# **PENERAPAN METODE KLASIFIKASI *K-NEAREST NEIGHBOR* DENGAN EKSTRAKSI FITUR *TF-IDF* UNTUK ANALISIS SENTIMEN PUBLIK BERBASIS LEKSIKON TERHADAP PEMINDAHAN IBU KOTA NEGARA**

**TUGAS AKHIR**

Sebagai syarat untuk memperoleh gelar sarjana S-1 di Program Studi Informatika, Jurusan

Informatika, Fakultas Teknik Industri, Universitas Pembangunan Nasional “Veteran” Yogyakarta



Disusun oleh:

Rifqi Maulana

123200128

**PROGRAM STUDI INFORMATIKA**

**JURUSAN INFORMATIKA**

**FAKULTAS TEKNIK INDUSTRI**

**UNIVERSITAS PEMBANGUNAN NASIONAL “VETERAN” YOGYAKARTA**

**2024**

# **HALAMAN PENGESAHAN PEMBIMBING**

**PENERAPAN METODE KLASIFIKASI *K-NEAREST NEIGHBOR* DENGAN EKSTRAKSI FITUR *TF-IDF* UNTUK ANALISIS SENTIMEN PUBLIK BERBASIS LEKSIKON TERHADAP PEMINDAHAN IBU KOTA NEGARA**

Disusun Oleh:

Rifqi Maulana

123200128



Telah di

periksa

dan

disetujui

oleh pembimbing

pada tanggal : .................................

Menyetujui,

Pembimbing

Rifki Indra Perwira, S.Kom., M.Eng.

NIP 19830708 202121 1001

Mengetahui,

Koordinator Program Studi

Wilis Kaswidjanti, S.Si., M.Kom NIDN. 0513047601

**DAFTAR ISI**

[**PENERAPAN METODE KLASIFIKASI *K-NEAREST NEIGHBOR* DENGAN EKSTRAKSI FITUR *TF-IDF* UNTUK ANALISIS SENTIMEN PUBLIK BERBASIS LEKSIKON TERHADAP PEMINDAHAN IBU KOTA NEGARA** 1](#_Toc181001507)

[**HALAMAN PENGESAHAN PEMBIMBING** 2](#_Toc181001508)

[**DAFTAR ISI** 3](#_Toc181001509)

[**DAFTAR GAMBAR** 6](#_Toc181001510)

[**DAFTAR TABEL** 6](#_Toc181001511)

[**BAB 1** 7](#_Toc181001512)

[**PENDAHULUAN** 7](#_Toc181001513)

[1.1 Latar Belakang 7](#_Toc181001514)

[1.2 Rumusan Masalah 9](#_Toc181001515)

[1.3 Batasan Masalah 9](#_Toc181001516)

[1.4 Tujuan Penelitian 9](#_Toc181001517)

[1.5 Manfaat Penelitian 9](#_Toc181001518)

[1.6 Metodologi Penelitian dan Pengembangan Sistem 10](#_Toc181001519)

[**1.6.1 Metode Pengumpulan Data** 10](#_Toc181001520)

[**1.6.2 Metode Pengembangan Sistem** 10](#_Toc181001521)

[1.7 Sistematika Penulisan 11](#_Toc181001522)

[**BAB II** 12](#_Toc181001523)

[**TINJAUAN LITERATUR** 12](#_Toc181001524)

[2.1 Analisis Sentimen 12](#_Toc181001525)

[2.2 Analisis *Term Frequency Inverse Document Frequency (TF-IDF)* 13](#_Toc181001526)

[2.3 *X (Twitter)* 13](#_Toc181001527)

[2.4 *Python* 14](#_Toc181001528)

[2.5 *Web Scraping* 14](#_Toc181001529)

[2.6 *SMOTE (Synthetic Minority Oversampling Technique)* 14](#_Toc181001530)

[2.7 *Chi Square* 15](#_Toc181001531)

[2.8 *Text Preprocessing* 15](#_Toc181001532)

[**2.8.1 *Cleansing*** 15](#_Toc181001533)

[**2.8.2 *Case Folding*** 16](#_Toc181001534)

[**2.8.3 *Tokenization*** 16](#_Toc181001535)

[**2.8.4 Normalisasi** 16](#_Toc181001536)

[**2.8.5 *Stopwords Removal*** 16](#_Toc181001537)

[**2.8.6 *Stemming*** 16](#_Toc181001538)

[2.9 *Lexicon Based* 17](#_Toc181001539)

[2.10 *K-Nearest Neighbors (KNN)* 18](#_Toc181001540)

[2.11 Validasi dan Pengujian 19](#_Toc181001541)

[2.12 Studi Pustaka 22](#_Toc181001542)

[**BAB III** 25](#_Toc181001543)

[**METODOLOGI PENELITIAN DAN PENGEMBANGAN SISTEM** 25](#_Toc181001544)

# **DAFTAR GAMBAR**

# **DAFTAR TABEL**

[**Tabel 2.1 Confusion Matrix** 20](#_Toc178763373)

[**Tabel 2.2 Studi Pustaka** 23](#_Toc178763374)

# **BAB 1**

# **PENDAHULUAN**

## 1.1 Latar Belakang

Keputusan Presiden Indonesia untuk memindahkan ibu kota negara ke luar Pulau Jawa menjadi salah satu proyek strategis yang tertuang dalam Rencana Pembangunan Jangka Menengah Nasional Tahun Anggaran 2020-2024. Pada 26 Agustus 2019, Presiden yang pada saat itu sedang menjabat mengumumkan ibu kota negara baru ini akan dibangun tepatnya di Kabupaten Penajam Paser Utara dan sebagian Kutai Kartanegara, Kalimantan Timur (Hadi, 2020). Pemindahan ibu kota negara Indonesia merupakan salah satu proyek mega infrastruktur yang paling signifikan dalam beberapa dekade terakhir. Keputusan berani ini tidak hanya memicu perdebatan di kalangan para pengambil kebijakan, tetapi juga memicu beragam reaksi dari masyarakat luas.

Transisi menuju ibu kota baru negara Indonesia merupakan hal yang sangat sensitif sehingga banyak dibicarakan di media sosial, tidak terkecuali pada media sosial *X* (sebelumnya dikenal dengan nama *Twitter*). *X* merupakan media sosial yang seringkali menjadi pusat *trending* mengenai isu di dunia baik itu skala nasional maupun internasional yang dijadikan warganet sebagai media untuk menyuarakan opini terkait sentimen terhadap apa pun yang terkini diperbincangkan di jejaring sosial yang begitu kompleks (Sandi et al., 2023). Sosial media *X* menjadi salah satu platform yang sering digunakan oleh masyarakat Indonesia untuk mengekspresikan pendapat dan respons terhadap peristiwa-peristiwa penting seperti pemindahan ibu kota negara baru. Indonesia menempati posisi ke-5 di dunia sebagai pengguna media sosial *X* terbanyak menurut data dari *Statista* pada bulan Januari 2023 (*Statista*, 2023). *Statista* merupakan salah satu situs *web* yang menyediakan data statistik yang dikenal di seluruh dunia.

Pada penelitian ini akan dilakukan analisa sentimen publik mengenai pemindahan ibu kota negara Indonesia di media sosial *X*. Teknik yang akan digunakan adalah *web scraping* untuk mengumpulkan data teks dari media sosial *X*. Penelitian ini akan dapat mengakses data teks yang mencakup berbagai macam opini, komentar, dan persepsi dari pengguna *X* terkait topik pemindahan ibu kota negara. Hal ini akan memungkinkan penulis untuk memiliki isi *dataset* dalam bentuk teks dan memadai untuk dilakukan analisis sentimen menggunakan metode klasifikasi *K-Nearest Neighbors (KNN)* dengan pelabelan sentimen *Lexicon Based*.

Metode *KNN* memberikan hasil yang kurang baik pada proses klasifikasi data karena terdapat fitur *noise*, namun terdapat beberapa penelitian yang menyebutkan bahwa performansi pada metode *machine learning* dapat menghasilkan performa yang baik ketika dikombinasikan dengan ekstraksi fitur dan seleksi fitur yang tepat (Pratomo et al., 2021). Penggabungan tersebut menawarkan pendekatan yang memadai dalam memproses dan menganalisis data teks untuk mengekstraksi informasi sentimen. Penerapan metode ini membuat kita dapat mengidentifikasi pandangan, sikap, dan respon yang diberikan publik terkait dengan topik pemindahan ibu kota negara tersebut.

Berdasarkan paparan sebelumnya, solusi yang penulis gunakan adalah metode klasifikasi *K-Nearest Neighbors* dengan pelabelan sentimen *Lexicon Based,* penyeimbang kelas dengan *SMOTE*,serta ekstraksi fitur dan seleksi fitur *TF-IDF* dan *Chi Square*. Tahapan awal pada penelitian ini adalah *data preprocessing* bertujuan agar data yang nanti diproses lebih terstruktur, selanjutnya proses pelabelan dan ekstraksi fitur serta seleksi fitur menggunakan metode leksikon, *TF-IDF*, dan *Chi Square*, tidak lupa juga menyeimbangkan kelas dengan *SMOTE*, lalu terakhir yaitu tahap klasifikasi. Penulis memilih beberapa teknologi yang telah disebutkan sebelumnyakarena penulis sudah melakukan tinjauan pustaka dari beberapa metode sejenis sebelumnya. Penelitian yang dilakukan oleh Setyo Adji Pratomo, dkk pada tahun 2021, dengan judul “Analisis Sentimen Pengaruh Kombinasi Ekstraksi Fitur TF-IDF dan LexiconPada Ulasan Film Menggunakan Metode KNN” Berkesimpulan bahwa penggabungan fitur ekstraksi *TF-IDF* dengan leksikon *SentiWordNet* memiliki hasil akurasi yang tidak lebih tinggi dibandingkan dengan hanya menggunakan fitur ekstraksi *TF-IDF* yaitu 73.31%, sedangkan dengan *TF-IDF* saja mendapatkan 81.04%, dan penggunaan fitur seleksi *Information Gain* (*IG*) dengan *threshold* yang tepat mampu mengoptimasi hasil performansi pada metode klasifikasi *KNN*. Lalu, penelitian yang dilakukan oleh Azhar pada tahun 2018, dengan judul “Analisis Kinerja Algoritma Naïve Bayes dan K-Nearest Neighbor Pada Sentimen Analisis dengan Pendekatan Lexicon di Media Twitter”. *Dataset* yang digunakan merupakan data dari sosial media *X* (*Twitter*) dengan menggunakan *Twitter API*. Proses *Natural Language Processing* yang digunakan adalah *case folding*, *filtering*, *tokenizing*, normalisasi, *stopwords*, dan *stemming*. Hasilnya nilai *KNN* pada *k*=5 dengan tingkat akurasi mencapai 77%. Terakhir, penelitian yang dilakukan oleh Muhammad Rayhan Elfansyah, dkk pada tahun 2024 dengan judul “Perbandingan Metode K-Nearest Neighbor (KNN) Dan Naïve Bayes Terhadap Analisis Sentimen Pada Pengguna E-Wallet Aplikasi Dana Menggunakan Fitur Ekstraksi TF-IDF” berkesimpulan bahwa metode *KNN* dan *Naïve Bayes* memiliki akurasi yang berbeda berdasarkan sumber label data. Pada data yang diberi label model *Lexicon*, akurasi *KNN* mencapai 78% dan *Naïve Bayes* 74%.

Berdasarkan latar belakang yang telah dipaparkan, penulis berharap penelitian ini dapat memberikan kontribusi dalam memahami bagaimana mengoptimalkan metode klasifikasi *K-Nearest Neighbors* dengan melakukan beberapa penyesuaian yang tepat dari mulai pra pengolahan dataset sampai ke tahap klasifikasi sentimen untuk memahami reaksi publik, khususnya warganet di *X* (*Twitter*), terhadap pemindahan ibu kota negara Indonesia, apakah lebih cenderung positif, netral, atau negatif. Hal ini dilakukan dengan mengintegrasikan data teks yang telah dibersihkan untuk diproses lebih lanjut oleh teknik pelabelan sentimen berbasis leksikon yang telah dikustomisasi, serta mencari nilai *k* yang optimal sebagai variabel penting dalam optimasi metode klasifikasi *KNN,* sebelum itu dilakukan penyeimbangan kelas dengan *SMOTE*,tidak lupa juga menyertakan ekstraksi fitur serta seleksi fitur *TF-IDF* dan *Chi Square* untuk memperoleh prediksi sentimen dengan akurasi tinggi. Selain itu, penulis berharap penelitian ini dapat memberikan kontribusi kecil dalam memahami bagaimana tanggapan serta dinamika sosial yang muncul saat terjadi perubahan besar, seperti pemindahan ibu kota. Hasil penelitian ini diharapkan dapat menjadi referensi bagi penelitian serupa di masa depan.

## 1.2 Rumusan Masalah

Berdasarkan uraian latar belakang sebelumnya, masalah yang dapat dirumuskan penulis mencakup bagaimana optimasi metode tersebut untuk mendapatkan hasil akhir akurasi yang tinggi serta pandangan dan sikap masyarakat atau publik terhadap keputusan Presiden terkait pemindahan ibu kota negara Indonesia?

## 1.3 Batasan Masalah

Agar masalah yang diteliti menghasilkan sasaran yang jelas, maka dibuatlah batasan masalah untuk menghindari adanya perluasan pembahasan kedepannya sebagai berikut:

1. Data penelitian yang digunakan merupakan data yang dihasilkan dari *web* *scraping* pada unggahan teks di media sosial *X* (*Twitter*), mengenai sentimen publik atas keputusan Presiden terkait pemindahan ibu kota negara.
2. Dataset yang digunakan dalam penelitian berjumlah total 5341 baris unggahan teks memakai format *.csv*.
3. Kategori sentimen dibagi menjadi 3 yaitu positif, negatif, serta netral.
4. Leksikon yang digunakan adalah leksikon berbahasa Indonesia yaitu *InSet* (*Indonesia Sentiment Lexicon*) dari literatur yang disusun oleh Fajri Koto dan Gemala Y. Rahmaningtyas pada tahun 2017 dengan judul “InSet Lexicon: Evaluation of a Word List for Indonesian Sentiment Analysis in Microblogs”. Leksikon *InSet* terdiri atas 3,609 kata positif dan 6,609 kata negatif dengan bobot antara -5 sampai +5.

## 1.4 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah mengimplementasikan serta mengoptimasi metode klasifikasi *K-Nearest Neighbors* dengan pelabelan sentimen *Lexicon Based* dalam melakukan analisa sentimen masyarakat berdasarkan unggahan teks di media sosial *X* (*Twitter*) mengenai keputusan Presiden terkait pemindahan ibu kota negara.

## 1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat bermanfaat untuk memberikan kontribusi dalam pengembangan metodologi analisis sentimen berbasis teks, khususnya dalam penerapan algoritma klasifikasi *K-Nearest Neighbors* dengan pelabelan sentimen *Lexicon Based* dalam konteks bahasa Indonesia. Penelitian ini juga diharapkan tidak hanya bermanfaat secara ilmiah sebagai referensi bagi penelitian-penelitian selanjutnya yang menggunakan pendekatan serupa, penelitian ini juga mendukung pengembangan algoritma *Natural Language Processing* (*NLP*) untuk teks berbahasa Indonesia. Selain itu, penelitian ini juga merupakan bagian dari pemenuhan salah satu syarat kelulusan strata satu (S1) Program Studi Informatika Fakultas Teknik Industri.

## 1.6 Metodologi Penelitian dan Pengembangan Sistem

Metode penelitian ini menggunakkan metode penelitian kuantitatif. Metode penelitian kuantitatif merupakan penelitian empiris dimana data dalam bentuk sesuatu yang dapat dihitung atau angka (Punch, 1988). Berikut merupakan tahapan-tahapan penelitian yang dilakukan:

### **1.6.1 Metode Pengumpulan Data**

Penelitian ini menggunakan teknik *web scraping* untuk mengumpulkan data teks yang berkaitan dengan pemindahan ibu kota negara dari platform media sosial *X* (*Twitter*).

### **1.6.2 Metode Pengembangan Sistem**

Metode *Waterfall* merupakan salah satu model *SDLC (Software Development Life Cycle)* yang sering digunakan dalam pengembangan sistem informasi atau perangkat lunak. Metode ini menggunakan pendekatan sistematis dan berurutan. Tahapan dalam model ini dimulai dari tahap perencanaan hingga tahap pengelolaan (*maintenance*) dan dilakukan secara bertahap (Abdul Wahid, 2020). Tahapan metode *waterfall* adalah sebagai berikut:

1. *Requirements Analysis and Definition*

Merupakan tahapan awal yang melibatkan identifikasi dan pemahaman yang mendalam terhadap kebutuhan. Tujuan utamanya yaitu mengumpulkan persyaratan fungsional dan non-fungsional yang nantinya akan menjadi dasar dari pengembangan sistem.

1. *System and Software Design*

Tahapan perancangan sistem ini mengalokasikan kebutuhan-kebutuhan sistem pada perangkat keras maupun perangkat lunak dengan membentuk arsitektur sistem secara keseluruhan.

1. *Implementation*

Pada tahap ini, perancangan perangkat lunak direalisasikan sebagai serangkaian program.

1. *System Testing*

Merupakan tahap pengujian terhadap sistem yang telah dibuat yang bertujuan untuk mengetahui apakah sistem yang dibuat udah siap digunakan atau belum.

## 1.7 Sistematika Penulisan

Sistematika penulisan pada penelitian ini adalah sebagai berikut:

|  |  |
| --- | --- |
| BAB I | PENDAHULUAN  Pada bagian pendahuluan membahas mengenai latar belakang masalah, perumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, metodologi penelitian, serta sistematika penulisan. |
| BAB II | TINJAUAN PUSTAKA  Tinjauan pustaka merupakan bagian yang memuat mengenai dasar teori yang digunakan untuk analisis serta perancangan sistem dan juga implementasi pada penelitian ini. Selain itu juga digunakan sebagai bahan referensi serta pondasi untuk memperkuat argumentasi pada penelitian ini. |
| BAB III | METODOLOGI PENELITIAN DAN PENGEMBANGAN SISTEM  Pada bab ini membahas mengenai metodologi penelitian, analisis sistem dan perancangan sistem analisis sentimen. |
| BAB IV | HASIL, PENGUJIAN, DAN PEMBAHASAN  Bab ini menyajikan hasil dari penelitian yang berisi hasil implementasi dari perancangan yang telah dibuat pada bab sebelumnya. Selain itu berisi pengujian terhadap hasil penelitian beserta pembahasannya. |
| BAB V | KESIMPULAN DAN SARAN  Bab ini berisi kesimpulan dari hasil penelitian serta saran yang diajukan oleh penulis untuk pengembangan pada penelitian selanjutnya. |

# **BAB II**

# **TINJAUAN LITERATUR**

## 2.1 Analisis Sentimen

Analisis sentimen adalah bidang studi yang menganalisis pendapat, sentimen, evaluasi, penilaian, sikap dan emosi seseorang terhadap sebuah produk, organisasi, individu, masalah, peristiwa atau topik (Liu, 2012). Tujuan dari analisis sentimen yaitu untuk memahami opini, perasaan, serta pandangan yang terkandung pada teks atau data unstruktural lainnya. Pengaruh dan manfaat dari analisis sentimen menyebabkan penelitian mengenai analisis sentimen berkembang pesat, serta kurang lebih 20-30 perusahaan di Amerika berfokus pada layanan analisis sentimen (Liu, 2012). Manfaat sentimen analisis dalam dunia usaha antara lain untuk melakukan pemantauan terhadap suatu produk. Secara cepat dapat digunakan sebagai alat bantu untuk melihat respon masyarakat terhadap suatu produk, sehingga dapat diambil langkah strategis berikutnya. Garis besar analisis sentimen itu sendiri bertujuan untuk mengekstrak atribut dan komponen dari beberapa komentar yang ada di media sosial dan sehingga dapat menentukan beberapa kelas positif, negatif dan netral (Permatasari et al., 2021).

Pada umumnya, sentimen analisis merupakan klasifikasi tetapi kenyataannya tidak semudah proses kualifikasi biasa karena terkait penggunaan bahasa, dimana terdapat ambigu dalam penggunaan kata serta perkembangan bahasa itu sendiri.

Menurut Liu (2012), analisis sentimen memiliki beberapa tahap untuk melakukan analisis sentimen, yaitu:

1. Level Dokumen

Level dokumen menganalisis satu dokumen penuh dan mengklasifikasikan dokumen tersebut memiliki sentimen positif atau negatif. Level analisis ini berasumsi bahwa keseluruhan dokumen hanya berisi opini tentang satu entitas saja. Level analisis ini tidak cocok diterapkan pada dokumen yang membandingkan lebih dari satu entitas.

1. Level Kalimat

Level kalimat menganalisis satu kalimat dan menentukan tiap kalimat sentimen bernilai positif, netral, atau negatif. Sentimen netral berarti kalimat tersebut bukan opini.

3. Level Aspek

Level aspek tidak melakukan analisis pada konstruksi bahasa (dokumen, paragraf, kalimat, klausa, atau frasa) melainkan melakukan langsung pada opini itu sendiri. Hal ini didasari bahwa opini terdiri dari sentimen (positif dan negatif) dan target dari opini tersebut. Tujuan level analisis ini adalah untuk menemukan sentimen entitas pada tiap aspek yang dibahas.

## 2.2 Analisis *Term Frequency Inverse Document Frequency (TF-IDF)*

*K-Nearest TF-IDF* atau *Term Frequency Inverse Document Frequency* merupakan metode pembobotan dengan menggabungkan metode *TF* dan *IDF*, metode ini memberikan bobot hubungan suatu kata terhadap dokumen (Wahyuni et al., 2017).

Proses Frekuensi dokumen yang mengandung kata tersebut menunjukkan seberapa umum kata tersebut. Sehingga bobot hubungan antara sebuah kata dan sebuah dokumen akan tinggi apabila frekuensi kata tersebut tinggi di dalam dokumen dan frekuensi keseluruhan dokumen yang mengandung kata tersebut yang rendah pada kumpulan dokumen.

**Nilai *TF* dapat dihitung dengan rumus:**

**Nilai *IDF* dapat dihitung dengan rumus**:

**Nilai *TF-IDF*:**

## 2.3 *X (Twitter)*

*X* adalah sebuah situs *web* yang dimiliki dan dioperasikan oleh Twitter Inc., yang menawarkan jaringan sosial berupa *microblog* sehingga memungkinkan penggunanya untuk mengirim dan membaca pesan dengan sebutan *tweet* atau kicauan(Akbar et al., 2013).

*Microblog* adalah jenis alat komunikasi daring dengan manfaat agar pengguna dapat memperbarui status tentang mereka yang sedang memikirkan dan melakukan sesuatu, apa pendapat mereka tentang suatu objek atau fenomena tertentu. *Tweet* atau kicauan adalah teks tulisan hingga 140 karakter (atau lebih jika berlangganan fitur khusus pada platform tersebut) yang ditampilkan pada halaman profil pengguna. *Tweet* bisa dilihat secara publik atau dapat dibatasi pengiriman pesan ke daftar pengguna lain tertentu saja. Pengguna dapat melihat *tweet* pengguna lain yang dikenal dengan sebutan pengikut atau *followers* (Ramadhon, 2020).

## 2.4 *Python*

*Python* adalah bahasa pemrograman tingkat tinggi yang ditafsirkan, berorientasi objek, dengan semantik dinamis (Stefana, 2022). Struktur data bawaan tingkat tinggi, dikombinasikan dengan pengetikan dan pengikatan dinamis membuatnya sangat menarik untuk *Rapid Application Development*, serta digunakan sebagai bahasa skrip atau lem untuk menghubungkan komponen yang ada bersama-sama.

Sintaks *python* yang sederhana dan mudah dipelajari menekankan keterbacaan dan karenanya mengurangi biaya pemeliharaan program. *Python* mendukung modul dan paket, yang mendorong modularitas program dan penggunaan kembali kode. Penerjemah *python* dan perpustakaan standar yang luas tersedia dalam bentuk sumber atau biner tanpa biaya untuk semua *platform* utama, dan dapat didistribusikan secara bebas (*Python – Wikipedia, 2017*).

## 2.5 *Web Scraping*

*Web scraping* merupakan sebuah teknik untuk mendapatkan informasi dari situs web secara otomatis tanpa harus menyalinnya secara manual (Ayani et al., 2019). Teknik *web scraping* memungkinkan konten utama yang terdapat pada situs dapat diekstraksi, dihimpun, kemudian dapat diproses. *Web scraping* akan melakukan ekstraksi data pada *World Wide Web* (*www)*, lalu data yang didapat akan disimpan pada *file system* atau basis data yang nantinya bisa diambil kembali atau dianalisis (Setiawan et al., 2020).

Cara kerja *web scraping* adalah dengan mengakses halaman pada *web*, menentukan data yang dalam halaman tersebut, melakukan ekstraksi, dan transformasi bila diperlukan, kemudian menyimpan data tersebut menjadi dataset terstruktur (Boeing, 2016).

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

Metode *SMOTE (Synthetic Minority Oversampling Technique)* adalah metode populer yang digunakan untuk menangani ketidakseimbangan kelas, teknik ini mensintesis sampel baru dari kelas minoritas untuk menyeimbangkan kumpulan data dengan melakukan *resampling* kelas minoritas (Siringoringo, 2018). Contoh proses *SMOTE* (Barus, 2022):

1. Ambil sampel acak contohnya P1
2. Kemudian terapkan algoritma *KNN* pada P1
3. Ambil jarak tetangga terdekat P1, contohnya P2
4. Lalu *generate* data baru P1’ = P1 + *rand*(0,1) x (P2 – P1)
5. Ulang terus proses tersebut hingga jumlah data minoritas sebanyak data mayoritas

## 2.7 *Chi Square*

*Chi Square* merupakan salah satu seleksi fitur yang biasa digunakan pada pengklasifikasi teks dimana seleksi fitur ini berguna dalam mereduksi fitur yang tidak relevan (Winata et al., 2022). Seleksi fitur dengan *Chi Square* memiliki keunggulan dalam meningkatkan keakuratan seperti dapat mereduksi dengan data yang besar dan tidak mengurangi akurasi sebelumnnya. Proses perhitungan yang dilakukan pada seleksi fitur *Chi Square* ini menggunakan teori statistika dalam pengujiannya dan mengevaluasi antar kata apakah saling berkaitan atau tidak (Winata et al., 2022). Adapun persamaan dari *Chi Square* adalah sebagai berikut:

**Keterangan:**

|  |  |
| --- | --- |
| t | : term atau kata |
| c | : kategori/kelas pada dokumen |
| N | : jumah keseluruhan dokumen |
| A | : banyaknya jumlah dokumen dalam kategori *c* yang mengandung kata *t* |
| B | : banyaknya jumlah dokumen yang bukan kategori *c* tetapi mengandung kata *t* |
| C | : banyaknya jumlah dokumen dalam kategori *c* tetapi tidak mengandung kata *t* |
| D | : banyaknya jumlah dokumen yang bukan kategori *c* dan tidak mengandung kata *t* |

Setelah proses penilaian pada tiap kata diketahui, kemudian mengurutkan kata berdasarkan nilai *Chi Square* dari nilai tertinggi hingga terendah dimana semakin besar nilainya maka semakin penting fitur tersebut untuk klasifikasi. Oleh karena itu, semakin kecil nilai taraf nyata (α) akan semakin besar nilai kritis. Jika nilai taraf nyata (α) semakin kecil yang diterapkan, maka banyak fitur yang akan terseleksi atau terbuang.

## 2.8 *Text Preprocessing*

Tahapan *preprocessing* merupakan tahapan awal untuk mempersiapkan dokumen agar lebih mudah untuk diproses, tahapan *preprocessing* sebelum proses klasterisasi meliputi *cleansing, case folding*, *tokenizing*, *filtering*, dan *stemming* (Amalia et al., 2018). Berdasarkan hal tersebut, mengubah data yang sebelumnya tidak terstruktur memerlukan proses pengubahan menjadi data yang terstruktur untuk diproses pada langkah berikutnya. Berikut adalah beberapa tahapan dari *preprocessing*:

### **2.8.1 *Cleansing***

*Cleansing* merupakan proses pembersihan data yang bertujuan untuk menghilangkan elemen-elemen yang tidak dibutuhkan pada sebuah dokumen. *Cleansing* bertujuan untuk memperbaiki kualitas data. Pada penelitian ini tahapan *cleansing* berfungsi untuk menghilangkan *mention*, *hashtag*, *url* dan *uri*, tanda baca, *emoji*, serta menghilangkan angka-angka.

### **2.8.2 *Case Folding***

*Case folding* merupakan tahapan yang bertujuan untuk mengubah semua huruf yang terdapat pada dokumen menjadi huruf kecil. Huruf ‘a’ sampai ‘z’ yang akan diterima, apabila pada dokumen terdapat karakter selain huruf maka akan dihilangkan dan dianggap *delimiter*. Tahap *case folding* akan menghasilkan kalimat yang sudah rapi dan akan memudahkan proses selanjutnya.

### **2.8.3 *Tokenization***

*Tokenization* atau tokenisasi merupakan tahapan pemisahan teks menjadi potongan-potongan berupa huruf, kata, atau kalimat menjadi kata yang tidak terhubung. Data teks yang masuk pada tahap *tokenization* akan diubah menjadi potongan-potongan kata. Pada umumnya, karakter spasi membedakan atau mengidentifikasi setiap kata satu sama lain. Sehingga, proses *tokenization* pada dokumen bergantung pada karakter spasi. Sebagai contoh, tokenisasi pada kalimat “ibu kota pindah” menghasilkan tiga *token*, yaitu “ibu”, “kota”, dan “pindah”.

### **2.8.4 Normalisasi**

Proses ini dilakukan untuk merubah kata-kata singkatan atau slangdalam bahasa Indonesia menjadi kata baku, serta mengontrol kata negasi.

### **2.8.5 *Stopwords Removal***

Penghapusan *stopwords* bertujuan untuk menghilangkan kata yang dianggap tidak memiliki makna atau tidak relevan yang terdapat pada dokumen. Contoh *stopwords* yang ada pada Bahasa Indonesia seperti “yang”, “dan”, “di”, “itu”, “adapun”, “agak” dan sebagainya. Kata-kata umum tersebut tidak mempunyai nilai pada sebuah dokumen, sehingga kata-kata yang termasuk dalam kamus *stopwords* dihilangkan sehingga ukuran data juga akan berkurang.

### **2.8.6 *Stemming***

*Stemming* merupakan proses yang bertujuan untuk menemukan kata dasar dari sebuah kata dengan menghapus imbuhan (*affixes*), termasuk awalan (*prefixes*), sisipan (*infixes*), akhiran (*suffixes*), serta kombinasi dari awalan dan akhiran (*confixes*) pada kata turunan. Tujuan utama dari proses *stemming* ini adalah mengubah bentuk kata menjadi kata dasar sesuai dengan bahasa Indonesia yang baik dan benar. Menurut algoritma Nazief & Adriani memiliki tahap-tahap sebagai berikut:

1. Mencari kata yang akan dilakukan *stemming* pada kamus, apabila ditemukan maka diasumsikan bahwa kata tersebut merupakan kata akar (*root word)*, jika terbukti maka proses diberhentikan pada tahap pertama.
2. Menghilangkan *inflection suffixes*, sebuah kata yang mengandung *inflection suffixes* yaitu apabila memiliki imbuhan “-lah”, “-ku”, “-kah”, “-mu”, atau “-nya”, apabila kata berupa *particles* atau dalam kata lain yang mengandung “-lah”, “-kah”, atau “-pun”, maka langkah untuk menghilangkan *inflection suffixes* diulangi agar dapat menghilangkan *possessive pronouns*, yang termasuk imbuhan *possessive pronouns* adalah “-ku”, “-mu”, atau “-nya”.
3. Menghapus *derivation suffixes* atau imbuhan turunan, yang termasuk pada kata imbuhan turunan adalah “-i”, “-an”, atau “-kan”.
4. Menghapus *derivational prefix* atau imbuhan yang berada pada awal kata, yang dimaksud dengan *derivational prefix* adalah “be-”, “di-”, “ke-”, “me-”, “pe-”, “se-”, dan “te-”.
5. Apabila 4 langkah tersebut telah dilakukan tetapi belum berhasil menemukan kata dasar maka algoritma ini akan melakukan analisis apakah kata tersebut termasuk ke dalam tabel ambiguitas kolom terakhir.
6. Apabila belum berhasil maka algoritma akan dikembalikan pada kata aslinya.

*Stemming* pada *phyton* dilakukan melalui kelas *StemmerFactory* yaitu sebuah kelas yang terdapat pada *library* yang bernama “*Sastrawi”* dan kompatibel dengan *input* berbahasa Indonesia, yang mana *library* ini akan lebih dulu di-*import* sebelum meng-*import* kelas *StemmerFactory*.

## 2.9 *Lexicon Based*

Metode berbasis *Lexicon* merupakan metode yang sederhana, layak, dan praktis untuk analisis sentimen dari data media sosial. Data yang cocok dengan metode *Lexicon Based* yaitu data kuesioner, data *Twitter*, data *Facebook*, atau media sosial lainnya yang berupa opini pelanggan tentang suatu produk atau pelayanan jasa (Matulatuwa et al., 2017).

Metode *Lexicon Based* didasarkan pada asumsi bahwa orientasi sentimen kontekstual adalah jumlah dari orientasi sentimen setiap kata atau frasa. Metode ini dapat digunakan untuk melakukan ekstraksi sentimen dari blog dengan mengombinasikan *lexical knowledge* dan klasifikasi teks. Metode *Lexicon* dapat dibuat secara manual atau diperluas secara otomatis dari *seed of words* (Matulatuwa et al., 2017).

Penentuan label sentimen dilakukan pada data teks berupa kalimat yang memiliki kata pada kamis *lexicon* yang terdiri dari kata negatif dan positif. Kata yang teridentifikasi dalam kamus *lexicon* akan dihitung skornya sesuai dengan jumlah kata pada setiap teks atau kalimat (Ismail et al., 2023).

*S positive* adalah bobot dari kalimat yang didapatkan melalui penjumlahan *n* skor polaritas kata opini positif dan *S negative* adalah bobot dari kalimat yang didapatkan melalui penjumlahan *n* skor polaritas kata opini negatif. Oleh karena itu, dari persamaan nilai sentimen dalam satu kalimat diperoleh persamaan untuk menentukan orientasi sentimen dengan perbandingan jumlah nilai positif, negatif, dan netral (Ismail et al., 2023).

Jika dalam suatu teks memiliki jumlah kata positif lebih banyak dari kata negatif, maka data teks tersebut akan dilabeli sentimen positif. Jika dalam suatu teks memiliki jumlah kata positif lebih sedikit dari kata negatif, maka data teks tersebut akan dilabeli sentimen negatif. Jika dalam suatu teks memiliki jumlah kata positif sama dengan kata negatif, maka data teks tersebut akan dilabeli sentimen netral (Ismail et al., 2023).

|  |
| --- |
| 2.10 *K-Nearest Neighbors (KNN)* |

*KNN* adalah algoritma *machine learning classifier* populer yang paling sederhana yang pertama kali diperkenalkan oleh T. Cover dan P. Hart pada tahun 1967 dimana algoritma ini mengklasifikasikan kelas sampel berdasarkan kelas tetangga terdekatnya (Fajri et al., 2020). *K-Nearest Neighbors* sendiri memiliki prinsip sederhana, bekerja berdasarkan jarak terpendek dari sampel uji ke sampel latih (Sari, 2020). Algoritma *KNN* bekerja dengan cara menghitung jarak tiap titik pada data tes dengan data latihan tiap kelas. Lalu, diurutkan dari jarak terdekat ke jarak terjauh dan akan dipilih jarak terdekat antara data tes dengan data latihan sejumlah *k*. Kelas yang memiliki jarak terdekat dengan data tes akan menjadi kelas data tes tersebut (Raschka, 2016). Adapun tahapan proses yang dilakukan pada *KNN* menurut Abdillah, (2023) adalah sebagai berikut:

1. Hitung jarak antara sampel yang tidak diketahui dengan semua sampel pada set data pelatihan menggunakan rumus jarak yang dipilih, didalam kasus ini digunakan *Cosine Similarity*.
2. Pilih *k* tetangga terdekat dari sampel yang tidak diketahui berdasarkan jarak yang telah dihitung.
3. Hitung label kelas mayoritas dari *k* tetangga terdekat. Dalam kasus klasifikasi biner, label mayoritas dapat dihitung dengan menghitung frekuensi masing-masing kelas pada *k* tetangga terdekat dan memilih kelas dengan frekuensi yang paling tinggi. Dalam kasus klasifikasi multikelas, label mayoritas dihitung dengan metode voting, yaitu dengan menghitung jumlah suara setiap kelas pada *k* tetangga terdekat dan memilih kelas dengan jumlah suara terbanyak.
4. Kembalikan label kelas mayoritas sebagai hasil klasifikasi untuk sampel yang tidak diketahui.

Berdasarkan langkah sebelumnya telah diketahui bobot tiap kata. Langkah selanjutnya adalah menghitung jarak atau tingkat kemiripan data dengan setiap data latih yang ada menggunakan rumus jarak *Cosine Similarity*. Lalu, sistem akan mengurutkan nilai jarak dari yang tertinggi sampai terendah. Kelebihan dari algoritma *Cosine Similarity* adalah tidak terpengaruh pada panjang pendeknya suatu dokumen dan memiliki tingkat akurasi yang tinggi. Tahapan pada *Cosine similarity* adalah sebagai berikut (Abdillah, 2023):

1. Kalikan bobot dari setiap *term* pada D1 dengan setiap *term* dari semua dokumen data latih yang ada.
2. Hasil perkalian D1 dengan setiap dokumen kemudian dijumlahkan.
3. Hitung hasil kuadrat dari masing-masing *term* dalam setiap dokumen (termasuk D1) kemudian jumlahkan lalu diakarkan.
4. Lakukan pembagian antara hasil dari langkah nomor 2 dengan langkah nomor 3. Maka, didapatkan nilai *Cosine Similarity*. Berikut adalah rumus *Cosine Similarity*:

**Keterangan:**

: kemiripan Q terhadap dokumen D

Q : data uji

D : data latih

n : jumlah data latih

## 2.11 Validasi dan Pengujian

Validasi dan pengujian sangat diperlukan untuk menilai kinerja dari sebuah sistem. Kinerja proses klasifikasi menggambarkan seberapa baik sistem dalam melakukan klasifikasi data. Pengukuran kinerja proses tersebut dapat dilakukan menggunakan *confusion matrix* (Abdillah, 2023).

*Confusion matrix* merupakan suatu matriks yang digunakan untuk menganalisa keakuratan dari model klasifikasi yang dibuat untuk mengidentifikasi data dengan kelas yang berbeda (Afrillia et al., 2022). Pengujian dengan *confusion matrix* ini digunakan untuk menghitung nilai *true positive*, *false positive*, *true negative*, serta *false negative* yang nantinya dapat digunakan untuk pengukuran nilai akurasi, presisi, serta *recall*. Dengan melakukan pengukuran Tingkat akurasi maka dapat mengetahui seberapa baik performa model klasifikasi tersebut. Bentuk dari *confusion matrix* adalah tabel dengan empat kombinasi yang berbeda antara nilai prediksi dan nilai aktual. Metode *confusion matrix* mempunyai empat kemungkinan yang merepresentasikan hasil dari proses klasifikasi, yaitu TP (*True Positive*), TN (*True Negative*), FP (*False Positive*), dan FN (*False Negative*).

Berikut merupakan tabel *confusion matrix*:

**Tabel 2.1 Confusion Matrix**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | ***Predicted*** | |  |
| **-1 *(Negative)*** | **+1 *(Positive)*** | **0 *(Neutral)*** |
| ***Actual*** | -1 *(Negative)* | TN | FP | FL2 |
| +1 *(Positive)* | FN | TP | FP2 |
| 0 (*Neutral*) | FN2 | FP2 | TL |

**Keterangan:**

|  |  |
| --- | --- |
| TN (*True Negative*) | : Prediksi benar bernilai negatif |
| TP (*True Positive*) | : Prediksi benar bernilai positif |
| TL (*True Neutral*) | : Prediksi benar bernilai netral |
| FN (*False Negative*) | : Negatif terprediksi positif |
| FP (*False Positive*) | : Positif terprediksi negatif |
| FL (*False Neutral*) | : Positif terprediksi netral |

FN2 (*False Negative* 2) : Netral terprediksi negatif

FP2 (*False Positive* 2) : Netral terprediksi positif

FL2 (*False Neutral* 2) : Negatif terprediksi netral

Tabel *confusion matrix* di atas dapat digunakan dalam perhitungan *performance matrix* yang bertujuan untuk mengukur model yang digunakan untuk dapat memperoleh nilai *accuracy*, *recall*, *precision,* dan *f1-score*.

1. *Accuracy*

Akurasi merupakan nilai yang menunjukan kedekatan antar nilai prediksi dan nilai aktual. Perhitungan akurasi dengan cara membagi jumlah data yang akan diklasifikasi secara tepat dengan totalsampel data *testing* yang diuji. Berikut merupakan rumus perhitungan nilai akurasi:

1. *Recall*

*Recall* merupakan perbandingan jumlah data yang dilakukan prediksi pada kelas positif yang benar dengan jumlah data yang diharapkan berada pada kelas positif. *Recall* dikatakan sebagai tingkat keberhasilan model dalam menemukan informasi. Berikut merupakan rumus perhitungan nilai *recall*:

1. *Precision*

*Precision* merupakan tingkat akurasi antar informasi yang diinginkan pengguna serta respon sistem. Nilai presisi menunjukkan data positif yang diklasifikasi dengan tepat kemudian dilakukan pembagian dengan jumlah data positif yang diklasifikasi. Berikut merupakan rumus perhitungan nilai presisi:

1. *F1-Score*

*F1-Score* merupakan perbandingan rata-rata *precision* dan *recall* yang telah dibobotkan. Nilai terbaik *F1-Score* adalah 1 dan nilai terburuknya yaitu 0. Perhitungan yang didapatkan berupa informasi bahwa model klasifikasi memiliki *precision* dan *recall* yang baik. Berikut merupakan rumus perhitungan nilai *F1-Score:*

## 2.12 Studi Pustaka

Penelitian di bawah ini merupakan penelitian-penelitian yang telah dilakukan sebelumnya dan berkaitan dengan penelitian tugas akhir ini, sehingga menjadi referensi dalam penelitian ini.

**Tabel 2. 2 Studi Pustaka**

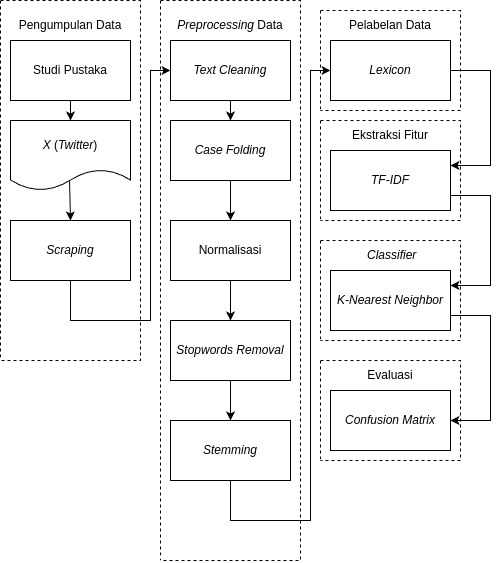
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No** | **Penulis** | **Judul** | **Metode** | **Hasil** |
| 1. | Elfansyah et al., 2024. | Perbandingan Metode K-Nearest Neighbor (KNN) Dan Naïve Bayes Terhadap Analisis Sentimen Pada Pengguna E-Wallet Aplikasi Dana Menggunakan Fitur Ekstraksi TF-IDF | Klasifikasi *KNN* dan *Naïve Bayes*. Pelabelan sentimen berbasis *Lexicon* dengan acuan kamus label oleh tenaga ahli bahasa (*expert*) dan *library Lexicon via Python*. Pembobotan kata menggunakan *TF-IDF*. | Data yang diberi label model *Lexicon*, akurasi *KNN* 78% dan *Naïve Bayes* 74%. Data yang diberi label oleh *expert*, akurasi kedua metode klasifikasi mencapai 96%.  Rasio pembagian data 70:30 memberikan akurasi terbaik sebesar 96%.  Rasio pembagian data 90:10 dan 80:20 masing-masing mencapai akurasi 95%.  Nilai *k = 1* mencapai akurasi 95.24%.  Nilai *k = 2, …, k = 20* mencapai akurasi yang sama yaitu 96.19%.  Terkait nilai *k*, keputusan akhir sebagai parameter final yang diambil menurut penulis adalah *k = 5* dikarenakan memberikan kinerja optimal tanpa penurunan akurasi yang signifikan. |
| 2. | Alamsyah & Mulyati, 2023. | Implementasi Algoritme K-Nearest Neighbour Dan Lexicon Based Untuk Analisis Sentimen Kepuasan Pengguna Aplikasi Gramedia Digital Pada Media Sosial Twitter | Klasifikasi *KNN*. Pelabelan sentimen berbasis *Lexicon* dengan acuan kamus label *InSet*. Pembobotan kata menggunakan *TF-IDF*. | *KNN* berhasil mencapai akurasi tertinggi sebesar 75.97% dengan nilai *k = 3* dengan rasio pembagian data 60:40. |
| 3. | Diwandanu & Wisudawati, 2023. | Analisis Sentimen Terhadap Twit Maxim Pada Twitter Menggunakan R Programming Dan K Nearest Neighbors | Klasifikasi *KNN*. Pelabelan sentimen berbasis *library* *Lexicon*. | Rasio pembagian data 80:20, 75:25, 70:30 dengan nilai *k = 1, …, k = 10*.  Hasil akurasi terbaik dengan rasio pembagian data 80:20 dan nilai *k = 1*, yaitu 95.43%. |
| 4. | Putri et al., 2023. | Analisis Sentimen dan Pemodelan Ulasan Aplikasi AdaKami Menggunakan Algoritma SVM dan KNN | Klasifikasi *KNN* dan *SVM*. Pelabelan sentimen berbasis *library Lexicon via Python* dan manual. Pembobotan kata menggunakan *TF-IDF*. | Rasio pembagian data 90:10, 80:20, 70:30, 60:40, 50:50.  Model analisis sentimen dengan performa paling optimal menggunakan algoritma *SVM* dengan metode pelabelan manual dan proporsi pembagian data 90:10 dengan akurasi sebesar 93%, presisi 93%, *recall* 93%, dan *f1-score* 92%.  Penelitian ini berkesimpulan bahwa *SVM* menghasilkan model dengan performa lebih optimal dibanding *KNN* |
| 5. | Arifin & Nugroho, 2023. | Uji Akurasi Penggunaan Metode KNN dalam Analisis Sentimen Kenaikan Harga BBM pada Media Twitter | Klasifikasi *KNN*. Pelabelan sentimen berbasis manual. | Dengan nilai *k = 5*, didapatkan akurasi, presisi, dan *recall* dari evaluasi terhadap algoritma klasifikasi *KNN* masing-masing 94.33%, 91.3%, dan 84%. |
| 6. | Handoko et al., 2024. | Analisis Sentimen Ulasan Penumpang Maskapai Penerbangan di Indonesia Dengan Algoritma Random Forest Dan KNN | Klasifikasi *KNN* dan *Random Forest*. Pelabelan sentimen berbasis *Lexicon Afinn*. Pembobotan kata menggunakan *TF-IDF*. | Hasil evaluasi model menunjukkan tingkat akurasi dari *Random Forest* sebesar 83% dan *KNN* mencapai 82%. |
| 7. | Rahayu et al., 2022. | Implementasi Metode K-Nearest Neighbor (K-NN) untuk Analisis Sentimen Kepuasan Pengguna Aplikasi Teknologi Finansial FLIP | Klasifikasi *KNN*. Pembobotan kata menggunakan *TF-IDF*. Pelabelan sentimen berbasis *Lexicon* dengan acuan kamus label *SentiWordnet*. | Rasio pembagian data 80:20 dengan algoritma klasifikasi *KNN* memperoleh akurasi sebesar 76.68%.  77.67% dari data uji, menurut penulis sudah benar terklasifikasi ke dalam kelas ulasan positif dengan nilai presisi dan *recall* masing-masing 82.67% dan 86.92%. |
| 8. | Angel et al., 2024. | Analisis Sentimen dan Emosi dari Ulasan Google Maps untuk Layanan Rumah Sakit di Palangka Raya Menggunakan Machine Learning | Klasifikasi *KNN, Logistic Regression,* dan *Decision Tree*. Pelabelan sentimen berbasis *Lexicon Vader via Python* dan *NRC Lexicon* (emosi)*.* Pembobotan kata menggunakan *TF-IDF*. | Akurasi tertinggi didapat oleh algoritma klasifikasi *Decision Tree* sebesar 92%, diikuti dengan *Logistic Regression* dengan akurasi 86%, dan *KNN* dengan akurasi 48%. |
| 9. | Sholeha et al., 2022. | Analisis Sentimen Pada Agen Perjalanan Online Menggunakan Naïve Bayes dan K-Nearest Neighbor | Klasifikasi *KNN* dan *Naïve Bayes*. Pelabelan sentimen berbasis manual. Penghitungan frekuensi kata dengan *TF*. | Menurut penulis, berdasarkan 1500 data komentar dari 3 *fanpage Facebook* agen perjalanan online, ditemukan bahwa algoritma *KNN* memiliki akurasi yang lebih baik pada rata-rata dibandingkan *Naïve Bayes* dengan akurasi tertinggi 52.35%.  Akurasi tertinggi kedua algoritma klasifikasi tersebut didapatkan saat seluruh data menggunakan huruf kecil. |
| 10. | Mustaqim et al., 2024. | Analisis Sentimen Ulasan Aplikasi PosPay untuk Meningkatkan Kepuasan Pengguna dengan Metode K-Nearest Neighbor (KNN) | Klasifikasi *KNN*. Pembobotan kata menggunakan *TF-IDF*. Pelabelan sentimen berbasis *library* *Lexicon*. | Menurut penulis, hasil akhir dari penelitian menunjukkan bahwa sentimen pengguna aplikasi Pospay cenderung positif. Model klasifikasi *KNN* menghasilkan akurasi sebesar 91%, presisi sebesar 90%, dan *recall* sebesar 99%. |

# **BAB III**

# **METODOLOGI PENELITIAN DAN PENGEMBANGAN SISTEM**

**3.1 Metodologi Penelitian**

Bagian ini membahas mengenai metodologi penelitian serta pengembangan sistem yang akan dilakukan dalam penelitian ini yaitu menganalisis sentimen publik terhadap topik pemindahan ibu kota negara berdasarkan unggahan dalam bentuk teks di sosial media *X (Twitter)* dengan metode klasifikasi *KNN* dan pelabelan sentimen berbasis leksikon serta ekstraksi fitur *TF-IDF*. Metodologi penelitian merupakan sub bab yang menggambarkan alur kerja serta tahapan pada penelitian ini. Tahapan metodologi penelitian pada tugas akhir ini yaitu pengumpulan data*, preprocessing* data, pelabelan data, pembobotan kata, pembuatan model, pengujian model, pengujian sistem, serta analisis dan visualisasi hasil. Berikut merupakan tahapan pada metodologi penelitian ini dapat dilihat pada gambar 3.1.



**3.1.1 Pengumpulan Data**

Pengumpulan data dilakukan untuk memperoleh data yang nantinya akan diolah menjadi informasi yang mudah dipahami dan dapat digunakan sebagai dasar pengambilan keputusan. Data yang terdapat dalam penelitian ini dikumpulkan dengan melakukan *web scraping* dari media sosial *X* (*Twitter*).

1. Studi Pustaka

Studi pustaka merupakan langkah awal dalam proses penelitian. Kegiatan ini bertujuan untuk mencari dan memperoleh informasi terkait metode, topik, serta masalah yang menjadi fokus penelitian. Studi pustaka ini didasarkan pada referensi dari jurnal, makalah, prosiding, buku, dan sumber terpercaya lainnya. Studi pustaka dilakukan untuk memperkuat argumentasi dalam penelitian yang sedang dilakukan

1. *X* (*Twitter*)

Data yang digunakan dalam penelitian ini diambil dari media sosial *X*. *X* atau yang sebelumnya dikenal dengan sebutan *Twitter*, merupakan media sosial yang memungkinkan pengguna untuk berbagi pesan singkat yang disebut *tweet* atau cuitan. Media sosial ini digunakan secara luas oleh berbagai kalangan, mulai dari individu hingga organisasi besar, untuk menyampaikan pendapat, berita, dan informasi lainnya. *Platform* ini memberikan peluang untuk dilakukannya analisis sentimen publik terhadap pemindahan ibu kota negara dan mengamati reaksi pengguna terhadap berbagai topik, termasuk ulasan tentang layanan tertentu.

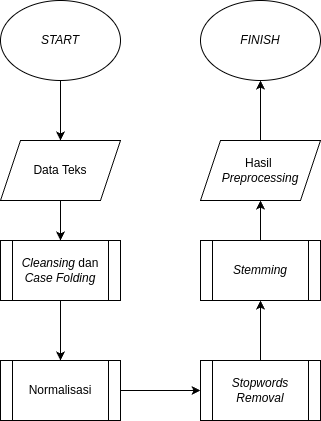
1. *Web Scraping*

Pengumpulan data dilakukan dengan cara *web scraping* terhadap media sosial *X* (*Twitter*). Data yang berhasil dikumpulkan akan dilakukan *preprocessing*. Data yang digunakan pada penelitian ini merupakan data berbahasa Indonesia dengan total data berjumlah 5341 baris *tweet* menggunakan 9 kata kunci yang tertera pada tabel 3.1.

|  |  |  |
| --- | --- | --- |
| **No.** | **Kata Kunci** | **Jumlah Data** |
| 1 | *ikn* | 1005 |
| 2 | *ibu kota baru* | 223 |
| 3 | *ibu kota nusantara* | 15 |
| 4 | *ibu kota pindah* | 1005 |
| 5 | *pemindahan ibu kota* | 1010 |
| 6 | *ibukota baru* | 212 |
| 7 | *ibukota nusantara* | 804 |
| 8 | *ibukota pindah* | 554 |
| 9 | *pemindahan ibukota* | 513 |

**3.1.2 *Preprocessing***

Tahap penting dalam pengolahan data yang bertujuan untuk membersihkan dan mempersiapkan data agar siap digunakan dalam analisis sentimen. Pada tahapan ini, berbagai langkah dilakukan, seperti menghapus data yang tidak relevan (misalnya tanda baca, angka, dan simbol khusus), mengubah teks menjadi format standar (seperti konversi huruf besar menjadi huruf kecil), menghilangkan kata-kata umum yang tidak memiliki makna khusus (*stopwords*), serta melakukan *stemming* untuk mengembalikan kata ke bentuk dasarnya. *Preprocessing* sangat penting untuk meningkatkan akurasi model dalam memahami dan menganalisis data secara efektif, terutama dalam tugas-tugas yang melibatkan analisis teks seperti analisis sentimen. Berikut merupakan *flowchart* dari *preprocessing* dapat dilihat pada gambar 3.2.



1. *Cleansing* dan *Case Folding*

Proses ini bertujuan untuk membersihkan data dengan menghapus elemen-elemen yang tidak relevan, seperti tanda baca, angka, emoji, *url*, atau karakter khusus lainnya yang tidak memiliki makna signifikan untuk analisis teks. *Cleansing* sangat membantu untuk fokus pada elemen-elemen yang relevan pada data untuk analisis lebih lanjut.

Tahap ini juga mengubah seluruh teks menjadi huruf kecil (*lower case*). Hal tersebut dilakukan untuk menghilangkan perbedaan antara huruf besar dan huruf kecil, karena dalam analisis berbasis teks, kata dengan huruf besar dan kecil dianggap berbeda.

1. Normalisasi
2. *Stopwords Removal*
3. *Stemming*

**3.1.3 *Sentiment Labelling***

**3.1.4 Pembobotan *TF-IDF***

**3.1.5 Analisis Sentimen dengan Model *KNN***

**3.1.6 Pembuatan Model Sentimen**

**3.1.7 Pengujian**

**3.2 Metode Pengembangan Sistem**

**3.2.1 *Requirements Analysis***

**3.2.2 *System and Software Design***

**3.2.3 *Implementation***

**3.2.4 *System Testing***