Berita Akademik: Integrasi  
Enhanced Conflx Stripping dan Chi-Squared pada  
Naïve Bayes Classifier  
Penulis : M. Naufal Muhadzib Al-Faruq  
NIM : 2108096034  
Jurusan : Teknologi Informasi

Telah diujikan dalam sidang tugas akhir oleh Dewan Penguji  
Fakultas Sains dan Teknologi UIN Walisongo dan dapat diterima sebagai  
salah satu syarat memperoleh gelar sarjana dalam Ilmu Teknologi  
Informasi.

Semarang, 02 Juli 2025

### DEWAN PENGUJI

Ketua Sidang

Mokhammad Ikhlil Mustofa, M.Kom  
NIP. 198808072019031010

Sekretaris Sidang

Dr. Wenty Dwi Yuniarti, S.Pd., M.Kom  
NIP. 197706222006042005

Penguji Utama I

Saiful Bakhri, M.MSI  
NIP. 198810302019031011

Penguji Utama II

Adzhal Arwani Mahfudh, M.Kom  
NIP. 199107032019031006

Pembimbing I

Dr. Wenty Dwi Yuniarti, S.Pd., M.Kom  
NIP. 197706222006042005

Pembimbing II

Dr. Khothibul Umam S.T., M. Kom  
NIP. 197908272011011007



## NOTA PEMBIMBING

Semarang, 18 Juni 2025

Yth. Ketua Program Studi Teknologi Informasi

Fakultas Sains dan Teknologi

UIN Walisongo Semarang

*Assalamu'alaikum Wr.Wb.*

Dengan ini diberitahukan bahwa saya telah melakukan bimbingan, arahan dan koreksi naskah skripsi dengan:

Judul : Model Hybrid Pra-Pemrosesan dan Seleksi Fitur untuk Klasifikasi Berita Akademik: Integrasi *Enhanced Confix Stripping* dan *Chi-Squared* pada *Naïve Bayes Classifier*

Penulis : **M. Naufal Muhadzib Al-Faruq**

NIM : 2108096034

Jurusan : Teknologi Informasi

Saya memandang bahwa naskah skripsi tersebut sudah dapat diajukan kepada Fakultas Sains dan Teknologi UIN Walisongo Semarang untuk diujikan dalam Sidang Munaqosah.

*Wassalamu'alaikum Wr.Wb.*

Pembimbing I,



Dr. Wenty Dwi Yuniarti S.Pd., M.Kom

NIP. 197706222006042005



## NOTA PEMBIMBING

Semarang, 18 Juni 2025

Yth. Ketua Program Studi Teknologi Informasi

Fakultas Sains dan Teknologi

UIN Walisongo Semarang

*Assalamu'alaikum Wr.Wb.*

Dengan ini diberitahukan bahwa saya telah melakukan bimbingan, arahan dan koreksi naskah skripsi dengan:

Judul : Model Hybrid Pra-Pemrosesan dan Seleksi Fitur untuk Klasifikasi Berita Akademik: Integrasi Enhanced Confix Stripping dan Chi-Squared pada Naïve Bayes Classifier

Penulis : **M. Naufal Muhadzib Al-Faruq**

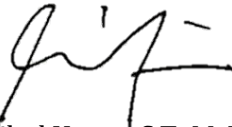
NIM : 2108096034

Jurusan : Teknologi Informasi

Saya memandang bahwa naskah skripsi tersebut sudah dapat diajukan kepada Fakultas Sains dan Teknologi UIN Walisongo Semarang untuk diujikan dalam Sidang Munaqosah.

*Wassalamu'alaikum Wr.Wb.*

Pembimbing II,



Dr. Khotibul Umam S.T., M. Kom

NIP/197908272011011007



## **LEMBAR PERSEMBAHAN**

Dengan rasa syukur yang mendalam, dengan telah diselesaikannya Tugas Akhir ini penulis mempersembahkan kepada:

1. Keluarga besar Penulis yang senantiasa mendukung dan selalu mendoakan penulis.
2. Segenap civitas akademik UIN Walisongo Semarang, staff, pengajar, karyawan, dan seluruh mahasiswa semoga selalu dalam keadaan sehat dan tetap semangat dalam beraktivitas mengisi hari-harinya di kampus tercinta UIN Walisongo semarang.
3. Teman-teman Penulis khususnya mahasiswa Teknologi Informasi angkatan 2021, yang selalu memberikan semangat dalam suka maupun duka selama pembuatan tugas akhir ini sehingga tugas akhir dapat terselesaikan dengan baik.



## **MOTTO**

“Aku adalah catatan dari perjuanganmu, bukti dari  
perlawananmu, dan hadiah dari penderitaanmu”

Sung Jin-Woo, Solo Leveling



## KATA PENGANTAR

Puji dan syukur saya panjatkan ke hadirat Allah SWT, atas segala nikmat-Nya sehingga penulis mampu menyelesaikan penelitian ini. Shalawat serta salam selalu dicurahkan kepada junjungan kita Nabi Muhammad SAW. Skripsi konversi dari artikel ilmiah saya yang berjudul **MODEL HYBRID PRA-PEMROSESAN DAN SELEKSI FITUR UNTUK KLASIFIKASI BERITA AKADEMIK: INTEGRASI *ENHANCED CONFIX STRIPPING* DAN *CHI-SQUARED* PADA *NAÏVE BAYES CLASSIFIER***. Pada kesempatan ini, penulis ingin menyampaikan ucapan terima kasih kepada semua pihak yang telah membantu, membimbing, memberikan arahan serta mendukung penulis sehingga dapat menyelesaikan penelitian dan penulisan skripsi ini.

1. Bapak Prof. Dr. Nizar, M.Ag, selaku Rektor Universitas Islam Negeri Walisongo Semarang.
2. Bapak Prof. Dr. H. Musahadi, M.Ag, selaku Dekan Fakultas Teknologi Informasi Universitas Islam Negeri Walisongo Semarang.
3. Bapak Dr. Khothibul Umam S.T., M. Kom selaku Ketua Program Studi Teknologi Informasi, Fakultas Sains dan Teknologi, UIN Walisongo Semarang sekaligus Wali Dosen dan Dosen Pembimbing II yang telah memberikan arahan,

saran dan dukungan dalam membimbing penulis sehingga dapat menyelesaikan skripsi.

4. Ibu Dr. Wenty Dwi Yuniarti S.Pd.,M.Kom selaku Pembimbing I yang telah memberikan arahan, saran dan dukungan dalam membimbing penulis sehingga dapat menyelesaikan skripsi.
5. Orang Tua dan keluarga atas semangat, didikan, motivasi, bimbingan yang telah diberikan sampai saat ini, semoga Allah senantiasa memberikan kesehatan dan keberkahan.
6. Ibu Hj. Maya Rini Handayani M.Kom. selaku dosen Mata Kuliah NLP, telah memberikan arahan dalam membimbing penulis, memberikan ide penelitian, saran dan dukungan dalam membimbing penulis.
7. Untuk Aas Aisyiah, terima kasih telah menjadi cahaya yang tak pernah padam di tengah segala proses dan perjuangan ini, Archon Team, Bisa IT team serta teman-teman Teknologi Informasi Kelas A Angkatan 2021 yang membantu penulis dalam penyusunan skripsi ini.
8. Semua pihak yang telah membantu hingga terselesaikan pembuatan skripsi yang tidak dapat disebutkan satu persatu.

Penulis menyadari bahwa dalam penyusunan Skripsi ini masih terdapat kekurangan dan penulis berharap semoga Skripsi ini bisa memberikan manfaat kepada para pembaca khususnya bagi penulis secara pribadi. Amin Ya Rabbal Alamin

Semarang, 18 Juni 2025

Penulis



**MODEL HYBRID PRA-PEMROSESAN DAN SELEKSI FITUR  
UNTUK KLASIFIKASI BERITA AKADEMIK: INTEGRASI  
*ENHANCED CONFIX STRIPPING* DAN *CHI-SQUARED* PADA  
*NAÏVE BAYES CLASSIFIER***

Oleh:

M. Naufal Muhadzib Al-Faruq

NIM.2108096034

**ABSTRAK**

Klasifikasi berita akademik pada situs web perguruan tinggi tetap menjadi tantangan karena meningkatnya volume konten dan kurangnya sistem kategorisasi yang efisien. Di UIN Walisongo Semarang, permasalahan ini menghambat mahasiswa, dosen, dan masyarakat umum dalam mengakses informasi yang relevan secara mudah. Penelitian ini bertujuan untuk mengembangkan sistem klasifikasi otomatis untuk berita akademik guna mengatasi permasalahan tersebut. Kami menerapkan model *Naïve Bayes Classifier* yang ditingkatkan dengan pembobotan *Term Frequency*, algoritma *Enhanced Confix Stripping Stemmer* untuk praproses bahasa Indonesia, serta seleksi fitur *Chi-Squared* untuk mengidentifikasi kata-kata yang paling informatif. Dataset terdiri dari 880 artikel berita akademik dari situs web UIN Walisongo, yang dibagi

menjadi 704 dokumen pelatihan dan 176 dokumen pengujian. Sistem ini mencapai akurasi sebesar 95% pada data uji. Untuk mengevaluasi kemampuan generalisasi, kami menggunakan 12 artikel baru sebagai data evaluasi dan memperoleh akurasi sebesar 83,3%. Tahapan praproses memainkan peran penting dalam mengurangi kompleksitas morfologis, sementara skor *Chi-Squared* meningkatkan relevansi fitur yang dipilih. Penelitian ini menyoroti pentingnya teknik klasifikasi teks yang andal dalam sistem informasi akademik, khususnya dalam konteks bahasa Indonesia yang memiliki tantangan morfologi tersendiri. Model yang diusulkan menunjukkan performa yang kuat, skalabilitas, dan potensi untuk diintegrasikan ke dalam portal akademik guna meningkatkan pencarian informasi. Studi ini memberikan kontribusi signifikan dalam bidang *Natural Language Processing* dan *machine learning* terapan di lingkungan akademik, khususnya untuk konten berbahasa Indonesia. Penelitian ini juga menawarkan solusi yang efektif untuk manajemen konten akademik otomatis dalam sistem informasi kelembagaan.

**Kata kunci:** *Berita Akademik, Chi-squared, Enhanced Confix Stripping Stemmer, Naive Bayes Classifier, Classification*

## DAFTAR ISI

|                                         |              |
|-----------------------------------------|--------------|
| <b>PERNYATAAN KEASLIAN NASKAH .....</b> | <b>iii</b>   |
| <b>LEMBAR PENGESAHAN .....</b>          | <b>v</b>     |
| <b>NOTA PEMBIMBING .....</b>            | <b>vii</b>   |
| <b>NOTA PEMBIMBING .....</b>            | <b>ix</b>    |
| <b>LEMBAR PERSEMBAHAN .....</b>         | <b>xi</b>    |
| <b>MOTTO .....</b>                      | <b>xiii</b>  |
| <b>KATA PENGANTAR.....</b>              | <b>xv</b>    |
| <b>ABSTRAK .....</b>                    | <b>xix</b>   |
| <b>DAFTAR ISI .....</b>                 | <b>xxi</b>   |
| <b>DAFTAR GAMBAR.....</b>               | <b>xxv</b>   |
| <b>DAFTAR TABEL .....</b>               | <b>xxvii</b> |
| <b>DAFTAR LAMPIRAN .....</b>            | <b>xxix</b>  |
| <b>BAB I.....</b>                       | <b>1</b>     |
| <b>PENDAHULUAN.....</b>                 | <b>1</b>     |
| <b>A. Latar Belakang Masalah .....</b>  | <b>1</b>     |
| <b>B. Identifikasi Masalah.....</b>     | <b>9</b>     |
| <b>C. Rumusan Masalah .....</b>         | <b>10</b>    |
| <b>D. Batasan Masalah .....</b>         | <b>11</b>    |
| <b>E. Tujuan Pengembangan.....</b>      | <b>11</b>    |
| <b>F. Manfaat Penelitian .....</b>      | <b>12</b>    |
| 1. Manfaat Akademis.....                | 12           |
| 2. Manfaat Praktis.....                 | 12           |

|                                                                   |           |
|-------------------------------------------------------------------|-----------|
| <b>BAB II .....</b>                                               | <b>13</b> |
| <b>LANDASAN PUSTAKA.....</b>                                      | <b>13</b> |
| <b>A. Portal Berita Online .....</b>                              | <b>13</b> |
| <b>B. UIN Walisongo .....</b>                                     | <b>15</b> |
| <b>C. <i>Naïve Bayes Classifier</i> .....</b>                     | <b>16</b> |
| <b>D. <i>Algoritma Enhanced Confix Stripping Stemmer</i> ....</b> | <b>17</b> |
| <b>E. <i>Chi – Squared</i> .....</b>                              | <b>18</b> |
| <b>BAB III.....</b>                                               | <b>21</b> |
| <b>METODOLOGI PENELITIAN .....</b>                                | <b>21</b> |
| <b>A. Pengumpulan Data .....</b>                                  | <b>22</b> |
| <b>B. <i>Text Preprocessing</i> .....</b>                         | <b>23</b> |
| <b>C. <i>Chi-Squared</i> .....</b>                                | <b>25</b> |
| <b>D. <i>Algoritma Enhanced Confix Stripping Stemmer</i> ....</b> | <b>27</b> |
| <b>E. <i>Term Frequency</i> .....</b>                             | <b>32</b> |
| <b>F. <i>Naive Bayes Classifier</i> .....</b>                     | <b>32</b> |
| <b>BAB IV .....</b>                                               | <b>37</b> |
| <b>HASIL DAN PEMBAHASAN .....</b>                                 | <b>37</b> |
| <b>A. <i>Scraping Data Berita</i> .....</b>                       | <b>38</b> |
| <b>B. <i>Preprocessing</i> .....</b>                              | <b>41</b> |
| <b>C. <i>Term Frequency</i> .....</b>                             | <b>51</b> |
| <b>D. <i>Chi-Squared Feature Selection</i> .....</b>              | <b>53</b> |
| <b>E. <i>Training Model Naive Bayes</i> .....</b>                 | <b>57</b> |
| <b>F. <i>Testing Model</i> .....</b>                              | <b>59</b> |
| <b>G. <i>Evaluation</i> .....</b>                                 | <b>63</b> |

|                                   |            |
|-----------------------------------|------------|
| <b>BAB V.....</b>                 | <b>73</b>  |
| <b>SIMPULAN DAN SARAN .....</b>   | <b>73</b>  |
| <b>A.    Simpulan .....</b>       | <b>73</b>  |
| <b>B.    Saran .....</b>          | <b>75</b>  |
| <b>DAFTAR PUSTAKA .....</b>       | <b>77</b>  |
| <b>LAMPIRAN - LAMPIRAN .....</b>  | <b>81</b>  |
| <b>DAFTAR RIWAYAT HIDUP .....</b> | <b>125</b> |



## DAFTAR GAMBAR

| Gambar       | Judul                                                                                   | Halaman |
|--------------|-----------------------------------------------------------------------------------------|---------|
| Gambar 3. 1  | Alur penelitian sistem<br>klasifikasi berita akademik                                   | 22      |
| Gambar 4. 1  | <i>Pseudocode scraping data<br/>berita</i>                                              | 39      |
| Gambar 4. 2  | <i>Pseudocode preprocessing</i>                                                         | 41      |
| Gambar 4. 3  | <i>Flowchart enhanced confix<br/>stripping stemmer</i>                                  | 46      |
| Gambar 4. 4  | <i>Pseudocode term frequency</i>                                                        | 51      |
| Gambar 4. 5  | <i>Pseudocode chi-squared</i>                                                           | 54      |
| Gambar 4. 6  | <i>Flowchart feature selection chi-<br/>squared</i>                                     | 55      |
| Gambar 4. 7  | <i>Pseudocode training model<br/>naive bayes</i>                                        | 58      |
| Gambar 4. 8  | <i>Pseudocode testing model</i>                                                         | 60      |
| Gambar 4. 9  | <i>Pseudocode evaluation</i>                                                            | 64      |
| Gambar 4. 10 | Perbandingan kategori prediksi<br>dan kategori aktual dalam hasil<br>klasifikasi berita | 70      |



## DAFTAR TABEL

| <b>Tabel</b> | <b>Judul</b>                                                     | <b>Halaman</b> |
|--------------|------------------------------------------------------------------|----------------|
| Tabel 3. 1   | Penambahan dan Modifikasi pada Aturan Awalan Kata                | 30             |
| Tabel 4. 1   | Contoh dokumen pelatihan                                         | 40             |
| Tabel 4. 2   | Contoh hasil dokumen dari proses case folding                    | 42             |
| Tabel 4. 3   | Contoh hasil dokumen dari proses tokenizing                      | 43             |
| Tabel 4. 4   | Contoh hasil dokumen dari proses filtering                       | 45             |
| Tabel 4. 5   | Contoh dokumen hasil penerapan enhanced confix stripping stemmer | 49             |
| Tabel 4. 6   | Contoh dokumen hasil penerapan term frequency                    | 52             |
| Tabel 4. 7   | Contoh dokumen hasil penerapan chi-squared                       | 57             |
| Tabel 4. 8   | Tabel hasil pelatihan model                                      | 59             |
| Tabel 4. 9   | Contoh dokumen pengujian                                         | 61             |
| Tabel 4. 10  | Contoh dokumen uji yang telah dikategorikan                      | 62             |
| Tabel 4. 11  | Hasil pengujian                                                  | 65             |
| Tabel 4. 12  | Evaluation metrics                                               | 71             |



**DAFTAR LAMPIRAN**

| <b>Lampiran</b> | <b>Judul</b>                | <b>Halaman</b> |
|-----------------|-----------------------------|----------------|
| Lampiran 1      | <i>Letter of Acceptance</i> | 82             |
| Lampiran 2      | Bukti Korespondensi         | 83             |
| Lampiran 3      | Artikel                     | 89             |
| Lampiran 4      | <i>Source Code</i>          | 109            |
| Lampiran 5      | <i>Dataset</i>              | 123            |



# **BAB I**

## **PENDAHULUAN**

### **A. Latar Belakang Masalah**

Perkembangan teknologi digital pada era modern telah mendorong transformasi dalam berbagai aspek kehidupan, termasuk dalam hal akses terhadap informasi. Salah satu bentuk transformasi tersebut adalah kehadiran berita online, yang kini menjadi aspek penting dalam mendukung proses penelitian dan pengembangan akademik di lingkungan perguruan tinggi.

Media berita online menyediakan informasi kepada masyarakat melalui platform digital yang dapat diakses kapan saja dan di mana saja (Lestari, 2020). Keunggulan ini menjadikan berita online sebagai sumber informasi yang efisien dan responsif terhadap kebutuhan pengguna di era digital. Selain menyajikan informasi dengan cepat, media online juga memungkinkan pengguna untuk mengakses berbagai topik secara real time, baik yang bersifat lokal maupun global. Fleksibilitas dalam akses ini turut mendukung penyebaran pengetahuan secara lebih luas, khususnya dalam ranah pendidikan dan akademik, karena mampu menjangkau audiens tanpa batasan geografis maupun waktu.

Universitas Islam Negeri (UIN) Walisongo Semarang sebagai institusi pendidikan tinggi yang berfokus pada pengembangan ilmu pengetahuan dan keilmuan Islam, tidak terlepas dari tantangan dalam menyediakan akses informasi yang efektif dan efisien bagi civitas akademika serta masyarakat umum. Seiring meningkatnya jumlah berita yang dipublikasikan melalui portal web resmi universitas, kebutuhan akan sistem pengelompokan informasi yang terstruktur menjadi semakin mendesak. Banyaknya berita dengan beragam topik menuntut adanya sistem kategorisasi yang mampu mengatur dan menyajikan informasi secara lebih sistematis, sehingga memudahkan pengguna baik mahasiswa, dosen, maupun masyarakat umum dalam menelusuri berita berdasarkan topik yang dibutuhkan.

Penerapan sistem kategorisasi berita yang tepat tidak hanya meningkatkan kemudahan dalam mengakses informasi, tetapi juga menunjang efisiensi dalam proses pencarian data yang relevan. Dengan pengelompokan berdasarkan kategori yang jelas, pengguna dapat menemukan konten yang sesuai dengan kebutuhan akademik dan informasional secara cepat dan tepat. Hal ini berkontribusi positif terhadap peningkatan kualitas layanan informasi kampus, sekaligus memperkuat peran

portal berita sebagai pusat informasi yang kredibel dan terstruktur.

Untuk mendukung proses kategorisasi berita secara lebih optimal, diperlukan pendekatan berbasis teknologi yang mampu mengolah volume data dalam jumlah besar secara otomatis. Salah satu pendekatan yang digunakan dalam penelitian ini adalah data *mining*, yaitu proses untuk mengekstraksi atau menggali informasi tersembunyi dari data yang sebelumnya belum diketahui (Purba et al., 2019). Data yang digunakan berasal dari portal web resmi UIN Walisongo, dengan teknik utama berupa klasifikasi, yakni proses membangun model yang dapat membedakan dan menggambarkan kelas data, sehingga memungkinkan prediksi terhadap objek yang belum diketahui labelnya (Widaningsih, 2019).

Sebelum data berita diklasifikasi, diperlukan tahapan awal yang disebut *preprocessing*, yang mencakup proses *case folding*, *tokenizing*, *filtering*, dan *stemming* (Nurdin et al., 2021). Salah satu tahap krusial dalam *preprocessing* adalah *stemming*, yaitu proses penghilangan imbuhan pada kata untuk memperoleh bentuk dasarnya. Dalam penelitian ini, digunakan algoritma *Enhanced Confix Stripping Stemmer*, yang merupakan pengembangan dari algoritma *Confix Stripping* dengan sejumlah perbaikan

aturan, sehingga mampu menurunkan tingkat kesalahan dalam proses stemming (Billah et al., 2023).

Metode klasifikasi yang digunakan adalah *Naïve Bayes Classifier*, yang telah terbukti efektif dalam berbagai bidang, termasuk dalam klasifikasi data medis, jaringan komputer, dan teks. Keunggulan utama metode ini terletak pada kesederhanaan, kecepatan komputasi, serta efektivitasnya dalam menghasilkan prediksi yang akurat (Sholekha et al., 2022). Oleh karena itu, metode ini dipandang ideal untuk diterapkan dalam proses klasifikasi berita online di lingkungan akademik.

Penelitian mengenai klasifikasi berita online telah banyak dilakukan oleh peneliti sebelumnya, baik dengan menggunakan pendekatan serupa maupun metode yang berbeda. Hal ini menunjukkan pentingnya pengembangan sistem klasifikasi berita berbasis teknologi untuk mendukung penyebaran informasi akademik secara lebih sistematis dan efisien.

Penelitian tentang klasifikasi berita daring telah banyak dilakukan oleh para peneliti terdahulu. Pramudita dkk. (2018) dalam penelitiannya menggunakan *Naïve Bayes Classifier* dengan *Enhanced Confix Stripping Stemmer* untuk mengklasifikasikan berita olahraga. *Data training* terdiri dari 18 berita olahraga yang dipilih secara acak oleh

pengguna. Analisis pengguna menunjukkan bahwa 14 dari 18 berita terklasifikasi dengan benar, sehingga menghasilkan akurasi sebesar 77%. Temuan ini menunjukkan bahwa metode yang diusulkan cukup menjanjikan untuk mengklasifikasikan berita olahraga ke dalam kategori yang telah ditentukan sebelumnya seperti Sepak Bola, Bola Basket, Raket, Formula 1, Motto GP, dan lain-lain (Pramudita et al., 2018).

Penelitian yang dilakukan oleh Dio Ariadi dan Kartika Fithriasari menggunakan dua metode klasifikasi, yaitu *Naïve Bayes Classifier* dan *Support Vector Machine*. Kedua metode ini dibandingkan untuk menentukan hasil klasifikasi yang paling baik. Hasil klasifikasi menggunakan *Naïve Bayes Classifier* menunjukkan performa yang cukup baik berdasarkan data pengujian dengan akurasi, *precision*, *recall*, dan *F-Measure* masing-masing se besar 82,2%, 83,9%, 82.2%, 82.4%. Sementara itu, metode *Support Vector Machine* dengan *kernel RBF* dan *kernel linier* pada *word vector* 10.000 menunjukkan kinerja yang sama baiknya dalam mengklasifikasikan berita Indonesia. Data pengujian menghasilkan akurasi, *precision*, *recall*, *F-Measure* masing-masing se besar 88,1%, 89,1%, 88,1%, dan 88.3%. Dari hasil perbandingan kedua metode ini dapat disimpulkan bahwa *Support Vector Machine* dengan

*kernel RBF* dan *linier* lebih unggul dibandingkan dengan *Naïve Bayes Classifier* (Sani et al., 2022).

Penelitian kedua yang dilakukan oleh Erwin Yudi Hidayat dan Muhammad Aditya Rizqi menggunakan 600 data latih yang diambil dari portal berita [www.jawapos.com](http://www.jawapos.com), Data tersebut dibagi menjadi 4 kategori, yaitu Olahraga, Teknologi, Ekonomi, dan lain-lain, dengan masing-masing kategori memiliki 150 data. Tahap awal penelitian meliputi proses *preprocessing* yang terdiri dari *case folding*, *tokenizing*, *filtering*, dan *stemming* agar data latih dan data uji pada dataset dapat diklasifikasikan. Hasil klasifikasi dokumen berita menunjukkan akurasi untuk kategori Olahraga, Teknologi, Ekonomi dan Lain-lain berturut-turut sebesar 90%, 90%, 100%, dan 100%. Rata-rata akurasi keempat kategori tersebut mencapai 95% (Hidayat & Rizqi, 2020) .

Penelitian selanjutnya dilakukan oleh Bobby Suryo Prakoso dan rekan-rekannya dengan menggunakan *Naïve Bayes Classifier* dengan seleksi fitur dan metode *boosting* berdasarkan kategori berita pada situs detik.com. Proses yang dilakukan mencakup pemodelan dan pengolahan data, dimulai dari *preprocessing*, seleksi fitur menggunakan *Information Gain*, dan penerapan *Bayesian Boosting*. Hasil dari penerapan seleksi fitur *Information*

*Gain* mencapai 69,5% sedangkan dengan model yang lebih sederhana yaitu model algoritma *Naïve Bayes Classifier* dengan *Bayesian Boosting* mendapatkan nilai evaluasi sebesar 73,2%. Selain itu, penerapan *Bayesian Boosting* untuk label yang bersifat *polynomial* menghasilkan peningkatan evaluasi sebesar 4,3%. Dari evaluasi yang dilakukan, dapat disimpulkan bahwa penerapan seleksi fitur *Information Gain* tidak memberikan pengaruh besar terhadap peningkatan performa terutama pada kondisi label *Polynomial* (Prakoso et al., 2019).

Berbeda dengan karya-karya sebelumnya, penelitian ini memperkenalkan penyempurnaan pada *Enhanced Confix Stripping Stemmer* dengan memodifikasi aturan penanganan awalan seperti untuk kata-kata dengan pola “mem+p”, “peng+k”, dan “penge+” dan menerapkan mekanisme pemulihan *sufiks* berulang. Penyempurnaan ini mengatasi kesalahan stemming yang umum ditemukan dalam teks-teks akademis Bahasa Indonesia, yang tidak ditangani secara memadai dalam penelitian sebelumnya.

Penelitian ini menggunakan metode *Chi-Squared* untuk mengidentifikasi kata-kata dengan skor *Chi-Squared* tertinggi dalam setiap kategori berita. Untuk mengurangi varian yang terinfleksi ke bentuk dasarnya, *Enhanced Confix Stripping Stemmer* kemudian diterapkan untuk

menyegmentasikan kata-kata dengan menghilangkan awalan dan akhiran. Setelah ini, algoritma *Naïve Bayes Classifier* diimplementasikan ke model untuk melatih prediksi kategori berita untuk mengklasifikasikan artikel berita dari situs web UIN Walisongo.

Selain itu, penelitian ini menyoroti sinergi antara *Enhanced Confix Stripping* dan pemilihan fitur *Chi-Squared*. Sementara stemmer mengurangi gangguan morfologi dan memastikan konsistensi bentuk kata, penilaian *Chi-Squared* menyaring istilah yang paling relevan secara statistik. Kombinasi ini secara signifikan meningkatkan kualitas linguistik dan statistik fitur yang digunakan dalam klasifikasi.

Penelitian ini menyelidiki efektivitas *Naïve Bayes Classifier* dalam mengkategorikan dokumen berita akademik secara akurat di situs web UIN Walisongo. Penelitian ini juga meneliti sejauh *mana Enhanced Confix Stripping Stemmer* meningkatkan proses stemming untuk teks Bahasa Indonesia dan mengevaluasi peran metode *Chi-Squared* dalam meningkatkan pemilihan fitur untuk model klasifikasi.

Berdasarkan latar belakang yang telah dipaparkan, penelitian ini dilakukan dengan judul **“MODEL HYBRID ENHANCED CONFIX STRIPPING STEMMER DAN CHI-**

**SQUARED FEATURE SELECTION PADA KLASIFIKASI BERITA AKADEMIK MENGGUNAKAN *NAÏVE BAYES CLASSIFIER***". Klasifikasi ini difokuskan pada website portal berita PPID UIN Walisongo Semarang.

## **B. Identifikasi Masalah**

Berdasarkan pengamatan terhadap situs web UIN Walisongo Semarang, ditemukan beberapa permasalahan terkait pengelolaan dan penyajian informasi berita akademik, antara lain:

1. Belum adanya sistem klasifikasi otomatis yang dapat mengelompokkan berita berdasarkan kategori tertentu seperti kampus, pendidikan, acara, dan pengumuman. Hal ini menyulitkan pengguna dalam menemukan informasi sesuai kebutuhan.
2. Volume berita yang terus bertambah setiap waktu tidak diimbangi dengan mekanisme pengelompokan yang efisien, sehingga menyebabkan akses terhadap informasi menjadi tidak optimal.
3. Bahasa dalam berita akademik memiliki kompleksitas morfologi khas bahasa Indonesia, yang memerlukan teknik *preprocessing* teks khusus agar dapat diproses secara akurat oleh sistem klasifikasi.

4. Belum diterapkannya metode seleksi fitur dan algoritma klasifikasi yang optimal dalam membangun sistem klasifikasi berbasis pembelajaran mesin untuk situs tersebut, padahal metode seperti *Enhanced Confix Stripping Stemmer*, *Chi-Squared Feature Selection*, dan *Naïve Bayes* telah terbukti efektif dalam penelitian terdahulu.

### C. Rumusan Masalah

Berdasarkan latar belakang masalah yang disampaikan, maka rumusan masalah adalah sebagai berikut:

1. Bagaimana merancang dan mengembangkan sistem klasifikasi otomatis untuk berita akademik pada situs web UIN Walisongo Semarang, yang selama ini masih belum memiliki fitur pengelompokan berita berbasis kategori?
2. Bagaimana mengimplementasikan *Model Hybrid* preprocessing teks menggunakan algoritma *Enhanced Confix Stripping Stemmer* serta menerapkan metode *Chi-Squared Feature Selection* dalam membangun model klasifikasi berita akademik pada situs web UIN Walisongo berbasis *Naïve Bayes Classifier* yang efektif dan efisien?

#### **D. Batasan Masalah**

Sebagai bentuk antisipasi agar materi yang dikaji dalam penelitian ini tidak terlalu luas, maka diberikan batasan masalah sebagai berikut :

1. Menghasilkan model klasifikasi otomatis untuk berita online pada portal website UIN Walisongo Semarang menggunakan *Model Hybrid Enhanced Confix Stripping Stemmer* dan *Chi-Squared Feature Selection* Menggunakan algoritma *Naïve Bayes Classifier*.
2. Menjelaskan tahapan penggunaan metode *Chi-Squared* dan *Enhanced Confix Stripping Stemmer* dalam proses preprocessing dan seleksi fitur untuk mendukung klasifikasi berita berdasarkan kategori yang telah ditentukan.

#### **E. Tujuan Pengembangan**

1. Menerapkan *Model Hybrid Enhanced Confix Stripping Stemmer* dan *Chi-Squared Feature Selection* menggunakan algoritma *Naïve Bayes Classifier* sebagai alat untuk mengklasifikasikan berita online berdasarkan kategori yang ada pada portal web UIN Walisongo Semarang.

## **F. Manfaat Penelitian**

### **1. Manfaat Akademis**

Hasil dari penerapan *Model Hybrid Enhanced Confix Stripping Stemmer* dan *Chi-Squared Feature Selection* menggunakan algoritma Naïve Bayes Classifier dalam mengklasifikasikan berita akademik dapat memberikan kontribusi teoritis dalam pengembangan ilmu pengetahuan di bidang *data mining* dan *text classification*, khususnya dalam pengelompokan informasi berbasis teks pada berita akademik.

### **2. Manfaat Praktis**

Skripsi ini dapat digunakan sebagai bahan referensi bagi peneliti lain yang tertarik untuk mengembangkan sistem klasifikasi otomatis berita akademik, khususnya dalam konteks pengelolaan portal informasi kampus menggunakan *Model Hybrid Enhanced Confix Stripping Stemmer* dan *Chi-Squared Feature Selection* menggunakan algoritma Naïve Bayes Classifier

## **BAB II**

### **LANDASAN PUSTAKA**

#### **A. Portal Berita Online**

Portal berita online merupakan situs atau halaman web yang menyajikan berbagai jenis informasi, mulai dari politik, ekonomi, sosial, budaya, hingga hiburan dalam bentuk *hard news* maupun *soft news*. Kehadiran portal berita online memberikan kemudahan akses informasi secara cepat dan real-time bagi masyarakat (Woro Harkandi Kencana et al., 2021). Namun, seperti yang dikemukakan oleh Romli dalam (Tisna Nugraha, 2020) mengatakan bahwa portal berita juga memiliki kelebihan dan kelemahan. Di satu sisi, kelebihanannya terletak pada kecepatan penyebaran informasi dan jangkauan yang luas. Di sisi lain, kelemahannya dapat berupa kurangnya validasi informasi, potensi penyebaran hoaks, serta kualitas konten yang terkadang tidak terverifikasi secara mendalam. Oleh karena itu, penting bagi pengguna untuk bersikap kritis dalam menyikapi informasi yang disajikan melalui media daring tersebut.

Sebagai konsekuensi dari kemudahan akses dan tingginya arus informasi yang disuguhkan oleh portal berita online, diperlukan adanya upaya sistematis untuk memastikan keakuratan dan kredibilitas konten yang

dipublikasikan. Hal ini menjadi semakin penting mengingat masyarakat kini tidak hanya menjadi konsumen informasi, tetapi juga berperan sebagai produsen yang turut menyebarkan berita melalui berbagai platform digital.

Untuk menciptakan ekosistem informasi akademik yang sehat dan berdaya guna, diperlukan strategi pengelolaan konten yang tidak hanya bersifat informatif, tetapi juga terstruktur dan mudah diakses. Salah satu pendekatan fundamental dalam hal ini adalah penerapan sistem klasifikasi atau pengelompokan berita berdasarkan kategori tertentu. Dalam konteks situs web akademik seperti milik UIN Walisongo, keberadaan klasifikasi berita bukanlah sekadar fitur tambahan, melainkan kebutuhan strategis untuk menunjang efisiensi distribusi informasi.

Berbeda dengan portal berita umum yang mengkategorikan konten ke dalam ranah seperti politik, ekonomi, atau hiburan, berita akademik memiliki segmentasi informasi yang lebih spesifik dan fungsional seperti pendidikan, kegiatan kampus, agenda ilmiah, dan pengumuman resmi institusi. Ketidakhadiran sistem klasifikasi menyebabkan pengguna baik mahasiswa, dosen, staf akademik, maupun masyarakat umum harus melakukan pencarian manual terhadap informasi yang

dibutuhkan. Hal ini bukan hanya mengurangi efisiensi waktu, tetapi juga berpotensi menurunkan daya jangkauan dan dampak dari konten penting yang seharusnya segera diketahui publik.

Lebih jauh lagi, dalam dunia akademik yang menekankan validitas dan akurasi informasi, sistem klasifikasi otomatis juga dapat membantu institusi menjaga keteraturan data, menganalisis tren konten, serta meningkatkan keterlibatan pengguna terhadap informasi yang disediakan. Dengan meningkatnya volume berita yang dipublikasikan secara berkala, pendekatan manual dalam pengelolaan berita jelas tidak lagi relevan. Oleh karena itu, penerapan metode klasifikasi otomatis berbasis teknologi dan pembelajaran mesin menjadi sangat relevan tidak hanya untuk meningkatkan kualitas sistem informasi akademik, tetapi juga untuk mendorong transformasi digital yang sejalan dengan prinsip keterbukaan dan akuntabilitas institusi pendidikan.

## **B. UIN Walisongo**

Universitas Islam Negeri (UIN) Walisongo Semarang merupakan salah satu perguruan tinggi negeri yang mengintegrasikan ilmu pengetahuan umum dengan nilai-nilai keislaman dalam setiap aspek pendidikan, penelitian, dan pengabdian masyarakat (Rumaisa, 2025). Sebagai

institusi pendidikan tinggi, UIN Walisongo memiliki komitmen dalam mengembangkan sistem informasi yang mendukung aktivitas akademik dan manajerial. Pemanfaatan teknologi informasi menjadi elemen penting dalam meningkatkan efisiensi dan kualitas layanan, termasuk dalam pengelolaan portal berita kampus.

Portal berita resmi UIN Walisongo tidak hanya menjadi media publikasi kegiatan akademik dan non-akademik, tetapi juga sebagai sarana komunikasi dan diseminasi informasi kepada sivitas akademika maupun masyarakat luas. Dengan semakin meningkatnya jumlah berita yang diterbitkan, maka diperlukan sistem yang mampu mengelompokkan berita secara otomatis berdasarkan topik atau kategori tertentu agar lebih mudah diakses dan dimanfaatkan. Dalam konteks ini, penerapan metode klasifikasi berbasis teknologi informasi, seperti *Naive Bayes Classifier* yang didukung oleh proses seleksi fitur dan pra-pemrosesan data teks, menjadi langkah strategis dalam menunjang transformasi digital di lingkungan kampus.

### C. *Naïve Bayes Classifier*

*Naive Bayes Classifier* atau disebut juga *Bayesian Classification* adalah sebuah teorema yang merujuk pada konsep probabilitas bersyarat yang digunakan untuk

memprediksi probabilitas keanggotaan suatu kelas (Apriliyani et al., 2024; Ni Wayan Ernawati I et al., 2023).

Metode *Naïve Bayes Classification* merupakan metode yang digunakan untuk mengklasifikasikan teks dalam penelitian ini karena Metode Naïve Bayes memiliki keunggulan efisiensi, memungkinkan prosedur analisis sentimen lebih cepat dan efisien serta algoritma *Naïve Bayes* memiliki kecenderungan untuk memperoleh akurasi tinggi dengan jumlah data pelatihan yang terbatas (Apriliyani et al., 2024).

#### **D. Algoritma *Enhanced Confix Stripping Stemmer***

Algoritma *Enhanced Confix Stripping Stemmer* merupakan hasil perbaikan dari algoritma *Confix Stripping Stemmer* yang merupakan algoritma pendahulunya. Perbaikan ini dilakukan setelah beberapa percobaan dan analisis. Di mana ditemukan beberapa kata yang tidak dapat di-*stemming* menggunakan *Confix Stripping Stemmer*.

Sebagai respons atas permasalahan tersebut, *Enhanced Confix Stripping Stemmer* dikembangkan berdasarkan studi terbarunya menunjukkan kemampuan lebih baik dalam memangkas *afiks*, khususnya *konfix*, serta dalam mengidentifikasi kata dasar secara akurat, evaluasi perbandingan algoritma *stemmer* Bahasa Indonesia tahun

2024 mengungkap bahwa *Enhanced Confix Stripping Stemmer* memiliki keunggulan signifikan dalam hal akurasi dan efisiensi dibandingkan algoritma-algoritma terdahulu (Billah et al., 2023) .

Dengan kemampuannya menangani kompleksitas morfologi Bahasa Indonesia secara lebih efektif, *Enhanced Confix Stripping Stemmer* terbukti relevan dan layak diimplementasikan sebagai metode utama dalam proses pra-pemrosesan teks seperti analisis berita dan dokumen digital.

#### **E. *Chi - Squared***

Metode *Chi-Squared* adalah salah satu statistik yang berguna untuk mengetahui hubungan antara dua variabel kategori (Yusliani et al., 2022). Dalam konteks klasifikasi teks, metode ini digunakan untuk memilih fitur-fitur yang paling relevan dalam membedakan antara berbagai kategori yang ada. *Chi-Squared Score* mengukur seberapa besar perbedaan antara frekuensi pengamatan aktual dan frekuensi yang diharapkan dari fitur-fitur dalam setiap kategori.

Metode *Chi-Squared* akan mengukur seberapa signifikan perbedaan antara frekuensi pengamatan aktual dan frekuensi yang diharapkan. Semakin besar nilai *Chi-Squared Score* yang didapat maka semakin signifikan hubungan antara fitur tersebut dengan kategori tertentu.

Dengan demikian, penerapan metode *Chi-Squared* dalam proses seleksi fitur menjadi langkah krusial untuk meningkatkan performa algoritma klasifikasi teks. Dengan hanya mempertahankan fitur-fitur yang memiliki hubungan signifikan terhadap kelas tertentu, sistem dapat bekerja lebih efisien dan akurat karena terbebas dari *noise* atau fitur yang tidak relevan. Proses ini tidak hanya mempercepat proses klasifikasi, tetapi juga membantu dalam meningkatkan interpretabilitas model, terutama ketika diterapkan pada data teks dalam skala besar seperti korpus berita digital.

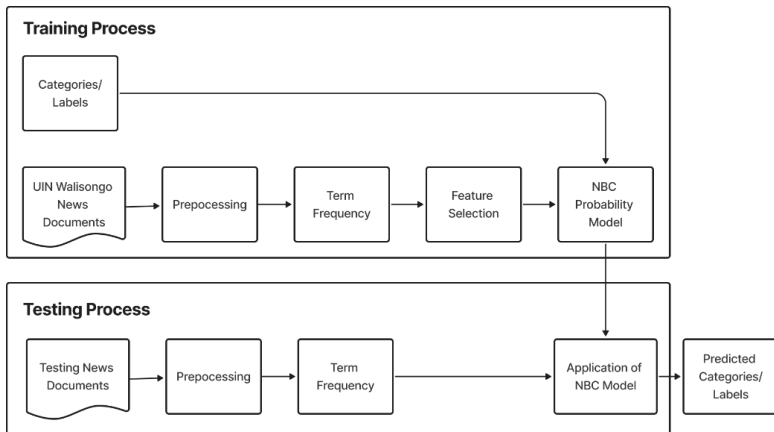


### **BAB III**

#### **METODOLOGI PENELITIAN**

Dalam pembelajaran mesin, klasifikasi mengacu pada tugas membangun model atau fungsi yang secara efektif dapat mengkategorikan titik data ke dalam kelas-kelas yang telah ditentukan. Tujuan utamanya adalah untuk memprediksi label kelas dari titik data baru yang tidak terlihat. Proses ini biasanya melibatkan dua tahap utama: pertama, model dilatih menggunakan sekumpulan titik data berlabel (data latih), dan kemudian dievaluasi dengan mengklasifikasikan sekumpulan titik data berlabel yang terpisah (data uji) untuk menilai keakuratannya (Asmiati, 2020).

Pada penelitian ini, tahap pertama dimulai dengan *preprocessing teks*, yang meliputi *case folding*, *tokenisasi*, *filtering*, dan *stemming* dengan menggunakan algoritma *Enhanced Confix Stripping Stemmer*. Selanjutnya, seleksi fitur dilakukan dengan menggunakan metode *Chi-Squared*, dilanjutkan dengan pelatihan model dengan algoritma *Naïve Bayes*. Alur penelitian secara lengkap divisualisasikan pada Gambar 1, yang menguraikan setiap tahap proses klasifikasi dari prapemrosesan data hingga evaluasi.



Gambar 3.1 Alur penelitian sistem klasifikasi berita akademik

Tahap kedua melibatkan pengujian model dengan data uji yang tidak terlihat, yang juga mengalami pra-pemrosesan (*case folding, tokenisasi, pemfilteran, dan stemming*). Data uji yang telah diproses sebelumnya kemudian diklasifikasikan menggunakan model yang telah dilatih. Terakhir, kinerja model dievaluasi dengan membandingkan prediksinya dengan label yang sebenarnya untuk menghitung keakuratannya, memastikan keandalannya untuk aplikasi di dunia nyata.

## A. Pengumpulan Data

Proses pengumpulan data dilakukan melalui portal berita [ppid.walisongo.ac.id](http://ppid.walisongo.ac.id), dengan total sebanyak 880 dokumen berita yang berhasil diperoleh. Seluruh dokumen kemudian dibagi menjadi dua bagian, yakni 80%

digunakan sebagai data latih dan 20% sisanya digunakan sebagai data uji. Pembagian ini bertujuan untuk mendukung proses pelatihan dan pengujian dalam analisis data selanjutnya.

Pengumpulan data dilakukan dengan teknik *web scraping* menggunakan bahasa pemrograman *Python*. Proses ini memanfaatkan fungsi *requests.get()* untuk mengambil konten *HTML* dari halaman Portal PPID UIN Walisongo, kemudian menggunakan *BeautifulSoup()* dari pustaka *bs4* untuk mem-parsing struktur *HTML*. Elemen-elemen yang memuat judul berita dan isi berita diekstraksi dengan *find\_all()* dan *get\_text(strip=True)*. Data hasil ekstraksi disusun ke dalam bentuk tabel menggunakan *pandas.DataFrame()* dan disimpan dalam format *XLSX*. Selanjutnya, data tersebut dikonversi ke format *CSV* dengan *df.to\_csv()* untuk digunakan dalam tahap pelatihan dan pengujian model klasifikasi

## **B. Text Preprocessing**

*Text Preprocessing* merupakan langkah dasar dalam jalur pemrosesan teks. Proses ini bertujuan untuk mengubah data teks mentah dan tidak terstruktur menjadi format yang memfasilitasi analisis komputasi yang efisien dan akurat. Proses ini melibatkan beberapa tahap (Nurdin

et al., 2021). Algoritma *stemming* yang disempurnakan seperti *Modified Enhanced Confix Stripping* sangat penting untuk menangani morfologi bahasa Indonesia (Lindrawati et al., 2023).

### 1. *Case Folding*

Langkah awal ini melibatkan pengubahan semua karakter dalam teks menjadi huruf kecil. Hal ini menormalkan data dengan menghilangkan ketidakkonsistenan yang timbul dari huruf besar, yang dapat menjadi tidak relevan untuk banyak tugas analisis teks.

### 2. *Tokenizing*

Teks kemudian disegmentasi menjadi unit-unit yang bermakna, biasanya berupa kata atau istilah. Proses ini menciptakan aliran token yang berfungsi sebagai elemen dasar untuk analisis lebih lanjut.

### 3. *Filtering*

Setelah tokenisasi, teknik penyaringan digunakan untuk memilih kata-kata atau token yang informatif. Tahap ini sering kali menggunakan algoritma penghilang kata henti, yang menghilangkan kata-kata yang umum dan tidak informatif (misalnya, “yang”, ‘sebuah’, “adalah”, dan lain-lain). Sebagai alternatif, daftar kata khusus dapat digunakan untuk

mempertahankan istilah yang spesifik atau relevan dengan domain.

#### 4. *Stemming*

Tujuan stemming adalah untuk mengurangi kata berimbuhan ke bentuk dasarnya. Proses ini bertujuan untuk mengidentifikasi akar morfologi (stem) dari setiap kata, sehingga mengelompokkan kata-kata yang terkait dan meningkatkan efektivitas keseluruhan tugas analisis selanjutnya.

### C. *Chi-Squared*

Uji *Chi-Squared*, sebuah metode statistik yang sudah mapan, terbukti bermanfaat dalam menentukan hubungan antara dua variabel kategorikal (Luna-Romera et al., 2019). Dalam domain klasifikasi teks, teknik ini digunakan untuk mengidentifikasi fitur yang paling relevan yang berkontribusi dalam membedakan antara kategori yang berbeda. Teknik pemilihan fitur seperti *Chi-Square* telah terbukti meningkatkan performa *Naïve Bayes* secara signifikan (Chen Sami et al., 2024). Skor *Chi-Squared* mengukur perbedaan antara frekuensi fitur yang diamati (O) dan yang diharapkan (E) dalam setiap kategori. Rumus spesifik untuk menghitung skor ini disajikan dalam persamaan (1).

$$X^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

Dalam persamaan (1),  $X^2$  merepresentasikan skor *Chi-Squared*, sementara  $O_i$  and  $E_i$  masing-masing menunjukkan frekuensi aktual (*observed*) dan ekspektasi (*expected*) dari sebuah kata dalam kategori tertentu. Parameter  $n$  mengacu pada jumlah kategori yang dianalisis. Persamaan ini menghitung jumlah dari selisih kuadrat antara frekuensi aktual dan ekspektasi  $(O_i - E_i)^2$  untuk setiap kategori, lalu dinormalisasi dengan frekuensi ekspektasi  $E_i$ , semakin tinggi skor *Chi-Squared*, semakin kuat hubungan antara suatu fitur (kata) dengan kategori tertentu. Oleh karena itu, fitur-fitur dengan skor *Chi-Squared* tertinggi dipilih sebagai fitur yang paling informatif, sehingga dapat meningkatkan akurasi model klasifikasi dengan menekankan fitur-fitur yang signifikan ini.

Frekuensi ekspektasi  $E_i$  yang sangat penting dalam perhitungan skor *Chi-Squared*, ditentukan berdasarkan distribusi keseluruhan fitur di seluruh kategori. Nilai ini dihitung menggunakan rumus yang ditampilkan dalam persamaan (2).

$$E_i = \frac{R \times C}{N} \quad (2)$$

Dalam persamaan (2),  $R$  merujuk pada total kemunculan fitur di semua kategori,  $C$  mewakili jumlah total kemunculan dalam kategori tertentu, dan  $N$  adalah jumlah total kemunculan semua fitur di semua kategori. Dengan membagi hasil kali  $R$  dan  $C$  dengan  $N$ , rumus ini memperkirakan frekuensi yang diharapkan dari suatu fitur dalam kategori tertentu dengan asumsi independensi. Perhitungan ini memastikan bahwa skor *Chi-Squared* secara efektif menangkap tingkat hubungan antara fitur dan kategori yang sesuai, memfasilitasi pemilihan fitur yang paling relevan untuk tugas klasifikasi.

#### **D. Algoritma *Enhanced Confix Stripping Stemmer***

Algoritma *Enhanced Confix Stripping Stemmer* dibangun berdasarkan pendahulunya, *Confix Stripping Stemmer*, yang bertujuan untuk mengatasi keterbatasan yang diidentifikasi selama eksperimen dan analisis. Keterbatasan ini termanifestasi dalam ketidakmampuan *Confix Stripping Stemmer* untuk secara efektif membendung kata-kata tertentu. Masalah-masalah

spesifik yang dihadapi adalah sebagai berikut (Billah et al., 2023).

1. Tidak adanya aturan pemenggalan awalan untuk kata-kata dengan format “mem+p...”. Hal ini mengakibatkan ketidakmampuan untuk membendung kata-kata seperti “mempromosikan” dan “memproteksi” dengan benar.
2. Keterbatasan lainnya adalah tidak adanya aturan pemenggalan awalan untuk kata-kata dengan format “men+s...”. Hal ini menyebabkan ketidakmampuan untuk melakukan stemming dengan benar pada kata-kata seperti “mensyaratkan” dan “mensyukuri”.
3. Ada keterbatasan dalam menangani kata kerja tertentu. Sebagai contoh, Stemmer gagal memisahkan awalan “menge+...” dari kata dasar kata kerja pada kata-kata seperti “mengerem”.
4. Stemmer tidak dapat menangani kata-kata dengan awalan “penge+...” yang diikuti oleh konsonan. Hal ini menyebabkan stemming yang tidak lengkap pada kata-kata seperti “pengeboman”.
5. Ketidakmampuan untuk menangani kata berawalan yang dimulai dengan “peng+k...” yang diikuti oleh konsonan, yang menyebabkan kesalahan stemming pada kata seperti “pengkajian”.

6. *Stemmer* salah mengidentifikasi sufiks tertentu “-an”, “-kan”, dan “-ku” pada kata-kata seperti “pelanggan”, ‘perpolitikan’, dan “pelaku” sebagai awalan, yang menyebabkan penghilangan elemen-elemen morfologi yang penting ini

Karena keterbatasan kemampuan penanganan awalan pada algoritma *Confix Stripping Stemmer* yang asli, versi *Enhanced* memperkenalkan beberapa modifikasi aturan untuk meningkatkan akurasi segmentasi awalan, terutama untuk formasi kata dalam bahasa Indonesia yang kompleks. Seperti yang dirangkum dalam Tabel 1, penyempurnaan ini mendefinisikan transformasi spesifik untuk struktur kata seperti “mem+V”, “peng+C”, dan “penge+V”. Setiap aturan bertujuan untuk memperbaiki kesalahan yang sebelumnya tidak ditangani dalam pendekatan standar. Dengan membuat penyempurnaan berbasis aturan ini menjadi eksplisit, *Enhanced Confix Stripping Stemmer* berkontribusi pada ekstraksi kata dasar yang lebih andal selama prapemrosesan - sebuah komponen penting dalam meningkatkan akurasi klasifikasi (Wardani & Nugraha, 2020).

Tabel 3.1 Penambahan dan Modifikasi pada Aturan Awalan Kata

| Aturan | Struktur Kata                       | <i>Prefix Segmentation</i>                                 |
|--------|-------------------------------------|------------------------------------------------------------|
| 14     | <i>men</i><br><i>{e c d j z}...</i> | <i>men-{e c d j z}...</i>                                  |
| 17     | <i>mengV</i>                        | <i>eng-V...   meng- kV...   (mengV-...<br/>If V= 'e')</i>  |
| 19     | <i>mempA...</i>                     | <i>mem-pA...where A!= 'e'</i>                              |
| 28     | <i>pengC...</i>                     | <i>peng-C...</i>                                           |
| 29     | <i>pengV...</i>                     | <i>peng-V...   peng- kV...   (pengV-...<br/>If V= 'e')</i> |

Seperti yang ditunjukkan pada tabel di atas, algoritma *Enhanced Confix Stripping Stemmer* menggabungkan beberapa peningkatan. Ini termasuk modifikasi aturan tertentu yang digunakan dalam pemenggalan kata dan pengenalan aturan baru, yang ditetapkan sebagai “untuk mengatasi kesalahan pemenggalan akhiran”. Rincian spesifik dari algoritma perulangan “Pengembalian Akhiran” didefinisikan sebagai berikut (Pramudita et al., 2018).

1. Setelah segmentasi awalan, algoritma mencoba merekonstruksi kata asli dengan mengembalikan awalan yang telah dihapus sebelumnya. Kata yang direkonstruksi ini, yang disebut sebagai “model kata”, kemudian menjalani proses pencarian kamus. Jika pencarian kamus berhasil mengidentifikasi model kata, maka proses akan berakhir. Sebaliknya, jika pencarian

kamus gagal menemukan model kata, algoritma akan melanjutkan ke langkah berikutnya.

2. Algoritma ini menggunakan proses berulang untuk mengembalikan sufiks yang telah dihapus sebelumnya. Pemulihan ini memprioritaskan akhiran dengan urutan sebagai berikut: *Sufiks* Turunan (“-i”, “-kan”, “-an”), Kata Ganti Kepemilikan (“-ku”, “-mu”, “-nya”), Imbuhan Partikel *Infleksi* (“-lah”, “-kah”, “-tah”, “-pun”). Untuk setiap jenis akhiran, algoritma mencoba merekonstruksi kata dengan menambahkan akhiran tersebut ke kata dasar saat ini dan memeriksa apakah kata yang dihasilkan ada dalam kamus kata dasar. Jika ditemukan kecocokan dalam kamus, proses akan berhenti. Jika tidak, algoritme melanjutkan ke aturan pemenggalan awalan untuk segmentasi lebih lanjut. Selain itu, jika pencarian kamus kata dasar tetap tidak berhasil setelah menambahkan semua akhiran yang disebutkan di atas, awalan yang sebelumnya dihapus akan dipulihkan kembali, diikuti dengan iterasi proses pemulihan akhiran.

### **E. *Term Frequency***

*Term Frequency* adalah teknik fundamental yang digunakan dalam pencarian informasi dan *text mining* untuk menilai pentingnya kata-kata individual dalam sebuah dokumen. Metode ini beroperasi dengan asumsi bahwa tingkat kepentingan sebuah kata berbanding lurus dengan frekuensi kemunculannya di dalam teks. Dalam istilah yang lebih sederhana, kata-kata yang lebih sering muncul dianggap memiliki bobot yang lebih besar dibandingkan dengan kata-kata yang lebih jarang muncul. Skema pembobotan ini didasarkan pada intuisi bahwa kata-kata yang diulang lebih sering cenderung lebih relevan atau informatif (bobot yang lebih tinggi) dalam menyampaikan tema sentral dokumen (Nurdin et al., 2021). Pemilihan fitur *Chi-Squared* telah diterapkan secara efektif untuk mengoptimalkan model klasifikasi teks (Yusliani et al., 2022).

### **F. *Naive Bayes Classifier***

*Naive Bayes Classifier*, juga dikenal sebagai *Bayesian Classification*, merupakan sebuah keluarga algoritma pembelajaran mesin yang memanfaatkan teorema Bayes untuk tujuan klasifikasi. *Teorema Bayes* adalah prinsip

dasar dalam teori probabilitas yang memungkinkan perhitungan probabilitas bersyarat dari suatu kejadian (keanggotaan kelas) yang terjadi mengingat kejadian lain (fitur tertentu). Intinya, algoritma memanfaatkan teorema ini untuk memprediksi kemungkinan sebuah titik data termasuk dalam kelas tertentu berdasarkan fitur-fitur yang diamati. Formulasi matematis dari *teorema Bayes* biasanya dinyatakan sebagai persamaan (3).

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \quad (3)$$

Probabilitas bersyarat dari  $A$  yang diberikan  $B$ , diwakili oleh  $P(A|B)$ , dan probabilitas bersyarat dari  $B$  yang diberikan  $A$  ( $P(B|A)$ ).  $P(A)$  adalah probabilitas kejadian  $A$ , dan  $P(B)$  adalah probabilitas kejadian  $B$ .

Penelitian ini menggunakan metode Klasifikasi *Naive Bayes* untuk klasifikasi teks karena kombinasi yang menarik antara kesederhanaan dan efisiensi. Prinsip utama algoritma ini terletak pada pendekatan probabilistiknya, di mana setiap dokumen direpresentasikan sebagai vektor fitur. Vektor ini dilambangkan sebagai “ $a_1, a_2, a_3, \dots, a_n$ ” yang menangkap ada atau tidaknya kata-kata individual ( $a_1$ ) hingga ( $a_n$ ) dalam dokumen.

Selama klasifikasi, algoritma *Naïve Bayes* berusaha mengidentifikasi kategori dengan probabilitas *posterior* tertinggi. Optimalisasi model klasifikasi teks menggunakan *Chi-Squared* dan *TF-IDF* telah dieksplorasi untuk meningkatkan akurasi klasifikasi (Falasari & Muslim, 2022). Probabilitas ini dihitung menggunakan  $V_{MAP}$ , yang dirinci dalam persamaan (4). Persamaan  $V_{MAP}$  diturunkan dari *teorema Bayes* dan mengasumsikan independensi bersyarat, sehingga memungkinkan perhitungan probabilitas *posterior* yang efisien untuk setiap kategori.

$$V_{MAP} = \underset{v_j \in v}{\operatorname{argmax}} P(v_j) \prod_i P(a_i|v_j) \quad (4)$$

$V_{MAP}$  menyajikan kategori dengan probabilitas *posterior* maksimum. Istilah  $P(v_j)$  menunjukkan probabilitas sebelumnya dari kategori  $v$ , sedangkan  $P(a_i|v_j)$  adalah probabilitas bersyarat dari fitur  $a_i$  mengingat bahwa data tersebut termasuk dalam kategori  $v_j$ . Hasil perkalian  $\prod_i P(a_i|v_j)$  menggabungkan probabilitas semua fitur dengan asumsi independen.

$$P(v_j) = \frac{|doc\ j|}{|training|} \quad (5)$$

Di mana  $|doc\ j|$  merepresentasikan jumlah dokumen (artikel berita) yang termasuk dalam kategori  $j$ , dibagi dengan jumlah total dokumen dalam training set. Selain itu, persamaan (6) mengatur perhitungan probabilitas kemunculan kata  $a_i$  di dalam setiap kategori  $P(a_i|v_j)$  selama proses pelatihan.

$$P(a_i|v_j) = \frac{|n_i + 1|}{|n + kosakata|} \quad (6)$$

$n_i$  merepresentasikan berapa kali kata  $a_i$  muncul dalam dokumen yang termasuk dalam kategori  $v_j$ , sedangkan  $n$  adalah jumlah total kata dalam dokumen yang termasuk dalam kategori  $v_j$ , dan kosakata adalah jumlah total kata dalam contoh dalam sampel pelatihan. Untuk mengukur akurasi klasifikasi, persamaan (7) dapat digunakan.

$$Accuracy = \frac{Number\ of\ Correct\ Test\ Documents}{Total\ Test\ Documents} \times 100\% \quad (6)$$



## BAB IV

### HASIL DAN PEMBAHASAN

Struktur klasifikasi yang digunakan untuk dokumen berita Universitas Islam Negeri Walisongo Semarang (UIN Walisongo) terdiri dari dua tahap utama: *training* dan *testing*. Pengumpulan data dilakukan melalui portal berita [ppid.walisongo.ac.id](http://ppid.walisongo.ac.id), yang menghasilkan korpus sebanyak 880 dokumen berita. Selanjutnya, dokumen-dokumen tersebut dibagi menjadi dua *subset* dengan rasio 80% untuk *training* dan 20% untuk *testing*. Selama tahap *training*, sebuah *probability* model dibangun menggunakan 704 dokumen *training*. Sisanya, sebanyak 176 dokumen, digunakan sebagai data uji untuk mengevaluasi efektivitas dari *probability* model yang telah dibentuk.

Sebelum proses *training* dimulai, dokumen *training* seperti yang ditunjukkan pada 4.1 terlebih dahulu melalui serangkaian proses *preprocessing*, termasuk analisis *term frequency* dan *feature selection*. Dokumen *testing* juga menjalani proses persiapan serupa sebelum diklasifikasikan menggunakan *probability* model yang telah dibentuk pada tahap pelatihan. Tahap *preprocessing* ini bertujuan untuk mengubah data teks tidak terstruktur menjadi format yang terstruktur dan sesuai untuk algoritma *machine learning*.

Tahapan ini mencakup beberapa teknik seperti *case folding*, *tokenization*, *filtering*, dan *stemming*.

Sebagaimana dijelaskan dalam Tabel 4.2, langkah pertama adalah *case folding*, di mana semua karakter diubah menjadi huruf kecil untuk memastikan konsistensi. Selanjutnya, tahap *tokenization* 4.3 memecah isi teks menjadi kata-kata individual. Setelah *tokenization* 4.4, proses *filtering* dilakukan untuk mempertahankan kata-kata yang relevan dan menghapus *stopwords*, yaitu kata-kata umum yang memiliki nilai semantik minimal.

Langkah terakhir dalam preprocessing adalah *stemming*, yang mengubah kata menjadi bentuk dasarnya. Penelitian ini menggunakan algoritma *Enhanced Confix Stripping Stemmer* untuk keperluan tersebut. Tabel 4.5 menunjukkan hasil dari proses ini pada sebuah dokumen.

#### **A. *Scraping Data Berita***

Pada tahap ini dilakukan proses pengambilan data berita dari situs resmi UIN Walisongo Semarang yang beralamat di <https://ppid.walisongo.ac.id>. Proses ini dilakukan secara otomatis menggunakan metode web scraping untuk memperoleh kumpulan berita beserta judul dan isi kontennya. Total data yang berhasil dikumpulkan sebanyak 880 berita, yang selanjutnya akan

digunakan dalam proses pelatihan dan pengujian sistem klasifikasi.

```

1  ALGORITMA Scraping_Berita_UIN
2  Deklarasi:
3      base_url, url      : string
4      data               : list of berita
5      response, soup     : objek HTML
6      article            : objek HTML
7      title, content     : string
8      kategori           : string
9      kata_kunci_pengumuman : array of string
10     kata_kunci_acara    : array of string
11     kata_kunci_pendidikan : array of string
12     kata_kunci_kampus  : array of string
13
14  Deskripsi:
15     Mulai
16     Inisialisasi base_url + "https://ppid.walisongo.ac.id/category/blog/"
17     Set url = base_url
18     Inisialisasi array kata kunci untuk setiap kategori
19     Selama jumlah data < 880 lakukan
20         Ambil halaman HTML dari url
21         Ambil semua elemen <article>
22         Untuk setiap artikel:
23             Ambil judul dan isi konten artikel
24             Tentukan kategori berdasarkan kata kunci
25             Simpan judul, isi, dan kategori ke dalam list data
26             Jika ada halaman berikutnya, pindah ke halaman tersebut
27     Simpan data ke dalam file CSV "berita_uin_walisongo_kategori.csv"
28     Selesai
29

```

Gambar 4.1 Pseudocode scraping data berita

Gambar 4.1 menunjukkan langkah-langkah utama dalam proses scraping. Program dimulai dengan mendeklarasikan *Uniform Resource Locator (URL)* utama situs berita UIN Walisongo. Selanjutnya dilakukan proses looping hingga terkumpul 880 berita. Pada setiap *iterasi*, sistem mengakses halaman *web*, mengambil elemen artikel, mengekstraksi judul dan isi konten, kemudian menentukan kategori berdasarkan kata kunci tertentu. Hasil scraping kemudian disimpan dalam file *Comma*

*Separated Values (CSV)* untuk keperluan pemrosesan lebih lanjut.

Tabel 4.1 Contoh dokumen pelatihan

| Judul                                                                                             | Isi                                                                                                                                      | Kategori |
|---------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------|----------|
| Walisongo Halal Center Terbitkan 275 Sertifikat Juru Sembelih Halal Se-eks Karesidenan Pekalongan | Walisongo Halal Center (WHC) bekerjasama dengan BI Tegal, Komite Nasional Ekonomi dan Keuangan Syariah ..... tegasnya                    | Acara    |
| FDK UIN Walisongo-PAMHU gelar Sertifikasi Pembimbing Manasik Haji Profesional Angkatan XV         | Fakultas Dakwah dan Komunikasi (FDK) UIN Walisongo Semarang menyelenggarakan sertifikasi pembimbing manasik haji profesional ..... PAHMU | Edukasi  |

Tabel 4.1 memperlihatkan sebagian data hasil *scraping* yang mencakup kolom Judul Berita dan Kategori. Judul-judul tersebut secara otomatis diklasifikasikan ke dalam salah satu dari empat kategori utama: Pendidikan, Pengumuman, Kampus, atau Acara, dengan menggunakan pencocokan kata kunci. Data yang telah dikumpulkan ini

nantinya akan digunakan untuk proses training dan pengujian model klasifikasi berita.

## B. *Preprocessing*

Setelah data berita berhasil dikumpulkan, tahap selanjutnya adalah *preprocessing*. *Preprocessing* bertujuan untuk membersihkan dan menyiapkan data agar dapat diproses oleh algoritma klasifikasi. Tahapan ini mencakup *case folding*, *tokenizing*, *stopword removal*, dan *stemming*. Proses ini dilakukan baik terhadap judul maupun isi berita.

```

1  ALGORITMA Preprocessing_Berita
2  Deklarasi:
3      df                : tabel data
4      stopwords_indo    : list of string
5      stemmer           : objek stemmer
6      judul, isi        : string
7      token, stemmed    : list of string
8
9  Deskripsi:
10     Mulai
11         Baca file CSV "berita_uin_walisongo_kategori.csv"
12         Ubah seluruh huruf menjadi lowercase
13         Tokenisasi setiap kata dalam judul dan isi
14         Hilangkan stopwords dari token
15         Stem setiap kata menggunakan Sastrawi
16         Simpan hasil ke dalam file "preprocessed_berita.csv"
17     Selesai
18

```

Gambar 4.2 Pseudocode preprocessing

Gambar 4.2 tersebut menunjukkan tahapan pembersihan teks berita. Data dibaca dari file hasil *scraping*, kemudian setiap karakter dikonversi menjadi huruf kecil. Kata-kata dalam judul dan isi kemudian dipisahkan menggunakan tokenisasi. *Stopword* yang tidak memiliki nilai informatif dihilangkan, dan kata-kata yang tersisa distem untuk mendapatkan bentuk dasar. Hasil akhir berupa teks bersih yang disimpan kembali ke file untuk tahap selanjutnya.

Tabel 4.2 Contoh hasil dokumen dari proses *case folding*

| Judul                                                                                                  | Isi                                                                                                                   | <i>Output</i><br>Judul                                                                                  | <i>Output</i> Isi                                                                                                        |
|--------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------|
| Walisongo Halal Center Terbi-tkan 275 Sertifi-kat Juru Semb-elih Halal Se-eks Kar-esidenan Pekalonga n | Walisongo Halal Center (WHC) bekerjasama dengan BI Tegal, Komite Nasional Ekonomi dan Keuangan Syariah ..... tegasnya | walisongo halal ce-nter terbi-tkan 275 sertifi-kat juru semb-elih halal se-eks kar-esidenan pekalonga n | walisongo halal center (whc) beker-jasama deng-an bi tegal, komite nasi-onal ekonomi dan keuangan syariah ..... tegasnya |
| FDK UIN Walisongo -PAMHU gelar                                                                         | Fakultas Dakwah dan Komunikasi (FDK) UIN                                                                              | fdk uin walisongo-pamhu gelar                                                                           | fakultas dakwah dan komunikasi (fdk) uin                                                                                 |

|                                                             |                                                                                                 |                                                             |                                                                                                 |
|-------------------------------------------------------------|-------------------------------------------------------------------------------------------------|-------------------------------------------------------------|-------------------------------------------------------------------------------------------------|
| Sertifikasi Pembimbing Manasik Haji Profesional Angkatan XV | Walisongo Semarang menyelenggarakan sertifikasi pembimbing manasik haji profesional ..... PAHMu | sertifikasi pembimbing manasik haji profesional angkatan xv | walisongo semarang menyelenggarakan sertifikasi pembimbing manasik haji profesional ..... pahmu |
|-------------------------------------------------------------|-------------------------------------------------------------------------------------------------|-------------------------------------------------------------|-------------------------------------------------------------------------------------------------|

Langkah selanjutnya dalam tahapan *preprocessing* adalah *case folding*, sebagaimana dijelaskan pada Tabel 4.2. Setelah proses *case folding* dilakukan, baik judul maupun isi berita dikonversi sepenuhnya menjadi huruf kecil. Langkah ini bertujuan untuk menghilangkan ketidak konsistenan yang disebabkan oleh kapitalisasi huruf, sehingga dapat mengurangi potensi duplikasi representasi data seperti yang ditampilkan pada Tabel 4.3.

Tabel 4.3 Contoh hasil dokumen dari proses *tokenizing*

| Judul                                                                            | Isi                                                                                            | Output Judul                                                             | Output Isi                                                                                   |
|----------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------|----------------------------------------------------------------------------------------------|
| walisongo halal center terbitkan 275 sertifikasi juru sembelih halal se-eks kar- | walisongo halal center (whc) bekerjasama dengan bi tegal, komite nasional ekonomi dan keuangan | walisongo halal center terbitkan 275 sertifikat juru sembelih halaleks', | walisongo halal center whc bekerjasama dengan bi tegal, komite nasional ekonomi dan keuangan |

| esidenan<br>pekalongan                                                                                                        | syariah .....<br>tegasnya                                                                                                                                                             | karesidenan<br>pekalongan                                                                                                | syariah .....<br>tegasnya                                                                                                                                                      |
|-------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| fdk uin<br>walisongo-<br>pamhu<br>gelar sert-<br>ifikasi pe-<br>mbimbing<br>manasik<br>haji profes-<br>ional angk-<br>atan xv | fakultas<br>dakwah dan<br>komunikasi<br>(fdk) uin<br>walisongo<br>semarang<br>menyelenggar-<br>akan sert-ifik-<br>asi pem-<br>bimbing ma-<br>nasik haji<br>profesional<br>..... pahmu | fdk uin<br>walisongo-<br>pamhu<br>gelar<br>sertifikasi<br>pembimb-<br>ing manasik<br>haji profes-<br>ional angk-<br>atan | fakultas<br>dakwah dan<br>komunikasi<br>fdk uin<br>walisongo<br>semarang<br>menyelenggar-<br>akan sertifik-<br>asi pembimb-<br>ing manasik<br>hajiprofesio-<br>nal ..... pahmu |

Setelah *case folding*, langkah berikutnya adalah *tokenization*. Proses ini melibatkan pemisahan teks menjadi unit-unit yang lebih kecil yang disebut *token*. *Tokenization* menyederhanakan teks dengan membaginya menjadi kata-kata atau sub-kata individual, sehingga memudahkan proses *natural language processing* selanjutnya.

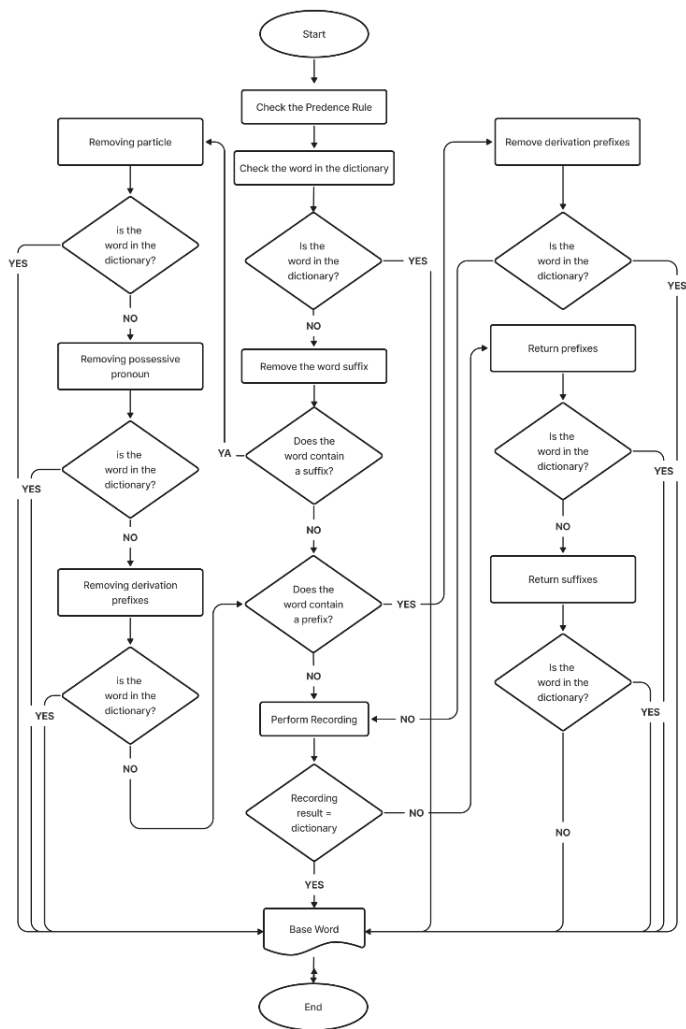
Setelah proses *tokenization*, di mana teks telah dipisahkan menjadi kata-kata individual, dilakukan proses *filtering* yang dapat dilihat pada Tabel 4.4. Tahapan ini bertujuan untuk memilih istilah-istilah yang informatif

dari korpus yang telah ditokenisasi. Sebaliknya, kata-kata yang dianggap tidak penting, yang disebut sebagai *stopwords*, dihapus menggunakan algoritma *stopwords removal*.

Tabel 4.4 Contoh hasil dokumen dari proses filtering

| Judul                                                                                              | Isi                                                                                                                 | Output Judul                                                                                                                      | Output Isi                                                                                                                             |
|----------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------|
| walisongo halal center terbitkan 275 sertifikat juru sembelih halal se-eks karresidenan pekalongan | walisongo halal center whc bekerjasama dengan bi tegal, komite nasional ekonomi dan keuangan syariah ..... tegasnya | 'walisongo', 'halal', 'center', 'terbitkan', '275', 'sertifikat', 'juru', 'sebelih', 'halal', 'eks', 'karresidenan', 'pekalongan' | 'walisongo', 'halal', 'center', 'whc', 'bekerjasama', 'tegal', 'komite', 'nasional', 'ekonomi', 'keuangan', 'syariah', ..... 'tegasya' |
| fdk uin walisongopamhu gelar sertifikasi pembimbing manasik haji profesional angkatan              | fakultas dakwah dan komunikasi fdk uin walisongo semarang menyelenggarakan sertifikasi pembimbing manasik haji      | 'fdk', 'uin', 'walisongo', 'pamhu', 'gelar', 'sertifikasi', 'pembimbing', 'manasik', 'haji', 'profesional', 'angkatan'            | 'fakultas', 'dakwah', 'komunikasi', 'fdk', 'uin', 'walisongo', 'semarang', 'menyelenggarakan', 'sertifikasi', 'pembimbing', 'manasik', |

|  |                             |  |                                            |
|--|-----------------------------|--|--------------------------------------------|
|  | professional<br>..... pahmu |  | 'haji',<br>'profesional',<br>..... 'pahmu' |
|--|-----------------------------|--|--------------------------------------------|



Gambar 4.3 Flowchart enhanced confix stripping stemmer

Gambar 4.3 menjelaskan tahapan-tahapan dalam algoritma *Enhanced Confix Stripping Stemmer*. Algoritma ini menggabungkan tiga komponen utama, yaitu *Enhanced*, *Confix Stripping*, dan *Stemmer*, yang masing-masing berfungsi untuk menangani aspek-aspek tertentu dalam proses normalisasi kata.

Komponen *Enhanced* merupakan penyempurnaan dari algoritma *Confix Stripping* yang asli. Pada tahap ini dilakukan pemeriksaan aturan prioritas (*precedence rule check*) serta penambahan aturan baru untuk menangani kombinasi afiks kompleks seperti "mem-", "meng-", dan "penge-". Penyempurnaan ini ditujukan untuk mengatasi keterbatasan pada algoritma sebelumnya, sehingga dapat memproses struktur kata dalam bahasa Indonesia dengan lebih akurat.

Proses *Confix Stripping* berfokus pada penghapusan afiks secara sistematis dari sebuah kata, yang dibagi ke dalam dua tahapan utama:

1. *Case Folding*: langkah awal ini mengubah seluruh karakter dalam teks menjadi huruf kecil.
2. *Suffix Removal*: tahap ini menghilangkan afiks partikel infleksional (misalnya "-lah", "-kah"), kata ganti

kepemilikan (misalnya "-ku", "-mu", "-nya"), serta sufiks derivatif (misalnya "-kan", "-an").

3. *Prefix Removal*: setelah penghapusan *sufiks*, awalan seperti "meng-", "pen-", dan "ber-" akan dihapus menggunakan aturan eliminasi prefiks derivatif. Hal ini memastikan bahwa kata hasilnya mendekati bentuk dasar. Tahap *Prefix Removal* bertujuan agar kata yang telah dihapus sufiksnya selanjutnya dapat dihilangkan awalan seperti "meng-", "pen-", dan "ber-" dengan menggunakan aturan derivatif, sehingga kata tersebut lebih sesuai dengan bentuk dasarnya.

Komponen *Stemmer* bertindak sebagai proses utama yang mengintegrasikan komponen *Enhanced* dan *Confix Stripping* untuk mengidentifikasi bentuk dasar sebuah kata. Jika sebuah kata tidak dapat dicocokkan dengan bentuk dasar setelah dilakukan penghapusan awalan dan akhiran, maka algoritma akan menjalankan proses *word recording*. Proses ini merekonstruksi kata dengan mengembalikan awalan atau akhiran yang sebelumnya telah dihapus, lalu mencocokkannya dengan basis data kata dasar. Jika tidak ditemukan kecocokan, algoritma akan melakukan iterasi melalui tahap Algoritma Pengembalian Akhiran untuk kembali menerapkan aturan *sufiks* guna menyempurnakan proses normalisasi kata.

Dengan pemisahan fungsi yang jelas dari masing-masing komponen tersebut, algoritma *Enhanced Confix Stripping Stemmer* mampu menangani tantangan dalam pemrosesan teks bahasa Indonesia secara efektif, serta menghasilkan hasil *stemming* yang akurat dan andal.

Tabel 4.5 Contoh dokumen hasil penerapan *enhanced confix stripping stemmer*

| Judul                                                                                                                                                                        | Isi                                                                                                                                                                  | Output Judul                                                                                                                                                    | Output Isi                                                                                                                                                         |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 'walisongo',<br>'halal',<br>'center',<br>'terbitkan',<br>'275',<br>'sertifikat',<br>'juru', 'sem-<br>belih', 'hal-<br>al', 'eks',<br>'ka-<br>residenan',<br>'pekalonga<br>n' | 'walisongo',<br>'halal',<br>'center', 'whc',<br>'bekerjasama',<br>'tegal',<br>'komite',<br>'nasional',<br>'ekonomi',<br>'keuangan',<br>'syariah', .....<br>'tegasya' | 'walisongo<br>, 'halal',<br>'center',<br>'terbit',<br>'275',<br>'sertifikat',<br>'juru',<br>'sembelih',<br>'halal',<br>'eks',<br>'karesiden<br>an',<br>'kalong' | 'walisongo',<br>'halal',<br>'center',<br>'whc',<br>'bekerjasama'<br>, 'tegal',<br>'komite',<br>'nasional',<br>'ekonomi',<br>'uang',<br>'syariah', .....<br>'tegas' |
| 'fdk', 'uin',<br>'walisongo',<br>'pamhu',<br>'gelar',<br>'sert-<br>ifikasi', 'pe-<br>mbimbing',<br>'manasik',<br>'haji',                                                     | 'fakultas',<br>'dakwah',<br>'komunikasi',<br>'fdk', 'uin',<br>'walisongo',<br>'semarang',<br>'menyelengga<br>rakan',<br>'sertifikasi',                               | 'fdk', 'uin',<br>'walisongo',<br>'pamhu',<br>'gelar',<br>'sertifikasi',<br>'pembimb<br>ing',<br>'manasik',<br>'haji',                                           | 'fakultas',<br>'dakwah',<br>'komunikasi',<br>'fdk', 'uin',<br>'walisongo',<br>'semarang',<br>'menyelenggar<br>akan',<br>'sertifikasi',                             |

|                                   |                                                                           |                                |                                                                          |
|-----------------------------------|---------------------------------------------------------------------------|--------------------------------|--------------------------------------------------------------------------|
| 'profe-<br>sional',<br>'angkatan' | 'pembimbing',<br>'manasik',<br>'haji',<br>'profesional',<br>..... 'pahmu' | 'profesional',<br>, 'angkatan' | 'pembimbing',<br>'manasik',<br>'haji',<br>'profesional'....<br>. 'pahmu' |
|-----------------------------------|---------------------------------------------------------------------------|--------------------------------|--------------------------------------------------------------------------|

Tabel 4.5 menyajikan hasil penerapan *Enhanced Confix Stripping Stemmer* pada dokumen berita akademik yang nyata. Kolom keluaran menunjukkan bagaimana kata-kata yang memiliki *afiks* kompleks dinormalisasi menjadi bentuk dasarnya. Proses ini secara signifikan meningkatkan konsistensi fitur dengan mengurangi variasi morfologis dalam teks bahasa Indonesia, yang sangat penting untuk mencapai akurasi yang lebih tinggi dalam klasifikasi teks. Kata-kata yang telah dinormalisasi memungkinkan model Naïve Bayes untuk melakukan generalisasi dengan lebih baik dan membedakan antara kategori seperti pendidikan, acara, kampus, dan pengumuman.

### C. Term Frequency

Setelah data berita dipreproses, tahap selanjutnya adalah menghitung *Term Frequency (TF)*. TF adalah ukuran frekuensi kemunculan suatu kata dalam dokumen. Tahapan ini dilakukan untuk mengetahui kata-kata yang paling sering muncul dalam setiap berita, baik di bagian judul maupun isi. Nilai TF akan digunakan sebagai dasar dalam pemilihan fitur pada tahap berikutnya.

```

1  ALGORITMA Hitung_Term_Frequency
2  Deklarasi:
3      df          : tabel data
4      TF_Judul    : dictionary
5      TF_Isi      : dictionary
6
7  Deskripsi:
8      Mulai
9      Baca file "preprocessed_berita.csv"
10     Hitung frekuensi setiap kata unik di judul dan isi per berita
11     Simpan hasil ke kolom baru dan gabungkan teks untuk corpus
12     Simpan ke file "tf_berita.csv"
13     Selesai
14

```

Gambar 4.4 Pseudocode term frequency

Gambar 4.4 ini Pseudocode ini menjelaskan proses menghitung *term frequency* dari data berita. Setiap berita akan dihitung kemunculan setiap kata unik dalam judul maupun isi. Hasil dari proses ini adalah dua kolom baru yaitu *TF\_Judul* dan *TF\_Isi* yang berisi pasangan kata dan

jumlah kemunculannya. Proses ini penting sebagai dasar pemilihan fitur.

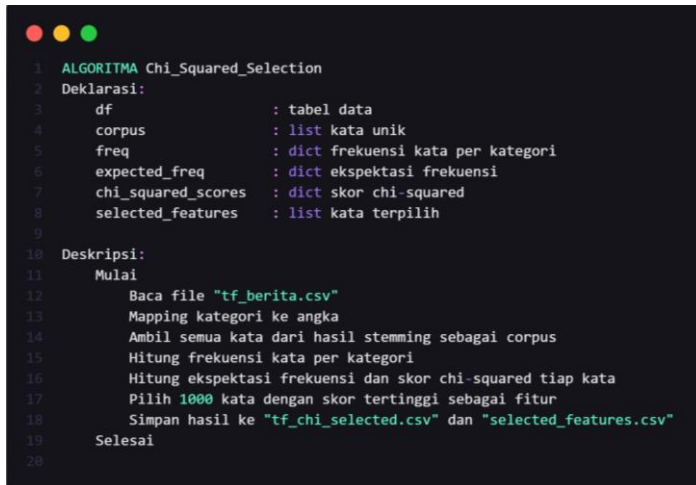
Tabel 4.6 Contoh dokumen hasil *penerapan term frequency*

| Judul                                                                                                                                                             | Isi                                                                                                                                                                                                       | Output Judul                                                                                                                                            | Output Isi                                                                                                                                                                                                      |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 'walisongo',<br>, 'halal',<br>'center',<br>'terbit',<br>'275',<br>'sertifikat',<br>'juru',<br>'sembelih',<br>'halal',<br>'eks',<br>'karesidenan',<br>'pekalongan' | 'walisongo',<br>'halal',<br>'center', 'whc',<br>'bekerjasama',<br>'tegal',<br>'komite',<br>'nasional',<br>'ekonomi',<br>'uang',<br>'syariah',<br>'tegas'                                                  | 'walisongo': 1, 'halal': 2, 'center': 1, 'terbit': 1, '275': 1, 'sertifikat': 1, 'juru': 1, 'sembelih': 1, 'eks': 1, 'karesidenan': 1, 'pekalongan': 1  | 'walisongo': 1, 'halal': 1, 'center': 1, 'whc': 1, 'bekerjasama': 1, 'tegal': 1, 'komite': 1, 'nasional': 1, 'ekonomi': 1, 'uang': 1, 'syariah': 1, 'tegas': 1                                                  |
| 'fdk', 'uin',<br>'walisongo',<br>'pamhu',<br>'gelar',<br>'sertifikasi',<br>'pembimbing',<br>'manasik',<br>'haji',<br>'profesional',<br>, 'angkatan'               | 'fakultas',<br>'dakwah',<br>'komunikasi',<br>'fdk', 'uin',<br>'walisongo',<br>'semarang',<br>'menyelenggarakan',<br>'sertifikasi',<br>'pembimbing',<br>'manasik',<br>'haji',<br>'profesional',<br>'pahmu' | 'fdk': 1, 'uin': 1, 'walisongo': 1, 'pamhu': 1, 'gelar': 1, 'sertifikasi': 1, 'pembimbing': 1, 'manasik': 1, 'haji': 1, 'profesional': 1, 'angkatan': 1 | 'fakultas': 1, 'dakwah': 1, 'komunikasi': 1, 'fdk': 1, 'uin': 1, 'walisongo': 1, 'semarang': 1, 'menyelenggarakan': 1, 'sertifikasi': 1, 'pembimbing': 1, 'manasik': 1, 'haji': 1, 'profesional': 1, 'pahmu': 1 |

Tabel 4.6 menampilkan perhitungan *term frequency* setelah data melalui proses *stemming* menggunakan *Enhanced Confix Stripping Stemmer*. Untuk setiap dokumen, dihitung jumlah kemunculan kata-kata yang telah dinormalisasi baik dalam judul maupun isi berita. Misalnya, pada Dokumen 1 bagian judul, kata “halal” muncul dua kali, sementara kata lainnya seperti “terbit” dan “sertifikat” masing-masing muncul satu kali. Proses ini bertujuan untuk membentuk representasi numerik awal sebelum dilakukan seleksi fitur menggunakan *Chi-Squared*.

#### **D. Chi-Squared Feature Selection**

Setelah perhitungan *term frequency* selesai, proses selanjutnya adalah pemilihan fitur menggunakan metode *Chi-Squared*. Tahapan ini bertujuan untuk memilih kata-kata yang paling berpengaruh dalam membedakan kategori berita. Dengan metode ini, hanya fitur (kata) yang memiliki nilai signifikansi tinggi yang akan dipertahankan untuk proses pelatihan model.



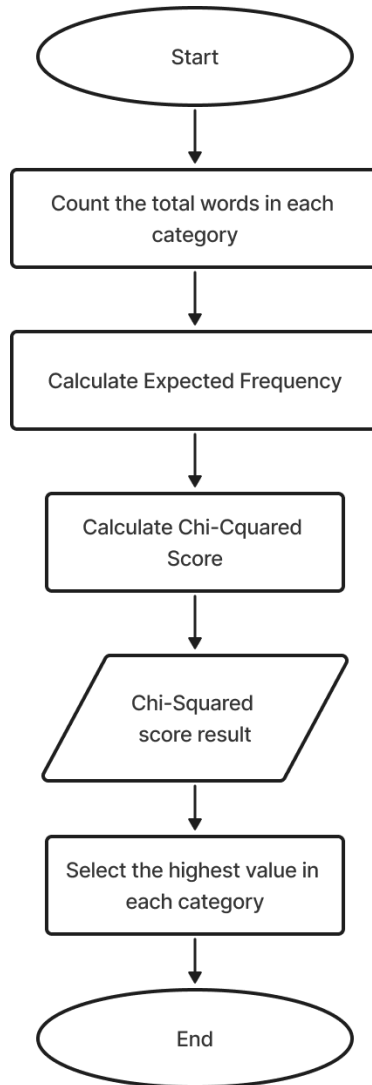
```

1  ALGORITMA Chi_Squared_Selection
2  Deklarasi:
3      df                : tabel data
4      corpus            : list kata unik
5      freq              : dict frekuensi kata per kategori
6      expected_freq     : dict ekspektasi frekuensi
7      chi_squared_scores : dict skor chi-squared
8      selected_features  : list kata terpilih
9
10 Deskripsi:
11     Mulai
12     Baca file "tf_berita.csv"
13     Mapping kategori ke angka
14     Ambil semua kata dari hasil stemming sebagai corpus
15     Hitung frekuensi kata per kategori
16     Hitung ekspektasi frekuensi dan skor chi-squared tiap kata
17     Pilih 1000 kata dengan skor tertinggi sebagai fitur
18     Simpan hasil ke "tf_chi_selected.csv" dan "selected_features.csv"
19     Selesai
20

```

Gambar 4.5 Pseudocode chi-squared

Gambar 4.5 ini menjelaskan langkah-langkah dalam memilih fitur menggunakan metode *Chi-Squared*. Proses dimulai dengan membangun korpus kata dari hasil *stemming*, kemudian menghitung frekuensi aktual dan ekspektasi kemunculan kata di setiap kategori. Skor *Chi-Squared* dihitung untuk mengetahui seberapa besar kontribusi masing-masing kata dalam membedakan kategori. Hanya kata-kata dengan skor tertinggi yang disimpan sebagai fitur penting.



*Gambar 4.6 Flowchart feature selection chi-squared*

Setelah tahap preprocessing, data pelatihan menjalani proses seleksi fitur menggunakan uji *Chi-Squared*. Metode ini bertujuan untuk mengidentifikasi fitur-fitur yang paling diskriminatif dalam pengkategorian berita, sehingga dapat meningkatkan performa *Naïve Bayes Classifier*. Prosesnya, yang divisualisasikan dalam Gambar 4.6, mencakup perhitungan jumlah total kata per kategori, penentuan *observed frequency* dan *expected frequency* untuk setiap *term* dalam setiap kategori, perhitungan skor *Chi-Squared*, dan pemilihan fitur berdasarkan nilai *Chi-Squared* tersebut.

*Observed frequency* merepresentasikan jumlah kemunculan aktual suatu term dalam kategori tertentu, sedangkan *expected frequency* merupakan jumlah kemunculan yang diprediksi secara teoritis dengan asumsi independensi antar variabel. Statistik *Chi-Squared* mengukur penyimpangan antara kedua frekuensi tersebut, yang mengindikasikan seberapa relevan term tersebut terhadap suatu kategori. Nilai *Chi-Squared* yang lebih tinggi menunjukkan hubungan yang lebih kuat antara term dengan kategori. Fitur-fitur dengan skor *Chi-Squared* tertinggi kemudian digunakan sebagai input

untuk *Naïve Bayes Classifier*. Tabel 4.7 menyajikan tiga *term* teratas dengan skor *Chi-Squared* tertinggi untuk masing-masing kategori. Skor *Chi-Squared* yang mendekati 1,0 menunjukkan adanya hubungan yang lebih kuat antara *term* dan kategori yang bersangkutan.

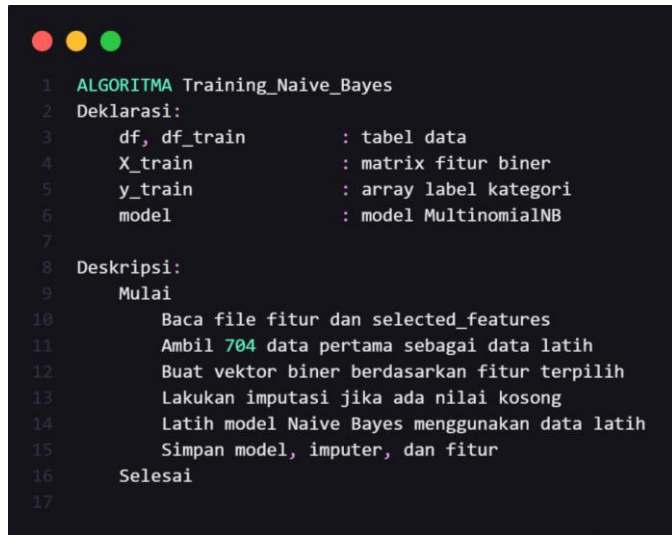
Tabel 4.7 Contoh dokumen hasil penerapan *chi-squared*

| Kata            | Chi-Squared Skor | Kategori     |
|-----------------|------------------|--------------|
| mahasiswa       | 1.0              | education    |
| dosen           | 0.98             | education    |
| kuliah          | 0.98             | education    |
| seminar         | 1.0              | event        |
| upacara         | 0.99             | event        |
| diskusi         | 0.98             | event        |
| ukt             | 1.0              | campus       |
| uin             | 0.99             | campus       |
| ptkin           | 0.99             | campus       |
| seleksi         | 1.0              | announcement |
| Pendaftar<br>an | 0.94             | announcement |
| informasi       | 0.94             | announcement |

### ***E. Training Model Naive Bayes***

Setelah fitur penting berhasil dipilih menggunakan metode *Chi-Squared*, proses berikutnya adalah pelatihan model klasifikasi. Pada tahap ini digunakan algoritma *Multinomial Naïve Bayes* yang dilatih menggunakan 704 data berita pertama dari total 880 data. Proses pelatihan

dilakukan dengan membentuk representasi *biner* dari fitur yang telah dipilih, kemudian digunakan untuk membangun model klasifikasi otomatis berdasarkan distribusi probabilitas tiap kategori.



```

1  ALGORITMA Training_Naive_Bayes
2  Deklarasi:
3      df, df_train      : tabel data
4      X_train           : matrix fitur biner
5      y_train           : array label kategori
6      model             : model MultinomialNB
7
8  Deskripsi:
9      Mulai
10     Baca file fitur dan selected_features
11     Ambil 704 data pertama sebagai data latih
12     Buat vektor biner berdasarkan fitur terpilih
13     Lakukan imputasi jika ada nilai kosong
14     Latih model Naive Bayes menggunakan data latih
15     Simpan model, imputer, dan fitur
16     Selesai
17

```

Gambar 4.7 Pseudocode training model naive bayes

Gambar 4.7 di atas menjelaskan proses pelatihan model klasifikasi. Sebanyak 704 data pertama digunakan untuk membangun model. Setiap data diubah menjadi *vektor* biner berdasarkan 1000 fitur terpilih. Jika terdapat nilai kosong, akan diisi dengan angka nol menggunakan metode imputasi. *Model Multinomial Naïve Bayes* dilatih

menggunakan data tersebut dan disimpan dalam format file model agar dapat digunakan pada tahap pengujian.

Tabel 4.8 Tabel hasil pelatihan model

| Kategori   | Jumlah Data Latih |
|------------|-------------------|
| Acara      | 178               |
| Pengumuman | 164               |
| Kampus     | 188               |
| Pendidikan | 174               |
| Total      | 704               |

Tabel 4.8 menunjukkan jumlah distribusi data latih dari masing-masing kategori berita. Model dilatih secara seimbang menggunakan keempat kategori, yaitu Acara, Pengumuman, Kampus, dan Pendidikan. Keseimbangan jumlah data ini penting untuk menghindari bias model terhadap salah satu kategori tertentu.

#### ***F. Testing Model***

Setelah model selesai dilatih, tahap berikutnya adalah melakukan pengujian terhadap 176 data berita yang tidak digunakan pada proses pelatihan. Pengujian ini bertujuan untuk mengevaluasi performa awal dari model klasifikasi terhadap data baru yang belum pernah dikenali sebelumnya, sehingga menjadi indikator seberapa baik model dapat mengklasifikasikan berita secara otomatis.

Proses pengujian ini mencakup beberapa langkah. Pertama, dokumen uji sebagaimana ditampilkan pada Tabel 8 diinput dan diproses melalui tahapan preprocessing. Selanjutnya, dilakukan perhitungan *term frequency* terhadap setiap dokumen. Sistem kemudian menghitung nilai  $V_j = P(v_j) \times P(a_i/v_j)^n$  untuk setiap kategori, menggunakan *probability model* yang telah dibentuk pada tahap pelatihan. Setelah itu, sistem secara otomatis menetapkan kategori untuk setiap dokumen uji. Evaluasi dilakukan menggunakan dokumen uji yang kategorinya tidak diketahui sebelumnya, sebagaimana tercantum pada Tabel 4.1.

```

1  ALGORITMA Testing_Model
2  Deklarasi:
3      df_test          : tabel data
4      X_test           : matrix fitur uji
5      y_test           : array label aktual
6      y_pred           : array hasil prediksi
7      acc              : float (akurasi)
8
9  Deskripsi:
10     Mulai
11         Baca file data dan model
12         Ambil 176 data dari indeks ke-704 sampai 879
13         Buat vektor fitur dan lakukan imputasi
14         Prediksi kategori menggunakan model Naive Bayes
15         Hitung akurasi prediksi
16         Tampilkan akurasi
17     Selesai
18

```

Gambar 4.8 Pseudocode testing model

Gambar 4.8 ini menjelaskan alur pengujian model terhadap data uji sebanyak 176 berita. Sama seperti data latih, fitur pada data uji dibentuk menjadi *vektor biner*. Setelah dilakukan imputasi, model yang sudah dilatih digunakan untuk melakukan prediksi. Akurasi dihitung berdasarkan kecocokan antara hasil prediksi dan label sebenarnya yang sudah diketahui.

Tabel 4.9 Contoh dokumen pengujian

| Judul                                                                                                     | Isi                                                                                                                                                                   | Kategori        |
|-----------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------|
| Moderasi Beragama, Hidup Damai di Tengah Masyarakat Plural                                                | Gagasan moderasi beragama ala Walisongo harus disebarkan kepada seluruh masyarakat Indonesia, karena sikap Walisongo dalam mengapresiasi local wisdom ..... pelatihan | Tidak diketahui |
| Sekretaris RMB UIN Walisongo mengajar Political Theology Moderasi Beragama di Universitas Vienna, Austria | Luthfi Rahman, diundang untuk mengisi mata kuliah Approaches of Islamic Theology to Dialogue di salah satu kelas mahasiswa jurusan ..... beragam                      | Tidak diketahui |

Tabel 4.9 menyajikan contoh dokumen berita akademik yang belum dikategorikan dari UIN Walisongo Semarang, yang terdiri dari judul dan isi berita. Dokumen ini akan digunakan sebagai kasus uji untuk mengevaluasi model probabilitas *Naïve Bayes* yang telah dibangun sebelumnya dan dilatih menggunakan *dataset* terpisah. Sebelum dilakukan klasifikasi, dokumen uji ini melalui tahapan *preprocessing*, yang meliputi *case folding*, *tokenizing*, *filtering*, dan *stemming*.

Tabel 4.10 Contoh dokumen uji yang telah dikategorikan

| Judul                                                                                                     | Isi                                                                                                                                                                   | Kategori |
|-----------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|
| Moderasi Beragama, Hidup Damai di Tengah Masyarakat Plural                                                | Gagasan moderasi beragama ala Walisongo harus disebarkan kepada seluruh masyarakat Indonesia, karena sikap Walisongo dalam mengapresiasi local wisdom ..... pelatihan | Edukasi  |
| Sekretaris RMB UIN Walisongo mengajar Political Theology Moderasi Beragama di Universitas Vienna, Austria | Luthfi Rahman, diundang untuk mengisi mata kuliah Approaches of Islamic Theology to Dialogue di salah satu kelas mahasiswa jurusan ..... beragam                      | Acara    |

Setelah tahap *preprocessing*, mesin secara otomatis mengategorikan dokumen uji menggunakan model probabilitas *Naïve Bayes* yang telah dibangun. Sebanyak 176 dokumen uji dievaluasi, dengan tingkat akurasi yang dicapai sebesar 95%. Seperti yang ditunjukkan pada Tabel 9, sebanyak 167 dokumen berhasil diklasifikasikan dengan benar.

#### **G. Evaluation**

Dalam konteks klasifikasi dokumen berita akademik dari UIN Walisongo Semarang, terdapat tiga kategori dokumen: pelatihan (*training*), pengujian (*testing*), dan dokumen baru. Pengumpulan data awal menghasilkan 880 dokumen berita dari situs [ppid.walisongo.ac.id](http://ppid.walisongo.ac.id). Dokumen-dokumen tersebut kemudian dibagi menjadi dua *subset*: 80% digunakan sebagai dokumen pelatihan dan 20% sebagai dokumen pengujian. Selama fase pelatihan, model probabilitas dibangun dengan menggunakan 704 dokumen pelatihan. Sisa 176 dokumen digunakan sebagai dokumen uji untuk mengevaluasi efektivitas model. Untuk menilai lebih lanjut akurasi sistem dalam pengategorian, dibuat satu set tambahan yang terdiri dari 12 berita

terbaru dan dibandingkan dengan kategori yang ditentukan oleh pengguna.

```

1  ALGORITMA Evaluasi_Berita_Baru
2  Deklarasi:
3      data_baru          : list berita baru
4      df                 : tabel data
5      X_eval             : matrix fitur berita baru
6      y_pred             : array hasil prediksi
7      gt_labels          : array label aktual (ground truth)
8      kategori_map       : mapping angka ke nama kategori
9
10 Deskripsi:
11     Mulai
12         Siapkan 12 data berita baru (judul dan isi)
13         Lakukan preprocessing: case folding, tokenisasi, stopword removal, stemming
14         Buat vektor fitur biner
15         Lakukan prediksi kategori
16         Tampilkan tabel hasil prediksi vs label user
17         Tampilkan metrik evaluasi: precision, recall, f1-score, support
18     Selesai
19

```

Gambar 4.9 Pseudocode evaluation

Gambar 4.9 ini menggambarkan alur proses evaluasi model terhadap 12 artikel berita baru. Artikel diinput secara manual, lalu diproses seperti data sebelumnya: melalui tahap *preprocessing* dan pembentukan *vektor* fitur. Model yang telah dilatih kemudian digunakan untuk memprediksi kategori dari setiap artikel, dan hasilnya dibandingkan dengan kategori aktual untuk dihitung metrik evaluasinya.

Tabel 4.11 Hasil pengujian

| No | Judul Berita                                                                      | Isi Berita                                                                                      | Kategori (Sistem) | Kategori (Pengguna) |
|----|-----------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------|-------------------|---------------------|
| 1  | Pengumuman Beasiswa KIP-K untuk Mahasiswa Baru Tahun 2024                         | UIN Walisongo memberikan pengumuman terkait program beasiswa KIP-K ..... kampus                 | Pengumuman        | Pengumuman          |
| 2  | Fakultas Sains dan Teknologi Adakah Pelatihan Kewirausahaan Bagi Mahasiswa        | 'Fakultas Sains dan Teknologi akan menyelenggarakan pelatihan kewirausahaan ..... mereka        | Acara             | Acara               |
| 3  | Civitas Akademik Melakukan Kerja Bakti Untuk Menjaga Kebersihan Lingkungan Kampus | Civitas akademik mengadakan kegiatan membersihkan lingkungan kampus. kegiatan ini ..... nyaman. | Acara             | Kampus              |

|   |                                                                                   |                                                                                                                 |         |         |
|---|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------|---------|---------|
| 4 | Fakultas Dakwah dan Komunikasi adakan Pelatihan Desain Grafis untuk Mahasiswa wa  | Kompetisi ini akan memberikan kesempatan bagi mahasiswa untuk menunjukkan ..... kemampuan desain grafis mereka. | Edukasi | Edukasi |
| 5 | UIN Walisongo Gelar Diskusi Etika dalam Kehidupan Sehari-hari                     | UIN Walisongo akan mengadakan diskusi tentang etika dan moralitas ..... kehidupan sehari-hari.                  | Kampus  | Acara   |
| 6 | UIN Walisongo Gelar Webinar Internasional Tentang Pendidikan Islam di Era Digital | Universitas Islam Negeri (UIN) Walisongo Semarang menggelar webinar internasional ..... mahasiswa               | Acara   | Acara   |

|    |                                                                                                                            |                                                                                                                      |                |                |
|----|----------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------|----------------|----------------|
| 7  | Kantin<br>FST UIN<br>Walisong<br>o<br>Longsor<br>Akibat<br>Hujan<br>Deras                                                  | Akibat hujan<br>deras selama<br>sehari penuh<br>kantin FST<br>longsor .....<br>perbaikan                             | Kampus         | Kampus         |
| 8  | PTIPD<br>Lakukan<br>Worksho<br>p<br>Pelatihan<br>Jurnalisti<br>k                                                           | Dalam<br>meningkatka<br>n kualitas<br>kepenulisan<br>mahasiswa<br>.....<br>jurnalistik                               | Edukasi        | Edukasi        |
| 9  | Mahasis<br>wa<br>Teknolog<br>i<br>Informasi<br>UIN<br>Walisong<br>o Raih<br>Prestasi<br>di<br>Tingkat<br>Internasi<br>onal | Tiga<br>mahasiswa<br>Teknologi<br>Informasi<br>UIN<br>Walisongo<br>Semarang<br>raih<br>penghargaan<br>..... Malaysia | Kampus         | Kampus         |
| 10 | Perubaha<br>n<br>Tanggal<br>Pelaksana<br>an<br>Wisuda                                                                      | Tanggal<br>Pelaksaan<br>Wisuda<br>Periode Juni<br>20024                                                              | Pengumu<br>man | Pengumuma<br>n |

|    |                                                                                                             |                                                                                                                |                |                |
|----|-------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------|----------------|----------------|
|    | Periode<br>Juni 2024                                                                                        | dirubah .....<br>mahasiswa                                                                                     |                |                |
| 11 | UIN<br>Walisong<br>o Buka<br>Jalur<br>Prestasi<br>untuk 34<br>Jurusan,<br>Ini<br>Jadwal<br>dan<br>Syaratnya | Jalur Prestasi<br>UIN<br>Walisongo<br>kembali<br>dibuka untuk<br>34 jurusan<br>..... syarat<br>kenentuan       | Pengumu<br>man | Pengumuma<br>n |
| 12 | Lantik<br>Pejabat<br>Struktura<br>l, Ini<br>Pesan<br>Rektor<br>UIN<br>Walisong<br>o                         | Dalam<br>rangka<br>pelantikan<br>pejabar<br>struktural<br>periode baru<br>rektor<br>berpesan .....<br>kualitas | Acara          | Ascara         |

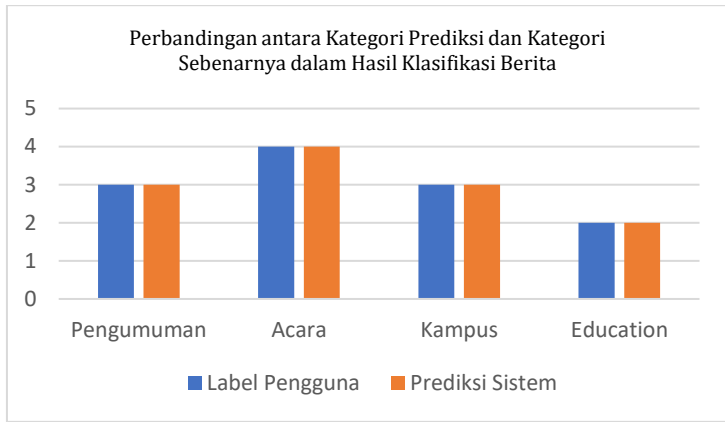
Sistem dievaluasi menggunakan dua belas dokumen berita terbaru. Hasil menunjukkan bahwa sepuluh berita berhasil dikategorikan secara akurat sesuai dengan penilaian pengguna, sebagaimana dijelaskan pada Tabel 10. Tingkat akurasi keseluruhan mencapai 83,33%, dengan tingkat kesalahan sebesar 16,67%.

$$Accuracy = \frac{10}{12} 100\% = 83,33\%$$

Pada berita ketiga, sistem memprediksi kategori “Acara” sementara pengguna memberikan label “Kampus”. Ketidaksesuaian ini kemungkinan terjadi karena artikel tersebut memuat frasa yang bersifat tindakan seperti “mengadakan kegiatan”, yang secara semantik lebih dekat dengan konteks acara meskipun sebenarnya merujuk pada kegiatan pemeliharaan kampus.

Berita kelima juga mengalami kesalahan klasifikasi. Sistem memberikan label “Kampus”, sementara pengguna menetakannya sebagai “Acara”. Kesalahan ini mungkin disebabkan oleh fokus sistem pada istilah institusional seperti “diskusi tentang etika dan moralitas”, yang mungkin diinterpretasikan sebagai konten terkait kampus alih-alih sebuah acara terjadwal.

Temuan ini menunjukkan bahwa model klasifikasi masih dapat disempurnakan lebih lanjut, khususnya dalam membedakan bahasa yang bernuansa, terutama ketika terdapat tumpang tindih semantik antara kategori seperti aktivitas kampus dan acara yang terorganisir.



*Gambar 4.10 Perbandingan kategori prediksi dan kategori aktual dalam hasil klasifikasi berita*

Gambar 4.8 menyajikan diagram batang yang membandingkan kategori hasil prediksi oleh sistem dan kategori aktual yang ditetapkan oleh pengguna untuk 12 dokumen berita uji. Distribusi tersebut menunjukkan kesesuaian yang kuat antara prediksi sistem dan label pengguna di keempat kategori. Tidak terdapat kecenderungan *overclassification* maupun *underclassification* yang signifikan, yang mengindikasikan bahwa model memiliki kemampuan pengenalan kategori yang seimbang.

Tabel 4.12 *Evaluation metrics*

| Kategori   | Precision | Recall | F1-Score | Support |
|------------|-----------|--------|----------|---------|
| Acara      | 0.75      | 0.75   | 0.75     | 4       |
| Kampus     | 0.67      | 0.67   | 0.67     | 4       |
| Edukasi    | 1.0       | 1.0    | 1.0      | 3       |
| Pengumuman | 1.0       | 1.0    | 1.0      | 2       |

Seperti yang ditunjukkan pada Tabel 4.12, metrik evaluasi menunjukkan bahwa sistem mencapai nilai presisi, *recall*, dan *F1-score* yang sempurna untuk kategori “Pengumuman” dan “Pendidikan”, yang mencerminkan bahwa tidak terdapat kesalahan klasifikasi pada kedua kelompok tersebut. Sebaliknya, kategori “Acara” dan “Kampus” mencatat nilai *F1-score* yang lebih rendah, masing-masing sebesar 0,75 dan 0,67, yang sesuai dengan dua kesalahan klasifikasi yang telah diidentifikasi sebelumnya. Hasil ini menunjukkan efektivitas model dalam mengenali sebagian besar kategori berita akademik, sekaligus menyoroti perlunya penyempurnaan lebih lanjut dalam menangani kategori yang memiliki tumpang tindih makna secara semantik seperti aktivitas yang berkaitan dengan kampus dan acara yang terorganisir.



## **BAB V**

### **SIMPULAN DAN SARAN**

#### **A. Simpulan**

Penelitian ini bertujuan untuk mengembangkan sistem klasifikasi dokumen berita akademik pada situs *web* resmi UIN Walisongo Semarang dengan memanfaatkan algoritma *Naïve Bayes Classifier* yang dipadukan dengan metode *Enhanced Confix Stripping Stemmer* dan seleksi fitur *Chi-Squared*. Berdasarkan hasil analisis dan evaluasi dari keseluruhan proses yang telah dilakukan, maka dapat disimpulkan beberapa hal sebagai berikut.

1. Model klasifikasi yang dibangun terbukti mampu mengkategorikan berita ke dalam empat kategori utama, yaitu Acara, Pengumuman, Kampus, dan Pendidikan, dengan tingkat akurasi sebesar 83.33% pada data artikel baru dan 95% pada data uji. Keberhasilan ini menunjukkan bahwa pendekatan yang digunakan cukup efektif dan layak untuk diterapkan pada klasifikasi teks berita akademik berbahasa Indonesia.

2. Penggunaan *Enhanced Confix Stripping Stemmer* memberikan kontribusi signifikan dalam proses stemming dengan meningkatkan akurasi pemrosesan kata berimbuhan kompleks dalam bahasa Indonesia. Hal ini membantu model dalam mengenali struktur kata yang lebih bersih dan konsisten.
3. Metode seleksi fitur *Chi-Squared* juga berhasil memilih fitur-fitur yang paling relevan dalam membedakan antar kategori berita. Hal ini berdampak langsung pada peningkatan performa model dalam mengklasifikasikan dokumen.
4. Secara umum, penelitian ini berkontribusi terhadap pengembangan sistem informasi cerdas yang dapat digunakan oleh institusi pendidikan dalam mengelola dan mengklasifikasikan informasi berita secara otomatis dan efisien. Sistem yang dibangun memiliki karakteristik ringan, mudah diimplementasikan, serta memiliki kemampuan interpretasi yang baik terhadap data teks akademik.

## B. Saran

Untuk pengembangan lebih lanjut, berikut beberapa saran yang dapat dipertimbangkan dalam penelitian atau implementasi selanjutnya:

1. Perbandingan performa model *Naïve Bayes* dengan algoritma lain seperti *Support Vector Machine (SVM)*, *Random Forest*, atau *Neural Network* untuk mendapatkan pemahaman yang lebih dalam terhadap keunggulan dan kelemahan masing-masing metode.
2. Penggunaan teknik *word embedding* atau representasi berbasis konteks seperti *BERT* untuk meningkatkan pemahaman semantik terhadap konteks kalimat.
3. Peningkatan jumlah dan variasi data agar sistem mampu belajar dari lebih banyak variasi bahasa dan struktur berita, sehingga meningkatkan akurasi dan generalisasi model.
4. Integrasi sistem klasifikasi ini ke dalam sistem manajemen konten *website* institusi agar proses kategorisasi berita menjadi otomatis dan efisien.
5. Penambahan fitur visualisasi seperti *dashboard* analitik untuk memudahkan administrator dalam memantau distribusi berita berdasarkan kategori.

6. Penelitian lebih lanjut terhadap efektivitas metode *stemming* dan *filtering* bahasa Indonesia lainnya untuk meningkatkan akurasi klasifikasi.

## DAFTAR PUSTAKA

- Apriliyani, M., Izzal Musyaffaq, M., Nur, S., Rini Handayani, M., Umam, K. (2024). Implementasi Analisis Sentimen pada Ulasan Aplikasi Duolingo di Google Playstore Menggunakan Algoritma Naïve Bayes. *AITI: Jurnal Teknologi Informasi*, 21(2), 298–311. <https://doi.org/https://doi.org/10.24246/aiti.v21i2.298-311>
- Asmiati, N. (2020). Penerapan Algoritma Naïve Bayes untuk Mengklasifikasi Pengaruh Negatif Game Online bagi Remaja Milenial. *JTIM: Jurnal Teknologi Informasi Dan Multimedia*, 2(3), 141–149. <https://doi.org/http://dx.doi.org/10.35746/jtim.v2i3.102>
- Billah, A. M., Wulandari, D. A. R., & Auliya, Y. A. (2023). Rancang Bangun Chatbot Pengaduan Kekerasan Perempuan Anak dengan Metode Fuzzy String Matching dan Enhanced Confix Stripping Stemmer. *INFROMAL: Informatics Journal*, 8(2), 101–109. <https://doi.org/https://doi.org/10.19184/isj.v8i2.42310>
- Chen Sami, D., Sugiharto, A., & Jie, F. (2024). Chi-Square Feature Selection for Improving Sentiment Analysis of News Data Privacy Threats. *JATIT: Journal of Theoretical and Applied Information Technology*, 102(18). [www.jatit.org](http://www.jatit.org)
- Falasari, A., & Muslim, M. A. (2022). Optimize Naïve Bayes Classifier Using Chi-Square and Term Frequency Inverse Document Frequency for Amazon Review Sentiment Analysis. *JOSCEX: Journal of Soft Computing Exploration*, 3(1), 31–36. <https://doi.org/https://doi.org/10.52465/joscex.v3i1.68>
- Hidayat, E. Y., & Rizqi, M. A. (2020). Klasifikasi Dokumen Berita Menggunakan Algoritma Enhanced Confix Stripping Stemmer dan Naïve Bayes Classifier. *Jurnal TEKNOLOGI: Nasional Teknologi Dan Sistem Informasi*, 6(2), 90–99.

<https://doi.org/https://doi.org/10.25077/TEKNOSI.v6i2.2020.90-99>

- Lestari, R. D. (2020). Jurnalisme Digital dan Etika Jurnalisme Media Sosial: Studi pada Akun Instagram @tempodotco dan @tribunjogja Digital Journalism and The Ethics of Social Media Journalism: Study in the Instagram Account @tempodotco and @tribunjogja. *Jurnal Ilmu Pengetahuan Dan Teknologi Komunikasi*, 22(2), 159–174. <https://doi.org/10.33164/iptekkom.22.2.2020.159-174>
- Lindrawati, E., Utami, E., & Yaqin, A. (2023). Comparison of Modified Nazief & Adriani and Modified Enhanced Confix Stripping Algorithms for Madurese Language Stemming. *INTENSIF: Jurnal Ilmiah Penelitian Dan Penerapan Teknologi Sistem Informasi*, 7(2), 276–289. <https://doi.org/https://doi.org/10.29407/intensif.v7i2.20103>
- Luna-Romera, J. M., Martínez-Ballesteros, M., García-Gutiérrez, J., & Riquelme, J. C. (2019). External Clustering Validity Index Based on Chi-Squared Statistical Test. *Information Sciences*, 487, 1–17. <https://doi.org/https://doi.org/10.1016/j.ins.2019.02.046>
- Ni Wayan Ernawati I, Nyoman Satya Kumara, & Widyadi Setiawan. (2023). Perbandingan Metode Klasifikasi Support Vector Machine dan Naïve Bayes pada Analisis Sentimen Kendaraan Listrik. *Jurnal SPEKTRUM*, 10(3).
- Nurdin, N., Suhendri, M., Afrilia, Y., & Rizal, R. (2021). Klasifikasi Karya Ilmiah (Tugas Akhir) Mahasiswa Menggunakan Metode Naïve Bayes Classifier (NBC). *SISTEMASI: Jurnal Sistem Informasi*, 10(2), 268–279. <https://doi.org/https://doi.org/10.32520/stmsi.v10i2.1193>

- Prakoso, B. S., Rosiyadi, D., Utama, H. S., & Aridarma, D. (2019). Klasifikasi Berita Menggunakan Algoritma Naïve Bayes Classifier dengan Seleksi Fitur dan Boosting. *Jurnal RESTI: Rekayasa Sistem Dan Teknologi Informasi*, 3(2), 227–232. <https://doi.org/https://doi.org/10.29207/resti.v3i2.1042>
- Pramudita, Y. D., Putro, S. S., & Makhmud, N. (2018). Klasifikasi Berita Olahraga Menggunakan Metode Naïve Bayes dengan Enhanced Confix Stripping Stemmer. *JTIK : Jurnal Teknologi Informasi Dan Ilmu Komputer*, 5(3), 269–276. <https://doi.org/https://doi.org/10.25126/jtiik.201853810>
- Purba, W., Siawin, W., Nababan, M. N. K., Dharshinni, N. P., & Aisyah, S. (2019). Implementasi Data Mining untuk Pengelompokan dan Prediksi Karyawan yang Berpotensi PHK dengan Algoritma K-Means Clustering. *JUSIKOM PRIMA : Jurnal Sistem Informasi Dan Ilmu Komputer Prima*, 2(2), 85–90. <https://doi.org/https://doi.org/10.34012/jusikom.v2i2.429>
- Rumaisa, F. (2025). Evaluation of Indonesian Language Stemmer Algorithms: A Comparative Analysis. *Brilliance: Research of Artificial Intelligence*, 5(1). <https://doi.org/10.47709/brilliance.v5i1.5679>
- Sani, R. R., Pratiwi, Y. A., Winarno, S., Udayanti, E. D., & Alzami, F. (2022). Analisis Perbandingan Algoritma Naïve Bayes Classifier dan Support Vector Machine untuk Klasifikasi Berita Hoax pada Berita Online Indonesia. *JMASIF : Jurnal Masyarakat Informatika*, 13(2), 85–98. <https://doi.org/https://doi.org/10.14710/jmasif.13.2.47983>
- Sholekha, I., Faqih, A., & Bahtiar, A. (2022). Sentiment Analysis of Public Opinion Covid-19 Vaccine Using Naïve Bayes and Random Forest Methods. *JTI : Jurnal Teknik Informatika*, 15(1), 34–43. <https://doi.org/https://doi.org/10.15408/jti.v15i1.24847>

- Tisna Nugraha, M. (2020). Integrasi Ilmu dan Agama: Praktik Islamisasi Ilmu Pengetahuan Umum di Perguruan Tinggi. *Al-Hikmah: Jurnal Agama Dan Ilmu Pengetahuan (AJAIP)*, 17(1). [https://doi.org/https://doi.org/10.25299/al-hikmah:jaip.2020.vol17\(1\).3927](https://doi.org/https://doi.org/10.25299/al-hikmah:jaip.2020.vol17(1).3927)
- Wardani, N. W., & Nugraha, P. G. S. C. (2020). Stemming Teks Bahasa Bali dengan Algoritma Enhanced Confix Stripping. *IJNSE : International Journal of Natural Science and Engineering*, 4(3), 103–113. <https://doi.org/https://doi.org/10.23887/ijnse.v4i3.30309>
- Widaningsih, S. (2019). Perbandingan Metode Data Mining untuk Prediksi Nilai dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika dengan Algoritma C4.5, Naïve Bayes, KNN dan SVM. *Jurnal Tekno Insentif*, 13(1), 16–25. <https://doi.org/https://doi.org/10.36787/jti.v13i1.78>
- Woro Harkandi Kencana, Ilona V Oisina Situmeang, Meisyanti Meisyanti, Khina Januar Rahmawati, & Herlin Nugroho. (2021). Penggunaan Media Sosial dalam Portal Berita Online. *IKRA-ITH HUMANIORA : Jurnal Social Dan Humaniora*, 6(2).
- Yusliani, N., Aruda, S. A. Q., Marieska, M. D., Saputra, D. M., & Abdiansah, A. (2022). The Effect of Chi-Square Feature Selection on Question Classification Using Multinomial Naïve Bayes. *Sinkron : Jurnal Dan Penelitian Teknik Informatika*, 7(4), 2430–2436. <https://doi.org/https://doi.org/10.33395/sinkron.v7i4.11788>

# **LAMPIRAN - LAMPIRAN**

# Lampiran 1 Letter of Acceptance



June 13, 2025

No : 4670/LoA/JUTIF/VI/2025  
About : Letter of Acceptance

To

1. Muhammad Naufal Muhadzib Al-Faruq, Walisongo State Islamic University, Indonesia
2. Wenty Dwi Yuniarti, Walisongo State Islamic University, Indonesia
3. Maya Rini Handayani, Walisongo State Islamic University, Indonesia
4. Khotibul Umam, Walisongo State Islamic University, Indonesia

**Jurnal Teknik Informatika (JUTIF)** is a national journal registered to BRIN with P-ISSN: 2723-3863 and e-ISSN: 2723-3871. JUTIF is accredited SINTA 2 based on the Decree of the Director General of Higher Education, Research, and Technology of the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia Number 177/E/KPT/2024.

We would like to thank you for submitting scientific article to the **Jurnal Teknik Informatika (JUTIF)** below:

|         |                                                                                                                                                                                                                                                                                                                                                                             |
|---------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Title   | <b>Implementation of Enhanced Confix Stripping Stemming and Chi-Squared Feature Selection on Classification UIN Walisongo Website With Naïve Bayes Classifier</b>                                                                                                                                                                                                           |
| Authors | <ol style="list-style-type: none"> <li>1. Muhammad Naufal Muhadzib Al-Faruq, Walisongo State Islamic University, Indonesia</li> <li>2. Wenty Dwi Yuniarti, Walisongo State Islamic University, Indonesia</li> <li>3. Maya Rini Handayani, Walisongo State Islamic University, Indonesia</li> <li>4. Khotibul Umam, Walisongo State Islamic University, Indonesia</li> </ol> |

Based on the review results, the article is **ACCEPTED** for publication in the **Jurnal Teknik Informatika (JUTIF)** Universitas Jenderal Soedirman (UNSOED) **Volume 6 Number 3, June 2025**.

We express our gratitude for your participation and good cooperation for publishing this article.

Chief Editor,

**Dr. Ir. Lasmedi Afuan, S.T., M.Cs., IPM.**  
NIP. 19850510 200812 1 002

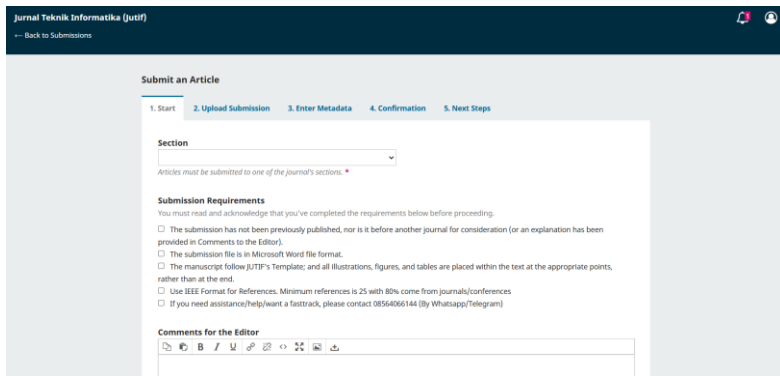
## Lampiran 2 Bukti Korespondensi

### History Jurnal:

Submit Artikel : 28 April 2025

Revisi : 26 Mei 2025

Publikasi : Juni 2025



**Jurnal Teknik Informatika (jutif)**

← Back to Submissions

### Submit an Article

1. Start   2. Upload Submission   3. Enter Metadata   4. Confirmation   5. Next Steps

**Section**

Articles must be submitted to one of the journal's sections. \*

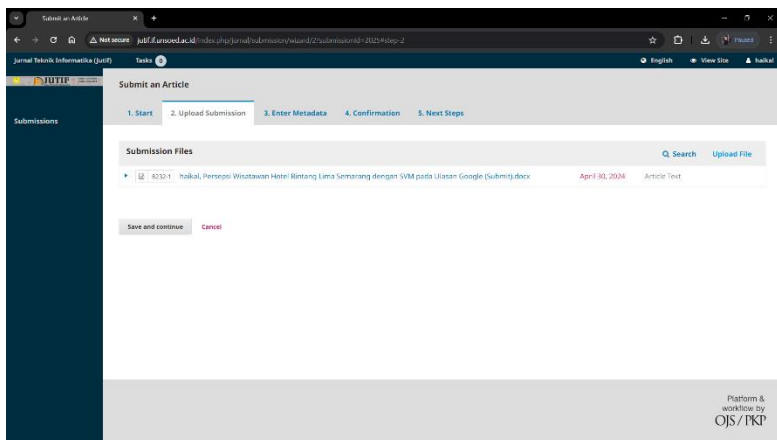
**Submission Requirements**

You must read and acknowledge that you've completed the requirements below before proceeding.

- ☐ The submission has not been previously published, nor is it before another journal for consideration (or an explanation has been provided in Comments to the Editor).
- ☐ The submission file is in Microsoft Word file format.
- ☐ The manuscript follow JUTIF's Template; and all Illustrations, figures, and tables are placed within the text at the appropriate points, rather than at the end.
- ☐ Use IEEE Format for References. Minimum references is 25 with 80% come from journals/conferences
- ☐ If you need assistance/help/want a fasttrack, please contact 0856406144 (By Whatsapp/Telegram)

**Comments for the Editor**

### Proses awal submit jurnal



**Jurnal Teknik Informatika (jutif)**

Submit an Article

1. Start   2. Upload Submission   3. Enter Metadata   4. Confirmation   5. Next Steps

**Submission Files**

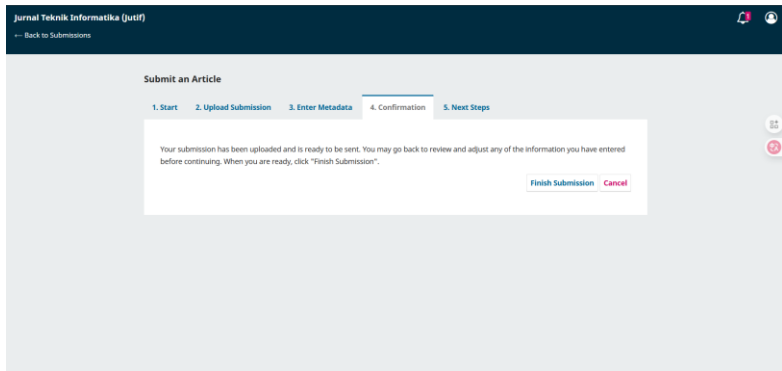
Search   Upload File

| File Name                                                                                        | Date           | Article Text |
|--------------------------------------------------------------------------------------------------|----------------|--------------|
| halal, Persepsi Wisatawan Hotel Bintang Lima Semarang dengan SVM pada Ujian Google (Submit).docx | April 30, 2024 | Article Text |

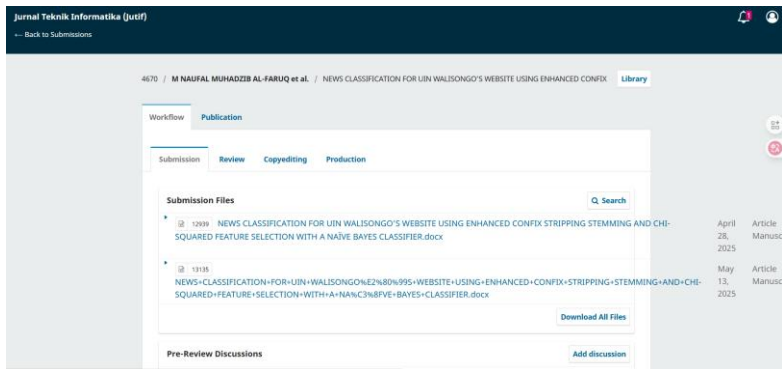
Save and continue   Cancel

Platform & workflow by OJS / PKP

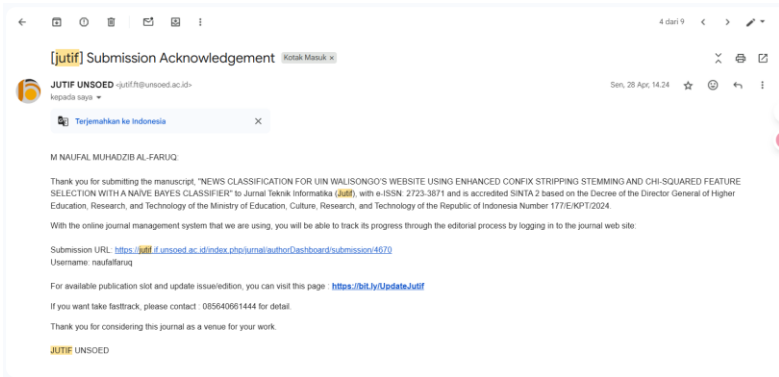
### Upload Naskah



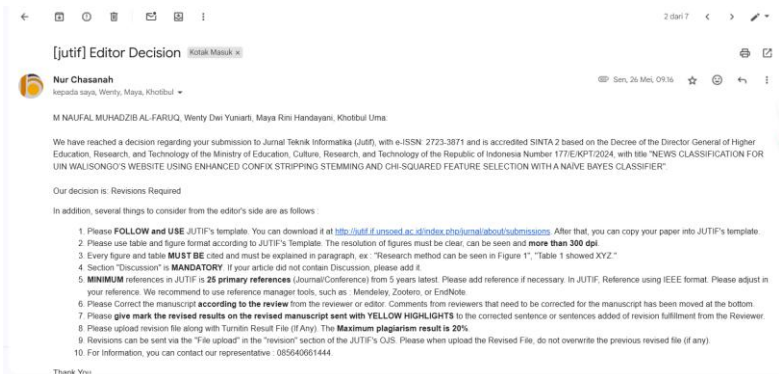
## Konfirmasi naskah



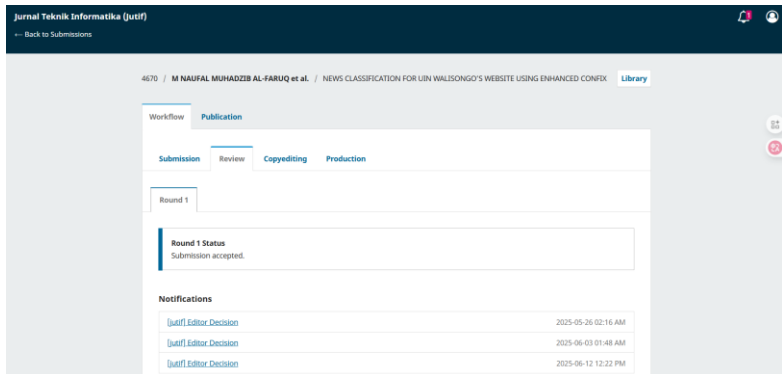
## Masuk tahap Submission



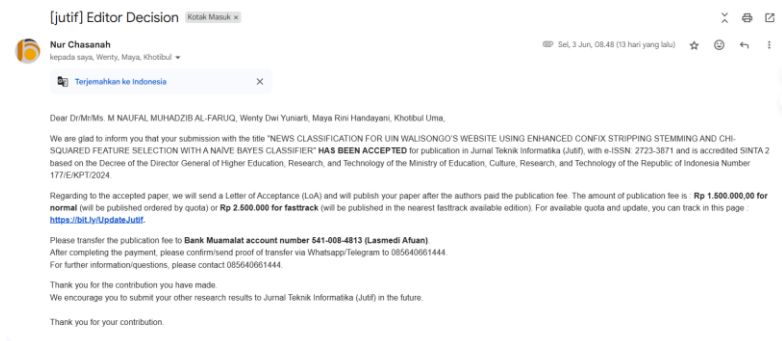
## Mendapat email balasan telah submit



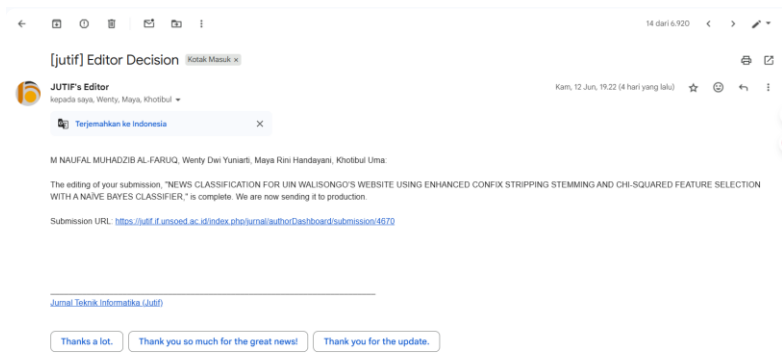
## Mendapat email untuk revisi artikel



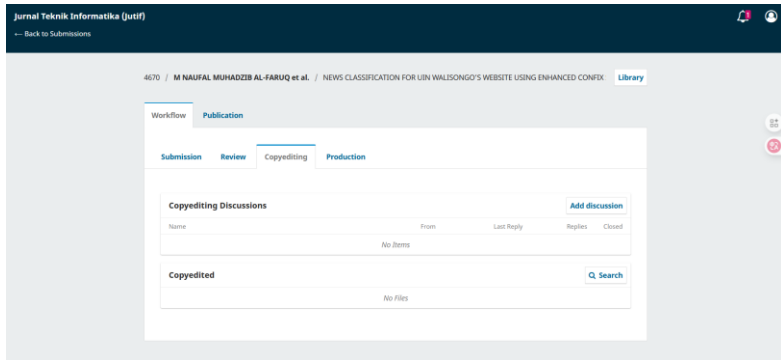
## Masuk ke tahap Review



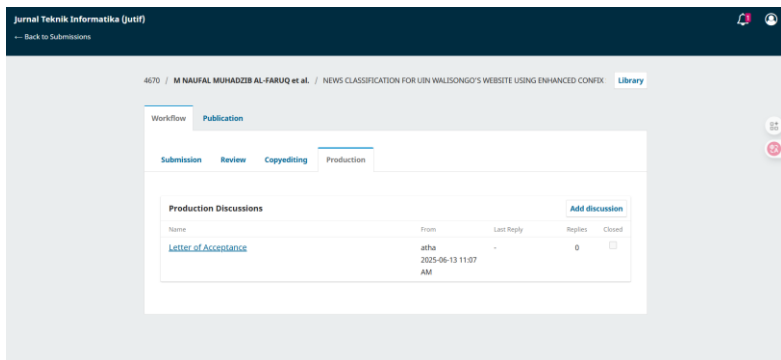
## Email informasi artikel telah diterima



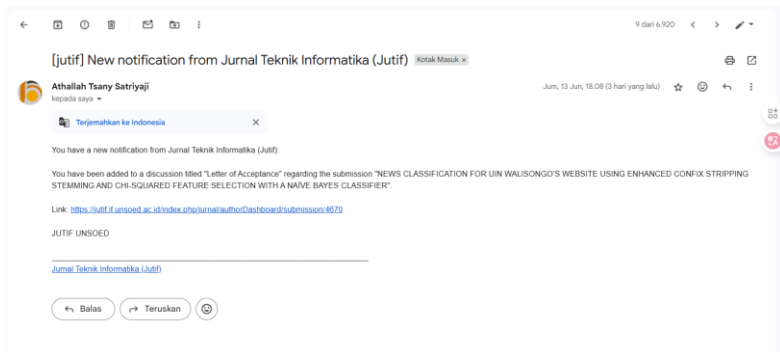
## Email informasi artikel masuk ke tahap production



## Tahap *Coppediting* dari pihak *publisher*



## Masuk tahap Produksi



Email informasi telah mendapatkan LOA



Sertifikat Jurnal

## Lampiran 3 Artikel

Jurnal Teknik Informatika (JUTIF)

P-ISSN: 2723-3863

E-ISSN: 2723-3871

Vol. 6, No. 6, December 2025, Page. 1530-1537

<https://jutif.if.unsoed.ac.id>

DOI: <https://doi.org/10.52436/1.jutif.6.6.3540>

### IMPLEMENTATION OF ENHANCED CONFIX STRIPPING STEMMING AND CHI-SQUARED FEATURE SELECTION ON CLASSIFICATION UIN WALISONGO WEBSITE WITH NAÏVE BAYES CLASSIFIER

Muhammad Naufal Muhadzib Al-Faruq<sup>1,\*</sup>, Wenty Dwi Yuniarti<sup>2</sup>, Maya Rini Handayani<sup>3</sup>, Khotibul Umam<sup>4</sup>

<sup>1,2,3,4</sup>Information Technology, Science and Technology,  
Walisongo State Islamic University, Indonesia

Email: <sup>1</sup>naufalfaruq082@gmail.com

Received : Jun 9, 2025; Revised : Nov 20, 2025; Accepted : Nov 22, 2025; Published : Dec 11, 2025  
Phone Number : 087722476038

#### Abstract

Academic news classification on university websites remains a challenge due to the growing volume of content and lack of efficient categorization systems. At UIN Walisongo Semarang, this problem hinders students, faculty, and the public from easily accessing relevant information. This study aims to develop an automated academic news classification system to address this issue. We applied a Naïve Bayes Classifier model, enhanced with Term Frequency weighting, the Enhanced Confix Stripping Stemmer for Indonesian language preprocessing, and Chi-Squared feature selection to identify the most informative terms. The dataset consisted of 880 academic news articles from UIN Walisongo's website, split into 704 training and 176 testing documents. The system achieved 95% accuracy on the test set. To evaluate generalizability, we used a separate evaluation set of 12 new articles, obtaining 83.3% accuracy. The preprocessing stage played a vital role in reducing morphological complexity, while Chi-Squared scoring improved the relevance of selected features. This research highlights the importance of robust text classification techniques in academic information systems, particularly in Indonesian language contexts where language morphology poses unique challenges. The proposed model demonstrates strong performance, scalability, and potential for integration into academic portals to improve information retrieval. This study contributes significantly to the field of Natural Language Processing and applied machine learning in academic settings, especially for Indonesian-language content. It provides an effective solution for automated academic content management in institutional information systems.

**Keywords :** Academic News, Chi-squared, Enhanced Confix Stripping Stemmer, Naive Bayes Classifier, classification

This work is an open access article and licensed under a Creative Commons Attribution-Non Commercial 4.0 International License



## 1. INTRODUCTION

The digital age has placed a premium on effortless access to information, a critical cornerstone for research and academic development within higher education institutions (HEIs). However, Walisongo State Islamic University Semarang (UIN Walisongo), despite its dedication to Islamic knowledge and science, faces challenges in ensuring efficient and effective information access for its diverse stakeholders, encompassing students, faculty, researchers, and broader community.

This article proposes a methodology for classifying news information on the UIN Walisongo's website. The ever-growing volume and diversified nature of news items necessitate the implementation of an automated news categorization system. Recent research has demonstrated the effectiveness of Naïve Bayes combined with Support Vector Machine in classifying user sentiment on mobile applications [1]. The objective of this study is to categorize news articles into four distinct categories:

campus, education, events, and announcements. This classification framework is designed to significantly enhance the accessibility and user experience of news content on the UIN Walisongo's website. To achieve this objective, data mining techniques are employed to analyze and automatically classify news articles.

This study addresses the gap in the availability of robust, domain-specific automated classification systems tailored for Indonesian academic news content. While prior studies have focused on general news classification, few have explored the challenges of classifying academic news articles, which often use formal, domain-specific language and unique structural patterns.

Data mining is a field of computer science to analyze and extract valuable information from previously unknown data [2]. Naïve Bayes classifiers have shown robust performance in hoax news classification in recent studies [3]. In this context, the data will be extracted from UIN Walisongo's website. Classification, another concept crucial to this study, refers to the process of assigning data points to specific categories based on a predefined set of criteria [4]. Naïve Bayes has also been successfully applied for classifying public complaints data, demonstrating its versatility across different text classification tasks [5]. The ultimate goal is to develop a classification model that can automatically categorize news articles into designated classes, such as campus news, educational announcements, upcoming events, and general university updates. This model will be built upon the analysis of the extracted dataset [6].

Prior to news classification, a preprocessing stage is implemented to prepare the data for analysis. This stage encompasses several essential steps, including case folding, tokenization/parsing, filtering, and stemming [7]. Stemming is a particularly important step within preprocessing, aiming to reduce inflected words to their base forms. This process is achieved through the Enhanced Confix Stripping Stemmer Algorithm, a development of the Confix Stripping Algorithm that addresses several limitations of its predecessor. The original Confix Stripping Algorithm presents certain limitations in processing Indonesian words, especially when dealing with complex affix combinations. This model often struggles to accurately remove prefixes and suffixes for specific word structures, which can lead to incorrect base forms. A comprehensive explanation of these limitations is provided in the methodology section [8].

This study explores the application of the Naïve Bayes Classifier for the automated categorization of online news articles. The Naïve Bayes Classifier has emerged as a prominent method for data classification across diverse domains, including medical diagnosis, computer network analysis, and text categorization [9]. Its enduring popularity stems from its inherent simplicity, effectiveness, and ability. The Naïve Bayes Classifier demonstrates competitive performance in terms of computational efficiency when compared to more complex classification algorithm [10].

Research on online news classification has been conducted extensively by previous researchers. Pramudita et al. (2018) in their study, a Naïve Bayes classifier with Enhanced Confix Stripping Stemmer was employed to classify sports news articles. Training data consisted of 18 randomly selected sports news articles chosen by the user. User analysis determined that 14 out of 18 articles were classified correctly, resulting in an accuracy 77%. This finding suggest that the proposed method holds promise for classifying sports news into predefined categories such as Football, Basketball, Racket, Formula 1, Motto GP, etc. [11].

Building upon the work of Ariadi and Fithiasari (2022) who compared Naïve Bayes (NB) and Support Vector Machine (SVM) classifier for Indonesian news classification, this study investigated the performance of these methods. Their finding indicated that the SVM classifier achieved superior performance using a Radial Basis Function (RBF) kernel and a linear kernel on a 10,000-word vector, with accuracy, precision, recall, and F-measure scores of 88.1%, 89.1%, 88.1%, and 88.3% respectively. In comparison the NB classifier achieved an accuracy of 82.2%, precision of 83.9%, recall of 82.2%,

and F-measure of 82.4%. These results suggest that SVM with RBF and linear kernel may be a more effective approach for classifying Indonesian news articles [12].

Hidayat and Rizqi (2020) investigated document classification using a dataset of 600 news articles obtained from the Jawa Pos news portal (<https://www.jawapos.com/>). The data was categorized into four balanced groups: Sports, Technology, Economy, and Miscellaneous (each containing 150 articles). The researchers employed a preprocessing stage involving case folding, tokenization, filtering, and stemming to prepare the data for classification. Their findings revealed high classification accuracy across all categories, achieving 90% for Sports, and Technology, 100% for Economy, and Miscellaneous. The overall average accuracy reached 95% [13].

Prakoso et al. (2019) explored Naïve Bayes classifier with feature selection and boosting techniques for news articles classification on the detik.com website. Their research involved data modelling and processing. They started with preprocessing, implemented Information Gain for feature selection, and using Bayesian Boosting for the algorithm. Information Gain feature selection fielded an accuracy of 69.5%, while a simpler model using Naïve Bayes Classifier with Bayesian Boosting achieved an accuracy of 73.2%. Furthermore, employing Bayesian Boosting for polynomial labels resulted in a 4.3% improvement in evaluation metrics. This finding suggest that Information Gain may not significantly enhance performance, particularly for dataset with polynomial labels [14].

Unlike previous works, this study introduces improvements to the Enhanced Confix Stripping Stemmer by modifying prefix-handling rules such as for words with the patterns "mem+*p*", "peng+*k*", and "penge+" and implementing an iterative suffix recovery mechanism. These enhancements address stemming errors commonly found in academic Indonesian texts, which are not sufficiently handled in prior research.

This study employs the Chi-Squared method to identify words with the highest Chi-Squared scores within each news category. To reduce inflected variants to their base form, the Enhanced Confix Stripping Stemmer is then applied to segment words by removing prefixes and suffixes. Following this, a Naïve Bayes Classifier algorithm is implemented to the model to train a news category prediction for classifying news articles from UIN Walisongo's website.

Additionally, this research highlights the synergy between Enhanced Confix Stripping and Chi-Squared feature selection. While the stemmer reduces morphological noise and ensures consistency of word forms, Chi-Squared scoring filters the most statistically relevant terms. This combination significantly improves both linguistic and statistical quality of features used in classification.

This study investigates the effectiveness of the Naïve Bayes Classifier in accurately categorizing academic news documents on UIN Walisongo's website. It also examines the extent to which the Enhanced Confix Stripping Stemmer improves the stemming process for Indonesian text and evaluates the role of the Chi-Squared method in enhancing feature selection for the classification model.

## 2. METHOD

In machine learning, classification refers to the task of constructing a model or function that can effectively categorize data points into predefined classes. The primary objective is to predict the class label of a new, unseen data point. This process typically involves two key stages: first, the model is trained using a set of labeled data points (training data), and then it is evaluated by classifying a separate set of labeled data points (test data) to assess its accuracy [15].

In machine learning, classification refers to the task of constructing a model or function that can effectively categorize data points into predefined classes. The primary objective is to predict the class label of a new, unseen data point. This process typically involves two key stages: first, the model is trained using a set of labeled data points (training data), and then it is evaluated by classifying a separate set of labeled data points (test data) to assess its accuracy.

In this study, the first stage begins with text preprocessing, which includes case folding, tokenization, filtering, and stemming using the Enhanced Confix Stripping Stemmer algorithm. Next, feature selection is performed using the Chi-Squared method, followed by model training with the Naïve Bayes algorithm. The complete research flow is visualized in Figure 1, outlining each stage of the classification process from data preprocessing to evaluation.

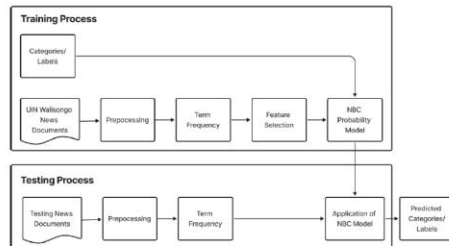


Figure 1. Research flow of the academic news classification system

The second stage involves testing the model with unseen test data, which also undergoes preprocessing (case folding, tokenization, filtering, and stemming). The preprocessed test data is then classified using the trained model. Finally, the model's performance is evaluated by comparing its predictions with the actual labels to calculate the accuracy, ensuring its reliability for real-world applications.

## 2.1. Text Preprocessing

Text Preprocessing constitutes the foundational step in text processing pipelines. It aims to transform raw, unstructured text data into a format that facilitates efficient and accurate computational analysis. This process involves several stages [7]. Enhanced stemming algorithms like Modified Enhanced Confix Stripping are crucial for handling Indonesian language morphology [16].

- Case Folding:** this initial step involves converting all characters in the text to lowercase. This normalizes the data by eliminating inconsistencies arising from capitalization, which can be irrelevant for many text analysis tasks.
- Tokenizing:** the text is then segmented into meaningful units, typically word or terms. This process creates a stream of token that serve as the basic elements for further analysis.
- Filtering:** following tokenization, filtering techniques are employed to select informative words or tokens. This stage often utilizes stopword removal algorithms, which eliminate common and uninformative words (e.g., "the", "a", "is" etc.). Alternatively, custom wordlist can be used to retain domain-specific or relevant terms.
- Stemming:** The goal of stemming is to reduce inflected words to their base forms. This process aims to identify the morphological root (stem) of each word, thereby grouping related words together and improving the overall effectiveness of subsequent analysis tasks.

## 2.2. Chi-Squared

The Chi-Squared test, a well-established statistical method, proves valuable in determining the association between two categorical variables [17]. Within the domain of text classification, this technique is employed to identify the most relevant features that contribute to distinguishing between different categories. Feature selection techniques such as Chi-Square have been proven to improve Naïve Bayes performance significantly [18]. The Chi-Squared score quantifies the discrepancy between the observed (O) and expected (E) frequencies of features within each category. The specific formula for calculating this score is presented in equation (1).

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

In equation (1),  $\chi^2$  represents the Chi-Squared score, while  $O_i$  and  $E_i$  denote the observed and expected frequencies of a word in a specific category, respectively. The parameter  $n$  refers to the number of categories being analyzed. This equation calculates the sum of the squared differences between observed and expected frequencies  $(O_i - E_i)^2$  for each category, normalized by the expected frequency  $E_i$ . A higher Chi-Squared score signifies a stronger association between a feature and a particular category. Therefore, features with the highest Chi-Squared scores are selected as the most informative, improving the accuracy of the classification model by emphasizing these significant features.

The expected frequency  $E_i$ , which is crucial for calculating the Chi-Squared score, is determined based on the overall distribution of features across categories. It is calculated using the formula presented in equation (2).

$$E_i = \frac{R \times C}{N} \quad (2)$$

In equation (2),  $R$  refers to the total occurrences of a feature across all categories,  $C$  represents the total number of occurrences in a specific category, and  $N$  is the total number of occurrences of all features across all categories. By dividing the product of  $R$  and  $C$  by  $N$ , the formula estimates the expected frequency of a feature within a given category under the assumption of independence. This calculation ensures that the Chi-Squared score effectively captures the degree of association between a feature and its corresponding category, facilitating the selection of the most relevant features for classification tasks.

## 2.3. Enhanced Confix Stripping Stemmer Algorithm

The Enhanced Confix Stripping Stemmer algorithm builds upon its predecessor, the Confix Stripping Stemmer, aiming to address limitations identified during experimentations and analysis. These limitations manifested in the inability the Confix Stripping Stemmer to effectively stem certain words. The specific issues encountered are as follows [8].

- The absence of prefix-breaking rules for words with the format "mem+p...". This results in the inability to correctly stem words like "mempromosikan" and "memproteksi".
- Other limitation is the absence of prefix-breaking rules for words following the format "men+s...". This leads to the incapability of correctly stemming words such as "mensyaratkan" and "mensyukuri".
- There are limitations in handling specific verb. For instance, the Stemmer fails to separate the prefix "menge+..." from the verb stem in words like "mengerem".
- The Stemmer cannot handle words with the prefix "penge+..." followed by consonant. This leads to incomplete stemming in words like "pengeboman".
- Inability to handle prefix words beginning with "peng+k..." followed by a consonant, leading to mis-stemmed words like "pengkajian".

- f. The Stemmer incorrectly identifies certain suffixes “-an”, “-kan”, and “-ku” in words like “pelanggan”, “perpolitikan”, and “pelaku” as prefixes, leading to the removal of these essential morphological elements

Due to limitations in the prefix-handling capabilities of the original Confix Stripping Stemmer algorithm, the Enhanced version introduces several rule modifications to improve prefix segmentation accuracy, especially for complex Indonesian word formations. As summarized in Table 1, these enhancements define specific transformations for word structures such as “mem+V”, “peng+C”, and “peng+V”. Each rule aims to correct errors previously unhandled in the standard approach. By making these rule-based refinements explicit, the Enhanced Confix Stripping Stemmer contributes to more reliable root word extraction during preprocessing—an essential component in improving classification accuracy. [19].

Table 1. Additions and Modifications to Word Prefix Rules in Enhance Confix Stripping Stemmer

| Rules | Word Structure     | Prefix Segmentation                            |
|-------|--------------------|------------------------------------------------|
| 14    | men {e c d j z}... | men-{e c d j z}...                             |
| 17    | mengV              | eng-V...   meng- kV...   (mengV-... If V='e')  |
| 19    | mempA...           | mem-pA...where A!=‘e’                          |
| 28    | pengC...           | peng-C...                                      |
| 29    | pengV...           | peng-V...   peng- kV...   (pengV-... If V='e') |

As shown in the table above, the Enhanced Confix Stripping Stemmer algorithm incorporates several enhancements. These include the modification of certain rules utilized in the decapitation of words and the introduction of a novel rule, designated as “to overcome suffix decapitating errors”. The specific details of loop “Pengembalian Akhiran” (Ending Return) algorithm is defined as follows [11].

- Following prefix segmentation, the algorithm attempts to reconstruct the original word by restoring the previously removed prefix. This reconstructed word, referred to as the “word model”, is then subjected to a dictionary lookup process. If the dictionary lookup successfully identifies the word model, the process terminates. Conversely, if the dictionary lookup fails to locate the word model, the algorithm proceeds to the next step.
- The algorithm employs an iterative process to reinstate previously removed suffixes. This restoration prioritizes suffixes in the following order: Derivational Suffixes (“-i”, “-kan”, “-an”), Possessive Pronoun (“-my”, “-mu”, “-nya”), Inflectional Particle Affixes (“-lah”, “-kah”, “-tah”, “-pun”). For each suffix type, the algorithm attempts to reconstruct the word by appending the suffix to the current stem and checks if the resulting form exists in a base word dictionary. If a match is found in the dictionary, the process terminates. Otherwise, the algorithm proceeds to the prefix beheading rules for further segmentation. Additionally, if the base word dictionary lookup remains unsuccessful after appending all the aforementioned suffixes, the previously removed prefixes are restored again, followed by iteration of the suffix restoration process.

#### 2.4. Term Frequency

Term Frequency is a fundamental technique used in information retrieval and text mining to assess the significance of individual words within a document. This method operates under the assumption that a word's importance is directly proportional to its frequency of occurrence within the text. In simpler terms, words that appear more frequently are considered to hold greater weight compared to those that appear less often. This weighting scheme is based on the intuition that words that are

repeated more frequently are likely to be more relevant or informative (higher weight) in conveying the document's central theme [7]. Chi-Square feature selection has been effectively applied for optimizing text classification models [20].

### 2.5. Naive Bayes Classifier

Naive Bayes Classifier, also known as Bayesian Classification, represent a family of machine learning algorithm that utilize Bayes' theorem for classification purposes. Bayes' theorem is a fundamental principle in probability theory that allows for the calculation of the conditional probability of an event (class membership) occurring given another event (specific features). In essence, the algorithms leverage the theorem to predict the likelihood of a data point belonging to a particular class based on its observed features. The mathematical formulation of Bayes' theorem is commonly expressed as equation (3).

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \quad (3)$$

The conditional probability of A given B, represented by  $P(A|B)$ , and the conditional probability of B given A ( $P(B|A)$ ).  $P(A)$  is the probability of event A, and  $P(B)$  is the probability of event B.

This study employs the Naive Bayes Classification method for text classification due to its appealing combination of simplicity and efficiency. The algorithm's core principle lies in its probabilistic approach, where each document is represented as a feature vector. This vector denoted as " $a_1, a_2, a_3, \dots, a_n$ " captures the presence or absence of individual words ( $a_1$ ) to ( $a_n$ ) within the document.

During classification, the Naive Bayes algorithm seeks to identify the category with the highest posterior probability. Optimization of text classification models using Chi-Square and TF-IDF has been explored to increase classification accuracy [21]. This probability is calculated using  $V_{MAP}$ , which is detailed in equation (4). The  $V_{MAP}$  equation is derived from Bayes' theorem and assumes conditional independence, enabling efficient computation of posterior probabilities for each category.

$$V_{MAP} = \operatorname{argmax}_{v_j \in v} P(v_j) \prod_i P(a_i|v_j) \quad (4)$$

$V_{MAP}$  presents the category with the maximum posterior probability. The term  $P(v_j)$  denotes the prior probability of category  $v$ , while  $P(a_i|v_j)$  is the conditional probability of feature  $a_i$  given that the data belongs to category  $v_j$ . The product  $\prod_i P(a_i|v_j)$  combines the probabilities of all features assuming independence.

$$P(v_j) = \frac{|\text{doc } j|}{|\text{training}|} \quad (5)$$

Where  $|\text{doc } j|$  represents the number of documents (news articles) belonging to category  $j$ , divided by the total number of documents in the training set. Additionally, equation (6) governs the calculation of the probability of word  $a_i$  occurring within each category  $P(a_i|v_j)$  during the training process.

$$P(a_i|v_j) = \frac{|n_i+1|}{|n+\text{kosakata}|} \quad (6)$$

The  $n_i$  represents the number of times word  $a_i$  appears in documents belonging to category  $v_j$ , while  $n$  is the total number of words in documents belonging to category  $v_j$ , and vocabulary is the total number of words in the examples in the training sample. To measure classification accuracy, equation (7) can be used.

$$\text{Accuracy} = \frac{\text{Number of Correct Test Documents}}{\text{Total Test Documents}} \times 100\% \quad (7)$$

### 3. RESULT

The classification structure employed for Walisongo State Islamic University Semarang (UIN Walisongo) news documents involves two distinct stages: training and testing. Data collection was

conducted through the [ppid.walisongo.ac.id](http://ppid.walisongo.ac.id) news portal, yielding a corpus of 880 news documents. Subsequently, the documents were partitioned into two subsets with a ratio of 80% designated for training and 20% reserved for testing purposes. During the training stage, a probability model was constructed utilizing 704 training documents. The remaining 176 documents were used as test documents to evaluate the efficacy of the constructed probability model.

Prior to the commencement of the training process, the training documents outlined in Table 2 undergo a series of preprocessing steps, including term frequency analysis and feature selection. Test documents are subjected to a similar preparatory process before classification based on probability model constructed during the learning phase. The preprocessing stage aims to transform unstructured text data into a structured format suitable for machine learning algorithms. This stage encompasses several techniques, such as case folding, tokenization, filtering, and stemming.

As detailed in Table 3, the first step involves case folding, where all characters are converted to lowercase to ensure consistency. Subsequently, the tokenization stage (Table 4) parses the text content into individual words. Following tokenization (Table 5), a filtering process is employed to retain only relevant words and eliminate stopwords, which are common words that hold minimal semantic value.

The final preprocessing step involves stemming, which reduces words to their root forms. This research employs the Enhanced Confix Stripping Stemmer algorithm for this purpose. Table 6 exemplifies the outcome of this process on a document.

Table 2. Sample training document

| Title                                                                                             | Content                                                                                                                                  | Category  |
|---------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------|-----------|
| Walisongo Halal Center Terbitkan 275 Sertifikat Juru Sembelih Halal Se-eks Karesidenan Pekalongan | Walisongo Halal Center (WHC) bekerjasama dengan BI Tegal, Komite Nasional Ekonomi dan Keuangan Syariah ..... tegasnya                    | Event     |
| FDK UIN Walisongo-PAMHU gelar Sertifikasi Pembimbing Manasik Haji Profesional Angkatan XV         | Fakultas Dakwah dan Komunikasi (FDK) UIN Walisongo Semarang menyelenggarakan sertifikasi pembimbing manasik haji profesional ..... PAHMU | Education |

Table 2 presents an illustrative example of an academic news document from UIN Walisongo Semarang. This documents comprises the title and content of the news article and will be utilized as a training document. The subsequent step in the preprocessing pipeline involves case folding, as detailed in Table 3.

Table 3. Example document result of case folding process

| Title                                                                                                 | Content                                                                                                               | Output Title                                                                                           | Output Content                                                                                                           |
|-------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------|
| Walisongo Halal Center Terbit-kan 275 Sertifi-kat Juru Semb-elih Halal Se-eks Kar-esidenan Pekalongan | Walisongo Halal Center (WHC) bekerjasama dengan BI Tegal, Komite Nasional Ekonomi dan Keuangan Syariah ..... tegasnya | walisongo halal cen-ter terbi-tkan 275 sertifi-kat juru semb-elih halal se-eks kar-esidenan pekalongan | walisongo halal center (whc) beker-jasama deng-an bi tegal, komite nasi-onal ekonomi dan keuangan syariah ..... tegasnya |
| FDK UIN Walisongo-PAMHU gelar Sertifikasi Pembimbing                                                  | Fakultas Dakwah dan Komunikasi (FDK) UIN Walisongo Semarang menyelenggarakan                                          | fdk uin walisongo-pamhu gelar sertifikasi pembimbing                                                   | fakultas dakwah dan komunikasi (fdk) uin walisongo semarang menyelenggarakan                                             |

|                                            |                                                                   |                                            |                                                                     |
|--------------------------------------------|-------------------------------------------------------------------|--------------------------------------------|---------------------------------------------------------------------|
| Manasik Haji<br>Profesional<br>Angkatan XV | sertifikasi pembimbing<br>manasik haji profesional<br>..... PAHMU | manasik haji profes-<br>ional angk-atan xv | sertifik-asi pembimb-ing<br>manasik hajiprofesio-nal<br>..... pahmu |
|--------------------------------------------|-------------------------------------------------------------------|--------------------------------------------|---------------------------------------------------------------------|

Following the case folding process, both the title and news content are converted entirely to lowercase letters. This step serves to eliminate inconsistencies arising from capitalization, thereby mitigating the potential for redundant data representations in Table 4.

Table 4. Example of tokenizing result document

| Title                                                                                                | Content                                                                                                                                   | Output Title                                                                                    | Output Content                                                                                                                          |
|------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------|
| walisongo halal center terbi-tkan 275 sertifi-kat juru sembelih halal se-eks kar-esidenan pekalongan | walisongo halal center (whc) beker-jasama deng-an bi tegal, komite nasi-onal ekonomi dan keuangan syariah ..... tegasnya                  | walisongo halal center terbitkan 275 sertifikat juru sembelih halaleks', karesidenan pekalongan | walisongo halal center whc beker-jasama deng-an bi tegal, komite nasi-onal ekonomi dan keuangan syariah ..... tegasnya                  |
| fdk uin walisongo-pamhu gelar sertifikasi pe-mbimbing manasik haji profesional angk-atan xv          | fakultas dakwah dan komunikasi (fdk) uin walisongo semarang menyelenggarakan sertifikasi-pem-bimbing manasik haji profesional ..... pahmu | fdk uin walisongo-pamhu gelar sertifikasi pembimbing manasik haji profesional angk-atan         | fakultas dakwah dan komunikasi fdk uin walisongo semarang menyelenggarakan sertifikasi-pembimb-ing manasik hajiprofesio-nal ..... pahmu |

Subsequent to case folding, the next step is tokenization. This process involves segmenting the text into smaller units known as tokens. Tokenization simplifies the text by dividing it into individual words or sub-words, facilitating subsequent natural language processing task.

Following tokenization, where the text is segmented into individual words, a filtering process is implemented. This stage aims to select informative terms from the tokenized corpus. Conversely, unimportant words, of the referred to as stopwords, are eliminated using a stopwords removal algorithm.

Table 5. Example of tokenizing result document

| Title                                                                                                | Content                                                                                                                | Output Title                                                                                                                         | Output Content                                                                                                                         |
|------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------|
| walisongo halal center terbi-tkan 275 sertifi-kat juru sembelih halal se-eks kar-esidenan pekalongan | walisongo halal center whc beker-jasama deng-an bi tegal, komite nasi-onal ekonomi dan keuangan syariah ..... tegasnya | 'walisongo', 'halal', 'center', 'terbitkan', '275', 'sertifikat', 'juru', 'sem-belih', 'hal-al', 'eks', 'ka-residenan', 'pekalongan' | 'walisongo', 'halal', 'center', 'whc', 'bekerjasama', 'tegal', 'komite', 'nasional', 'ekonomi', 'keuangan', 'syariah', ..... 'tegasya' |
| fdk uin walisongo-pamhu gelar sertifikasi pe-mbimbing                                                | fakultas dakwah dan komunikasi fdk uin walisongo semarang menyelenggarakan sertifikasi-pem-bimbing                     | 'fdk', 'uin', 'walisongo', 'pamhu', 'gelar', 'sertifikasi', 'pe-mbimbing',                                                           | 'fakultas', 'dakwah', 'komunikasi', 'fdk', 'uin', 'walisongo', 'semarang', 'menyelenggarakan',                                         |

manasik haji profes-  
ional angk-atan

ifik-asi pem-bimbing man-  
asik haji profesional .....  
pahmu

'manasik', 'haji',  
'profe-sional',  
'angkatan'

'sertifikasi',  
'pembimbing', 'manasik',  
'haji', 'profesional', .....  
'pahmu'

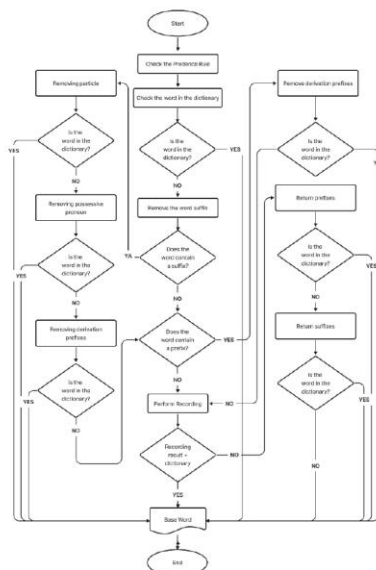


Figure 2. Flowchart enhanced confix stripping stemmer

Figure 2 outlines the stages of the Enhanced Confix Stripping Stemmer Algorithm. This algorithm integrates three distinct components: Enhanced, Confix Stripping, and Stemmer, each addressing specific aspects of word normalization.

The Enhanced component represents improvements over the original Confix Stripping Algorithm. This stage begins with a precedence rule check, introducing new rules to handle complex affix combinations such as "mem-", "meng-", and "penge-". These enhancements address limitations in the original algorithm, allowing it to process a broader range of Indonesian word structures accurately.

The Confix Stripping process focuses on systematically removing affixes from words. It is divided into two main stages:

- Case Folding:** this initial step involves converting all characters in the text to lowercase. This **Suffix Removal:** This stage eliminates inflectional particle affixes (e.g., "-lah", "-kah"), possessive pronouns (e.g., "-ku", "-mu", "-nya"), and derivational suffixes (e.g., "-kan", "-an").
- Prefix Removal:** After suffix removal, prefixes such as "meng-", "pen-", and "ber-" are removed using derivation prefix elimination rules. This ensures that the resulting word aligns more closely with its base form.

**Prefix Removal:** After suffix removal, prefixes such as "meng-", "pen-", and "ber-" are removed using derivation prefix elimination rules. This ensures that the resulting word aligns more closely with its base form.

The Stemmer serves as the overarching process that integrates the Enhanced and Confix Stripping components to identify the base word. If a word cannot be matched to a base word after prefix and suffix removal, the algorithm employs the word recording process. This process reconstructs the word by restoring previously removed prefixes or suffixes and checking the reconstructed word against a base word database. If no match is found, the algorithm iterates through the Ending Return Algorithm stage, reapplying suffix rules to refine the normalization process.

By clearly delineating these components, the Enhanced Confix Stripping Stemmer Algorithm effectively addresses the challenges of processing Indonesian text, ensuring robust and accurate stemming results.

Table 6. Document example of enhanced confix stripping stemmer implementation result

| Title                                                                                                                                                   | Content                                                                                                                                                                                          | Output Title                                                                                                                                 | Output Content                                                                                                                                                                                 |
|---------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 'walisongo', 'halal',<br>'center', 'terbitkan',<br>'275', 'sertifikat',<br>'juru', 'sem-belih',<br>'hal-al', 'eks', 'ka-<br>residenan',<br>'pekalongan' | 'walisongo', 'halal',<br>'center', 'whc',<br>'bekerjasama', 'tegal',<br>'komite', 'nasional',<br>'ekonomi', 'keuangan',<br>'syariah', ..... 'tegasya'                                            | 'walisongo', 'halal',<br>'center', 'terbit',<br>'275', 'sertifikat',<br>'juru', 'sembelih',<br>'halal', 'eks',<br>'karesidenan',<br>'kalong' | 'walisongo', 'halal',<br>'center', 'whc',<br>'bekerjasama', 'tegal',<br>'komite', 'nasional',<br>'ekonomi', 'uang',<br>'syariah', ..... 'tegas'                                                |
| 'fdk', 'uin',<br>'walisongo', 'pamhu',<br>'gelar', 'sertifikasi',<br>'pe-mbimbing',<br>'manasik', 'haji',<br>'profe-sional',<br>'angkatan'              | 'fakultas', 'dakwah',<br>'komunikasi', 'fdk', 'uin',<br>'walisongo', 'semarang',<br>'menyelenggarakan',<br>'sertifikasi',<br>'pembimbing', 'manasik',<br>'haji', 'profesional', .....<br>'pahmu' | 'fdk', 'uin',<br>'walisongo', 'pamhu',<br>'gelar', 'sertifikasi',<br>'pembimbing',<br>'manasik', 'haji',<br>'profesional',<br>'angkatan'     | 'fakultas', 'dakwah',<br>'komunikasi', 'fdk', 'uin',<br>'walisongo', 'semarang',<br>'menyelenggarakan',<br>'sertifikasi',<br>'pembimbing', 'manasik',<br>'haji', 'profesional'.....<br>'pahmu' |

Table 6 presents the results of applying the Enhanced Confix Stripping Stemmer to real academic news documents. The output columns demonstrate how complex affixed words are normalized into their root forms. This process significantly improves feature consistency by reducing morphological variation in Indonesian text, which is essential for achieving higher accuracy in text classification. The normalized terms allow the Naïve Bayes model to better generalize and distinguish between categories such as education, event, campus, and announcements.

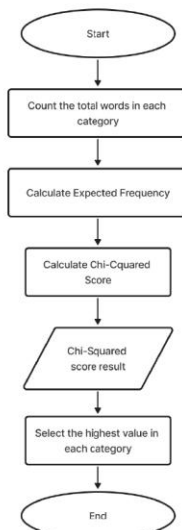


Figure 3. Flowchart feature selection chi-squared

Following preprocessing, the training data underwent feature selection using the Chi-Square test. This method aimed to identify the most discriminative features for news categorization, thereby enhancing the performance of the Naïve Bayes classifier. The process, visualized in Figure 3, involved calculating the total word count per category, determining observed and expected frequencies for each term within each category, computing the Chi-Square score, and selecting features based on their associated Chi-Square values.

Observed frequency represents the actual occurrences of a term within a specific category, while expected frequency is the theoretically predicted occurrence under the assumption of independence. The Chi-squared statistic measures the deviation between these frequencies, indicating the term's relevance to the category. Higher Chi-Square values correspond to stronger associations between terms and categories. The features with the highest Chi-Square scores were subsequently employed as input for the Naïve Bayes classifier. Table 7 presents the top three terms with the highest Chi-Square values for each category. A Chi-Square score approaching 1.0 signifies a more pronounced association between term and its corresponding category.

Table 7. Sample document of chi-squared implementation result

| Word        | Chi-Squared Score | Category     |
|-------------|-------------------|--------------|
| mahasiswa   | 1.0               | education    |
| dosen       | 0.98              | education    |
| kuliah      | 0.98              | education    |
| seminar     | 1.0               | event        |
| upacara     | 0.99              | event        |
| diskusi     | 0.98              | event        |
| ukt         | 1.0               | campus       |
| uin         | 0.99              | campus       |
| ptkin       | 0.99              | campus       |
| seleksi     | 1.0               | announcement |
| Pendaftaran | 0.94              | announcement |
| informasi   | 0.94              | announcement |

The final stage involved training the Naïve Bayes classifier using the preprocessed training documents. This process encompassed several steps. Firstly, test documents, as outlined in Table 8, were inputted and subjected to preprocessing. Subsequently, term frequency calculations were performed. The system then calculated  $P_j = P(v_j) \cdot n \cdot P(a_i|v_j)$  for each category, utilizing the probability model established during the training phase. Upon completion, the system automatically assigned categories to the test documents. Evaluation was conducted using test documents with unknown categories, as listed in Table 8.

Table 8. Sample test document

| Title                                                                                                     | Content                                                                                                                                                               | Category |
|-----------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|
| Moderasi Beragama, Hidup Damai di Tengah Masyarakat Plural                                                | Gagasan moderasi beragama ala Walisongo harus disebarkan kepada seluruh masyarakat Indonesia, karena sikap Walisongo dalam mengapresiasi local wisdom ..... pelatihan | Unknow   |
| Sekretaris RMB UIN Walisongo mengajar Political Theology Moderasi Beragama di Universitas Vienna, Austria | Luthfi Rahman, diundang untuk mengisi mata kuliah Approaches of Islamic Theology to Dialogue di salah satu kelas mahasiswa jurusan .... beragam                       | Unknow   |

Table 8 presents an example of an uncategorized academic news document from UIN Walisongo Semarang. Comprising a headline and news content. This document will serve as a test case for evaluating the previously constructed Naïve Bayes probability model, which was trained using a separate dataset. Prior to classification, the test document undergoes preprocessing, including case folding, tokenizing, filtering, and stemming.

Table 9. Example of categorized test document

| Title                                                                                                     | Content                                                                                                                                                                 | Category  |
|-----------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|
| Moderasi Beragama, Hidup Damai di Tengah Masyarakat Plural                                                | Gagasan moderasi beragama ala Walisongo harus disebarakan kepada seluruh masyara-kat Indonesia, karena sikap Walisongo dalam mengapresiasi local wisdom ..... pelatihan | Education |
| Sekretaris RMB UIN Walisongo mengajar Political Theology Moderasi Beragama di Universitas Vienna, Austria | Luthfi Rahman, diundang untuk mengisi mata kul-iah Approaches of Islamic Theology to Dialogue di sa-lah satu kelas ma-hasiswa jurusan ..... beragam                     | Event     |

Following the preprocessing stage, the machine automatically categorized the test documents using the constructed Naïve Bayes probability model. A total of 176 test documents were evaluated, achieving the accuracy rate of 95%. As detailed in Table 9, 167 documents were classified correctly.

In the context of classifying academic news documents from UIN Walisongo Semarang, three documents categories exist: training, testing, and new documents. The initial data collection retrieved 880 news documents from [ppid.walisongo.ac.id](http://ppid.walisongo.ac.id). these documents were subsequently divided into two subsets: 80% designated as training documents and 20% reserved for testing purposes. During the training phase, the probability model was built utilizing 704 training documents. The remaining 176 documents were employed as test documents to evaluate the efficacy of the model. To further assess the system's categorization accuracy, an additional set of 12 latest news was created and compared against user-assigned categories.

Table 10. Testing results

| No | News Title                                                                        | News Content                                                                                                    | Category (System) | Category (User) |
|----|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------|-------------------|-----------------|
| 1  | Pengumuman Beasiswa KIP-K untuk Mahasiswa Baru Tahun 2024                         | UIN Walisongo memberikan pengumuman terkait program beasiswa KIP-K ..... kampus                                 | Announcement      | Announcement    |
| 2  | Fakultas Sains dan Teknologi Adakah Pelatihan Kewirausahaan Bagi Mahasiswa        | Fakultas Sains dan Teknologi akan menyelenggarakan pelatihan kewirausahaan ..... mereka                         | Event             | Event           |
| 3  | Civitas Akademik Melakukan Kerja Bakti Untuk Menjaga Kebersihan Lingkungan Kampus | Civitas akademik mengadakan kegiatan membersihkan lingkungan kampus. kegiatan ini ..... nyaman.                 | Event             | Campus          |
| 4  | Fakultas Dakwah dan Komunikasi adakan Pelatihan Desain Grafis untuk Mahasiswa     | Kompetisi ini akan memberikan kesempatan bagi mahasiswa untuk menunjukkan .... kemam-puan desain grafis mereka. | Education         | Education       |

|    |                                                                                    |                                                                                                  |              |              |
|----|------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------|--------------|--------------|
| 5  | UIN Walisongo Gelar Diskusi Etika dalam Kehidupan Sehari-hari                      | UIN Walisongo akan mengadakan diskusi tentang etika dan moralitas ..... kehidupan sehari-hari.   | Campus       | Acara        |
| 6  | UIN Walisongo Gelar Webinar Internasional Tentang Pendidikan Islam di Era Digital  | Universitas Islam Negeri (UIN) Walisongo Semarang menggelar webinar internasional .... mahasiswa | Event        | Event        |
| 7  | Kantin FST UIN Walisongo Longsor Akibat Hujan Deras                                | Akibat hujan deras selama sehari penuh kantin FST longsor ..... perbaikan                        | Campus       | Kampus       |
| 8  | PTIPD Lakukan Workshop Pelatihan Jurnalistik                                       | Dalam meningkatkan kualitas kepenulisan mahasiswa ..... jurnalistik                              | Education    | Education    |
| 9  | Mahasiswa Teknologi Informasi UIN Walisongo Raih Prestasi di Tingkat Internasional | Tiga mahasiswa Teknologi Informasi UIN Walisongo Semarang raih penghargaan ..... Malaysia        | Campus       | Campus       |
| 10 | Perubahan Tanggal Pelaksanaan Wisuda Periode Juni 2024                             | Tanggal Pelaksanaan Wisuda Periode Juni 2024 dirubah ..... mahasiswa                             | Announcement | Announcement |
| 11 | UIN Walisongo Buka Jalur Prestasi untuk 34 Jurusan, Ini Jadwal dan Syaratnya       | Jalur Prestasi UIN Walisongo kembali dibuka untuk 34 jurusan ..... syarat ketentuan              | Announcement | Announcement |
| 12 | Lantik Pejabat Struktural, Ini Pesan Rektor UIN Walisongo                          | Dalam rangka pelantikan pejabat struktural periode baru rektor berpesan ..... kualitas           | Event        | Event        |

The system was evaluated using twelve novel news documents. Results indicate that ten news stories were accurately categorized according to user assessments, as detailed in Table 10. The overall accuracy rate was 83.33%, with an error rate of 16.67%.

$$Accuracy = \frac{10}{12} \times 100\% = 83,33\%$$

Two misclassifications were identified in the test dataset. In the third news item, the system predicted "Event" while the user labeled it as "Campus". This discrepancy likely occurred because the article included action-oriented phrases such as "mengadakan kegiatan" (holding an activity), which semantically align with event contexts despite referring to campus maintenance.

The fifth news item was also misclassified. The system labeled it as “Campus” whereas the user labeled it as “Event”. This error may have been caused by the system focusing on institutional terms like “diskusi tentang etika dan moralitas” (discussion on ethics and morality), which may have been interpreted as campus-related content rather than a scheduled event.

These findings indicate that the classification model could benefit from further refinement in distinguishing nuanced language, particularly when semantic overlap exists between categories such as campus activities and organized events.

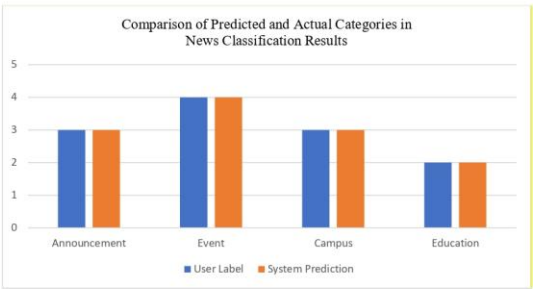


Figure 4. Comparison of Predicted and Actual Categories in News Classification Results

Figure 4 presents a bar chart comparing the predicted categories by the system and the actual categories labeled by users for the 12 test news documents. The distribution shows a strong alignment between system predictions and user labels across all four categories. No significant overclassification or underclassification was observed, indicating balanced category recognition by the model.

| Table 11. Evaluation Metrics |           |        |          |         |
|------------------------------|-----------|--------|----------|---------|
| Category                     | Precision | Recall | F1-Score | Support |
| Event                        | 0.75      | 0.75   | 0.75     | 4       |
| Campus                       | 0.67      | 0.67   | 0.67     | 4       |
| Education                    | 1.0       | 1.0    | 1.0      | 3       |
| Announcement                 | 1.0       | 1.0    | 1.0      | 2       |

As shown in Table 11, the evaluation metrics indicate that the system achieved perfect precision, recall, and F1-score for the “Announcement” and “Education” categories, reflecting that no misclassifications occurred in those groups. In contrast, the “Event” and “Campus” categories recorded lower F1-scores of 0.75 and 0.67 respectively, which is consistent with the two classification errors identified earlier. These results demonstrate the model’s effectiveness in recognizing most academic news categories, while highlighting the need for further refinement in handling semantically overlapping categories such as campus-related activities and organized events.

#### 4. DISCUSSIONS

The findings of this study demonstrate that the integration of the Enhanced Confix Stripping Stemmer, Chi-Squared feature selection, and Naïve Bayes Classifier achieves a robust level of performance in categorizing academic news documents from UIN Walisongo's website. The proposed model attained an accuracy of 83.33% on newly collected data, indicating the model's generalizability and adaptability in real-world academic contexts. Aspect-based sentiment analysis using SVM has been successfully implemented on e-commerce review datasets [22].

Compared to prior studies, the performance of this model is promising. For instance, Pramudita et al. [8] applied a similar approach for sports news classification and achieved an accuracy of 77%. The higher accuracy observed in this research suggests that improvements made to the stemming process and feature selection significantly enhance classification performance.

Ariadi and Fithiasari [9] compared Naïve Bayes and Support Vector Machine (SVM) for Indonesian news classification and reported that SVM outperformed Naïve Bayes, with SVM achieving 88.1% accuracy while Naïve Bayes achieved 82.2%. Despite this, Naïve Bayes remains a favorable choice due to its computational simplicity and interpretability, which align with the objectives of this study focusing on scalable academic categorization.

In addition, the study by Hidayat and Rizqi [10], which utilized the Enhanced Confix Stripping Stemmer and Naïve Bayes, obtained a 95% accuracy in a general news classification task. Although this is higher than the result in the current study, it is important to note that their dataset consisted of standard news content, whereas this study focused on academic news with specific linguistic patterns and domain terms.

Overall, the classification accuracy obtained confirms that the selected approach is well-suited for academic environments. The Enhanced Confix Stripping Stemmer effectively handled morphological complexities in the Indonesian language, while Chi-Squared feature selection allowed for the identification of discriminative terms across categories. Recent comparative studies highlight the advantage of combining FastText and Chi-Square for improving sentiment analysis models [23]. The Naïve Bayes Classifier, despite its simplicity, proved capable of maintaining competitive accuracy while offering efficient computational performance, making it an appropriate choice for institutional applications. Social media data poses significant challenges for text classification, especially during events like the COVID-19 pandemic [24]. Hoax news detection leveraging Naïve Bayes and SVM continues to be an important research trend [25].

Despite the promising results, this study has several limitations. The classification model was trained on a relatively small dataset, which may limit its ability to generalize across a broader and more diverse range of academic news. In addition, although the probabilistic classification method used in this study is computationally efficient and easy to implement, it lacks the capability to understand contextual meaning within the text, especially when compared to more advanced deep learning-based classification models.

Future research could focus on expanding the dataset and evaluating the performance of the classification model in more dynamic environments. Another important direction would be to implement this system in real-time academic content platforms to enable automatic classification of news upon publication. Furthermore, integrating this system with more advanced language models that are capable of understanding word context and sentence structure could significantly enhance classification accuracy, particularly in academic texts that are longer and more complex.

## 5. CONCLUSION

To provide a clearer understanding of the outcomes of this research, the key conclusions have been structured into the following points. These conclusions are derived from the analysis and evaluation of the classification model applied to academic news documents from UIN Walisongo's website, highlighting the effectiveness of the techniques used and the contributions made to the field of academic text classification.

- The Naïve Bayes Classifier, integrated with the Enhanced Confix Stripping Stemmer and Chi-Squared feature selection, successfully categorized academic news documents on UIN Walisongo's website, achieving an accuracy of 83.33% and an error rate of 16.67%.
- The Enhanced Confix Stripping Stemmer improved the stemming process by addressing challenges in processing complex Indonesian word structures, contributing to the overall performance of the model.
- The Chi-Squared method effectively selected informative features, enhancing the model's ability to differentiate between news categories.
- These findings support the advancement of academic text classification, especially for Indonesian-language educational content, by offering a lightweight and interpretable model suitable for institutional use.
- Future research is encouraged to compare this model with other classification algorithms such as Support Vector Machine and neural networks, to explore improvements in contextual understanding and accuracy.
- Overall, this study contributes meaningfully to the field of Informatics, particularly in the development of intelligent information systems for academic environments.

## REFERENCES

- [1] M. D. Rizkiyanto, M. D. Parbolaksono, and W. Astuti, "Sentiment Analysis Classification on PLN Mobile Application Reviews Using Random Forest Method and TF-IDF Feature Extraction," *INTEK: Jurnal Penelitian*, vol. 11, no. 1, pp. 37–43, Apr. 2024, doi: <https://doi.org/10.31963/intek.v11i1.4774>.
- [2] W. Purba, W. Siawin, M. N. K. Nababan, N. P. Dharshinni, and S. Aisyah, "Implementasi Data Mining untuk Pengelompokan dan Prediksi Karyawan yang Berpotensi PHK dengan Algoritma K-Means Clustering," *JUSIKOM PRIMA: Jurnal Sistem Informasi Dan Ilmu Komputer Prima*, vol. 2, no. 2, pp. 85–90, 2019, doi: <https://doi.org/10.34012/jusikom.v2i2.429>.
- [3] R. Qubra and R. A. Saputra, "Classification of Hoax News Using the Naïve Bayes Method," *IJSECS: International Journal Software Engineering and Computer Science*, vol. 4, no. 1, pp. 40–48, Apr. 2024, doi: <https://doi.org/10.35870/ijsecs.v4i1.2068>.
- [4] N. Buslim, L. K. Oh, M. H. Athallah Hardy, and Y. Wijaya, "Comparative Analysis of KNN, Naïve Bayes and SVM Algorithms for Movie Genres Classification Based on Synopsis," *JTI: Jurnal Teknik Informatika*, vol. 15, no. 2, pp. 169–177, Dec. 2022, doi: <https://doi.org/10.15408/jti.v15i2.29302>.
- [5] K. Wabang, Oky Dwi Nurhayati, and Farikhin, "Application of the Naïve Bayes Classifier Algorithm to Classify Community Complaints," *Jurnal RESTI: Rekayasa Sistem dan Teknologi Informasi*, vol. 6, no. 5, pp. 872–876, Nov. 2022, doi: <https://doi.org/10.29207/resti.v6i5.4498>.
- [6] S. Widaningsih, "Perbandingan Metode Data Mining untuk Prediksi Nilai dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika dengan Algoritma C4.5, Naïve Bayes, KNN

- dan SVM," *Jurnal Tekno Insentif*, vol. 13, no. 1, pp. 16–25, 2019, doi: <https://doi.org/10.36787/jti.v13i1.78>.
- [7] N. Nurdin, M. Suhendri, Y. Afrilia, and R. Rizal, "Klasifikasi Karya Ilmiah (Tugas Akhir) Mahasiswa Menggunakan Metode Naïve Bayes Classifier (NBC)," *SISTEMASI : Jurnal Sistem Informasi*, vol. 10, no. 2, pp. 268–279, 2021, doi: <https://doi.org/10.32520/stmsi.v10i2.1193>.
  - [8] A. M. Billah, D. A. R. Wulandari, and Y. A. Auliya, "Rancang Bangun Chatbot Pengaduan Kekerasan Perempuan Anak dengan Metode Fuzzy String Matching dan Enhanced Confix Stripping Stemmer," *INFROMAL : Informatics Journal*, vol. 8, no. 2, pp. 101–109, 2023, doi: <https://doi.org/10.19184/isj.v8i2.42310>.
  - [9] I. Sholekha, A. Faqih, and A. Baltiar, "Sentiment Analysis of Public Opinion Covid-19 Vaccine Using Naïve Bayes and Random Forest Methods," *JTI : Jurnal Teknik Informatika*, vol. 15, no. 1, pp. 34–43, Jun. 2022, doi: <https://doi.org/10.15408/jti.v15i1.24847>.
  - [10] N. Agustina, E. Sutinah, and M. Martini, "Kolaborasi Metode Naïve Bayes dan MPE dalam Pengambilan Keputusan Pemilihan Supplier Ban Motor," *Jurnal Media Informatika Budidarma*, vol. 8, no. 2, pp. 1097–1108, 2024, doi: <https://doi.org/10.30865/mib.v8i2.7538>.
  - [11] Y. D. Pramudita, S. S. Putro, and N. Makhmud, "Klasifikasi Berita Olahraga Menggunakan Metode Naïve Bayes dengan Enhanced Confix Stripping Stemmer," *JTIK : Jurnal Teknologi Informasi Dan Ilmu Komputer*, vol. 5, no. 3, pp. 269–276, 2018, doi: <https://doi.org/10.25126/jtiik.201853810>.
  - [12] R. R. Sani, Y. A. Pratiwi, S. Winarno, E. D. Udayanti, and F. Alzami, "Analisis Perbandingan Algoritma Naïve Bayes Classifier dan Support Vector Machine untuk Klasifikasi Berita Hoax pada Berita Online Indonesia," *JMASIF : Jurnal Masyarakat Informatika*, vol. 13, no. 2, pp. 85–98, 2022, doi: <https://doi.org/10.14710/jmasif.13.2.47983>.
  - [13] E. Y. Hidayat and M. A. Rizqi, "Klasifikasi Dokumen Berita Menggunakan Algoritma Enhanced Confix Stripping Stemmer dan Naïve Bayes Classifier," *Jurnal TEKNOLOGI : Nasional Teknologi dan Sistem Informasi*, vol. 6, no. 2, pp. 90–99, 2020, doi: <https://doi.org/10.25077/TEKNOSI.v6i2.2020.90-99>.
  - [14] B. S. Prakoso, D. Rosiyadi, H. S. Utama, and D. Aridarma, "Klasifikasi Berita Menggunakan Algoritma Naïve Bayes Classifier dengan Seleksi Fitur dan Boosting," *Jurnal RESTI : Rekayasa Sistem dan Teknologi Informasi*, vol. 3, no. 2, pp. 227–232, 2019, doi: <https://doi.org/10.29207/resti.v3i2.1042>.
  - [15] N. Asmiati, "Penerapan Algoritma Naïve Bayes untuk Mengklasifikasi Pengaruh Negatif Game Online bagi Remaja Milenial," *JTIM : Jurnal Teknologi Informasi Dan Multimedia*, vol. 2, no. 3, pp. 141–149, 2020, doi: <http://dx.doi.org/10.35746/jtim.v2i3.102>.
  - [16] E. Lindrawati, E. Utami, and A. Yaqin, "Comparison of Modified Nazief & Adriani and Modified Enhanced Confix Stripping Algorithms for Madurese Language Stemming," *INTENSIF : Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi*, vol. 7, no. 2, pp. 276–289, Aug. 2023, doi: <https://doi.org/10.29407/intensif.v7i2.20103>.
  - [17] J. M. Luna-Romera, M. Martínez-Ballesteros, J. García-Gutiérrez, and J. C. Riquelme, "External Clustering Validity Index Based on Chi-Squared Statistical Test," *Information Sciences*, vol. 487, pp. 1–17, 2019, doi: <https://doi.org/10.1016/j.ins.2019.02.046>.
  - [18] D. Chen Sami, A. Sugiharto, and F. Jie, "Chi-Square Feature Selection for Improving Sentiment Analysis of News Data Privacy Threats," *JATIT : Journal of Theoretical and Applied Information Technology*, vol. 102, no. 18, 2024, [Online]. Available: [www.jatit.org](http://www.jatit.org)
  - [19] N. W. Wardani and P. G. S. C. Nugraha, "Stemming Teks Bahasa Bali dengan Algoritma Enhanced Confix Stripping," *LINSE : International Journal of Natural Science and*

*Engineering*, vol. 4, no. 3, pp. 103–113, Dec. 2020, doi:

<https://doi.org/10.23887/ijnse.v4i3.30309>.

- [20] N. Yusliani, S. A. Q. Aruda, M. D. Marieska, D. M. Saputra, and A. Abdiansah, "The Effect of Chi-Square Feature Selection on Question Classification Using Multinomial Naïve Bayes," *Sinkron : Jurnal dan Penelitian Teknik Informatika*, vol. 7, no. 4, pp. 2430–2436, Oct. 2022, doi: <https://doi.org/10.33395/sinkron.v7i4.11788>.
- [21] A. Falasari and M. A. Muslim, "Optimize Naïve Bayes Classifier Using Chi-Square and Term Frequency Inverse Document Frequency for Amazon Review Sentiment Analysis," *JOSCEX : Journal of Soft Computing Exploration*, vol. 3, no. 1, pp. 31–36, Mar. 2022, doi: <https://doi.org/10.52465/josce.v3i1.68>.
- [22] D. A. Wulandari, F. A. Bachtar, and I. Indriati, "Aspect-Based Sentiment Analysis on Shopee Application Reviews Using Support Vector Machine," *Lontar Komputer : Jurnal Ilmiah Teknologi Informatika*, vol. 15, no. 02, p. 99, Jan. 2025, doi: <https://doi.org/10.24843/LKJITL.2024.v15i02.p03>.
- [23] R. Fajriah and D. Kurniawan, "Optimalisasi Model Klasifikasi Naïve Bayes dan Support Vector Machine dengan FastText dan Chi-Square," *Faktor Exacta*, vol. 17, no. 4, pp. 1979–276, 2024, doi: 10.30998/faktorexacta.v17i4.24751.
- [24] F. J. Damanik and D. B. Setyohadi, "Analysis of Public Sentiment About COVID-19 in Indonesia on Twitter Using Multinomial Naïve Bayes and Support Vector Machine," in *IOP Conference Series: Earth and Environmental Science*, IOP Publishing Ltd, Apr. 2021, doi: 10.1088/1755-1315/704/1/012027.
- [25] N. E. Febriyanti, M. A. Hariyadi, and C. Crysdiand, "Hoax Detection News Using Naïve Bayes and Support Vector Machine Algorithm," *IJADIS : International Journal of Advances in Data and Information Systems*, vol. 4, no. 2, pp. 191–200, Oct. 2023, doi: <https://doi.org/10.25008/ijadis.v4i2.1306>.

*Lampiran 4 Source Code*

```

import requests

from bs4 import BeautifulSoup

import pandas as pd

base_url = "https://ppid.walisongo.ac.id/category/blog/"

url = base_url

data = []

# Kata kunci kategori

kata_kunci_pengumuman = ["pengumuman", "pendaftaran",
"seleksi", "informasi"]

kata_kunci_acara = ["acara", "seminar", "workshop",
"lomba", "kegiatan"]

kata_kunci_pendidikan = ["pendidikan", "kurikulum",
"mahasiswa", "dosen"]

kata_kunci_kampus = ["kampus", "uin walisongo",
"universitas", "prestasi"]

# Looping scraping (akan dipotong ke 880 data)

while len(data) < 880:

    response = requests.get(url)

    soup = BeautifulSoup(response.text, "html.parser")

    articles = soup.find_all("article")

    for article in articles:

        if len(data) >= 880:

            break

        title_tag = article.find("h4", class_="entry-
title")

        title = title_tag.text.strip() if title_tag else
"Title not found"

```

```

paragraphs = [p.text.strip() for p in
content_tag.find_all("p")] if content_tag else []

    content = " ".join(paragraphs)

    if any(k in title.lower() or k in content.lower()
for k in kata_kunci_pengumuman):

        kategori = "Pengumuman"

    elif any(k in title.lower() or k in
content.lower() for k in kata_kunci_acara):

        kategori = "Acara"

    elif any(k in title.lower() or k in
content.lower() for k in kata_kunci_pendidikan):

        kategori = "Pendidikan"

    elif any(k in title.lower() or k in
content.lower() for k in kata_kunci_kampus):

        kategori = "Kampus"

    else:

        kategori = "Lainnya"

    data.append({"Judul": title, "Isi": content,
"Kategori": kategori})

    next_page = soup.find("a", class_="next page-
numbers")

    if next_page:

        url = next_page["href"]

    else:

        break

```

```
# Simpan data ke CSV

df = pd.DataFrame(data)

df.to_csv("berita_uin_walisongo_kategori.csv",
index=False)

print("✅ Data berhasil disimpan sebanyak", len(df),
"berita.")
```

### *Source Code Scraping 880 Data Berita*

```
import pandas as pd

import re

import nltk

from nltk.corpus import stopwords

from Sastrawi.Stemmer.StemmerFactory import
StemmerFactory

nltk.download('stopwords')

stopwords_indo = set(stopwords.words('indonesian'))

stemmer = StemmerFactory().create_stemmer()

df = pd.read_csv("berita_uin_walisongo_kategori.csv")

df['Judul'] = df['Judul'].fillna('').str.lower()

df['Isi'] = df['Isi'].fillna('').str.lower()

df['Tokenized_Judul'] = df['Judul'].apply(lambda x:
re.findall(r'\b[\w]{3,}\b', x))

df['Tokenized_Isi'] = df['Isi'].apply(lambda x:
re.findall(r'\b[\w]{3,}\b', x))

df['Filtered_Judul'] = df['Tokenized_Judul'].apply(lambda
x: [w for w in x if w not in stopwords_indo])

df['Filtered_Isi'] = df['Tokenized_Isi'].apply(lambda x:
[w for w in x if w not in stopwords_indo])
```

```

df['Stemmed_Judul'] = df['Filtered_Judul'].apply(lambda
x: [stemmer.stem(w) for w in x])

df['Stemmed_Isi'] = df['Filtered_Isi'].apply(lambda x:
[stemmer.stem(w) for w in x])

df.to_csv("preprocessed_berita.csv", index=False)

print("✅ Preprocessing selesai.")

```

### *Source Code Preprocessing (Case Folding, Tokenizing, Stemming)*

```

df = pd.read_csv("preprocessed_berita.csv")

df['TF_Judul'] =
df['Stemmed_Judul'].apply(eval).apply(lambda x: {term:
x.count(term) for term in set(x)})

df['TF_Isi'] = df['Stemmed_Isi'].apply(eval).apply(lambda
x: {term: x.count(term) for term in set(x)})

df['Teks'] = df['Judul'] + ' ' + df['Isi']

df.to_csv("tf_berita.csv", index=False)

print("✅ Term Frequency dihitung.")

```

### *Source Code Term Frequency*

```

import pandas as pd

import numpy as np

df = pd.read_csv("tf_berita.csv")

kategori_map = {"Acara": 0, "Pengumuman": 1, "Kampus": 2,
"Pendidikan": 3}

df['Kategori_Numerik'] = df['Kategori'].map(kategori_map)

# ✅ Gunakan hasil stemming sebagai corpus

corpus = list(set(

    word

    for doc in df['Stemmed_Judul'].apply(eval).tolist() +
df['Stemmed_Isi'].apply(eval).tolist()

    for word in doc

))

# Bangun struktur frekuensi kata per kategori

freq = {k: {w: 0 for w in corpus} for k in
df['Kategori_Numerik'].unique()}

for i, row in df.iterrows():

    if not pd.isnull(row['Kategori_Numerik']):

        for word in eval(row['TF_Judul']).keys():

            if word in freq[row['Kategori_Numerik']]:

                freq[row['Kategori_Numerik']][word] += 1

        for word in eval(row['TF_Isi']).keys():

            if word in freq[row['Kategori_Numerik']]:

                freq[row['Kategori_Numerik']][word] += 1

```

```

# Chi-Squared computation

epsilon = 1e-10

total_words_cat = {k: sum(freq[k].values()) for k in
freq}

expected_freq = {
    cat: {
        word: total_words_cat[cat] *
sum(freq[other][word] for other in freq if other != cat)
/ sum(total_words_cat.values())

        for word in freq[cat]
    }
    for cat in freq
}

chi_squared_scores = {
    cat: {
        word: ((freq[cat][word] -
expected_freq[cat][word]) ** 2) /
(expected_freq[cat][word] + epsilon)

        for word in freq[cat]
    }
    for cat in freq
}

# Pilih fitur terbaik

num_features = 1000

selected_features = sorted(set(
    word for cat in chi_squared_scores

```

```

        for word, _ in
sorted(chi_squared_scores[cat].items(), key=lambda x:
x[1], reverse=True)[:num_features]

    ))

df.to_csv("tf_chi_selected.csv", index=False)

pd.Series(selected_features).to_csv("selected_features.csv",
index=False)

print("✅ Chi-Squared selesai dan aman dari KeyError.
Total fitur:", len(selected_features))

```

### *Source Code Chi-Squared Feature Selection*

```

import pandas as pd

import joblib

from sklearn.naive_bayes import MultinomialNB

from sklearn.impute import SimpleImputer

# Load data dan fitur

df = pd.read_csv("tf_chi_selected.csv")

selected_features = pd.read_csv("selected_features.csv",
header=None)[0].tolist()

kategori_map = {"Acara": 0, "Pengumuman": 1, "Kampus": 2,
"Pendidikan": 3}

df['Kategori_Numerik'] = df['Kategori'].map(kategori_map)

# Ambil 704 data pertama dan pastikan tidak ada NaN

df_train =
df.iloc[:704].dropna(subset=["Kategori_Numerik"])

# Bangun fitur biner

def build_feature_vector(row):

    words = set(eval(row['Stemmed_Judul']) +
eval(row['Stemmed_Isi']))

```

```

        return [1 if f in words else 0 for f in
selected_features]

X_train = df_train.apply(build_feature_vector, axis=1,
result_type='expand')

y_train = df_train['Kategori_Numerik'].astype(int) #
pastikan integer

# Imputasi

imputer = SimpleImputer(strategy='constant',
fill_value=0)

X_train_imputed = imputer.fit_transform(X_train)

# Train model

model = MultinomialNB()

model.fit(X_train_imputed, y_train)

# Save model

joblib.dump(model, "nb_model.pkl")

joblib.dump(imputer, "imputer.pkl")

joblib.dump(selected_features, "selected_features.pkl")

print("✅ Model Naive Bayes berhasil dilatih (704
data).")

```

### *Source Code Training Model*

```

import pandas as pd

import joblib

from sklearn.metrics import accuracy_score

# Load model dan data

model = joblib.load("nb_model.pkl")

imputer = joblib.load("imputer.pkl")

selected_features = joblib.load("selected_features.pkl")


# Baca 176 data testing (baris ke-704 sampai 879)

df_test =
pd.read_csv("tf_chi_selected.csv").iloc[704:880]

df_test['Kategori_Numerik'] = df_test['Kategori'].map({

    "Acara": 0,

    "Pengumuman": 1,

    "Kampus": 2,

    "Pendidikan": 3

})

# Fungsi membuat vektor fitur

def build_feature_vector(row):

    words = set(eval(row['Stemmed_Judul']) +
eval(row['Stemmed_Isi']))

    return [1 if f in words else 0 for f in
selected_features]

# Bangun X_test dan y_test

X_test = df_test.apply(build_feature_vector, axis=1,
result_type='expand')

y_test = df_test['Kategori_Numerik'].astype(int)

X_test_imputed = imputer.transform(X_test)

```

```
# Prediksi model

y_pred = model.predict(X_test_imputed)

# Evaluasi akurasi

acc = accuracy_score(y_test, y_pred)

print(f"🎯 Akurasi asli pada 176 data uji: {acc *
100:.2f}%")
```

### *Source Code Testing Model*

```
import pandas as pd

import re

import joblib

from sklearn.metrics import accuracy_score,
classification_report

from nltk.corpus import stopwords

from Sastrawi.Stemmer.StemmerFactory import
StemmerFactory

import nltk

# Inisialisasi

nltk.download('stopwords')

stopwords_indo = set(stopwords.words('indonesian'))

stemmer = StemmerFactory().create_stemmer()

# Load model dan fitur

model = joblib.load("nb_model.pkl")

imputer = joblib.load("imputer.pkl")

selected_features = joblib.load("selected_features.pkl")
```

```
# Data uji (12 artikel baru)

data_baru = [

    {"Judul": "Pengumuman Beasiswa KIP-K untuk Mahasiswa Baru Tahun 2024", "Isi": "UIN Walisongo memberikan pengumuman terkait program beasiswa KIP-K kepada mahasiswa"},

    {"Judul": "Fakultas Sains dan Teknologi Adakan Pelatihan Kewirausahaan Bagi Mahasiswa", "Isi": "Fakultas Sains dan Teknologi akan menyelenggarakan pelatihan kewirausahaan bagi mahasiswa"},

    {"Judul": "Civitas Akademik Melakukan Kerja Bakti Untuk Menjaga Kebersihan Lingkungan Kampus", "Isi": "Civitas akademik mengadakan kegiatan membersihkan lingkungan kampus. Kegiatan ini untuk menjaga kenyamanan"},

    {"Judul": "Fakultas Dakwah dan Komunikasi adakan Pelatihan Desain Grafis untuk Mahasiswa", "Isi": "Kompetisi ini akan memberikan kesempatan bagi mahasiswa untuk menunjukkan kemampuan desain grafis mereka"},

    {"Judul": "UIN Walisongo Gelar Diskusi Etika dalam Kehidupan Sehari-hari", "Isi": "UIN Walisongo akan mengadakan diskusi tentang etika dan moralitas dalam kehidupan sehari-hari"},

    {"Judul": "UIN Walisongo Gelar Webinar Internasional Tentang Pendidikan Islam di Era Digital", "Isi": "Universitas Islam Negeri (UIN) Walisongo Semarang menggelar webinar internasional tentang pendidikan islam di era digital"},

    {"Judul": "Kantin FST UIN Walisongo Longsor Akibat Hujan Deras", "Isi": "Akibat hujan deras selama sehari penuh kantin FST longsor dan membutuhkan perbaikan"},

    {"Judul": "PTIPD Lakukan Workshop Pelatihan Jurnalistik", "Isi": "Dalam meningkatkan kualitas kepenulisan mahasiswa, PTIPD mengadakan workshop jurnalistik"},


```

```

        {"Judul": "Mahasiswa Teknologi Informasi UIN
Walisongo Raih Prestasi di Tingkat Internasional", "Isi":
"Tiga mahasiswa Teknologi Informasi UIN Walisongo,
Semarang, raih penghargaan di Malaysia"},

        {"Judul": "Perubahan Tanggal Pelaksanaan Wisuda
Periode Juni 2024", "Isi": "Tanggal Pelaksanaan Wisuda
Periode Juni 2024 dirubah dan diumumkan kepada
mahasiswa"},

        {"Judul": "UIN Walisongo Buka Jalur Prestasi untuk 34
Jurusan, Ini Jadwal dan Syaratnya", "Isi": "Jalur
Prestasi UIN Walisongo kembali dibuka untuk 34 jurusan
sesuai syarat dan ketentuan"},

        {"Judul": "Lantik Pejabat Struktural, Ini Pesan
Rektor UIN Walisongo", "Isi": "Dalam rangka pelantikan
pejabat struktural, rektor berpesan tentang peningkatan
kualitas"},

    ]

# Ground truth (disesuaikan manual berdasarkan pemahaman
masing-masing artikel)

gt_labels = [

    3, # Announcement

    0, # Event

    2, # Campus

    0, # Event

    0, # Event

    0, # Event

    2, # Campus

    3, # Education (jika dianggap pelatihan)

    2, # Campus

    3, # Announcement

    3, # Announcement

    0, # Event

]

```

```

# Mapping label
kategori_map = {0: "Event", 1: "Campus", 2: "Education",
3: "Announcement"}

# Preprocessing fungsi
def preprocess(text):
    text = text.lower()

    tokens = re.findall(r'\b[\w]{3,}\b', text)

    filtered = [w for w in tokens if w not in
stopwords_indo]

    return [stemmer.stem(w) for w in filtered]

df = pd.DataFrame(data_baru)

df['Stemmed_Judul'] = df['Judul'].apply(preprocess)
df['Stemmed_Isi'] = df['Isi'].apply(preprocess)

# Buat fitur
def build_feature_vector(row):
    words = set(row['Stemmed_Judul'] +
row['Stemmed_Isi'])

    return [1 if f in words else 0 for f in
selected_features]

X_eval = df.apply(build_feature_vector, axis=1,
result_type='expand')

X_eval_imputed = imputer.transform(X_eval)

# Prediksi model
y_pred = model.predict(X_eval_imputed)

```

```

# Evaluasi

from IPython.display import display

result_df = pd.DataFrame({

    "No": list(range(1, 13)),

    "News Title": df["Judul"],

    "News Content": df["Isi"],

    "Category (System)": [kategori_map[p] for p in
y_pred],

    "Category (User)": [kategori_map[t] for t in
gt_labels],

})

accuracy = accuracy_score(gt_labels, y_pred)

print("📊 Tabel Evaluasi (12 Artikel Baru - Hasil Asli):")

display(result_df[["No", "News Title", "News Content",
"Category (System)", "Category (User)"]])

print(f"\n🎯 Akurasi asli terhadap 12 artikel baru:
{accuracy * 100:.2f}%")

print("\n📈 Metrik Evaluasi Asli (Precision, Recall, F1-
Score, Support):")

print(classification_report(gt_labels, y_pred,
target_names=list(kategori_map.values()), digits=2))

```

### *Source Code Evaluation (12 Artikel Baru)*

Lampiran 5 Dataset

berita\_uin\_walisongo\_kategori.csv X

| Judul                                                                       | Isi                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | Kategori   |
|-----------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|
| PENGUMUMAN PERPANJANGAN WAKTU PEMBAYARAN UKT JALUR SNBP DAN SPAN-PTKN 2025  | PENGUMUMAN PERPANJANGAN WAKTU PEMBAYARAN UKT JALUR SNBP DAN SPAN-PTKN 2025 SELENGKAPNYA UNDIH DISINI                                                                                                                                                                                                                                                                                                                                                                                    | Pengumuman |
| PERUBAHAN JADWAL PENGUMUMAN & PEMBAYARAN UKT JALUR SNBP DAN SPAN-PTKN 2025  | PERUBAHAN JADWAL PENGUMUMAN DAN PEMBAYARAN UKT JALUR SNBP DAN SPAN-PTKN 2025 PENGUMUMAN UNDIH DISINI                                                                                                                                                                                                                                                                                                                                                                                    | Pengumuman |
| WPRC Ajak Perempuan Bersuara, Perempuan Merdeka                             | UIN Walisongo Online, Semarang- Hari Kartini menjadi momentum penting untuk peningkatan kebangkitan semangat perempuan dalam bersuara maupun berkarya. Walisongo Public Relations Community (WPRC) menggelar peringatan Bersuara, perempuan merdeka dalam seminar Suara Perempuan dengan tema Ekspresi dalam Kata dan Karya, di Laboratorium Dakwah (Jadid) Fakultas Dakwah dan Komunikasi (FDK), pada Kamis (24/4/2025) dan diikuti 90 peserta mahasiswa Komunikasi dan Penyiaran FDK. | Acara      |
| UIN Walisongo Resmikan Fasilitas Kesehatan, Siap Cetak Dokter Berbakat      | UIN Walisongo Online, Semarang – Universitas Islam Negeri (UIN) Walisongo Semarang resmi meluncurkan Fakultas Kedokteran pada Rabu (23/4/2025). Prosesi peluncuran berlangsung di Ruang Teater Gedung KH. Soleh Darat, Kampus 3 UIN Walisongo, diakhiri oleh para pimpinan perguruan tinggi undangan, perwakilan rumah sakit, serta pejabat internal kampus. Ketua panitia, Dr. Ahmad Ismail, M.Ag., dalam laporannya menyampaikan ungkapan syukur atas pemberian fasilitas.            | Kampus     |
| Akhir Program Desa Binaan, GenBil UIN Walisongo Gelar Closing Ceremony      | UIN Walisongo Online, Kendal- Generasi Baru Indonesia (GenBil) menggelar closing ceremony dengan penuh semangat pada Senin (21/04/2025) di Balai Desa Gedas, Kec. Singorojo, Kab. Kendal. Acara ini dimulai pada pukul 09.00 WIB dan dihadiri oleh sekitar 150 peserta dengan mengundang Wakil Rektor 3 UIN Walisongo, Perangkul Desa, Perwakilan Guru dari Sekolah Dasar, Guru Madrasah, Perwakilan Warga, serta Mahasiswa GenBil. Program.                                            | Acara      |
| UIN Walisongo Terima Hibah Tanah 5,4 Hektar untuk Pembangunan Kampus Terasu | UIN Walisongo Online, Semarang – Universitas Islam Negeri (UIN) Walisongo mendapatkan hibah tanah seluas 5,4 hektar dari Kementerian Agama RI. Penyerahan aset ini berlangsung pada Senin (10/2/2025) di Ruang Theater lantai 4 Gedung KH Soleh Darat UIN Walisongo Semarang. Acara serah terima tersebut dipimpin oleh Dr. Saiful Mujib, M.A., Kepala Kantor Wilayah Kementerian Agama Provinsi Jawa Tengah, yang menyerahkan secara simbolis.                                         | Acara      |
| Dari Kursi Roda ke Panggung Wisuda: Rizal                                   | UIN Walisongo Online, Semarang – Pada prosesi wisuda UIN Walisongo Semarang periode Februari 2025, Muhammad Rizal Kan Nabhan, mahasiswa prodi                                                                                                                                                                                                                                                                                                                                           | Acara      |

Variables Terminal 1:19PM Python 3

Dataset Berita Akademik PPID UIN Walisongo



## **DAFTAR RIWAYAT HIDUP**

### **A. Identitas Diri**

1. Nama Lengkap : M. Naufal Muhadzib Al-Faruq
2. Tempat & Tanggal Lahir : Pati, 08 September 2004
3. Alamat Rumah : Desa Cengkalsewu RT 004 RW 004, Kecamatan, Sukolilo, Kabupaten Pati
4. HP : 087722476038
5. E-mail : naufalfaruq082@gmail.com

### **B. Riwayat Pendidikan**

1. Pendidikan Formal:
  - a. MI Iqbalul Athfal Cengkalsewu
  - b. MTs Miftahut Thullab Cengkalsewu
  - c. MA NU Tasywiquth Thullab Salafiyah (TBS) Kudus
2. Pendidikan Non-Formal:
  - a. Asrama Pendidikan Islam (API) Darussalam Kudus
  - b. Pondok Pesantren Al-Firdaus Semarang

### **C. Prestasi Akademik**

1. Juara 1 Desain Web Rafatech 2022 UIN Raden Fatah Palembang
2. Silver Medal World Invention Competition and Exhibition (WICE) 2022 SEGi University
3. Silver Medal Pitch Battle Santri Fest 2022 Kementrian Ketenagakerjaan Republik Indonesia (KEMNAKER)
4. Juara 1 Lomba Essay Nasional Pekan Santri Literasi 2022 Universitas Nahdlatul Ulama Sunan Giri

5. Juara 3 UI/UX Design Switfest 2022 UIN Walisongo Semarang
6. Juara 2 Design Prototype Malisic Competition III UIN Maulana Malik Ibrahim Malang
7. Gold Medal International Science and Invention Fair (ISIF) 2022 Ganesha University of Education
8. Gold Medal World Young Invention Exhibition (WYIE) 2023 MINDS (Malaysian Invention and Design Society)
9. Gold Medal and Special Award Tokyo Best Innovation di ajang Japan Design, Idea and Expo (JDIE) 2024
10. Core Team UI/UX Design Google Developer Student Clubs (GDSC) UIN Walisongo Semarang

#### **D. Karya Ilmiah**

1. PERANCANGAN UI/UX SEMARANG VIRTUAL TOURISM DENGAN FIGMA
2. AUGMENTED REALITY FOR UNIVERSITAS ISLAM NEGERI WALISONGO'S PROFILE USING AGILE METHODOLOGY
3. PERANCANGAN UI/UX APLIKASI MINATANI SISTEM INFORMASI AGRICULTURE TECHNOLOGY MENGGUNAKAN METODE DESIGN THINKING
4. IMPLEMENTATION OF ENHANCED CONFIX STRIPPING STEMING AND CHI-SQUARED FEATURE SELECTION ON CLASSIFICATION UIN WALISONGO WEBSITE WITH NAÏVE BAYES CLASSIFIER