**PERBANDINGAN SELEKSI FITUR UNTUK KLASIFIKASI SENTIMEN BERITA PARIWISATA MENGGUNAKAN METODE *NAÏVE BAYES***

****

**Disusun oleh:**

**MOH. SA’ID**

**19.04.111.00045**

**Dosen Pembimbing 1 : Husni, S.Kom., M.T.**

**Dosen Pembimbing 2 : Prof. Dr. Arif Muntasa, S.Si., M.T.**

**PROGRAM STUDI TEKNIK INFORMATIKA**

**JURUSAN TEKNIK INFORMATIKA**

**FAKULTAS TEKNIK**

**UNIVERSITAS TRUNOJOYO MADURA**

**2023**

# **ABSTRAK**

Pariwisata telah menjadi salah satu penyumbang devisa utama bagi Indonesia yang harus dipertahankan dan dikembangkan. Pulau Madura merupakan salah satu wilayah yang memiliki berbagai objek wisata beraneka ragam. Destinasi tempat wisata di Pulau Madura dapat diakses informasinya melalui media online. Informasi yang bersentimen negatif mengenai objek wisata menyebabkan menurunnya tingkat kunjungan wisatawan sehingga memerlukan anilisis tentang sentimen pada berita informasi tersebut. Analisis sentimen pada informasi berita yang ada pada media online menggunakan perhitungan dari sebuah metode. Penelitian ini menggunakan perhitungan dari metode Naïve Bayes dengan membandingkan empat metode seleksi fitur yaitu *information gain, chi square, sequential forward selection*, dan *sequential backward elimination*. Metode seleksi fitur digunakan dengan tujuan untuk mengurangi vektor fitur yang tidak terlalu berpengaruh untuk meningkatkan performa dari metode klasifikasi. Data yang digunakan pada penelitian ini sebanyak 200 berita pariwisata madura yang diambil secara manual.

**Kata Kunci**: Pariwisata Madura, Analisis Sentimen, Seleksi Fitur*, Naïve Bayes*

# **DAFTAR ISI**

[**ABSTRAK** ii](#_Toc137636322)

[**DAFTAR ISI** iii](#_Toc137636323)

[**DAFTAR TABEL** v](#_Toc137636324)

[**DAFTAR GAMBAR** vi](#_Toc137636325)

[**BAB I PENDAHULUAN** 1](#_Toc137636326)

[**1.1.** **Latar Belakang** 1](#_Toc137636327)

[**1.2.** **Rumusan Masalah** 5](#_Toc137636328)

[**1.2.1.** **Permasalahan** 5](#_Toc137636329)

[**1.2.2.** **Metode Usulan** 5](#_Toc137636330)

[**1.2.3.** **Pertanyaan Penelitian** 5](#_Toc137636331)

[**1.3.** **Tujuan dan Manfaat Penelitian** 6](#_Toc137636333)

[**1.3.1.** **Tujuan** 6](#_Toc137636334)

[**1.3.2.** **Manfaat** 6](#_Toc137636340)

[**1.4.** **Batasan Penelitian** 6](#_Toc137636341)

[**1.5.** **Sistematika Tugas Akhir** 7](#_Toc137636347)

[**BAB II KAJIAN PUSTAKA** 8](#_Toc137636348)

[**2.1.** **Landasan Teori** 8](#_Toc137636349)

[**2.1.1.** ***Data Mining*** 8](#_Toc137636350)

[**2.1.2.** ***Text Mining*** 9](#_Toc137636355)

[**2.1.3.** **Analisis Sentimen** 10](#_Toc137636356)

[**2.1.4.** **Portal Berita *Online*** 11](#_Toc137636357)

[**2.1.5.** ***Synthetic Minority Oversampling Technique* (SMOTE)** 11](#_Toc137636365)

[**2.1.6.** **Seleksi Fitur** 12](#_Toc137636366)

[**2.1.7.** **Klasifikasi *Naïve Bayes*** 15](#_Toc137636376)

[**2.1.8.** **Penelitian Terkait** 16](#_Toc137636377)

[**BAB III METODE USULAN** 18](#_Toc137636378)

[**3.1.** **Arsitektur Sistem** 18](#_Toc137636379)

[**3.2.** **Dataset Penelitian** 19](#_Toc137636380)

[**3.3.** **Pelabelan Data** 20](#_Toc137636381)

[**3.4.** **Metode** 20](#_Toc137636382)

[**3.4.1.** ***Text Preprocessing*** 20](#_Toc137636383)

[**3.4.2.** ***Extraction TF*** 22](#_Toc137636392)

[**3.4.3.** **Metode Information Gain** 25](#_Toc137636393)

[**3.4.4.** **Metode Chi Square** 28](#_Toc137636394)

[**3.4.5.** **Metode Sequential Forward Selection (SFS)** 32](#_Toc137636406)

[**3.4.6.** **Metode Sequential Backward Elimination (SBE)** 35](#_Toc137636407)

[**3.4.7.** **SMOTE** 38](#_Toc137636408)

[**3.4.8.** **Metode Naïve Bayes** 41](#_Toc137636409)

[**3.8.** **Skenario Pengujian** 44](#_Toc137636413)

[**REFERENSI** 45](#_Toc137636414)

# **DAFTAR TABEL**

[Tabel 2. 1 Penelitian Terkait 16](#_Toc137319121)

[Tabel 3. 1 Data Judul Berita 20](#_Toc137319184)

[Tabel 3. 2 Text Preprocessing 21](#_Toc137319185)

[Tabel 3. 3 Data yang Sudah Melalui Tahapan Text Preprocessing 22](#_Toc137319186)

[Tabel 3. 4 Data yang diekstrak dengan pembobotan TF 23](#_Toc137319187)

[Tabel 3. 5 Contoh vektor fitur data dengan pembobotan TF 26](#_Toc137319188)

[Tabel 3. 6 Total Sentimen dari Judul Berita 26](#_Toc137319189)

[Tabel 3. 7 Hasil Information Gain pada tiap fitur 27](#_Toc137319190)

[Tabel 3. 8 Tabel Kontingensi 28](#_Toc137319191)

[Tabel 3. 9 Jumlah variable A, B, C dan D pada tiap fitur 30](#_Toc137319192)

[Tabel 3. 10 Nilai chi pada tiap fitur 31](#_Toc137319193)

[Tabel 3. 11 Tabel chi square 32](#_Toc137319194)

[Tabel 3. 12 Hasil akurasi pada penggunaan satu fitur yang terurut dari tertinggi 34](#_Toc137319195)

[Tabel 3. 13 Hasil akurasi terbaik pada tiap penambahan fitur 35](#_Toc137319196)

[Tabel 3. 14 Hasil akurasi pada pengurangan satu fitur yang terurut dari tertinggi 37](#_Toc137319197)

[Tabel 3. 15 Hasil akurasi terbaik pada tiap pengurangan fitur 38](#_Toc137319198)

[Tabel 3. 16 Data Minoritas 38](#_Toc137319199)

[Tabel 3. 17 Hasil Pembentukan Data Sintetis 40](#_Toc137319200)

[Tabel 3. 18 Data Sintetis yang telah terbentuk 41](#_Toc137319201)

[Tabel 3. 19 Nilai prior tiap term terhadap kelas 43](#_Toc137319202)

# **DAFTAR GAMBAR**

[Gambar 2. 1 Arsitektur Pemrosesan Data Menggunakan CRISP-DM 8](file:///D:\Said\SKRIPSI%20SAID.docx#_Toc137319122)

[Gambar 3. 1 Arsitektur Sitem 18](file:///D:\Said\SKRIPSI%20SAID.docx#_Toc137319123)

[Gambar 3. 2 Flowchart seleksi fitur Information Gain 25](file:///D:\Said\SKRIPSI%20SAID.docx#_Toc137319124)

[Gambar 3. 3 Flowchart seleksi fitur Chi Square 29](file:///D:\Said\SKRIPSI%20SAID.docx#_Toc137319125)

[Gambar 3. 4 Flowchart seleksi fitur Sequential Forward Selection (SFS) 33](file:///D:\Said\SKRIPSI%20SAID.docx#_Toc137319126)

[Gambar 3. 5 Flowchart seleksi fitur Sequential Backward Elimination 36](file:///D:\Said\SKRIPSI%20SAID.docx#_Toc137319127)

[Gambar 3. 6 Flowchart SMOTE 39](file:///D:\Said\SKRIPSI%20SAID.docx#_Toc137319128)

[Gambar 3. 7 Flowchart klasifikasi metode Naïve Bayes 42](file:///D:\Said\SKRIPSI%20SAID.docx#_Toc137319129)

# **BAB I PENDAHULUAN**

## **Latar Belakang**

Sektor pariwisata adalah salah satu sektor yang berkembang dan menjadi salah satu devisa utama bagi Indonesia dan negara lain. Bahkan, sektor pariwisata berpengaruh dalam meningkatkan lapangan pekerjaan dan dapat mengaktifkan sektor produksi lain dalam negara. Beberapa ahli berpendapat perkembangan sektor pariwisata sangat sulit diprediksi dan terlalu komplek serta akan terus berubah dan akan selalu beradaptasi dengan perkembangan zaman. Indonesia sendiri memiliki banyak tempat wisata yang sudah dikelola dan tempat wisata yang belum tergarap [1]. Tempat yang belum tergarap memiliki potensi untuk meningkatkan jumlah wisatawan yang datang ke tempat tersebut jika dikelola dengan baik.

Wisatawan mancanegara memiliki dampak besar terhadap pertumbuhan ekonomi negara di Indonesia, diantaranya adalah inflasi dan nilai tukar rupiah [2]. Jumlah wisatawan mancanegara yang berkunjung pada bulan November tahun 2022 mencatat peningkatan yang sangat signifikan dibandingkan dengan bulan November tahun 2021, yakni sebanyak 336,50%. Dalam bulan November tahun 2022, tercatat ada 657.269 wisatawan mancanegara, sedangkan pada bulan November tahun 2021 hanya tercatat 150.557 wisatawan mancanegara [3]. Analisis mengenai faktor apa saja yang mempengaruhi peningkatan dan penurunan jumlah wisatawan mancanegara perlu dilakukan setiap tahunya.

Salah satu wilayah yang berpotensi untuk membantu dalam peningkatan pertumbuhan ekonomi melalui sektor pariwisata adalah Pulau Madura. Pulau Madura memiliki berbagai objek wisata yang beraneka ragam, meliputi wisata religi, wisata alam, dan wisata budaya. Namun, beberapa tempat wisata masih memiliki keterbatasan fasilitas, sarana dan prasarana dalam pembangunan tempat wisata tersebut [4]. Maka dari itu pengembangan objek wisata di Pulau Madura harus dioptimalkan agar menjadi pondasi utama dalam membangun pertumbuhan ekonomi, khususnya di Pulau Madura.

Salah satu objek wisata yang ada di Pulau Madura adalah Desa Wisata Keris Aeng Tongtong, Kabupaten Sumenep yang memperoleh rekor MURI sebagai desa wisata terbanyak jumlah empu keris di dunia [5]. Kemudian Pulau Gili Iyang yang terletak di Kabupaten Sumenep, memiliki kadar oksigen terbaik di dunia. Hal tersebut dibuktikan dari penelitian yang dilakukan oleh Lembaga Penelitian Antariksa Nasional (LAPAN) pada tahun 2006. Namun, Pulau Gili Iyang masih menempati nomor 2 setelah laut mati di Yordania untuk kadar oksigen terbaik di dunia [6]. Melalui banyaknya penghargaan yang diperoleh dari skala nasional dan internasional, diharapkan pulau madura menjadi salah satu pondasi terbesar dalam meningkatkan pertumbuhan ekonomi di Indonesia melalui sektor pariwisata.

Meskipun Pulau Madura memiliki banyak wisata potensial. Namun, masih ada beberapa wilayah yang belum layak untuk dijadikan tempat destinasi wisata karena beberapa halangan. Salah satunya yaitu Pulau Mandingan yang terletak di Kabupaten Sampang terhalang oleh faktor kebersihan. Tempat masuk wisatawan ke Pulau Mandingan terhalang sampah karena masyarakat setempat tidak tahu cara mengatasi sampah-sampah yang telah menggunung dan tersebar ke beberapa pantai yang ada di Pulau Mandingan [7]. Kemudian destinasi wisata Bukit Kapur Arosbaya yang terletak di Kabupaten Bangkalan yang terhalang oleh faktor keamanan. Ada Insiden pembegalan pengendara motor di salah satu jalur masuk ke Destinasi wisata Bukit Kapur Arosbaya yaitu di Desa Ombul, Kecamatan Arosbaya, Kabupaten Bangkalan pada bulan Februari tahun 2022 [8]. Beberapa faktor halangan yang terjadi pada tempat ataupun sekitar tempat destinasi wisata tersebut akan berdampak dalam penurunan jumlah wisatawan.

Saat ini untuk mengakses informasi mengenai destinasi tempat wisata di Pulau Madura yang ada di Internet sangat mudah didapatkan hanya bermodalkan jaringan internet dan telepon genggam. Penyebaran informasi yang sangat mudah tersebut pada objek tertentu akan sangat berdampak terhadap pemikiran individu atau perorangan maupun kelompok terhadap objek tersebut yaitu pariwisata madura. Maka dari itu, pemilihan media informasi sangatlah penting untuk mencari informasi yang terpercaya dan faktual.

Banyaknya media informasi yang ada di internet seperti, surat kabar online, media sosial, youtube dan lain sebagainya. Surat kabar online merupakan salah satu media yang terpercaya dan aktual, seperti *detiknews.com,* *tribunnews.com*, *kompas.com* dan lain-lain. Penelitian yang dilakukan Vika Syafriyenti yang berjudul “Analisis Isi Akurasi Berita Covid-19 di Media Online Detik.com Edisi 30 Maret – 06 April 2020” menyebutkan bahwa akurasi berita mencapai 81%. Bahkan dalam penelitian tersebut menyebutkan bahwa berita tersebut tergolong akurat [9]. Maka dari hal itu, data penelitian ini menggunakan berita yang ada di media surat kabar online tentang pariwisata madura dari pertengahan tahun 2021 ke tahun 2023.

Data yang diambil secara manual tersebut merupakan text mentah yang perlu diolah lagi menjadi data penelitian. Setelah itu, yang perlu dilakukan terlebih dahulu melabelkan data tersebut menjadi data yang bersentimen negatif atau positif. Cara melabelkan data tersebut ada berbagai cara, bisa dengan cara melabelkan secara otomatis atau melakukan pelabelan secara manual. Dalam penelitian ini menggunakan pelabelan secara manual oleh ahli di bidang sastra indonesia. Karena penulisan dari isi teks berita berdasarkan kaidah-kaidah penulisan jurnalistik [10].

Hasil yang didapatkan dari analisis pelabelan sentimen yang dilakukan dalam penelitian ini menunjukkan adanya hasil ketidakseimbangan antara sentimen negatif dan positif. Masalah tersebut akan mengakibatkan hasil klasifikasi pada kelas minoritas akan sering salah dikarenakan kurangnya informasi mengenai kelas minoritas [11]. Namun dari masalah yang terjadi dapat di evaluasi menggunakan metode *SMOTE*. *Metode Synthetic Minority Oversampling Technique* (*SMOTE*) merupakan salah satu metode yang bertujuan untuk mengatasi ketidak timpangan pelabelan tersebut dengan cara menambahkan data sintetik untuk label yang minoritas. Pada penelitian yang dilakukan oleh Muladi yang berjudul ‘*Predicting high school graduates using Naive Bayes in State University Entrance Selections*’ diperoleh nilai akurasi yang lebih besar dengan metode *SMOTE* yaitu sebesar 95,49% daripada klasifikasi naive bayes yang tidak menggunakan metode SMOTE yaitu nilai akurasi sebesar 90,1 % [12].

Data dengan jumlah atribut yang besar akan mengurangi akurasi klasifikasi [11]. Maka dari itu penelitian ini akan melakukan seleksi fitur pada dataset penelitian. Metode seleksi fitur dibagi menjadi 3 kategori, yaitu*: filter method*, *wrapper method*, dan *embedded method*. Penelitian ini akan mengeksplorasi beberapa metode *feature selection*, seperti *information gain, chi-square, forward selection, dan backward elimination*.

Metode *information gain* atau *mutual information maximum* merupakan metode perhitungan entropi yang dalam penyeleksian fiturnya sangat bergantung kuat hubungan sebuah fitur dengan labelnya dan setiap fitur tidak bergantung sama lain. Metode *feature selection chi-square* merupakan metode penyeleksian atribut atau fitur berdasarkan uji statistik untuk menghitung tingkat dependensi pada suatu kelas atau label. Dengan demikian, kedua metode, yaitu *chi-square* dan *information gain* termasuk dalam kategori seleksi fitur yang menggunakan metode Filter. Kekurangan dari metode filter kurang memperhitungkan interaksi antar fitur dan kurangnya mengenal pola dengan baik selama proses pembelajaran [13].

*Forward selection* adalah metode seleksi fitur dalam kategori metode *Wrapper*. Proses perhitungan dimulai dengan set fitur kosong dan secara bertahap menambahkan satu fitur pada set fitur yang sudah teridentifikasi sebelumnya. Fitur baru akan ditambahkan ke dalam set fitur apabila dianggap penting dan memiliki pengaruh terhadap generalisasi. Sedangkan, metode *backward elimination* adalah metode seleksi fitur yang perhitungannya dimulai dengan seluruh fitur sebagai set awal yang secara iteratif menghapus fitur-fitur yang tidak penting hingga mencapai performa terbaik pada suatu fungsi objektif. Dikarenakan *forward selection* dan *backward elimination* termasuk dalam metode Wrapper, maka dari itu kedua metode tersebut memiliki kelemahan *resiko overfitting* yang lebih tinggi dikarenakan terlalu kompleks dalam perhitungannya dan beberapa fitur mungkin tidak dipertimbangkan untuk evaluasi ketika dihapus pada tahap awal [13].

Dalam tahapan klasifikasi pada penelitian ini yaitu metode Naive Bayes. Metode ini dipilih dikarenakan penelitian sebelumnya yang dilakukan Rian Ardianto yang berjudul ‘*Sentiment Analysis On E-Sports for Education Curriculum Using Naive Bayes and Support Vector Machine’* metode naive bayes dengan SMOTE menjadi metode yang memperoleh nilai akurasi terbaik yaitu sebesar 70,32% daripada metode SVM dengan SMOTE. Dalam penelitian ini, kami menggunakan algoritma Naive Bayes Multinomial untuk mengklasifikasikan data. Pemilihan metode ini didasarkan pada hasil pengujian sebelumnya yang berjudul ‘*Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification’* menunjukkan bahwa metode multinomial memberikan hasil akurasi sebesar 73% lebih baik dibandingkan metode Bernoulli yang memberikan hasil 69% [14].

## **Rumusan Masalah**

Berdasarkan dari uraian dari latar belakang yang telah disebutkan, maka rumusan masalah dapat digambarkan sebagai berikut:

### **Permasalahan**

Berdasarkan latar belakang yang telah diuraikan permasalahan yang bisa didapatkan bahwa sentimen surat kabar berita online terhadap pariwisata madura perlu dianalisa dengan baik. Maka dari itu, dibutuhkan sistem analis sentimen yang dapat mengklasifikasikan sentimen dengan metode yang sesuai.

### **Metode Usulan**

Solusi yang ditawarkan dalam penelitian ini adalah penggunaan teknik *oversampling SMOTE (Synthetic Minority Oversampling Technique*) sebagai penyeimbang data yang tidak seimbang, beberapa metode fitur seleksi dengan metode (*information gain, chi-square, forward selection, backward* *elimination)* sebagai pengurangan fitur yang tidak diperlukan, dan metode *naïve bayes* sebagai metode klasifikasi sentimen terhadap surat kabar berita online pariwisata.

### **Pertanyaan Penelitian**

Dalam penelitian ini ada pertanyaan yang diperoleh dari permasalahan dan metode usulan yang telah disebutkan:

* + - * 1. Bagaimana pengaruh dari 4 metode fitur seleksi (*information gain*, *chi-square*, *forward selection*, dan *backward* *elimination*) dan teknik *oversampling* SMOTE *(Synthetic Minority Oversampling Technique*) dalam meningkatkan performa akurasi dan efesiensi waktu klasifikasi sentimen berita online pariwisata dengan algoritma *naïve bayes classifier*?

## **Tujuan dan Manfaat Penelitian**

Tujuan dan manfaat dari penelitian ini sebagai berikut:

### **Tujuan**

Penelitian ini bertujuan untuk mengetahui pengaruh dari 4 metode fitur seleksi (*information gain*, *chi-square*, *forward selection*, dan *backward* *elimination*) dan teknik oversampling SMOTE *(Synthetic Minority Oversampling Technique*) dalam meningkatkan performa akurasi dan efesiensi waktu klasifikasi sentimen berita online pariwisata dengan algoritma *naïve bayes classifier*.



### **Manfaat**

Penelitian ini berisi beberapa manfaat yang dapat diperoleh, diantaranya sebagai berikut:

* + - * 1. Membantu pemerintah daerah untuk mengetahui sentimen terhadap beberapa pariwisata yang ada di daerahnya bersentimen negatif atau positif di surat kabar online yang ada di Internet.
        2. Dapat digunakan untuk menganalisis seberapa efektif metode-metode yang digunakan efektif dalam pengklasifikasian sentimen.
        3. Sebagai bahan rujukan untuk penelitian selanjutnya.

## **Batasan Penelitian**

Dalam penelitian ini ada batasan-batasan yang diterapkan, diantaranya sebagai berikut:

1. Dataset yang digunakan berjumlah 200 dalam bentuk csv.
2. Dataset yang diambil dari surat kabar online (*Detiknews.com, Kompas.com, Inews.com, Tribunnews.com, liputan6.com, tvonenews.com, dll)*. Dataset yang diambil mengenai sentiment surat kabar online terhadap pariwisata madura secara manual.
3. Dataset yang digunakan memiliki 2 label yaitu label positif dan label negatif.
4. Pelabelan dataset dilakukan dengan cara manual oleh peneliti dan divalidasi oleh ahli.

## **Sistematika Tugas Akhir**

Sistematika penulisan penelitian ini terdiri dari beberapa bab, diantaranya sebagai berikut:

* + 1. Bab I: Pendahuluan

Bab ini berisi latar belakang penelitian dilakukan, rumusan masalah pada penelitian, tujuan dan manfaat penelitian, batasan penelitian dan sistematis untuk mempersiapkan tugas akhir.

* + 1. BAB II: Kajian Pustaka

Bab ini berisi tentang referensi atau landasan teori mengenai materi ataupun metode yang digunakan dalam penelitian ini, diantaranya meliputi rujukan paper yang digunakan terkait data mining, text mining, analisis sentimen, portal berita online, seleksi fitur (*information gain*, *chi-square*, *forward selection*, dan *backward* *elimination),* teknik *oversampling SMOTE (Synthetic Minority Oversampling Technique*) dan *Naive Bayes.*

* + 1. BAB III: Metodologi Penelitian

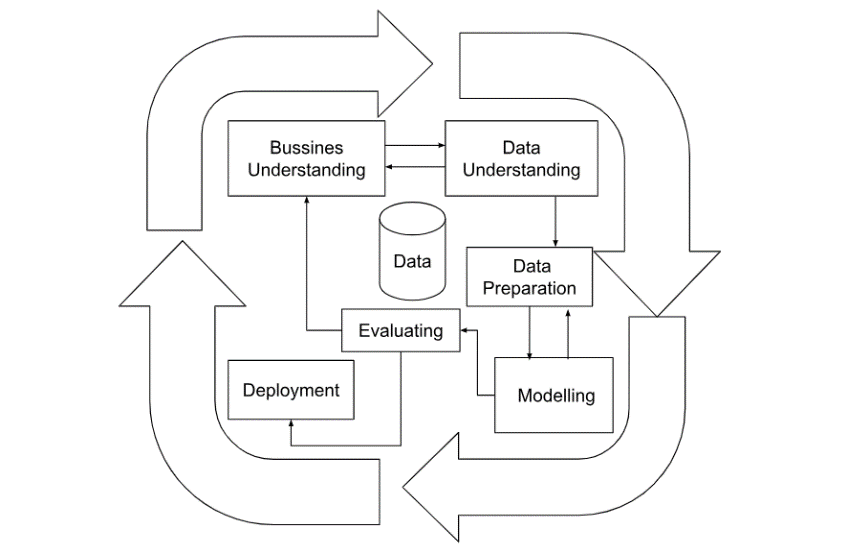
Bab ini menjelaskan proses penelitian serta mengenai data yang digunakan, keseluruhan rancangan sistem dan analisis mengenai penelitian yang dilakukan termasuk diantaranya arsitektur sistem, serta arsitektur pelatihan dan pengujian.

# **BAB II KAJIAN PUSTAKA**

## **Landasan Teori**

### ***Data Mining***

Data mining merupakan teknik pengolahan otomatis yang bertujuan untuk menganalisis suatu data informasi dalam menemukan pola tertentu dalam tempat penyimpanan data yang besar. Secara spesifik data mining menggunakan konsep dan metodologi dari dua bidang pokok, yaitu bidang statistika dan *machine learning*. Dalam bidang statistika, diantaranya proses pengambilan contoh, estimasi, dan pengujian hipotesis. Sementara dari bidang *machine learning* data mining memanfaatkan algoritma pencarian, teknik pemodelan, dan teori pembelajaran dari *AI* (*Artificial intelligence*), pengenalan pola algoritma, dan konsep machine learning lainnya [15].

Data mining memiliki standar pemrosesan data yang bertujuan untuk menghasilkan data yang lebih efisien yakni menggunakan pendekatan *CRISP-DM* (*Cross Industry Standard Process for Data Mining*). Proses dalam pendekatan *CRISP-DM* ditunjukkan pada gambar 2.1 [16].

Gambar 2. 1 Arsitektur Pemrosesan Data Menggunakan CRISP-DM

Berikut penjelasan mengenai tahapan-tahapan yang ada dalam pendekatan pemrosesan data menggunakan *CRISP-DM*, di antaranya:

* + 1. *Business Understanding*, tahapan ini bertujuan untuk memahami konsep dan tujuan dari proyek yang dikerjakan dari sudut pandang bisnis. Kemudian informasi yang dihasilkan diolah menjadi permasalahan data mining untuk mencapai tujuan dari proyek yang telah ditentukan sebelumnya.
    2. *Data Understanding*, tahapan ini untuk pengumpulan dan mengidentifikasi data yang bertujuan untuk mencari masalah yang ada pada data serta mendeteksi subset yang menarik untuk menghasilkan hipotesis mengenai informasi yang tersembunyi pada data tersebut.
    3. *Data Preparation*, tahapan ini bertujuan untuk persiapan untuk menyusun kumpulan data akhir ke dalam tahapan *modelling*. Persiapan yang dikerjakan dalam tahapan ini diantaranya transformasi ataupun pemilihan tabel, baris (*record),* dan kolom (*atribut*) serta pembersihan data untuk tahapan *modelling.*
    4. *Modelling*, Dalam tahapan ini dilakukan beberapa teknik dan parameter yang dibutuhkan dalam prosesnya untuk menghasilkan model model tertentu yang akan berperan penting dalam tahapan evaluasi. Dari beberapa teknik yang dilakukan tahapan ini mengharuskan kembali ke dalam tahapan *data preparation* untuk menghasil nilai yang lebih optimal.
    5. *Evaluation*, Dalam tahapan *evaluation* bertujuan untuk menentukan apakah ada permasalahan dalam penentuan tujuan dan konsep dalam tahap *Business Understanding.* Tahapan *evaluation* harus menghasilkan keputusan untuk penggunaan hasil dari penambangan data.
    6. *Deployment*, Dalam tahapan ini bertujuan untuk membangun sebuah model untuk digunakannya dalam penyajian tertentu. Bahkan dalam banyak kasus, tahapan ini dibangun langsung oleh *customer* bukan *data analyst* yang melakukan tahapan ini.



### ***Text Mining***

*Text mining* merupakan turunan dari data mining itu sendiri. *Text mining* hanya mengolah sebuah data yang berupa teks yang bertujuan untuk menghasilkan suatu model tertentu. Data mentah yang berupa teks memiliki bentuk yang tidak terstruktur dan tidak terdefinisi. Maka dari itu, teknik text mining dapat memberikan keuntungan seperti pemrosesan bahasa alami, klasifikasi dan pengelompokan teks, pencarian informasi, dan ringkasan teks [17].

*Text mining* secara umum diterapkan dalam menganalisa sentimen, penyaringan spam, meringkas dokumen, mengukur preferensi pelanggan, pengelompokan topik penelitian. Beberapa penerapan text mining untuk menunjang efisiensi serta performa yang akan dihasilkan sangat tergantung dari preprocessing teks, penggunaan algoritma tertentu untuk mendapatkan pola-pola tertentu, serta elemen presentasi untuk menunjukkan hasil dari pola yang didapatkan [18].

### **Analisis Sentimen**

Analisis sentimen atau yang sering dikenal *opinion mining* merupakan proses identifikasi dari dalam penentuan sentimen pada subjek tertentu menggunakan data yang berupa teks. Proses identifikasi tersebut bertujuan untuk membantu membuat keputusan yang tepat dari opini publik. Pengelompokan atau pengekspresian sentimen dibagi menjadi dua ekspresi yaitu sentimen negatif atau sentimen positif [19].

Analisis sentimen menjadi studi kasus yang banyak diteliti karena beberapa alasan, diantaranya berkembangnya bidang NLP (*natural language processing),* tersedia banyaknya opini yang tersebar di berbagai media, kemajuan teknologi dan relevansi praktis. Secara umum, analisis sentimen berdasarkan konsep pengklasifikasiannya dibagi menjadi tiga tingkatan, diantaranya sebagai berikut [20]:

1. Tingkat dokumen

Tingkatan dokumen merupakan tingkatan dalam mengklasifikasikan dokumen menjadi sentimen negatif atau positif. Tingkatan ini menjadikan dokumen menjadi entitas tunggal yang diklasifikasikan tanpa membandingkan opini beberapa entitas didalamnya.

1. Tingkat kalimat

Tingkatan kalimat merupakan tingkatan dalam mengklasifikasikan kalimat atau sebuah pernyataan yang dibuat oleh individu tertentu, diantaranya seperti komentar atau tweet mengenai subjek tertentu.

1. Tingkat aspek

Tingkat aspek atau yang sering disebut tingkat fitur merupakan tingkat terendah dari tingkat sebelumnya karena tingkat ini mengklasifikasikan suatu objek berdasarkan fitur atau atribut yang ada pada subjek tersebut.

### **Portal Berita *Online***

Surat kabar yang sering disebut koran merupakan salah satu media cetak utama yang dibuat menggunakan kertas berbiaya rendah dan berisi berita-berita terkini mengenai berbagai topik. Bahkan surat kabar ada menyebutkan sama dengan pers. Persepsi ini diambil karena pers berhubungan khusus dengan media yang bercetak secara langsung. Maka dari itu pers sering disebut persuratkabaran karena merupakan hasil elaborasi dari ilmu publisistik dan ilmu jurnalisitik [21]. Akan tetapi dalam artian luas pers meliputi media cetak maupun media elektronik. Salah satu dari media elektronik adalah portal berita online.

Perkembangan dari zaman mengharuskan surat kabar tidak hanya menggunakan media konvensional tapi beralih ke media *online* seperti halnya, portal berita online *kompas.com, detiknews.com, tribunnews.com,* dan lain-lain. Selain itu media online memiliki keunggulan dibandingkan dengan media konvensional diantaranya, media online menyampaikan berita secara singkat dan cepat serta dianggap lebih murah dan mudah dalam mengaksesnya dibandingkan dengan media konvensional [22].

Pers memiliki hak, kewajiban, dan peranan untuk mendukung kebebasan dalam berekspresi di dalam bermasyarakat. Sehingga diperlukan kebebasan pers untuk memberikan informasi yang aktual. Wartawan yang mempunyai tugas mengumpulkan dan menyediakan informasi yang aktual untuk dipublikasikan oleh pers memerlukan landasan dan etika profesi untuk menjaga kepercayaan publik. Salah satunya wartawan indonesia tidak diperkenankan untuk menyebarkan informasi hoax[23].



### ***Synthetic Minority Oversampling Technique* (SMOTE)**

*Synthetic Minority Oversampling Technique* (*SMOTE*) merupakan salah satu teknik *oversampling* atau penambahan data minoritas dengan cara menambahkan data sintetis yang bertujuan untuk menyeimbangkan jumlah data antara data minoritas dan data mayoritas [24]. Teknik ini mampu meningkatkan performa dari proses klasifikasi. Karena performa proses klasifikasi akan menurun ketika data tidak seimbang antara data minoritas dan dan tidak seimbang [25]. Pembuatan data sintesis dibuat berdasarkan *k* tetangga terdekat (*k-nearest neighbor*). Data yang berskala numerik berbeda dengan data yang berskala kategorik dalam pembuatan data sintesis.

Data yang berskala numerik dibuat dengan cara perhitungan jarak antar vektor fitur, salah satunya *Euclidean* yang dinyatakan dalam persamaan 2.1.

(2.1)

Untuk menghasilkan vektor sintetis kelas minoritas baru dengan cara memilih pilihan terbanyak yang digunakan dari fitur tetangga terpilih, seperti halnya sebagai berikut:

Dari ketiga vektor fitur tersebut maka akan terbentuk vektor fitur sintetis seperti berikut:

### **Seleksi Fitur**

Seleksi fitur dibagi menjadi 3 kategori, yaitu diantaranya kategori *filter*, kategori *wrapper,* dan kategori *embeded.* Dua contoh seleksi fitur kategorti *filter* terdapat *chi square* dan *information gain*. Sedangkan dua contoh seleksi fitur kategori wrapper terdapat *sequantial forward selection* dan *sequantial backward elimination*. Penjelasan dari keempat seleksi fitur sebagai berikut:

#### **Chi-Square**

Seleksi fitur *Chi-Square* adalah teknik seleksi fitur yang digunakan untuk memilih fitur yang paling berpengaruh dalam masalah klasifikasi dengan menggunakan statistik Chi-Square. Seleksi fitur *chi square* termasuk dalam seleksi fitur berkategori filter*.* Metode seleksi fitur *chi square* berdasarkan pada asumsi bahwa fitur yang paling relevan adalah fitur yang paling tidak independen dengan variabel target [26].

Pencarian atribut atau fitur terpenting dimulai dengan caraperhitungan *degrees of freedom* dari tabel kontingensi, seperti berikut:

(2.2)

Kemudian cari nilai *chi square* tiap *term* menggunakan persamaan berikut [27]:

(2.3)

Dimana:

N = total dari jumlah dokumen

A = total jumlah dokumen yang memuat fitur *t* pada dokumen kelas *c*

B = total jumlah dokumen yang memuat fitur *t* selain pada dokumen kelas *c*

C = total jumlah dokumen yang tidak memuat fitur *t* pada dokumen kelas c

D = total jumlah dokumen yang tidak memuat fitur *t* selain pada dokumen kelas c

Setelah proses perhitungan nilai *chi square* pada kata (*term*)selesai. Maka tahapan selanjutnya membandingkan *critical value* dengan nilai chi square yang didapatkan. Jika nilai chi square lebih tinggi daripada *critical value* maka fitur itu diterima untuk proses selanjutnya [28].

#### **Information Gain (Mutual Information Maximum)**

Seleksi fitur *information gain* atau *mutual information maximum* merupakan salah satu seleksi fitur yang simpel dan cepat dan yang sering digunakan dalam aplikasi klasifikasi teks [29]. Seleksi fitur adalah salah satu seleksi fitur berkategori *filter* yang mampu dalam menghindari masalah *overfitting.*  Fitur terbaik bisa digunakan dalam perhitungan entropy fitur. *Entropy* merupakan ukuran ketidakpastian fitur yang dapat digunakan untuk menyimpulkan distribusi fitur dalam bentuk yang ringkas. Maka untuk persamaan 2.4 merupakan cara untuk menghitung *entropy* total[30].

Dimana c merupakan jumlah nilai pada kelas klasifikasi dan merupakan jumlah sampel atau untuk kelas . Setelah mendapatkan nilai *entropy* pada persamaan 2.5 kemudian dilanjutkan dalam perhitungan nilai information gain 2.6.

(2.5)

(2.4)

Dimana merupakan atribut, merupakan himpunan nilai yang memungkin untuk , merupakanjumlah sampel untuk , merupakan jumlah data sampel, sedangkan merupakan *entropy* untuk sampel yang bernilai **.** Kemudian untuk memilih atribut yang terpenting dengan cara memilih atribut dengan nilai *information gain* yang lebih besar daripada *threshold* yang sudah ditentukan.

#### **Forward Selection**

Forward selection merupakan seleksi fitur yang berbasis *wrapper* dengan menambahkan fitur satu-persatu dalam proses penyeleksiannya dan pemodelan awal yang digunakan tanpa penggunaan fitur [31]. Proses awal Langkah-langkah proses forward selection sebagai berikut [29]:

1. Fitur paling penting = dipilih pertama kali menggunakan suatu kriteria. Kriteria yang digunakan bisa menggunakan nilai Akurasi tertinggi dalam pemodelan data dengan fitur yang digunakan satu persatu atau performa matrik yang telah ditentukan.
2. Kemudian pasangan fitur dibentuk dengan dan himpunan terbaik dipilih sebagai = . Fitur yang sudah dipilih sebelumnya tidak boleh dimasukkan kembali.
3. Kemudian pilih himpunan dengan kriteria terbaik. Kriteria terbaik juga berdasarkan suatu performa pemodelan yang ditentukan sebelumnya.
4. Proses ini diulangi sampai jumlah fitur yang telah ditentukan terpilih.

#### **Backward Elimination**

Backward Elimination merupakan seleksi fitur yang berbasis *wrapper* dengan menghapus fitur satu-persatu yang dirasa tidak penting dalam proses penyeleksiannya dan pemodelan awal yang digunakan dengan menggunakan semua fitur [31]. Langkah-langkah proses backward elimination sebagai berikut [29]:

* 1. Fitur paling tidak penting = dipilih pertama kali menggunakan suatu kriteria. Kriteria yang digunakan bisa menggunakan nilai Akurasi tertinggi dalam pemodelan data dengan mengeliminasi fitur tersebut [32].
  2. Kemudian eliminasi satu fitur lagi sehingga membentuk himpunan dan himpunan terbaik dipilih sebagai = . Fitur yang sudah tereliminasi sebelumnya tidak boleh dimasukkan kembali.
  3. Kemudian pilih himpunan dengan kriteria terbaik. Kriteria terbaik juga berdasarkan suatu performa pemodelan yang ditentukan sebelumnya.
  4. Proses ini diulangi sampai jumlah fitur yang telah ditentukan terpilih.



### **Klasifikasi *Naïve Bayes***

Klasifikasi Naive bayes merupakan metode klasifikasi yang termasuk dalam *supervised learning* yang mengasumsikan setiap atribut atau fitur bersifat independen tidak berkaitan satu sama lain antar fiturnya, asumsi ini berdasarkan teori dari *bayes* dengan pengambilan kata *naive* [33]. Rumus dasar yang dapat digunakan dalam pengklasifikasian naive bayes sebagai berikut:

(2.6)

Dimana:

*c* : kategori atau kelas yang didapatkan

: Himpunan dari kategori yang ada pada data

: *Prior probability* dari kelas *c*

: *Posterior probability* dari kelas c terhadap dokumen testing

: Probabilitas fitur pada kelas c

Perbedaan dari klasifikasi naive bayes yaitu terdapat dalam perhitungan mengenai distribusi fitur terhadap kategori pada data sesuai dengan algoritma yang digunakan. Diantara beberapa algoritma yang ada pada pada klasifikasi *naive bayes*, *Multinomial Naive Bayes* yang sering digunakan pada klasifikasi teks [34][35].

Multinomial naive bayes merupakan algoritma yang dikembangkan dari naive bayes dengan menghitung kemungkinan setiap kata atau fitur yang muncul pada dokumen. Sehingga algoritma ini sangat cocok dengan pengklasifikasian teks atau dokumen. Pada multinomial naive bayes perhitungan mengenai distribusi menggunakan persamaan sebagai berikut:

(2.7)

Dimana:

: Jumlah kemunculan fitur *i* pada data *training* di kategori *c*

: Jumlah frekuensi seluruh fitur pada kelas *c*

: nilai konstanta yang bernilai 1 yang didapatkan dari teknik *Laplace*

*smoothing*

: jumlah seluruh kata pada *vocabulary* (kosa kata)

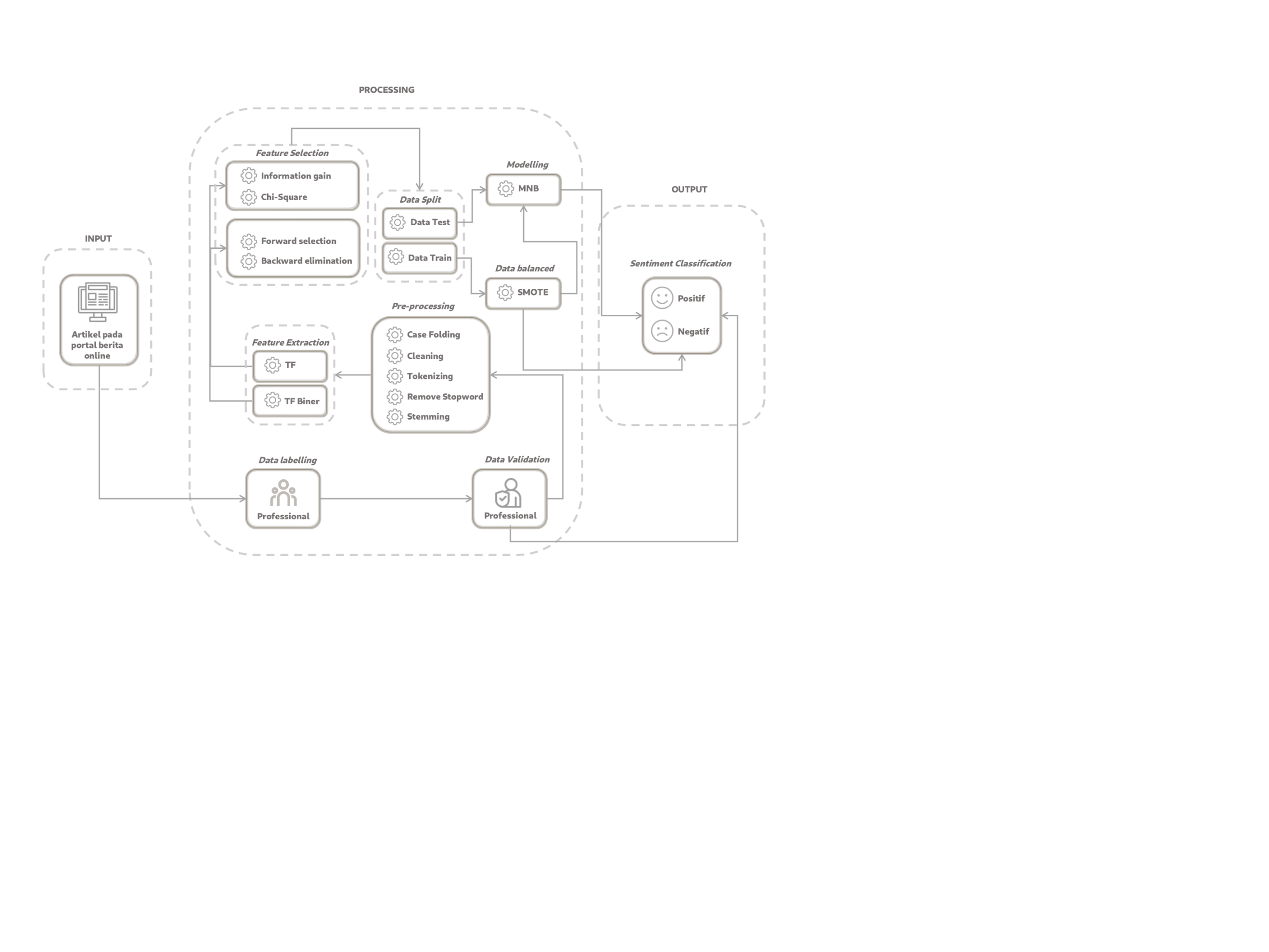
Penambahan teknik *laplace smoothing* pada persamaan ini untuk mengatasi perhitungan menggunakan perhitungan *maximum likelihood* yang memiliki kelemahan ketika frekuensi suatu kosa kata dalam kelas bernilai 0 maka nilai yang didapatkan akan bernilai 0.

### **Penelitian Terkait**

Tabel 2. 1 Penelitian Terkait

| **Judul** | **Banyak Data** | **Metode** | **Akurasi** |
| --- | --- | --- | --- |
| *Empirical study of feature selection methods over classification algorithms* | 299.285 | Chi Square, Naive Bayes | 86.2% |
| *A study on credit scoring modelling with different feature selection and machine learning approaches* | 1000 | Chi Square, Naive Bayes | 69,71% |
| *A study on credit scoring modelling with different feature selection and machine learning approaches* | 1000 | Information Gain, Naive Bayes | 70,59% |
| *Twitter Sentiment Analysis of Movie Reviews Using Information Gain and Naïve Bayes Classifier* | 8231 | Information Gain, Naive Bayes | 82,19% |
| Penggabungan Forward Selection untuk Pemilihan Fitur pada Prediksi Bimbingan Konseling Siswa dengan Menggunakan Algoritma Naive Bayes | 329 | *Forward Selection, Naive Bayes* | 94,84% |
| Implementasi Metode Forward Selection pada Algoritma Support Vector Machine (SVM) dan Naive Bayes Classifier Kernel Density (Studi Kasus Klasifikasi Jalur Minat SMA) | 280 | *Forward Selection, Naive Bayes* | 98,21% |
| *Naïve Bayes Algorithm Based On Backward Elimination For Predicting Cervical Cancer* | 72 | Backward Elimination, Naive Bayes | 94,44% |
| Optimasi Backward Elimination untuk Klasifikasi Kepuasan Pelanggan Menggunakan Algoritma k-Nearest Neighbor (k-NN) dan Naïve Bayes | 129.880 | Backward Elimination, Naive Bayes | 99,04% |
| *Improved Accuracy of Naive Bayes Classifier for Determination of Customer Churn Uses SMOTE and Genetic Algorithms* | 1000 | SMOTE, Naive Bayes | 80,948% |
| *Analyzing the Application of SMOTE on Machine Learning Classifiers* | 131 | SMOTE, Naive Bayes | 69.61% |

# **BAB III METODE USULAN**

1. 

## **Arsitektur Sistem**

Gambar 3. 1 Arsitektur Sitem

Arsitektur sistem pada penelitian ini dimulai dari pengumpulan data berita pariwisata madura yang diambil dari beberapa sumber dan dilanjutkan dengan pelabelan berita yang dilakukan oleh peneliti dan divalidasi oleh ahli. Tahapan selanjutnya adalah *preprocessing* yang didalamnya terdapat beberapa proses yaitu *case folding, filtering*, *tokenizing, stopword removal*, dan *stemming*. Setelah melalui tahap *preprocessing*, selanjutnya ekstraksi fitur, dimana ekstraksi fitur dibagi menjadi 2 yaitu TF dan binary.

Hasil ekstraksi fitur TF akan digunakan pada seleksi fitur *forward selection* dan *backward elimination*, sedangkan hasil ekstraksi fitur binary digunakan pada seleksi fitur *information gain* dan *chi square*. Selanjutnya adalah tahap pembagian data. Dataset akan dibagi menjadi dua kelompok yaitu data latih dan data uji. Pada data latih berjumlah 80% dari data keseluruhan sedangkan data latih sebanyak 20% dari data keseluruhan. Karena data mengalami ketidakseimbangan label sentimen sehingga diperlukan metode SMOTE untuk menyeimbangkan data latih dengan label minoritas. Setelah data seimbang, tahap selanjutnya adalah klasfikasi sentimen dengan metode *Naïve Bayes Multinomial*.

## **Dataset Penelitian**

Dataset yang digunakan dalam penelitian ini adalah dalam format CSV dengan jumlah total data sebanyak 200 entri dan terdapat 7 kolom diantaranya *url*, tanggal terbit, isi, judul, objek, kabupaten, dan label. Dataset diambil dari beberapa laman internet diantaranya seperti: *detik.com, regional.kompas.com, radarmadura.jawapos.com,* dan lain-lain. Dataset yang diambil berfokus pada topik pariwisata madura dan berita terbitan dari bulan mei tahun 2021 sampai terbitan terbaru. Data dalam penelitian ini telah dikategorikan ke dalam dua kelompok berdasarkan sentimen isi beritanya, yaitu berita sentimen negatif dan berita sentimen positif. Berita yang bersentimen negatif berjumlah 69 berita sedangkan berita yang bersentimen positif berjumlah 131 berita.

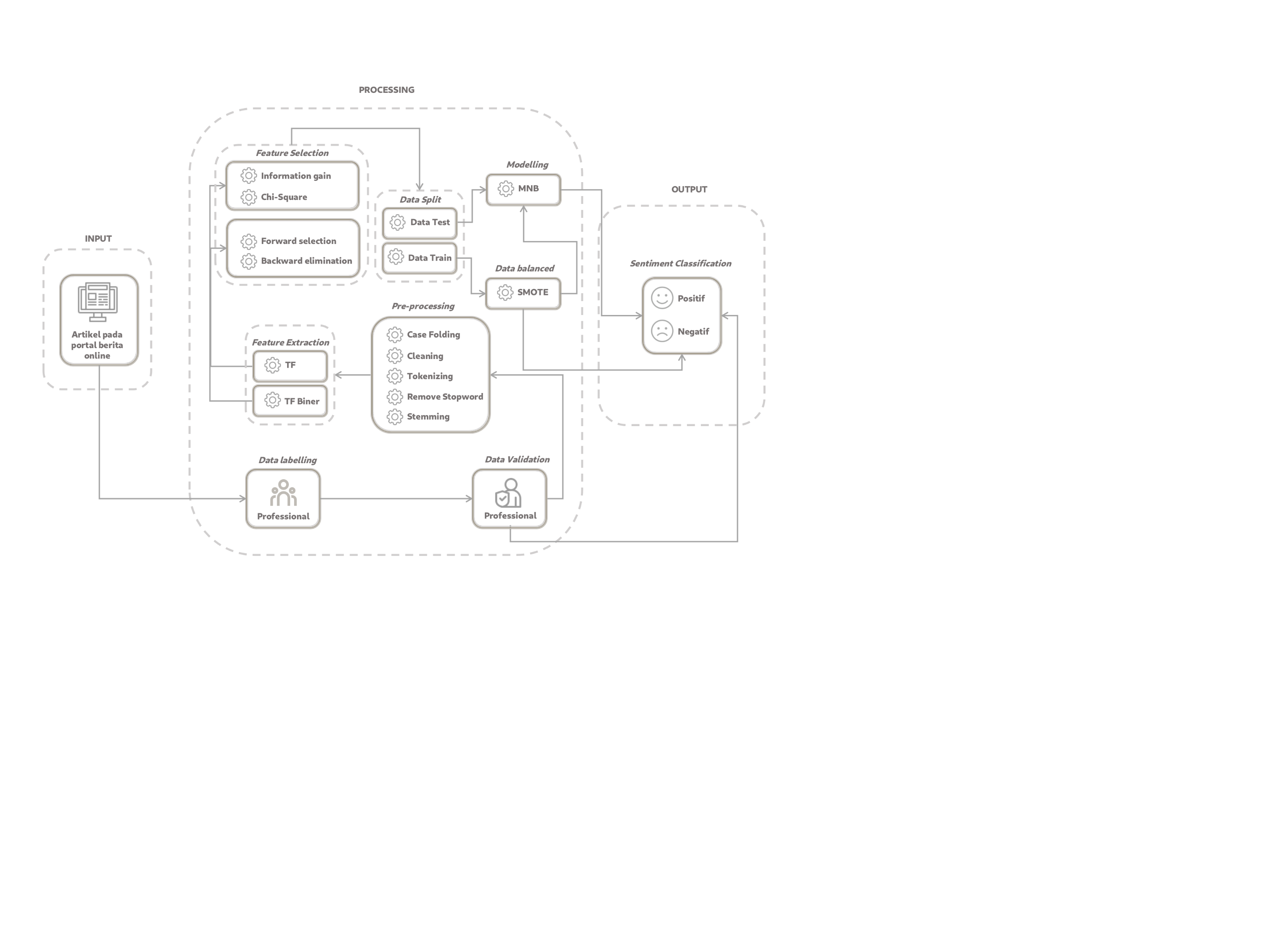
## **Pelabelan Data**

Pelabelan data dilakukan secara manual oleh peneliti dengan melakukan pembacaan dan analisis terhadap isi teks berita. Data dibagi menjadi 2 label yaitu data yang berlabelkan positif dan data yang berlabelkan negatif. Kemudian setelah data dilabelkan akan dilakukan tahapan validasi data dengan sukarelawan dosen bahassa indonesia di kampus Kariman Wirayudha (INKADHA SUMENEP) status dosen bahasa Indonesia yang bernama bapak Amir Chalis Andriyansyah. S.Pd, M.Pd.

## **Metode**

Berikut adalah beberapa metode yang digunakan dalam penelitian ini:

### ***Text Preprocessing***

Data yang akan digunakan dalam contoh text preprocessing yaitu 10 judul berita pada dataset penelitian, seperti yang ditunjukkan pada tabel 3.1

Tabel 3. 1 Data Judul Berita

| **No** | **Judul** | **Label** |
| --- | --- | --- |
| 1 | “Sampah Berserakan di Objek Wisata Pantai Lombang-lombang Jadi Sorotan Pengunjung” | Negatif |
| 2 | “Mayat Tanpa Identitas Hanya Tinggal Tengkorak Ditemukan Warga Sumenep” | Negatif |
| 3 | “Mayat Pria Tanpa Identitas Ditemukan di Perairan Kangean Sumenep” | Negatif |
| 4 | “Desa Wisata Keris di Sumenep Raih Penghargaan Bergengsi” | Positif |
| 5 | “Desa Wisata Aeng Tong Tong raih penghargaan dari Kemenparekraf” | Positif |
| 6 | “Sandiaga Tetapkan Aeng Tong-tong Sumenep Masuk 50 Desa Wisata Terbaik” | Positif |
| 7 | “Mantap! Keris Aeng Tongtong dari Sumenep Jadi Suvenir Resmi KTT G20” | Positif |
| 8 | "6 Desa Wisata Terbaik di Indonesia, Asri dan Mengagumkan!” | Positif |
| 9 | “Mengenal Desa Wisata Aeng Tong-tong, Asal Pembuat Keris Suvenir untuk Delegasi G20” | Positif |
| 10 | “Pengunjung Wisata Pantai Camplong Sampang Membludak, Personel Gabungan Beri Pengamanan Ekstra” | Positif |

Tabel 3.2 menunjukkan salah satu data pada tabel 3.1 yang melalui 5 tahapan *text preprocessing*, yaitu sebagai berikut: *case fold, filtering, tokenizing, remove stopword,* dan *stemming*. Sedangkan pada tabel 3.3 merupakan hasil text preprocessing dari semua data yang ada pada tabel 3.1

Tabel 3. 2 Text Preprocessing

|  |  |
| --- | --- |
| **Tahapan** | **Sesudah** |
| Data awal | “6 Desa Wisata Terbaik di Indonesia, Asri dan Mengagumkan!” |
| Case Fold | “6 desa wisata terbaik di indonesia, asri dan mengagumkan!” |
| Cleaning | “desa wisata terbaik di indonesia asri dan mengagumkan” |
| Tokenizing | “‘desa’, ‘wisata’, ‘terbaik’, ‘di’, ‘indonesia’, ‘asri’, ‘dan’, ‘mengagumkan’” |
| Remove Stopword | “‘desa’, ‘wisata’, ‘terbaik’, ‘indonesia’, ‘asri’, ‘mengagumkan’” |
| Stemming | “‘desa’, ‘wisata’, ‘baik’, ‘indonesia’, ‘asri’, ‘kagum’” |

Tabel 3. 3 Data yang Sudah Melalui Tahapan Text Preprocessing

|  |  |  |
| --- | --- | --- |
| **No** | **Judul** | **Label** |
| 1 | “sampah sera objek wisata pantai lombanglombang sorot unjung” | Negatif |
| 2 | “mayat identitas tinggal tengkorak temu warga sumenep” | Negatif |
| 3 | “mayat pria identitas temu air kangean sumenep” | Negatif |
| 4 | “desa wisata keris sumenep raih harga gengsi” | Positif |
| 5 | “desa wisata aeng tong tong raih harga kemenparekraf” | Positif |
| 6 | “sandiaga tetap aeng tongtong sumenep masuk desa wisata baik” | Positif |
| 7 | “mantap keris aeng tongtong sumenep suvenir resmi ktt” | Positif |
| 8 | “desa wisata baik indonesia asri kagum” | Positif |
| 9 | “kenal desa wisata aeng tongtong buat keris suvenir delegasi” | Positif |
| 10 | “unjung wisata pantai camplong sampang membludak personel gabung aman ekstra” | Positif |



### ***Extraction TF***

Pada tahapan pertama *feature extraction* menggunakan TF yaitu mengubah korpus dokumen pada tabel 3.3 menjadi vektor fitur dengan pembobotan *term frequency* (tf) seperti yang ditunjukkan pada tabel 3.4.

Tabel 3. 4 Data yang diekstrak dengan pembobotan TF

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Term*** | **D1** | **D2** | **D3** | **D4** | **D5** | **D6** | **D7** | **D8** | **D9** | **D10** |
| aeng | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 |
| air | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| aman | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| asri | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| baik | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| buat | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| camplong | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| delegasi | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| desa | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 |
| ekstra | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| gabung | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| gengsi | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| harga | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| identitas | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| indonesia | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| kagum | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| kangean | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| kemenparekraf | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| kenal | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| keris | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| ktt | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| lombanglombang | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| mantap | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| masuk | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| mayat | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| membludak | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| objek | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| pantai | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| personel | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| pria | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| raih | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| resmi | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| sampah | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| sampang | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| sandiaga | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| sera | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| sorot | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| sumenep | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 |
| suvenir | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| temu | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| tengkorak | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| tetap | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| tinggal | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| tong | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| tongtong | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| unjung | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| warga | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| wisata | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |

### **Metode Information Gain**

Gambar 3. 2 Flowchart seleksi fitur Information Gain

Pada tahapan seleksi fitur menggunakan metode *information gain*, hal pertama yang dilakukan dalam penelitian yaitu mempersiapkan data. Data yang dipersiapkan yaitu data pada tabel 3.4 yang telah diubah dengan pembobotan biner yang sebelumnya *term frequency*. Pembobotan biner *term* atau fitur ada dalam dokumen akan bernilai 1, dan jika tidak ada dalam dokumen akan bernilai 0, seperti yang ditunjukkan pada tabel 3.5 dengan 2 contoh term pada 10 dokumen.

Tabel 3. 5 Contoh vektor fitur data dengan pembobotan TF

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **term** | **D1** | **D2** | **D3** | **D4** | **D5** | **D6** | **D7** | **D8** | **D9** | **D10** |
| … | .. | .. | .. | .. | .. | .. | .. | .. | .. | .. |
| tong | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| … | .. | .. | .. | .. | .. | .. | .. | .. | .. | .. |
| wisata | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |

Setelah pembobotan biner pada data dilakukan, berikutnya menghitung total dokumen pada data di tabel 3.3 berdasarkan sentimen berita seperti yang ditunjukkan pada tabel 3.6.

Tabel 3. 6 Total Sentimen dari Judul Berita

| **Sentimen** | **Total Sentiment** |
| --- | --- |
| Positif | 7 |
| Negatif | 3 |
| Total Data | 10 |

Menghitung entropy total dengan 5 kelas positif dan 5 kelas negatif, seperti yang ditunjukkan pada tabel 3.2 menggunakan persamaan 2.4 sebagai berikut:

Selanjutnya menghitung dari sebuah fitur atau kata dengan persamaan 2.4, sebagai contoh hitung entropy nilai pada kata *wisata* yang terdapat pada 2 kelas yaitu positif dan negatif. Dimana pada setiap kelas mempunyai 2 nilai yaitu 1 dan 0 dan jumlah dari setiap nilainya ditunjukkan pada tabel 3.4. Maka perhitungan entropynya sebagai berikut:

Setelah perhitungan tiap nilai telah dihitung, maka tahapan selanjutnya menghitung nilai information gain dari kata atau fitur *wisata* dengan menggunakan persamaan 2.5 sebagai berikut:

Langkah terlahir memilih kata sesuai dengan information gain yang diperoleh dengan cara membandingkan dengan nilai bobot minimum atau *threshold*. Jika threshold sama dengan 0.2 maka 5 dari 48 fitur yang digunakan untuk proses klasifikasi seperti yang ditunjukkan pada tabel 3.7.

Tabel 3. 7 Hasil Information Gain pada tiap fitur

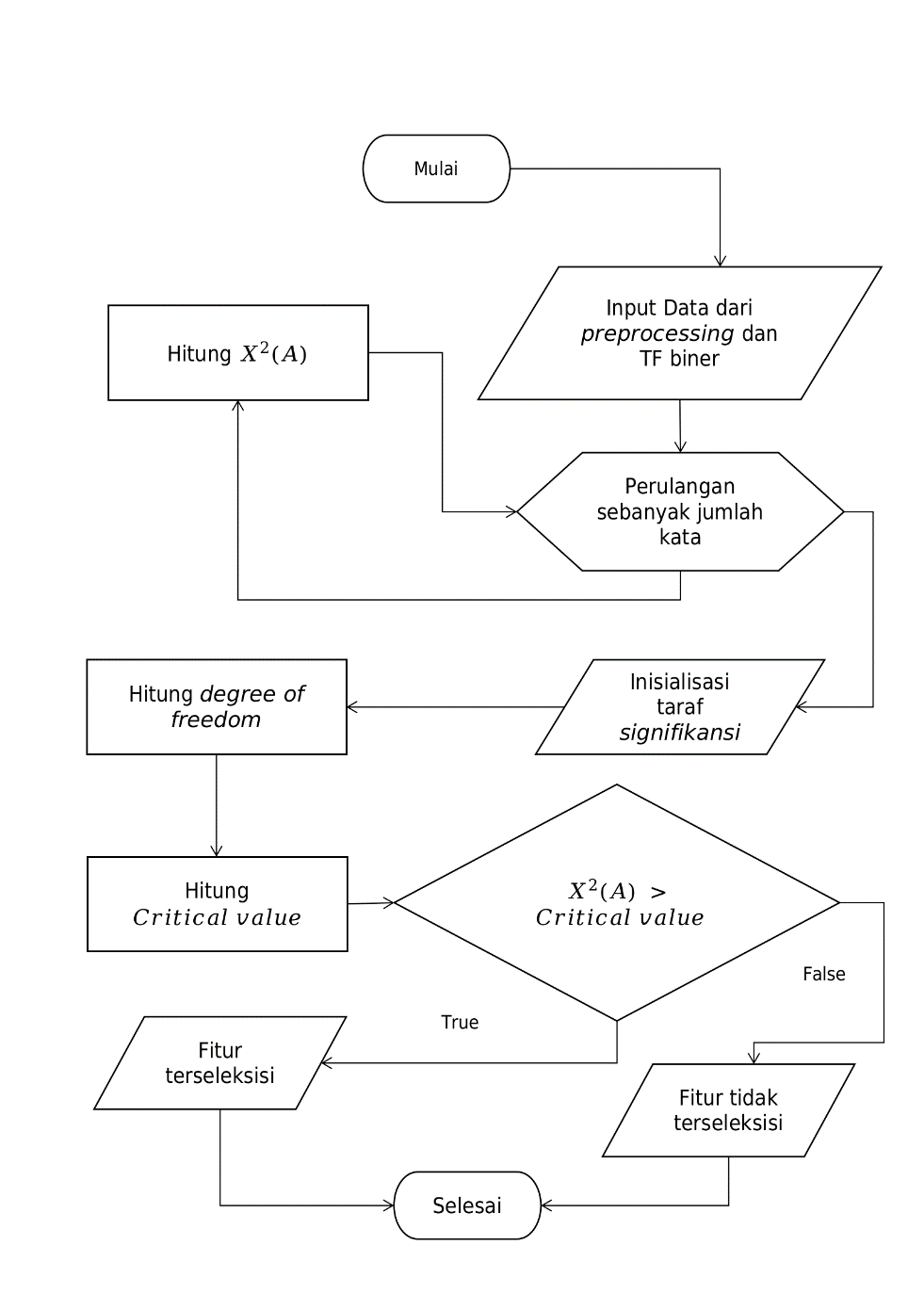
| **Fitur** | **Nilai IG** | **Fitur** | **Nilai IG** | **Fitur** | **Nilai IG** |
| --- | --- | --- | --- | --- | --- |
| pantai | 0.0323 | mantap | 0.0548 | air | 0.194 |
| unjung | 0.03.23 | masuk | 0.0548 | kangean | 0.194 |
| sumenep | 0.0349 | membludak | 0.0548 | lombanglombang | 0.194 |
| aman | 0.0548 | personel | 0.0548 | objek | 0.194 |
| asri | 0.0548 | resmi | 0.0548 | pria | 0.194 |
| buat | 0.0548 | sampang | 0.0548 | sampah | 0.194 |
| camplong | 0.0548 | sandiaga | 0.0548 | sera | 0.194 |
| delegasi | 0.0548 | tetap | 0.0548 | sorot | 0.194 |
| ekstra | 0.0548 | tong | 0.0548 | tengkorak | 0.194 |
| gabung | 0.0548 | baik | 0.118 | tinggal | 0.194 |
| gengsi | 0.0548 | harga | 0.118 | warga | 0.194 |
| indonesia | 0.0548 | raih | 0.118 | aeng | 0.281 |
| kagum | 0.0548 | suvenir | 0.118 | desa | 0.396 |
| kemenparekraf | 0.0548 | keris | 0.192 | identitas | 0.446 |
| kenal | 0.0548 | tongtong | 0.192 | mayat | 0.446 |
| ktt | 0.0548 | wisata | 0.192 | temu | 0.446 |

### **Metode Chi Square**

Pada tahapan seleksi fitur menggunakan metode *chi square*, hal pertama yang dilakukan sama dengan metode information gain yaitu membangun vektor fitur data seperti tabel 3.5. Karena data yang digunakan sama dengan metode information gain yaitu data yang ditunjukkan tabel 3.3. Maka dari itu, tahapan selanjutnya membuat tabel kontingensi dari data tersebut seperti yang ditunjukkan tabel 3.8.

Tabel 3. 8 Tabel Kontingensi

|  | **Positif = 1** | **Negatif = 0** |
| --- | --- | --- |
| ***Term*** | *A* | B |
| ***Not Term*** | C | D |

Kemudian tahapan selanjutnya, perhitungan nilai chi-square tiap *term* dihitung menggunakan persamaan 2.3 dan variabel-variabel yang diperlukan untuk persamaannya tersedia di tabel 3.7, sebagai contoh menggunakan kata *nelayan* seperti berikut:

Gambar 3. 3 Flowchart seleksi fitur Chi Square

Tabel 3. 9 Jumlah variable A, B, C dan D pada tiap fitur

| **Kata** | **Jumlah** | | | | **Kata** | **Jumlah** | | | | **Kata** | **Jumlah** | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **A** | **B** | **C** | **D** | **A** | **B** | **C** | **D** | **A** | **B** | **C** | **D** |
| aeng | 0 | 4 | 3 | 3 | kangean | 1 | 0 | 2 | 7 | sampah | 1 | 0 | 2 | 7 |
| air | 1 | 0 | 2 | 7 | kemenparekraf | 0 | 1 | 3 | 6 | sampang | 0 | 1 | 3 | 6 |
| aman | 0 | 1 | 3 | 6 | kenal | 0 | 1 | 3 | 6 | sandiaga | 0 | 1 | 3 | 6 |
| asri | 0 | 1 | 3 | 6 | keris | 0 | 3 | 3 | 4 | sera | 1 | 0 | 2 | 7 |
| baik | 0 | 2 | 3 | 5 | ktt | 0 | 1 | 3 | 6 | sorot | 1 | 0 | 2 | 7 |
| buat | 0 | 1 | 3 | 6 | lombanglombang | 1 | 0 | 2 | 7 | sumenep | 2 | 3 | 1 | 4 |
| camplong | 0 | 1 | 3 | 6 | mantap | 0 | 1 | 3 | 6 | suvenir | 0 | 2 | 3 | 5 |
| delegasi | 0 | 1 | 3 | 6 | masuk | 0 | 1 | 3 | 6 | temu | 2 | 0 | 1 | 7 |
| desa | 0 | 5 | 3 | 2 | mayat | 2 | 0 | 1 | 7 | tengkorak | 1 | 0 | 2 | 7 |
| ekstra | 0 | 1 | 3 | 6 | membludak | 0 | 1 | 3 | 6 | tetap | 0 | 1 | 3 | 6 |
| gabung | 0 | 1 | 3 | 6 | objek | 1 | 0 | 2 | 7 | tinggal | 1 | 0 | 2 | 7 |
| gengsi | 0 | 1 | 3 | 6 | pantai | 1 | 1 | 2 | 6 | tong | 0 | 1 | 3 | 6 |
| harga | 0 | 2 | 3 | 5 | personel | 0 | 1 | 3 | 6 | tongtong | 0 | 3 | 3 | 4 |
| identitas | 2 | 0 | 1 | 7 | pria | 1 | 0 | 2 | 7 | unjung | 1 | 1 | 2 | 6 |
| indonesia | 0 | 1 | 3 | 6 | raih | 0 | 2 | 3 | 5 | warga | 1 | 0 | 2 | 7 |
| kagum | 0 | 1 | 3 | 6 | resmi | 0 | 1 | 3 | 6 | wisata | 1 | 6 | 2 | 1 |

Tabel 3. 10 Nilai chi pada tiap fitur

| **Fitur** | **Nilai** | **Fitur** | **Nilai** | **Fitur** | **Nilai** |
| --- | --- | --- | --- | --- | --- |
| aman | 0.476 | pantai | 0.476 | kangean | 2.59 |
| asri | 0.476 | personel | 0.476 | lombanglombang | 2.59 |
| buat | 0.476 | resmi | 0.476 | objek | 2.59 |
| camplong | 0.476 | sampang | 0.476 | pria | 2.59 |
| delegasi | 0.476 | sandiaga | 0.476 | sampah | 2.59 |
| ekstra | 0.476 | sumenep | 0.476 | sera | 2.59 |
| gabung | 0.476 | tetap | 0.476 | sorot | 2.59 |
| gengsi | 0.476 | tong | 0.476 | tengkorak | 2.59 |
| indonesia | 0.476 | unjung | 0.476 | tinggal | 2.59 |
| kagum | 0.476 | baik | 1.07 | warga | 2.59 |
| kemenparekraf | 0.476 | harga | 1.07 | wisata | 2.74 |
| kenal | 0.476 | raih | 1.07 | aeng | 2.85 |
| ktt | 0.476 | suvenir | 1.07 | desa | 4.29 |
| mantap | 0.476 | keris | 1.84 | identitas | 5.83 |
| masuk | 0.476 | tongtong | 1.84 | mayat | 5.83 |
| membludak | 0.476 | air | 2.59 | temu | 5.83 |

Setelah semua tiap *term* dihitung nilai chi squarenya, maka tahapan selanjutnya mencari nilai kritis yang diperoleh dari tabel chi square yang ada di tabel 3.9. Untuk mencari nilai kritis, parameter yang dibutuhkan adalah nilai derajat kebebasan yang dihitung menggunakan persamaan 2.6 dan nilai taraf signifikansi, sebagai contoh nilai signifikansi yang digunakan 0.2. Berikut adalah cara menghitung nilai derajat kebebasan atau *degree of freedom* dengan tabel kontingensi yang diacu adalah tabel 3.6 dan persamaan 2.2:

= 1

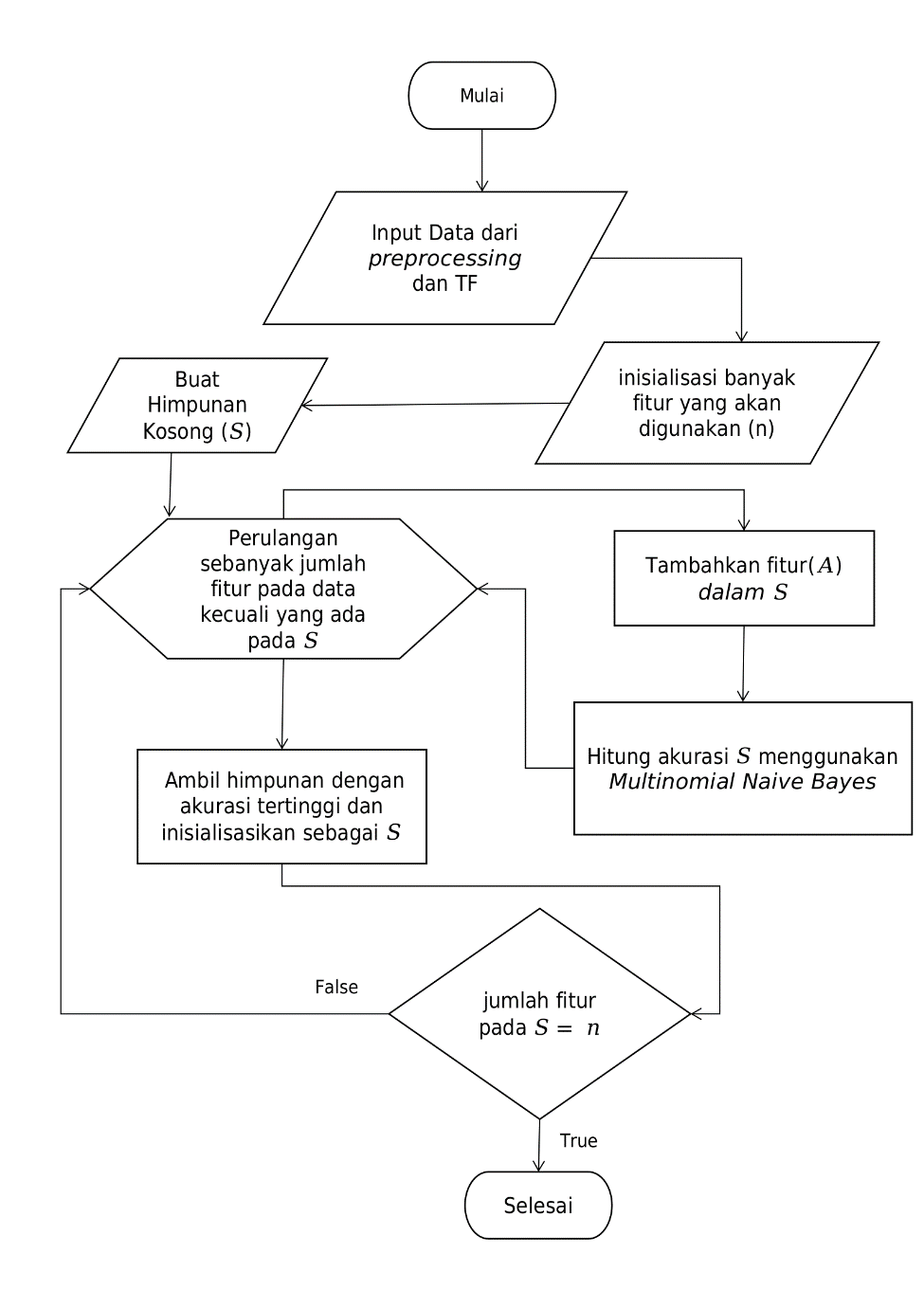
Tabel 3. 11 Tabel chi square

| ***degree of freedom (df)*** | **Taraf Signifikansi** | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **0.5** | **0.3** | **0.2** | **0.1** | **0.05** | **0.01** |
| **1** | 0.455 | 1.074 | 1.642 | 2.706 | 3.481 | 6.635 |
| **2** | 0.139 | 2.408 | 3.219 | 3.605 | 5.591 | 9.210 |
| **3** | 2.366 | 3.665 | 4.642 | 6.251 | 7.815 | 11.341 |
| **4** | 3.357 | 4.878 | 5.989 | 7.779 | 9.488 | 13.277 |
| **5** | 4.351 | 6.064 | 7.289 | 9.236 | 11.070 | 15.086 |

Dengan contoh penggunaan taraf signifikansi 0.05 dan nilai derajat kebebasan 1 maka akan didapatkan nilai kritis sama dengan 3.814. Setelah itu bandingkan nilai chi square setiap fitur dengan nilai kritis, ketika lebih kecil dari nilai kritis maka fitur tersebut ditolak dan tidak akan digunakan di proses selanjutnya. Maka dari itu fitur yang diambil hanya yang memiliki nilai yang lebih besar sama dengan nilai kritis yang didapatkan. Dari contoh perhitungan, fitur yang diterima hanya 4 dari 48 fitur.



### **Metode Sequential Forward Selection (SFS)**

Pada tahapan seleksi fitur dengan metode *SFS*, hal pertama yang dipersiapkan terlebih dahulu dalam penelitian ini yaitu vektor fitur data pada tabel 3.4*.* Kemudian, Inisialisasi atau banyak fitur yang akan digunakan, sebagai contoh akan digunakan fitur sebanyak 2 fitur dari 48 fitur untuk digunakan dalam proses selanjutnya. Dikarenakan metode ini dimulai dengan mengosongkan data tanpa fitur apapun sama sekali, maka dari itu buat himpunan kosong (S) untuk menampung fitur yang terpilih seperti berikut:

Gambar 3. 4 Flowchart seleksi fitur Sequential Forward Selection (SFS)

Kemudian tambahkan satu fitur dan hitung nilai evaluasi model sebagai contoh menggunakan pemodelan *naive bayes multinomial*. Dalam perhitungan nilai akurasi pada pemodelan menggunakan *library* *sklearn* di *python*.

….

Setelah mendapatkan akurasi dari semua himpunan fitur, maka pilih himpunan fitur dengan akurasi tertinggi, seperti yang ditunjukkan pada tabel 3.12. Fitur yang terpilih akan ditambahkan ke himpunan kosong (S), seperti berikut:

Tabel 3. 12 Hasil akurasi pada penggunaan satu fitur yang terurut dari tertinggi

| **Term** | **Accuracy** |
| --- | --- |
| aeng | 0.57 |
| mayat | 0.43 |
| identitas | 0.43 |
| temu | 0.28 |
| … | … |
| unjung | 0.14 |

Kemudian ulangi proses di atas dan tambahkan fitur satu persatu ke dalam himpunan tanpa menambahkan fitur yang ada pada himpunan sendiri, contohnya seperti berikut:

…

Maka pilih himpunan atau penambahan fitur yang mempunyai akurasi tertinggi. Ulangi Proses tersebut sampai himpunan terisi sebanyak . Pada tabel 3.12 menunjukkan nilai himpunan terbaik pada setiap iterasinya. Maka dari itu terpilih 2 himpunan fitur terbaik dari 48 fitur yang akan digunakan pada proses klasifikasi.

Tabel 3. 13 Hasil akurasi terbaik pada tiap penambahan fitur

| **No** | **Himpunan fitur yang diterima** | **Akurasi** |
| --- | --- | --- |
| 1 | aeng | 0.67 |
| 2 | mayat, aeng | 0.83 |

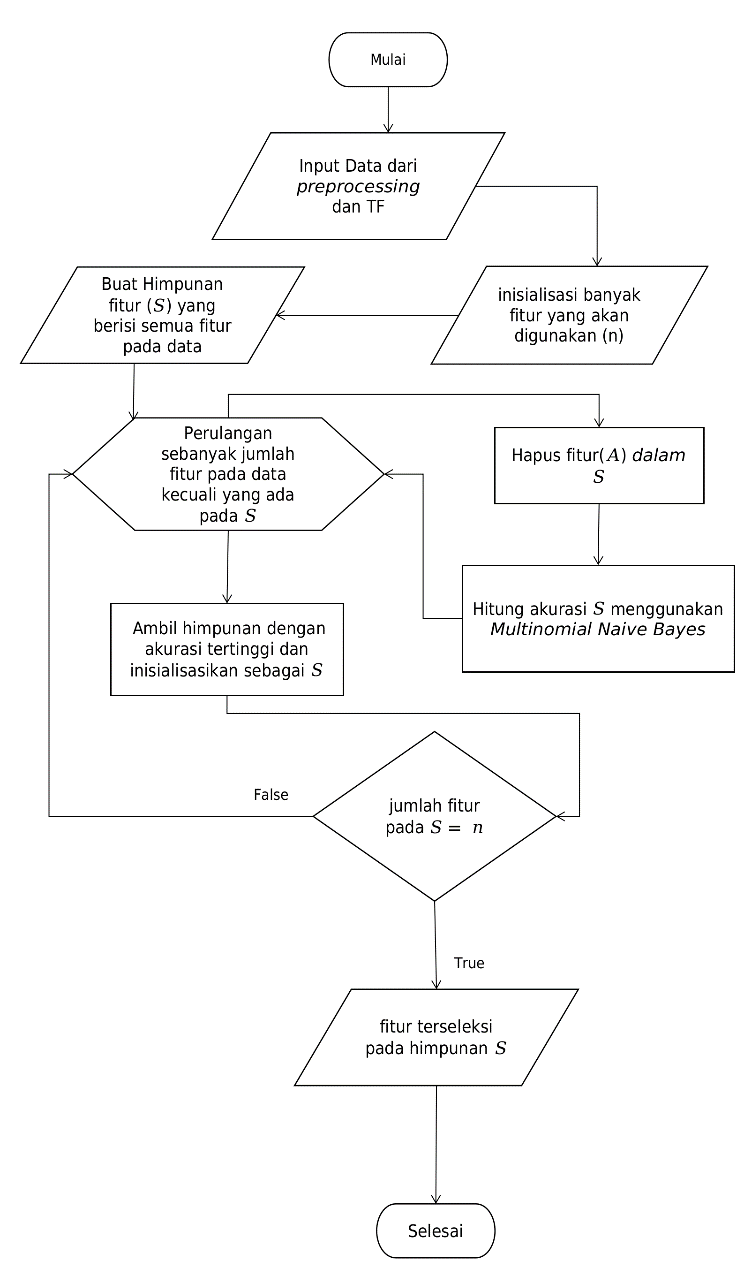
### **Metode Sequential Backward Elimination (SBE)**

Pada tahapan seleksi fitur dengan metode *Sequential Backward Elimination’*, hal pertama yang dipersiapkan terlebih dahulu dalam penelitian ini sama dengan metode *sequential backward elimination.* yaitu data pada tabel 3.9 dengan banyak dokumen 175 berita. *Term* yang digunakan untuk sampel perhitungan manual yaitu 62 dari 1489 *term* dengan IDF tertinggi.

Inisialisasi atau banyak fitur yang akan digunakan, sebagai contoh akan digunakan fitur sebanyak 46 fitur atau 2 fitur ditolak untuk digunakan untuk proses selanjutnya. Selanjutnya dimulai dengan membuat himpunan untuk menampung fitur yang akan digunakan untuk proses selanjutnya yang berisi semua fitur yang ada, contohnya sebagai berikut:

Kemudian kurangi satu fitur dan hitung nilai evaluasi model. Sebagai contoh menggunakan pemodelan *naive bayes multinomial* dengan data testing 50% dan data training 50% agar hasil yang didapatkan lebih bervariatif dengan contoh 10 dokumen yang digunakan. Dalam perhitungan nilai akurasi pada pemodelan menggunakan *library* *sklearn* di *python*.

….

Setelah mendapatkan akurasi dari semua himpunan fitur, maka pilih himpunan fitur dengan akurasi tertinggi atau penurunan akurasi paling minimal, seperti yang ditunjukkan pada tabel 3.14.

Gambar 3. 5 Flowchart seleksi fitur Sequential Backward Elimination

Tabel 3. 14 Hasil akurasi pada pengurangan satu fitur yang terurut dari tertinggi

| **Term** | **Accuracy** |
| --- | --- |
| pantai | 0.86 |
| unjung | 0.86 |
| buat | 0.71 |
| ekstra | 0.71 |
| … | … |
| mayat | 0.57 |

Fitur yang terpilih akan dihapus dari himpunan (S), maka himpunan akan berisi 47 fitur yang sebelumnya 48 fitur, seperti berikut:

Kemudian ulangi proses di atas dan kurangi fitur satu persatu pada himpunan tanpa menambahkan fitur yang ada pada himpunan sendiri, seperti berikut:

Maka pilih himpunan atau pengurangan fitur yang mempunyai akurasi tertinggi. Ulangi Proses tersebut sampai himpunan terisi sebanyak 46 fitur atau 2 fitur ditolak. Pada tabel 3.15 menunjukkan nilai akurasi himpunan terbaik atau penurunan nilai akurasi paling minimal pada setiap iterasinya. Maka dari itu terpilih 2 fitur yang ditolak dan 46 fitur untuk di proses pada tahapan selanjutnya.

Tabel 3. 15 Hasil akurasi terbaik pada tiap pengurangan fitur

|  |  |  |
| --- | --- | --- |
| **No** | **Himpunan fitur yang ditolak** | **Akurasi** |
| 1 | pantai | 0.86 |
| 2 | buat, pantai | 0.86 |

### **SMOTE**

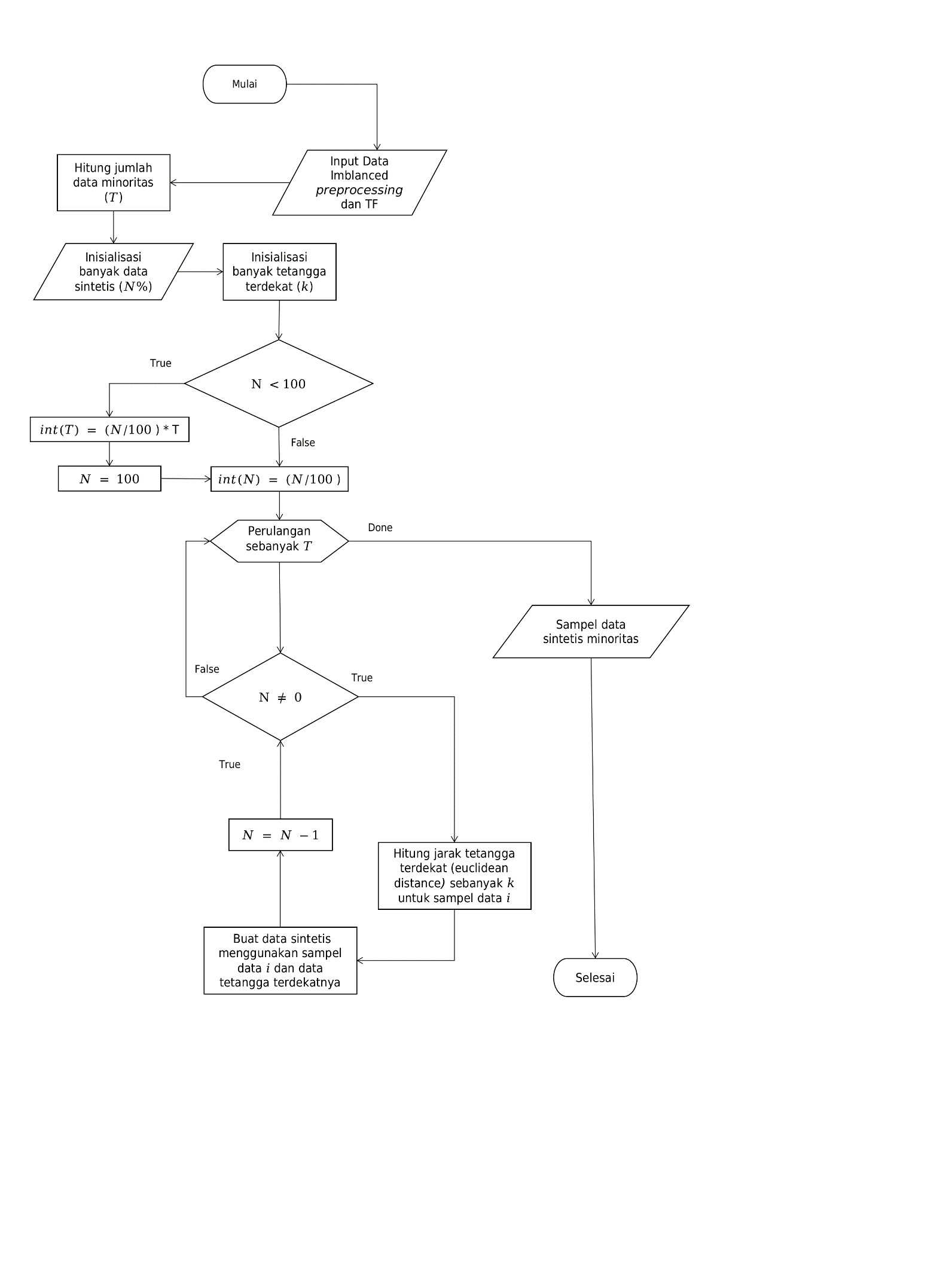
Persiapan awal dalam menerapkan metode SMOTE pada data yang tidak seimbang pada penelitian ini merupakan pembobotan teks pada data. Pada tabel 3.16 menggunakan semua fitur yang ada pada tabel 3.4.

Tabel 3. 16 Data Minoritas

| **NO** | **aeng** | **desa** | **identitas** | **…** | **wisata** |
| --- | --- | --- | --- | --- | --- |
| **1** | 0 | 0 | 0 | … | 1 |
| **2** | 0 | 0 | 0 | … | 0 |
| **3** | 0 | 1 | 0 | … | 0 |

Hitung jumlah data minoritas dan simpan dalam variabel () dan inisialisasikan berapa banyak data sintetis yang akan dibuat dalam bentuk persen, contohnya seperti berikut:

Karena nilai lebih kecil dari 100%, maka:



Gambar 3. 6 Flowchart SMOTE

Setelah itu pilih secara acak, data yang akan dijadikan tolak ukur pembuatan data sintetis, sebagai contoh data dokumen 2 pada tabel 3.14. Kemudian dari data tersebut cari tetangga terdekat atau *nearest neighbors*. karena data minoritas pada contoh hanya 3 dan *k* yang diinisialisasikan bernilai 2, maka tidak perlu dicari tetangga terdekatnya. Akan tetapi jika data yang digunakan lebih dari 3, maka diperlukan mencari jarak antar vektor data. Salah satu mencari tetangga terdekat dengan cara *euclidean distance* dengan persamaan 2.1, seperti berikut:

Setelah menghitung semua jarak antar vektor fitur, maka bandingkan semua jarak antar vektor dan pilih hasil nilai yang paling minimal. Pemilihan banyak data sesuai dengan nilai k yang diinisialisasikan. Karena *k* bernilai 2 maka pembuatan data sintetis dibuat dari 3 data minoritas yang ada pada tabel 3.16. Cara pembuatan data sintetis dengan cara memilih nilai fitur yang paling mayoritas dari 3 data tersebut. Berikut contoh pembuatan data sintetis dengan menggunakan 5 fitur yang ada pada data di tabel 3.17:

Tabel 3. 17 Hasil Pembentukan Data Sintetis

| Dokumen | air | mayat | pantai | tinggal | tong |
| --- | --- | --- | --- | --- | --- |
| Dokumen-2 | 0 | 1 | 0 | 1 | 0 |
| Dokumen-1 | 0 | 0 | 1 | 0 | 0 |
| Dokumen-3 | 1 | 1 | 0 | 0 | 0 |
| Data sintetis | 0 | 1 | 0 | 0 | 0 |

Dari contoh pembuatan data sintetis tersebut, maka dokumen yang berlabel negatif akan bertambah 4 dokumen, seperti berikut:

Tabel 3. 18 Data Sintetis yang telah terbentuk

| **No** | **Judul** | **Label** |
| --- | --- | --- |
| 1 | “sampah sera objek wisata pantai lombanglombang sorot unjung” | Negatif |
| 2 | “mayat identitas tinggal tengkorak temu warga sumenep” | Negatif |
| 3 | “mayat pria identitas temu air kangean sumenep” | Negatif |
| 4 | “identitas mayat sumenep temu” | Negatif |

Dari tabel 3.18 dokumen yang bernilai negatif masih tidak imbang dengan dokumen yang berlabel positif. Setelah pembuatan data sintetis selesai, maka ulangi proses tersebut sebanyak . Karena maka dibulatkan menjadi 1. Sehingga data sintetis yang akan terbentuk dalam proses sebanyak 1.

### **Metode Naïve Bayes**

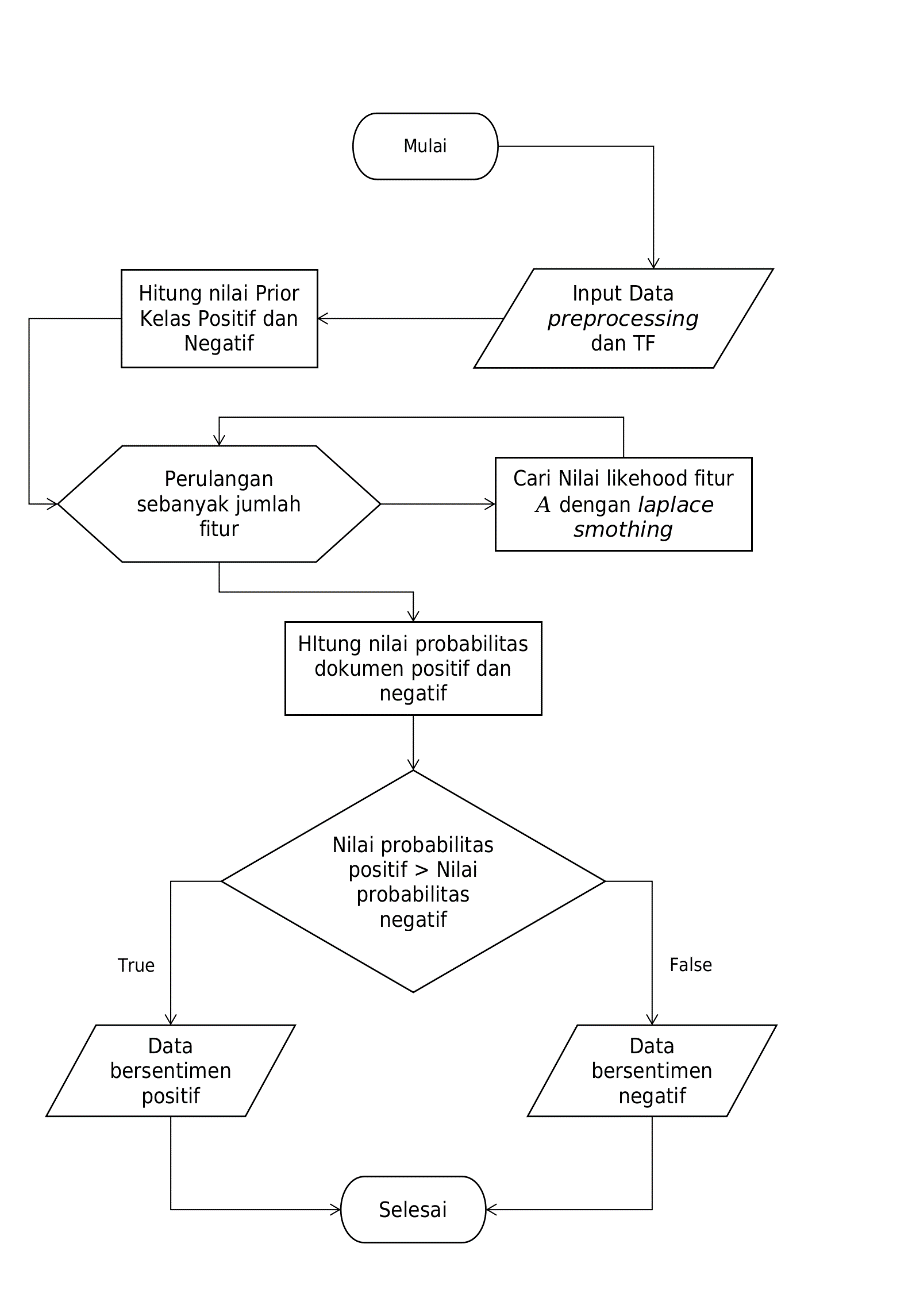
Pada tahapan proses pengklasifikasian menggunakan metode naive bayes. Hal pertama yang dilakukan adalah penginputan data yang telah di bobotkan dengan *term frequency* seperti pada tabel 3.4. Semua kolom pada tabel 3.4 digunakan kecuali kolom *IDF* dan ditambah label dari dokumen tersebut*.* Kemudian data tersebut dibagi menjadi dua bagian, yaitu bagian data testing dan bagian data train, sebagai contoh 1 dokumen menjadi data testing yaitu dokumen ke 8 dan 9 dokumen yang lain menjadi data training.

Kemudian menghitung nilai *likelihood* tiap *term x* di data testingterhadap kelas *c* dengan *laplace smoothing* seperti pada persamaan 2.7:

= 0.0877

…

= 0.0175



Gambar 3. 7 Flowchart klasifikasi metode Naïve Bayes

Ketika semua atribut telah dihitung pada kelas *c* maka nilai *maximum likelihood* tiap *term x* terhadap kelas c selanjutnya. Karena pada penelitian ini hanya mempunyai dua kelas maka hanya terdiri kelas positif dan kelas negatif, seperti yang ada pada tabel 3.19.

Tabel 3. 19 Nilai prior tiap term terhadap kelas

|  |  |  |
| --- | --- | --- |
| ***term*** | **positif** | **negatif** |
| desa | 0.0877 | 0.0175 |
| wisata | 0.105 | 0.0351 |
| baik | 0.0351 | 0.0175 |
| indonesia | 0.0175 | 0.0175 |
| asri | 0.0175 | 0.0175 |
| kagum | 0.0175 | 0.0175 |

Setelah menghitung prior atribut, maka proses selanjutnya menghitung nilai *likehood* pada setiap kelas. Nilai prior pada setiap kelas diperlukan pada persamaan 2.6 seperti berikut:

Kemudian kalikan hasil nilai prior pada setiap kelas dan nilai *likehood* tiap *term x* terhadap kelasnya masing masing untuk memperoleh probabilitas superior yang akan diprediksi seperti berikut:

Karena nilai probabilitaspada kelas positif lebih besar daripada kelas negatif maka dokumen pada *data testing* tersebut termasuk kelas positif.

## **Skenario Pengujian**

Skenario pengujian bertujuan untuk mengetahui apakah sistem telah berjalan sesuai dengan yang diharapkan. Pengujian dilakukan dengan beberapa tahap yaitu:

Tabel 3. 20 Skenario Pengujian

|  |  |
| --- | --- |
| **No** | **SkenarionPengujian** |
| 1 | Pengujian pertama yang dilakukan adalah pengujian komputasi waktu dalam klasifikasi naive bayes menggunakan seleksi fitur *information gain, chi square*, *sequential forward selection, sequential backward elimination* dan tanpa menggunakan seleksi fitur sama sekali. |
| 2 | Pengujian kedua yang dilakukan adalah pengujian nilai akurasi dalam klasifikasi naive bayes menggunakan seleksi fitur *information gain, chi square*, *sequential forward selection,* dan *sequential backward elimination* dan tanpa menggunakan seleksi fitur sama sekali. |
| 3 | Pengujian ketiga yang dilakukan adalah pengujian nilai akurasi dalam dalam klasifikasi naive bayes menggunakan *SMOTE* untuk penyeimbangan data yang berlabel negatif dan positif. Dalam pengujian akan menggunakan data yang telah melalui tahapan seleksi fitur dan tanpa seleksi fitur. |

# **REFERENSI**

[1] N. Fairuuz, F. Nofrian, and D. Desmintari, “Peranan Jumlah Wisatawan Asing, Nilai Tukar, dan PMDN dalam Sektor Pariwisata terhadap Pendapatan Devisa Pariwisata Indonesia,” *J. Indones. Sos. Sains*, vol. 3, no. 4, pp. 694–707, 2022, doi: 10.36418/jiss.v3i4.570.

[2] Z. Mukaffi and T. Haryanto, “Faktor-Faktor Penentu Pariwisata yang Mempengaruhi Pertumbuhan Ekonomi: Tinjauan Sistematis,” *J. Ilm. Univ. Batanghari Jambi*, vol. 22, no. 3, p. 1598, 2022, doi: 10.33087/jiubj.v22i3.2590.

[3] A. Sucipto *et al.*, “Tabel . Perkembangan Kunjungan Wisatawan Mancanegara Ke Indonesia Pintu Masuk Utama,” vol. 2017, no. Irts 2008, pp. 2017–2019, 2018.

[4] U. A. H. (2021) Supriyadi1, Azibur Rahman, Fauzan, “Optimalisasi Pariwisata Syariah Dalam,” vol. 3, no. 2, pp. 56–66, 2021.

[5] F. Taselan, “Wisata Keris di Sumenep Raih Penghargaan Bergengsi,” *mediaindonesia.com*, 2022. https://mediaindonesia.com/nusantara/535016/desa-wisata-keris-di-sumenep-raih-penghargaan-bergengsi

[6] P. K. C. Media, “Mengenal Gili Iyang di Sumenep, Pulau dengan Kadar Oksigen Terbaik Nomor 2 di Dunia,” *Kompas.com*, 2022.

[7] A. A. M. Ramadan, “Lautan Sampah di Pulau Mandangin Madura, Warganet: Pandangan Kurang Mengenakkan Bagi Wisatawan,” *Suarajatim.id*, 2022. https://jatim.suara.com/read/2022/05/14/110141/potret-lautan-sampah-di-pulau-mandangin-madura-warganet-pandangan-kurang-mengenakkan-bagi-wisatawan

[8] R. N. Laily, “Fakta Baru Guru di Bangkalan Jadi Korban Begal, Motif Pelaku Bikin Geleng Kepala,” *Merdeka.com*, 2022. https://www.merdeka.com/jatim/fakta-baru-guru-di-bangkalan-jadi-korban-begal-motif-pelaku-bikin-geleng-kepala.html

[9] E. Pelaksanaan *et al.*, *Diajukan Kepada Fakultas Dakwah Dan Ilmu Komunikasi Universitas Islam Negri Sultan Syarif Kasim Riau Untuk Memenuhi Sebagian Syarat Memperoleh Gelar Sarjana Starta Satu Sosial (S.Sos )*, no. April 2020. 2021.

[10] N. Putu, Y. Anggreswari, G. Ika, and P. Puteri, “Analisis Kaidah Jurnalistik pada Situs Berita Suara.com,” *Ganaya J. Ilmu Sos. dan Hum.*, vol. 4, no. 1, pp. 207–223, 2021, [Online]. Available: https://jayapanguspress.penerbit.org/index.php/ganaya/article/view/1259

[11] A. R. Safitri and M. A. Muslim, “Improved Accuracy of Naive Bayes Classifier for Determination of Customer Churn Uses SMOTE and Genetic Algorithms,” *J. Soft Comput. Explor.*, vol. 1, no. 1, pp. 70–75, 2020, doi: 10.52465/joscex.v1i1.5.

[12] Muladi, U. Pujianto, and U. Qomaria, “Predicting high school graduates using Naive Bayes in State University Entrance Selections,” *4th Int. Conf. Vocat. Educ. Training, ICOVET 2020*, pp. 155–159, 2020, doi: 10.1109/ICOVET50258.2020.9230336.

[13] B. Venkatesh and J. Anuradha, “A review of Feature Selection and its methods,” *Cybern. Inf. Technol.*, vol. 19, no. 1, pp. 3–26, 2019, doi: 10.2478/CAIT-2019-0001.

[14] G. Singh, B. Kumar, L. Gaur, and A. Tyagi, “Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification,” *2019 Int. Conf. Autom. Comput. Technol. Manag. ICACTM 2019*, no. May 2020, pp. 593–596, 2019, doi: 10.1109/ICACTM.2019.8776800.

[15] R. S. Wahono, *Data Mining Data mining*, vol. 2, no. January 2013. 2005. [Online]. Available: https://www.cambridge.org/core/product/identifier/CBO9781139058452A007/type/book\_part

[16] F. Schäfer, C. Zeiselmair, and J. Becker, “2022 IEEE International Conference on Technology Management, Operations and Decisions, ICTMOD 2022,” *2022 IEEE Int. Conf. Technol. Manag. Oper. Decis. ICTMOD 2022*, pp. 190–195, 2022.

[17] R. Ferreira-Mello, M. André, A. Pinheiro, E. Costa, and C. Romero, “Text mining in education,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 9, no. 6, 2019, doi: 10.1002/widm.1332.

[18] T. Jo, *Text Categorization: Approaches*, vol. 45. 2019. doi: 10.1007/978-3-319-91815-0\_6.

[19] I. Chaturvedi, E. Cambria, R. E. Welsch, and F. Herrera, “Distinguishing between facts and opinions for sentiment analysis: Survey and challenges,” *Inf. Fusion*, vol. 44, pp. 65–77, 2018, doi: 10.1016/j.inffus.2017.12.006.

[20] P. Mehta and S. Pandya, “A review on sentiment analysis methodologies, practices and applications,” *Int. J. Sci. Technol. Res.*, vol. 9, no. 2, pp. 601–609, 2020.

[21] E. Y. Ritonga, “Teori Agenda Setting dalam Ilmu Komunikasi,” *J. SIMBOLIKA Res. Learn. Commun. Study*, vol. 4, no. 1, p. 32, 2018, doi: 10.31289/simbollika.v4i1.1460.

[22] I. V. Situmeang, “Media Konvensional dan Media Online,” *Media Konvensional dan Media Online*, p. 118, 2020.

[23] S. Bekti Nugroho, “Pers Berkualitas, Masyarakat Cerdas,” *Dewan Pers*, pp. 1–345, 2013, [Online]. Available: https://dewanpers.or.id/assets/ebook/buku/822-Buku Pers berkualitas masyarakat Cerdas\_final.pdf

[24] V. Rattan, R. Mittal, J. Singh, and V. Malik, “Analyzing the application of SMOTE on machine learning classifiers,” *2021 Int. Conf. Emerg. Smart Comput. Informatics, ESCI 2021*, pp. 692–695, 2021, doi: 10.1109/ESCI50559.2021.9396962.

[25] V. García, J. S. Sánchez, and R. A. Mollineda, “On the effectiveness of preprocessing methods when dealing with different levels of class imbalance,” *Knowledge-Based Syst.*, vol. 25, no. 1, pp. 13–21, 2012, doi: 10.1016/j.knosys.2011.06.013.

[26] Nurhayati, A. E. Putra, L. K. Wardhani, and Busman, “Chi-Square Feature Selection Effect on Naive Bayes Classifier Algorithm Performance for Sentiment Analysis Document,” *2019 7th Int. Conf. Cyber IT Serv. Manag. CITSM 2019*, no. November, 2019, doi: 10.1109/CITSM47753.2019.8965332.

[27] O. Al-Harbi, “A Comparative Study of Feature Selection Methods for Dialectal Arabic Sentiment Classification Using Support Vector Machine,” vol. 19, no. 1, pp. 167–176, 2019, [Online]. Available: http://arxiv.org/abs/1902.06242

[28] M. L. McHugh, “The Chi-square test of independence,” *Biochem. Medica*, vol. 23, no. 2, pp. 143–149, 2012, doi: 10.11613/BM.2013.018.

[29] M. A. Hall and G. Holmes, “Benchmarking Attribute Selection Techniques for Discrete Class Data Mining,” *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 6, pp. 1437–1447, 2003, doi: 10.1109/TKDE.2003.1245283.

[30] Kurniabudi, D. Stiawan, Darmawijoyo, M. Y. Bin Bin Idris, A. M. Bamhdi, and R. Budiarto, “CICIDS-2017 Dataset Feature Analysis with Information Gain for Anomaly Detection,” *IEEE Access*, vol. 8, pp. 132911–132921, 2020, doi: 10.1109/ACCESS.2020.3009843.

[31] D. T. Larose, *Data Mining Methods and Models*. 2006. doi: 10.1002/0471756482.

[32] S. Shafiee, L. M. Lied, I. Burud, J. A. Dieseth, M. Alsheikh, and M. Lillemo, “Sequential forward selection and support vector regression in comparison to LASSO regression for spring wheat yield prediction based on UAV imagery,” *Comput. Electron. Agric.*, vol. 183, no. 1432, p. 106036, 2021, doi: 10.1016/j.compag.2021.106036.

[33] H. Zhang, “The optimality of Naive Bayes,” *Proc. Seventeenth Int. Florida Artif. Intell. Res. Soc. Conf. FLAIRS 2004*, vol. 2, pp. 562–567, 2004.

[34] R. M. Maertens, A. S. Long, and P. A. White, “Performance of the in vitro transgene mutation assay in MutaMouse FE1 cells: Evaluation of nine misleading (‘False’) positive chemicals,” *Environ. Mol. Mutagen.*, vol. 58, no. 8, pp. 582–591, 2017, doi: 10.1002/em.22125.

[35] S. Ruan, H. Li, C. Li, and K. Song, “Class-specific deep feature weighting for naïve bayes text classifiers,” *IEEE Access*, vol. 8, pp. 20151–20159, 2020, doi: 10.1109/ACCESS.2020.2968984.