BAB II
TINJAUAN PUSTAKA

2.1. Desa

Dikutip dari wikipedia Indonesia, istilah desa adalah pembagian wilayah administratif di bawah kecamatan dalam pemerintahan kabupaten, yang dipimpin oleh kepala desa. Menurut para ahli, pengertian desa adalah sebagai berikut

1. Bambang Utoyo

   Desa merupakan tempat sebagian besar penduduk yang bermata pencarian di bidang pertanian dan menghasilkan bahan makanan.

2. R. Bintarto

   Desa adalah perwujudan geografis yang ditimbulkan oleh unsur-unsur fisiografis, sosial, ekonomis politik, kultural setempat dalam hubungan dan pengaruh timbal balik dengan daerah lain.

3. Sutarjo Kartohadikusumo

   Desa merupakan kesatuan hukum tempat tinggal suatu masyarakat yang berhak menyelenggarakan rumah tangganya sendiri merupakan pemerintahan terendah di bawah camat.

2.2. Desa Margo Mulyo

Desa Margo Mulyo berada di Kecamatan Tumijajar, Kabupaten Tulang Bawang Barat, Provinsi Lampung. Dengan jumlah penduduk saat ini kurang lebih 4.637 jiwa, sekitar 75 % masyarakatnya berprofesi sebagai Petani/pekebun.

Menjadi desa transmigrasi sejak tahun 1982, Desa Margo Mulyo memiliki luas 1.006 Ha dengan sekitar 830 Ha nya adalah lahan pertanian dan perkebunan, sehingga menjadikan masyarakatnya mengandalkan sektor pertanian sebagai sumber penghidupan. Selain itu, nuansa pedesaan yang asri dengan mempertahankan kehidupan sosial yang tinggi masih terjaga di Desa Margo Mulyo.

Dengan adanya dana desa yang diberikan pemerintah sejak ahun 2015 merupakan tahun pertama kalinya Indonesia mengucurkan Dana Desa sesuai amanat UU No. 6 Tahun 2014 tentang Desa. Dana ini diharapkan agar dimanfaatkan oleh pemerintah desa untuk membiayai penyelenggaraan pemerintahan, pembangunan, dan pemberdayaan masyarakat desa sehingga menuntut Desa Margo Mulyo untuk dapat menciptakan perubahan, terutama kemajuan pada Desanya.

Dalam rangka mengetahui sejauh mana dampak dari adanya dana desa, Pemerintah melalui Kementerian Desa, Pembangunan Daerah Tertinggal dan Transmigrasi (PDTT) melakukan penilaian pada Indeks Ketahanan Lingkungan/ Ekologi terdiri dari Dimensi Ekologi (indikator kualitas lingkungan dan potensi rawan bencana dan tanggap bencana). PDTT meluncurkan Indeks Desa Membangun (IDM) pada Oktober 2016 dan wajib untuk di isi setiap tahunnya oleh

semua desa yang mendapatkan dana desa. Berdasarkan nilai IDM, desa dibagi menjadi lima kategori, yaitu: Desa mandiri, jika memiliki nilai IDM >0,8155. Desa maju, jika memiliki nilai IDM ≤ 0,8155 dan >0,7072. Desa berkembang, jika memiliki nilai IDM ≤ 0,7072 dan >0,5989.

Dan pada tahun 2023 Desa Margo Mulyo mendapat skor 0,8270 atau mendapat predikat Desa Mandiri, dan salah satu aspek penting yang menjadi tolak ukurnya adalah pelayanan dan informasi publik yang saat ini sudah digitalisasi dan dapat diakses pada laman www.margomulyo-tubaba.desa.id.

2.3. Pelayanan Digital

Pelayanan publik menjadi tolak ukur keberhasilan pelaksanaan tugas dan pengukuran kinerja pemerintah melalui birokrasi. Pelayanan publik sebagai penggerak utama juga dianggap penting oleh semua aktor dari unsur *good governance*. [1] Subarsono mendefenisikan pelayanan publik sebagai serangkaian aktivitas yang dilakukan oleh birokrasi publik untuk memenuhi kebutuhan warga pengguna.

Pada saat ini, perkembangan teknologi informasi dan komunikasi sangat dibutuhkan dalam semua aspek kehidupan manusia. Dengan adanya sistem informasi global dapat menghasilkan keterbukaan informasi publik. [6] Di era Revolusi Industri 4.0, digitalisasi menjadi proses manufaktur dengan memanfaatkan teknologi komputer serta internet. Penelitian ini bertujuan melihat sejauh mana respon masyarakat terhadap pengembangan dan penerapan sistem

sebagai wadah untuk mendigitalisasi surat sebagai bentuk arsip digital. Dan juga, diharapkan dapat meminimalisir kesalahan-kesalahan yang dapat terjadi dalam pembuatan surat dan laporan juga dapat memaksimalkan kinerja pelayanan desa diwaktu yang akan datang agar lebih transparan dan mudah diawasi.

2.4. Sistem Informasi Desa

Sistem Informasi Desa (SID) adalah seperangkat alat dan proses pemanfaatan data dan informasi untuk mendukung pengelolaan sumber daya berbasis komunitas di tingkat Desa. Setidaknya ada dua hal yang menjadikan kehadiran SID menjadi penting. Pertama, keinginan untuk mewujudkan partisipasi, transparansi dan akuntabilitas pemerintahan Desa, yang berarti bahwa SID selain sebagai perangkat pemroses informasi juga menjadi perangkat demokrasi. Kedua, banyaknya data Desa yang berserakan dan tidak terkumpul secara rapi di arsip pemerintahan Desa yang berarti bahwa SID merupakan perangkat teknokratis yang membuat penyelenggaraan pemerintahan Desa menjadi lebih efisien dan efektif.

Untuk mendukung pengelolaan dan pemanfaaatan data Desa yang diatur dalam bagian ketiga UU Desa Pasal 86, pemerintah memberikan platform tata kelola Desa. Meskipun demikian, pada praktiknya masih banyak Desa yang belum menjalankan karena keterbatasan oleh berbagai faktor, salah satunya adalah kurangnya sumber daya manusia yang paham tentang Sistem Informasi Desa. Selain itu terdapat pula faktor teknis terkait dengan SID seperti proses adminstrasi yang cukup panjang, fitur yang kurang, dukungan komunitas dan informasi yang masih sedikit, dan

lainnya. Salah satu sistem informasi alternatif selain SiDeKa adalah OpenSID yang berbasis *open source* (sumber terbuka) dengan dukungan komunitas yang besar.[7]

Saat ini Desa Margo Mulyo menggunakan layanan OpenSID yang terintegrasi didalam website, sehingga pelayanan dan juga informasi publik dapat diakses melalui laman website www.margomulyo-tubaba.desa.id.

2.5. Analisis Sentimen

Analisis sentimen adalah cabang ilmu *Data mining* yaitu *text mining* yang mempelajari sentimen yang ada pada suatu teks opini. Prinsip dasar dari analisis sentimen adalah mengklasifikasikan sebuah teks apakah teks tersebut bernilai positif, negatif atau netral. Analisis sentimen mengacu pada bidang yang luas dari pengolahan bahasa alami, komputasi *linguistic* dan *text mining* yang bertujuan menganalisa pendapat, sentimen, evaluasi, sikap, penilaian, dan emosi seseorang.[8] Analisis sentimen merupakan cara mengumpulkan pendapat khalayak umum menggunakan jejaring sosial yang didalamnya terdapat mengandung pelayanan umum, serta isu terkini. Disisi lain, analisis sentimen dapat digunakan untuk mengevaluasi suatu kinerja, pelayanan dan sebagainya.[4]

Analisis Sentimen atau opini mining adalah studi komputasi tentang pendapat, sentimen dan emosi yang dinyatakan dalam teks [5], Langkah-langkah umum pada analisis sentimen klasifikasi teks adalah.

1. Definisikan domain dataset

Mengumpulkan dataset seperti opini pelayanan pembuatan surat, opini pelayanan konsultasi, review produk dan lain-lain

2. *Pre-processing*

   Pada tahapan ini biasanya dilakukan *Tokenization, Stopwords Removal, dan Stemming.*

3. *Transformation*

   Pembobotan dari data tekstual, proses yang sering digunakan adalah TF-IDF.

4. *Feature Selection*

   Membuat pengklasifikasi lebih efisien dengan mengurangi jumlah data yang dianalisa.

5. *Classification*

   Pengklasifikasi teks biasanya menggunakan metode *Naive Bayes, K-Nearest Neighbor*, SVM dan lain-lain.

6. *Interpretation/Evaluation*

   Biasanya evaluasi untuk menghitung nilai akurasi

Analisis sentimen dianggap sebagai masalah pengelompokan. Sama seperti dalam laporan besar, nilai sentimen dapat dikomunikasikan dalam berbagai cara dan ditandai dengan adanya sentimen didalamnya. jika ada sentimen dalam opini Masyarakat, mengandung *polar word* atau kata-kata berlawanan maka itu ditetapkan positif atau negatif, jika tidak dianggap Netral. Langkah analisis sentimen sebagai berikut:

1. Level 1: Mencari sentimen negatif, positif, serta netral pada setiap baris.
2. Level 2: Analisa sentimen seluruh dokumen sebagai negatif, positif, serta netral.

3. Level 3: Menerapkan pengelompokan dimana mengumpulkan semua atribut yang ada dengan hasil sentimen yang sama.
4. Level 4: Memanfaatkan visualisasi data dari analisis sentimen untuk interaksi antar *user*.

2.6. *Google Form*

Keakraban masyarakat dengan berbagai produk teknologi seperti Komputer, Tablet dan *Smartphone*, serta tersedianya koneksi internet yang semakin murah juga menjadi peluang untuk pemanfaatan teknologi informasi dan komunikasi dalam pelaksanaan sistem pelayanan masyarakat. Salah satu software yang mudah diakses, gratis digunakan, sederhana dalam pengoperasiannya, dan cukup baik untuk dikembangkan sebagai alat evaluasi kinerja adalah *Google Form*.

Google Form merupakan salah satu komponen layanan *Google Docs*, aplikasi ini sangat cocok untuk mahasiswa, guru, dosen, pegawai kantor dan professional yang senang membuat *quiz, form dan survey online*. Fitur dari *Google Form* dapat di bagi ke orang-orang secara terbuka atau khusus kepada pemilik akun Google dengan pilihan aksesibilitas, seperti: *read only* (hanya dapat membaca) atau *editable* (dapat mengedit dokumen). Selain itu, *Google docs* juga dapat menjadi alternatif bagi orang-orang yang tidak memiliki dana untuk membeli aplikasi berbayar untuk menggunakan program gratisan dibandingkan membajak program berbayar seperti *Microsoft Office*, karena kita tahu bahwa membajak program itu adalah tidak baik.[9]

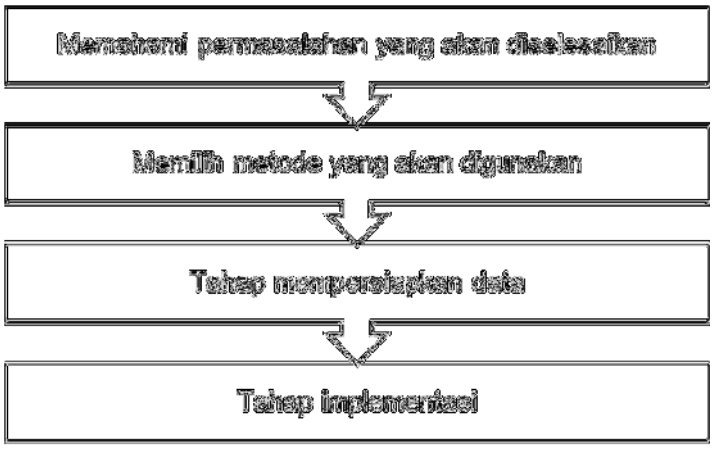
Adapun beberapa fungsi Google Form untuk dunia pendidikaan adalah sebagai berikut.

1. Memberikan tugas latihan/ ulangan online melalui laman website,
2. Mengumpulkan pendapat orang lain melalui laman website.
3. Mengumpulkan berbagai data siswa/ guru melalui halaman website.
4. Membuat formulir pendaftaran *online* untuk sekolah.
5. Membagikan kuesioner kepada orang-orang secara o*nline.*

2.7. *Machine Learning*

Teknik untuk melakukan inferensi terhadap data dengan pendekatan matematis. Inti Machine Learning adalah untuk membuat model (matematis) yang merefleksikan pola-pola data. Infersi yang dimaksud lebih menitikberatkan ranah hubungan variabel. Misalnya, apakah data dengan dua arah bisa diklasifikasikan kedalam masing-masing plot. Tujuan dari machine learning minimal ada dua, memprediksi masa depan dan/atau memperoleh ilmu pengetahuan.[10]

Adapun Langkah membuat machine learning sebagai tertera didiagram berikut.



Gambar 2.1 Langkah Membuat Machine Learning

Beberapa permasalahan yang dapat diselesaikan dengan pendekatan *machine learning* diantaranya.

1. Klasifikasi

   Digunakan untuk mengidentifikasi kategori yang dimiliki suatu objek. Misalnya, apakah itu spam? Atau apakah itu kanker?

2. Regresi

   Digunakan untuk memprediksi aspek bernilai numerik berkelanjutan yang terkait dengan suatu objek. Misalnya, kemungkinan pengguna mengklik iklan atau prediksi harga saham.

3. Kesamaan / Anomali

   Digunakan untuk mengambil objek serupa atau untuk menemukan anomali dalam perilaku. Misalnya, mencari gambar serupa atau mendeteksi penipuan dalam perilaku pengguna.

4. Peringkat

   Digunakan untuk mengurutkan data yang relevan menurut input tertentu. Misalnya, Peringkat Halaman Google

5. *Sequence Prediction*

   Digunakan untuk memprediksi elemen berikutnya dalam serangkaian data. Misalnya, memprediksi kata berikutnya dalam sebuah kalimat.

Machine Learning terbagi menjadi 3 kategori yaitu *Supervised learning, unsupervised Machine Learning, Semi-Supervised learning-Based*.

1. *Supervised learning*

   Supervised learning adalah jenis machine learning yang melibatkan data yang telah terlabel sebelumnya. Dalam *Supervised learning*, algoritma belajar dari contoh-contoh data yang diklasifikasikan sebelumnya dan mencoba untuk mempelajari pola atau korelasi dalam data tersebut. Dalam proses ini, algoritma menggunakan label yang diberikan pada setiap contoh data sebagai bimbingan untuk membuat prediksi yang akurat pada data baru yang belum terlihat sebelumnya. Contoh dari *Supervised learning* termasuk regresi linier, klasifikasi, dan deteksi anomali.

2. *Unsupervised Machine Learning*

   Sebuah teknik dalam pembelajaran mesin di mana model atau algoritma tidak diberikan label atau tanda pengenal pada data yang dikumpulkan, melainkan dipercayakan untuk menemukan korelasi dan pola sendiri dari data tersebut. Dalam *unsupervised machine learning*, tujuan utama adalah menemukan

struktur atau kelompok dalam data tanpa mengharuskan model memiliki informasi sebelumnya tentang hasil yang diinginkan atau klasifikasi tertentu. Teknik ini biasanya digunakan untuk analisis klaster, reduksi dimensi, dan pemilihan fitur. Contoh penerapan dari *unsupervised machine learning* adalah pemasaran terarah dan pengelompokan komentar pelanggan pada media sosial.

3. *Semi-Supervised learning-Based*

Teknik pembelajaran mesin yang menggabungkan metode pembelajaran supervisi dan non-supervisi. Pada pembelajaran supervisi, model dilatih dengan data berlabel, sedangkan pada pembelajaran non-supervisi, model belajar dengan menggunakan data tidak berlabel.

Dalam *semi-supervised learning*, sebagian data diawali dengan label, sementara sebagian lagi tanpa label. Dengan memanfaatkan data yang berlabel, model dilatih untuk mempelajari pola yang ada dalam data tersebut. Kemudian, model ini digunakan untuk memprediksi label pada data yang tidak berlabel. Dengan demikian, teknik ini dapat meningkatkan akurasi model, terutama ketika ketersediaan data yang berlabel terbatas.

Semi-Supervised learning-based banyak digunakan dalam berbagai aplikasi seperti identifikasi wajah, pengenalan suara, klasifikasi teks, dan sebagainya. Namun, teknik ini membutuhkan pemilihan data berlabel yang tepat dan lebih banyak waktu untuk melatih model.

2.8. *Natural Language Processing*

Natural Language Processing (NLP) adalah bidang ilmu yang mengkaji tentang bagaimana komputer dapat memahami, menganalisis, dan menghasilkan teks berdasarkan bahasa alami manusia.[11]

Penerapan NLP menurut para ahli dapat dilakukan dalam berbagai bidang diantaranya:

1. Analisis sentimen
   NLP dapat digunakan untuk menganalisis sentimen dalam berbagai macam teks, seperti review produk, media sosial, atau komen di forum diskusi.
2. Pemrosesan bahasa alami
   NLP dapat digunakan untuk mengolah teks, menganalisis struktur kalimat, mengenali frasa frasa umum yang digunakan dan melakukan klasifikasi teks.
3. Penerjemahan Bahasa
   NLP dapat digunakan untuk menerjemahkan bahasa dari satu bahasa ke bahasa lain dengan menggunakan terjemahan otomatis.
4. Pengenalan suara
   NLP dapat digunakan untuk memproses suara manusia dengan menggunakan sistem pengenalan suara untuk mengubah suara menjadi teks.
5. *Chatbot* dan *asisten virtual*
   NLP juga dapat digunakan untuk menciptakan chatbot atau asisten virtual yang dapat melakukan dialog dengan pengguna menggunakan bahasa alami.
6. Ekstraksi informasi
   NLP dapat digunakan untuk mengekstrak informasi penting dari dokumen berbahasa alami, seperti daftar nama, tanggal, lokasi atau nomor telepon.
7. Pengenalan entitas bernama
   NLP dapat digunakan untuk mengenali entitas bernama seperti orang, tempat, dan organisasi di dalam sebuah teks.

Dalam keseluruhan penerapannya, NLP memiliki potensi besar untuk meningkatkan efisiensi dan presisi dalam pengolahan bahasa alami secara otomatis.

2.9. *Text Mining*

Text Mining memiliki definisi menambang data yang berupa teks, dimana sumber data biasanya didapatkan dari sebuah dokumen dan tujuannya untuk mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen. [12]

Text mining memiliki sifat yang sama seperti *Data mining*, tetapi lebih fokus pada pengolahan teks daripada data yang terstruktur. Analisa *text mining* ini merupakan langkah pertama yang dilakukan dalam penambangan teks, yaitu mengatur dan menyusun data dengan cara tertentu sehingga dapat menjadi sasaran analisis kualitatif dan kuantitatif. Melakukannya secara khusus melibatkan penggunaan teknologi NLP, yang menerapkan prinsip-prinsip linguistik komputasional untuk menguraikan dan menginterpretasikan set data. Seperti pengkategorian, pengelompokan dan teks penandaan; meringkas set data; menciptakan taksonomi; dan mengekstraksi informasi tentang hal-hal seperti frekuensi kata dan hubungan antar entitas data. Model analitik kemudian dijalankan untuk menghasilkan temuan yang dapat membantu mendorong strategi bisnis dan tindakan operasional.

2.10. *Data Preparation*

Data Preparation atau bisa disebut prapemrosesan teks bertujuan untuk membersihkan teks dari *noise* yang mengganggu agar teks lebih mudah dianalisis untuk kemudian mengembalikan token yang telah dibersihkan. Token adalah kata tunggal atau kelompok kata yang dihitung berdasarkan frekuensinya dan berfungsi sebagai fitur analisis. Tahapan–tahapan yang dilakukan berguna untuk menstandardisasi data, sehingga mengurangi jumlah dimensi dalam kumpulan data teks. Ada keseimbangan antara informasi yang disimpan dan pengurangan kompleksitas dalam pilihan yang dibuat selama proses.

Detail dalam pembersihan dan preprocessing membuat proses analisis lebih lancar. Preprocessing itu sendiri memiliki beberapa tahapan diantaranya.

1. *Case Folding*

   Case Folding adalah mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf a sampai z yang diterima. Karakter selain huruf dihilangkan dan dianggap *delimiter*.

2. *Cleansing*

   Dalam langkah data *cleansing,* data mentah akan dibersihkan melalui beberapa proses seperti mengisi nilai yang hilang, menghaluskan *noisy data*, dan menyelesaikan inkonsestensi yang ditemukan.

3. *Tokenizing*

   Tahap *tokenizing* adalah tahap pemotongan string input berdasarkan setiap kata yang menyusun sebuah kalimat.

4. *Filtering/ Stopword Removal*

   Filtering adalah tahap mengambil kata-kata penting dari hasil token. Terdapat beberapa algoritma dalam *filtering* yaitu *stop-list* dan *word-list*. Algoritma *stop- word* merupakan algoritma yang digunakan untuk mengeliminasi kata-kata yang tidak deskriptif. Algoritma *word-list* adalah algoritma yang digunakan menyimpan kata-kata memiliki nilai deskriptif.

5. *Normalization*

   Normalization adalah langkah untuk mengubah data ke dalam skala yang teratur sehingga dapat dibandingkan dengan lebih akurat.

6. *Stemming*

   Stemming adalah proses untuk memecahkan setiap varian-varian suatu kata menjadi kata dasar. Proses stemming pada kata Bahasa Indonesia berbeda dengan *stemming* bahasa Inggris. Proses *stemming* pada kata bahasa Inggris adalah proses untuk mengeliminasi sufiks pada kata, sementara proses *stemming* bahasa Indonesia adalah proses untuk mengeliminasi sufiks, prefiks, infiks, dan konfiks.

2.11. Pelabelan

Pada penelitian ini pelabelan dilakukan untuk memisahkan data opini apakah termasuk ke dalam opini positif, negatif, atau netral, pelabelan dilakukan dengan menggunakan metode *Lexicon,* metode ini secara otomatis mengklasifikasikan teks opini dengan menguji data latih berupa teks opini yang sebelumnya diklasifikasikan secara manual.

2.12. Algoritma *Stemming*

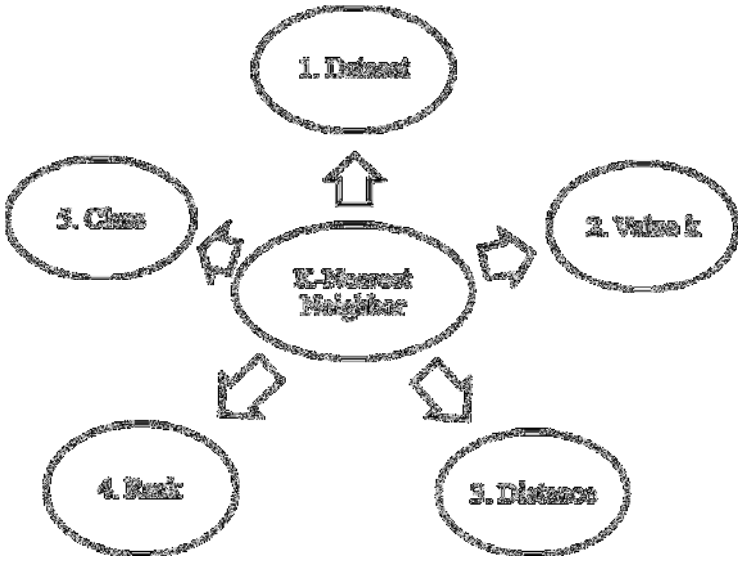
Dalam tahapan *preprocessing* terdapat tahapan lain yang lebih spesifik, salah satunya yaitu *stemming*. *Stemming* merupakan suatu proses untuk menemukan kata dasar dari sebuah kata dengan menghilangkan semua imbuhan (*affixes*) baik yang terdiri dari awalan (*prefixes*), sisipan (*infixes*), akhiran (*suffixes*) dan kombinasi dari awalan dan akhiran (*confixes*) pada kata turunan. *Stemming* adalah inti dari teknik NLP untuk mendapatkan informasi kembali (*information retrieval*) yang efektif, efisien, dan secara luas dapat diterima oleh pengguna. *Stemming* digunakan untuk mengganti bentuk dari suatu kata menjadi kata dasar dari kata tersebut yang sesuai dengan struktur morfologi bahasa Indonesia yang baik dan benar.

2.13. Algoritma *K-Nearest Neighbor*

K-Nearest Neighbor atau yang biasa disebut dengan KNN merupakan algoritma klasifikasi yang mengelompokan data baru berdasarkan jarak data baru ke beberapa data atau tetangga terdekat. Algoritma ini bekerja dengan cara mencari sejumlah k pola (diantara semua pola latih yang ada di semua kelas) yang terdekat dengan pola masukan, kemudian menentukan kelas keputusan berdasarkan jumlah pola terbanyak diantara k pola tersebut. [3]

Berdasarkan penelitian sebelumnya nilai akurasi menggunakan k=2, Dihasilkan nilai k yang dalam melakukan proses pengujianpada klasifikasi algoritma KNN mencapai tingkat akurasi pada nilai k =2 dengan nilai kurasi 88%.

KNN dengan klasifikasi berdasarkan pembelajaran menggunakan *training samples*. Setiap sample mewakili poin data dalam n-*dimensional space*. KNN menempatkan bobot pada setiap atribut. KNN dapat juga digunakan untuk prediksi, untuk mengembalikan nilai real prediksi terhadap unknown sample. KNN termasuk kelompok *instancebased learning*. [13]Algoritma ini juga merupakan salah satu teknik *lazy learning* KNN dilakukan dengan mencari kelompok k obyek dalam data *training* yang paling dekat (mirip) dengan obyek pada data baru atau data *testing*. Tahapan metode KNN ditampilkan pada gambar.



Gambar 2.2 KNN Model

Berdasarkan Gambar , KNN Model terdiri dari: Dataset, Value k, Distance, Rank, and Class.

1. *Dataset* terdiri dari *Data Training* dan *Data Testing*.
2. *Value k* digunakan untuk menentukan kemiripan (*similarity*) dan jarak data (*distance*).

3. *Distance* (jarak data) mengggunakan *Euclidean Distance* dengan rumus.

$$d(X,Y) = \sqrt{\sum_{i-1}^{n}(x_i - y_i)^{\ 2}}$$

Dimana *Euclidean Distance* (d) antara dua titik, X = (x1, x2, ..., xn) dan Y = (y1, y2,...., yn) dilambangkan dengan d (X, Y).

4. *Rank* Urutan data (rank) berdasarkan nilai tertinggi dalam dataset.
5. *Class* ditentukan dari hasil proses klasifikasi berdasarkan nilai mayoritas dari dataset (*determine by majority*).

2.14. Algoritma TF-IDF

Berdasarkan hasil penelitian sebelumnya, [13] analisa sentimen terhadap transportasi online dengan menggunakan metode pembobotan Hybrid TF-IDF dapat disimpulkan bahwa, Klasifikasi Hasil terbaik dari pengujian cross validation pada uji variable k adalah k=5 dengan nilai akurasi 70%, presisi kelas positif 68%, presisi kelas negatif 75%, recall kelas positif 82%, recall kelas negatif 59%, f-measur kelas positif 74% dan f-measure kelas negatif 65%.

Metode *Term Frequency-Inverse Document Frequency* (TF-IDF) adalah cara pemberian bobot hubungan suatu kata (term) terhadap dokumen. TF-IDF ini adalah sebuah ukuran statistik yang digunakan untuk mengevaluasi seberapa penting sebuah kata di dalam sebuah dokumen atau dalam sekelompok kata, TF-

IDF digunakan rumus untuk menghitung bobot (W) masing masing dokumen terhadap kata kunci dengan rumus yaitu.

Wdt = tfdt * Idft

Dimana:

Wdt = bobot dokumen ke-d terhadap kata ke-t

tfdt = banyaknya kata yang dicari pada sebuah dokumen

Idft = *Inversed Document Frequency* (log (N/df)) N = total dokumen

Df = banyak dokumen yang mengandung kata yang dicari.

Term frequency (TF) adalah metode yang paling sederhana dalam pembobotan kata yang bergantung pada jumlah kemunculan term dalam dokumen dengan menghitung skor antara term dan dokumen berdasarkan bobot (*weight*) term di dalam dokumen. [13]

$$TF\ (d,t) = f(d,t)$$

Keterangan :

F(d,t) : frekuensi kemunculan term t pada dokumen d.

Inverse Document Frequency (IDF) adalah metode yang memperhatikan kemunculan kata di dalam dokumen dengan cara menghitung bobot kemunculan kata pada kumpulan dokumen.

$$ID\ (t) = \log(Nd)\, df(t)$$

Keterangan:

Nm : jumlah seluruh dokumen.

df(t) : jumlah dokumen yang mengandung term t.

Persamaan adalah menggabungkan konsep perhitungan TF-IDF dapat dijabarkan sebagai berikut.

$$W\ (t,d) = TF\ (d,t) x\ IDF = TF\ (d,t) x (\log(d))\ df(t)$$

Keterangan :

W(d,t) : bobot term t pada dokumen d.

TF(d,t) : total kemunculan term t pada dokumen d

Nd : total seluruh dokumen.

df(t) : total dokumen yang memiliki term.

2.15. *Confusion Matrix*

Confusion Matrix adalah pengukuran performa untuk masalah klasifikasi machine learning dimana keluaran dapat berupa dua kelas atau lebih. *Confusion Matrix* adalah tabel dengan 4 kombinasi berbeda dari nilai prediksi dan nilai aktual. [14]

Berikut adalah struktur *Confusion Matrix* 2x2

Actual \ Predicted	Positive (1)	Negative (0)
Positive (1)	True Positive (TP)	False Negative (FN)
Negative (0)	False Positive (FP)	True Negative (TN)

Tabel 2.1 *Confusion Matrix*

Penjelasan tabel *Confusion Matrix* di atas adalah sebagai berikut:

1. *True Positive* (TP) : Jumlah prediksi yang benar, dimana nilai kelas aktual positif dan diprediksi sebagai positif.
2. *False Negative* (FN) : Jumlah prediksi yang salah, dimana nilai kelas aktual positif namun diprediksi sebagai negatif.
3. *False Positive* (FP) : Jumlah prediksi yang salah, dimana nilai kelas aktual negatif namun diprediksi sebagai positif.
4. *True Negative* (TN) : Jumlah prediksi yang benar, dimana nilai kelas aktual negatif dan diprediksi sebagai negatif.

Berikut adalah *performance measure* yang dapat diukur oleh *Confusion Matrix*.

1. *Accuracy*

   Accuracy adalah perbandingan antara data yang terdeteksi benar dengan seluruh data hasil prediksi.

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

2. *Precision*

   Precision adalah perbandingan antara data yang terdeteksi benar dengan seluruh data prediksi pada suatu kelas atau tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem.

   $$Precision = \frac{TP}{TP + FP}$$

3. *Recall*

   Recall adalah perbandingan hasil klasifikasi dengan kelas sesungguhnya atau tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi.

   $$Recall = \frac{TP}{TP + FN}$$

4. *F1-Score*

   F1-Score menggambarkan perbandingan rata-rata precision dan recall yang dibobotkan.

   $$FI - Score = \frac{2 * recall * precision}{recall + precision}$$

2.16. *Google Colabs*

Google Collaboratory atau *Google Colab* dan biasa disebut dengan *colab* adalah platform berbasis cloud untuk menulis, menjalankan, dan berbagi kode *Python* melalui web browser. *Python* merupakan bahasa pemrograman tingkat tinggi yang mendukung pemrograman berorientasi objek. *Python* memiliki perbedaan dengan bahasa pemrograman lain yaitu dalam penulisan *sintaks*. Dalam bahasa pemrograman *python* terdapat berbagai *library* dan *framework* yang digunakan

untuk melakukan analisis data. Dalam penelitian ini menggunakan *library* sebagai berikut.

1. *Natural Language Tool Kit (*NLTK) *Library* ini digunakan pada tahapan *tokenization* yaitu memisahkan kalimat menjadi per kata yang didapat untuk diolah.
2. *Sastrawi Library* digunakan dalam tahapan *stop removal* untuk menghapus kata-kata yang dianggap tidak berhubungan.
3. *Scikit-learn* merupakan proyek *open source*, artinya bebas untuk menggunakan dan mendistribusikan, dan siapapun dapat dengan mudah mendapatkan kode sumber. *Scikit-learn* berisi sejumlah algoritma *machine learning* yang canggih, serta dokumentasi komprehensif tentang masing-masing algoritma. scikit-learn adalah alat yang sangat populer, dan merupakan *library python* yang paling menonjol untuk *machine learning. scikit-learn* banyak digunakan di industri dan akademisi dan banyak tutorial serta cuplikan kode tersedia secara *online*[15]
4. *Pandas* adalah sebuah *library di Python* yang berlisensi BSD dan *open source* yang menyediakan struktur data dan analisis data yang mudah digunakan. Pandas biasa digunakan untuk membuat tabel, mengubah dimensi data, mengecek data, dan lain sebagainya. Struktur data dasar pada Pandas dinamakan DataFrame, yang memudahkan kita untuk membaca sebuah file dengan banyak jenis format seperti file .txt, .csv, dan .tsv.

2.17. Studi Literatur

No.	Judul	Penulis	Tahun	Hasil
1.	Analisis Sentimen Pada Agen Perjalanan Online Menggunakan Naïve Bayes dan K-Nearest Neighbor	Eka Wahyu Sholeha, Selviana Yunita, Rifqi Hammad, Veny Cahya Hardita, Kaharuddin	2022	- Menggunakan data Twitter berbahasa Indonesia -Pelabelan dengan Transformers
2.	Perbandingan Algoritma *K-Nearest Neighbor* (KNN) dan Naive Bayes Classifier (NBC) dengan pelabelan Transformers serta Ektraksi Fitur TF-IDF dan N-Gram untuk Analisis Sentimen Terhadap Penundaan Pemilu	MP Firdaus	2023	Dari penelitian ini diperoleh hasil akurasi tertinggi terhadap data uji mencapai 82,86% menggunakan Algoritma KNN sementara NBC sebesar 78,56%