BAB II
TINJAUAN PUSTAKA

2.1 Analisis Sentimen

Analisis Sentimen adalah suatu metode yang digunakan untuk mengekstrak data opini, memahami, dan secara otomatis mengolah data tekstual untuk melihat opini yang terkandung dalam suatu opini. (Sari & Wibowo, 2019)

Analisis Sentimen adalah proses otomatis untuk mengekstrak, memproses, dan memahami data dalam bentuk teks tidak terstruktur untuk memperoleh informasi sentimental yang terkandung dalam sebuah opini atau kalimat opini. (Brahimi, Touahria dan Tari, 2019)

2.2 *Hate Speech* (Ujaran Kebencian)

Ujaran kebencian menjadi bukti merusak fungsi bahasa. Sekarang semakin banyak orang yang tidak tertarik dengan bahasa Indonesia yang baik.(Syafyahya, 2018). Ujaran kebencian adalah tindakan bertujuan untuk menghasut khalayak ramai dalam melakukan seperti provokasi, hinaan dan tindakan celaan terutama hal yang mengandung SARA (Suku, Agama, Ras dan Antar golongan) yang ditunjukkan kepada seseorang, kelompok ataupun organisasi.

2.2 *Python*

Ujaran kebencian menjadi bukti merusak fungsi bahasa. Sekarang semakin banyak orang yang tidak tertarik dengan bahasa Indonesia yang baik.(Syafyahya, 2018). Ujaran kebencian adalah tindakan bertujuan untuk menghasut khalayak ramai dalam melakukan seperti provokasi, hinaan dan tindakan celaan terutama hal yang mengandung SARA yang ditunjukkan kepada seseorang, kelompok ataupun organisasi.

2.3 NLP(Natural Language Processing)

NLP atau *Natural Language Processing* adalah bidang ilmu komputer yang bertujuan untuk memahami konsep dan tujuan bahasa manusia. Manusia memiliki pemahaman yang sangat baik tentang sintaksis linguistik dan tata bahasa dan hubungan spasial implisit, tetapi komputer sangat sulit untuk menangani pertanyaan bahasa alami (Allen, 1995).

Python TextBlob adalah pustaka Python (versi 2 dan 3) yang digunakan untuk memproses data teks. TextBlob menyediakan API yang dapat digunakan untuk pemrosesan bahasa alami (NLP). seperti frase kata benda, analisis sentimen, klasifikasi, terjemahan, dll. Hasil dari objek TextBlob digunakan untuk memproses pembelajaran bahasa alami, dan perpustakaan TextBlob hanya dapat mengenali bahasa Inggris.(Parlika et al., 2020)

2.4 Machine Learning

Machine Learning adalah kecerdasan buatan (artificial intelligence) yang belajar mengembangkan sistem, menghasilkan data, dan membuat algoritma yang dapat melakukan tugas sendiri tanpa ada arahan dari pengguna, dan memungkinkan pemrogram untuk 'belajar'.(Luis & Moncayo, n.d.)

2.5 Rapidminer

RapidMiner adalah perangkat lunak yang berguna untuk jembatan ilmu Pendidikan khususnya ilmu data mining. Platform dibesarkan oleh industri yang ditujukan untuk seluruh masyarakat. Langkah-langkah dengan sejumlah besar data di perusahaan Perdagangan, penelitian, pendidikan, pelatihan, pembelajaran. RapidMiner diperkirakan mempunyai 100 solusi pembelajaran Pengelompokan, klasifikasi, dan analisis regresi. RapidMiner .xls, .csv, dan lain-lain.(Prasetyo et al., 2021)

2.6 Pembobotan Kata TF-IDF

Pembobotan kata atau *Term Weight* adalah metode pembobotan kata (*term*) yang memberikan bobot atau nilai pada kata yang terdapat dalam dokumen. Bobot nilai ini merupakan ukuran jumlah dan derajat kontribusi kata terhadap keputusan kelas atau kategori dalam sebuah dokumen. Ada beberapa metode pembobotan term, antara lain TF, TF-IDF, WIDF, dan TF-RF. (Deolika et al., 2019)

Data yang sudah menyelesaikan tahap preprocessing sebelumnya haruslah berbentuk numerik, untuk mengubah data yang berupa kata-kata atau String menjadi berbentuk numerik adalah menggunakan metode Pembobotan kata atau *Term Weight* yaitu TF-IDF. Metode TF-IDF (Term Frequency Invers Document Frequency) adalah langkah yang dipakai dalam memastikan sejauh mana kemiripan suatu *term* atau kata mengenai suatu kumpulan data melalui pemberian bobot pada setiap kata. Metode TF-IDF ini mengkombinasikan lebih dari satu konsep adalah frekuensi kemunculan sebuah kata dalam sebuah dokumen dan frekuensi kebalikan dari dokumen yang mengandung kata tersebut.(Deolika et al., 2019)

Rumus pembobotan kata TF-IDF adalah sebagai berikut:

$$\text{Wt} = \left(log \frac{n}{df}\right).Tf$$

Keterangan:

Wt : Pembobotan TF-IDF

n : Jumlah total dokumen atau data

df : Jumlah kata yang muncul dalam seluruh dokumen

Tf : Jumlah kemunculan kata pada masing-masing dokumen

2.7 K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) merupakan langkah pengelompokan yang dapat dikatakan sederhana untuk memisahkan suatu citra dengan melihat kedekatan dengan citra tetangganya. (Farokhah, 2020)

Algoritma K-Nearest Neighbor sangat cocok untuk memperkirakan peluang apa yang akan terjadi selanjutnya menggunakan kasus-kasus yang sudah ada. Dengan metode K-Nearest Neighbor maka akan sangat cocok dalam pengambilan keputusan berdasarkan kemiripan dengan kasus-kasus terdahulunya.

Tahapan Metode K-Nearest Neighbor, Langkah yang digunakan dalam metode *K-Nearest Neighbor*:

a. Tentukan parameter *K* (jumlah tetangga paling dekat).
b. Hitung kuadrat jarak euclid masing – masing objek terhadap data sampel yang diberikan.
c. Urutkan objek – objek kedalam kelompok yang memiliki jarak terkecil.
d. Kumpulkan kategori *Y* (Klasifikasi nearest neighbor).
e. Dengan kategori nearest neighbor yang paling banyak, maka dapat diprediksikan nilai query instance yang telah dihitung.

K-Nearest Neighbor dirumuskan sebagai berikut:

$$\cos(\theta ij) = \frac{\sum_k (dik.dij)}{\sqrt{\sum_k d^2 ik}\sqrt{\sum_k d^2 jk}}$$

Keterangan:

Cos(Øij) : *Similiarity K-Nearest Neighbor*
dik : bobot Data dokumen *Testing*
dij : bobot Data dokumen *Training*
d^2 : Panjang Vektor dokumen

Bobot Label pada Dataset

1 = Positif
2 = Netral
3 = Negatif

2.8 Pengujian Kota Hitam (Black Box Testing)

Pengujian black box adalah metode pengujian perangkat lunak yang mengevaluasi sistem atau aplikasi tanpa mengetahui implementasinya atau kode sumbernya. Dalam pengujian black box, tester hanya mengevaluasi sistem atau aplikasi melalui user interface yang disediakan tanpa mengetahui bagaimana sistem atau aplikasi bekerja di dalamnya. Pengujian black box dapat digunakan untuk mengevaluasi berbagai aspek sistem atau aplikasi, termasuk keandalan, kinerja, kompatibilitas, dan kepatuhan. Penguji dapat mengevaluasi sistem atau aplikasi dengan menjalankan serangkaian pengujian yang dirancang untuk menguji berbagai aspek sistem atau aplikasi.(Hidayat & Muttaqin, 2018)

Berbagai jenis pengujian yang biasa dilakukan dalam pengujian kotak hitam meliputi:

a. Functional Testing, yaitu pengujian yang dirancang untuk menilai apakah suatu sistem atau aplikasi dapat menjalankan fungsinya sesuai dengan yang diharapkan.
b. Compatibility Testing, yaitu pengujian yang dirancang untuk menilai apakah suatu sistem atau aplikasi dapat berinteraksi dengan perangkat keras atau perangkat lunak lain seperti yang diharapkan.
c. Performance testing, yaitu pengujian untuk mengevaluasi kinerja suatu sistem atau aplikasi dalam berbagai kondisi.
d. Usability Testing, bertujuan untuk mengevaluasi seberapa mudah user dalam menggunakan suatu sistem atau aplikasi.

2.9 Penelitian Sebelumnya

(Yustanti, 2012) Penelitian yang dilakukan oleh Wiyli Yustanti pada Tahun 2012 dengan judul “Algoritma K-Nearest Neighbour Untuk Memprediksi Harga Jual Tanah”. Peneliti menggunakan algoritma K-Nearest Neighbour dalam memperkirakan harga jual tanah menggunakan beberapa faktor tertentu. Dalam penelitian ini dapat disimpulkan bahwa pemerosesan cukup lama, hal ini dikarenakan data lama akan dibandingkan dengan data baru.

Penelitian ini mendapatkan tingkat akurasi sekitar 80% menggunakan algoritma K-Nearest Neighbour.

(Farokhah, 2020) Penelitian ini dilakukan oleh Lia Farokhah pada tahun 2019 dengan judul "Implementasi K-Nearest Neighbor Untuk Klasifikasi Bunga Dengan Ekstraksi Fitur Warna Rgb". Peneliti menggunakan algoritma K-Nearest Neighbour. Dalam mengklasifikasikan bunga dan Data yang digunakan adalah sebanyak 320. Yang terdiri dari bunga matahari, bunga daisy, bunga dandelion dan Bungan coltsfoot. Tingkat akurasi yang didapatkan dibagi menjadi 3, pada percobaan pertama, jika K=1 maka akurasi 57%, jika K=3 maka akurasi 64%, jika K=5 maka akurasi 70%. Pada percobaan kedua semua bagian meningkat menjadi 90%-100%.

(Yahya & Puspita Hidayanti, 2020) Peneliti menggunakan algoritma K-Nearest Neighbour dengan judul "Penerapan Algoritma K-Nearest Neighbor Untuk Klasifikasi Efektivitas Penjualan Vape (Rokok Elektrik) Pada "Lombok Vape On" " dan dataset yang digunakan adalah dataset penjualan pada "Lombok Vape On" dan mendapatkan hasil akurasi diperoleh sebesar 86.48% dan AUC sebesar 0.874%.
(Suwirmayanti, 2017) Penelitian dengan judul "Penerapan Metode K-Nearest Neighbor Untuk Sistem Rekomendasi Pemilihan Mobil" ini dapat membantu pengguna dalam memberikan pilihan dalam membeli mobil berdasarkan beberapa variable seperti tujuan pembelian mobil, harga mobil, tahun pembuatan mobil, kapasitas penumpang, warna, kapasitas mesin, jenis transmisi. Penelitian ini menggunakan algoritma KNN dan penelitian ini mengambil data melalui 3 cara yaitu studi pustaka, wawancara dan observasi.

(Setianto et al., 2019) Penelitian ini dilakukan oleh Yuni Ambar S[1], Kusrini[2], Henderi[3] pada tahun 2019 dengan judul "Penerapan Algoritma K-Nearest Neighbour Dalam Menentukan Pembinaan Koperasi Kabupaten Kotawaringin Timur". Penelitian ini bertujuan untuk mengelompokan data

kriteria penderita Thalassaemia berdasarkan umur, Hb level dan kebutuhan jumlah darah dengan pendekatan data mining menggunakan algoritma K-means. Penelitian ini menggunakan data dari 100 koperasi yang diambil dari Dinas Koperasi dan UKM Kabupaten Kotawaringin Timur, dan Menghasilkan akurasi tertinggi pada K = 7 yaitu sebesar 96,33%.

(FADILAH, 2021) Penelitian ini dilakukan oleh Aidil Fadilah pada tahun 2021 dengan judul "Penerapan Algoritma K-Nearest Neighbor Untuk Mendeteksi Ujaran Kebencian Dan Bahasa Kasar Pada Twitter Bahasa Indonesia". Data yang dipakai sebanyak 13.127 yang berasal dari aplikasi twitter. Dan Penelitian ini membagi akurasi menjadi 3, yaitu pada hate speech 79,13%, pada abusive 83,54% dan pada kelas level 73,56%.

(Windania Purba1, Fando Marehitno Salim 2, Antoni3 , Yuni Suhendrik4, n.d.) Penelitian dengan judul "Klasifikasi Komentar Bullying Pada Instagram Menggunakan Metode'k-Nearest'neighbor" ini digunakan untuk mengklasifikasi komentar bullying pada instagram menggunakan metode K-Nearest Neighbor. Untuk memprediksi tingkat keakuratan dengan menggunakan Teknik K-Nearest Neighbor pada pengkatagorian ulasan perundungan di media sosial Instagram. Data ulasan yang dipakai sebanyak 1000 komentar dengan analogi data sampel berjumlah 70:30, 80:20 dan 90:10, dan dihasilkan keakuratan sebanyak 87,07% dari penelitian ini sebesar.

Tabel 2.7.1 Penelitian Terkait

No	Judul	Algoritma	Akurasi	Dataset
1.	Algoritma K-Nearest Neighbour Untuk Memprediksi Harga Jual Tanah (Yustanti,2012)	K-Nearest Neighbour	80%	-

2.	Implementasi K-Nearest Neighbor Untuk Klasifikasi Bunga Dengan Ekstraksi Fitur Warna Rgb (Farokhah, 2020)	K-Nearest Neighbour	K = 1, akurasi 57% K = 3, akurasi 64% K = 5, akurasi 70%	320 data
3.	Penerapan Algoritma K-Nearest Neighbor Untuk Klasifikasi Efektivitas Penjualan Vape (Rokok Elektrik) Pada "Lombok Vape On" (Yahya & Puspita Hidayanti, 2020)	K-Nearest Neighbour	86,48% Dan AUC sebesar 0.874%	10 data
4.	Penerapan Metode K-Nearest Neighbor Untuk Sistem Rekomendasi Pemilihan Mobil (Suwirmayanti, 2017)	K-Nearest Neighbour	-	studi pustaka, wawancara dan observasi
5.	Penerapan Algoritma K-Nearest Neighbour Dalam Menentukan Pembinaan Koperasi Kabupaten Kotawaringin Timur (Setianto et al.,2019)	K-Nearest Neighbour	K = 7, akurasi 96,33%	100 data
6.	Penerapan Algoritma K-Nearest Neighbor Untuk Mendeteksi Ujaran Kebencian Dan Bahasa Kasar Pada Twitter Bahasa	K-Nearest Neighbour	hate speech 79,13%, pada abusive	13.127 tweet

	Indonesia (FADILAH, 2021)		83,54% dan pada kelas level 73,56%	
7.	Klasifikasi Komentar Bullying Pada Instagram Menggunakan Metode'k-Nearest'neighbor(Windania Purba1, Fando Marehitno Salim 2, Antoni3 , Yuni Suhendrik4, n.d.)	K-Nearest Neighbour	87,07%	1000 data
8.	Analisis Sentimen Pada Postingan *Hate Speech* Di Media Sosial Menggunakan Metode *K-Nearest Neighbor*	K-Nearest Neighbour & NLP (Natural Language Processing)	56,74%	703 data

Dalam penelitian ini hasil dari metode algoritma K-Nearest Neighbor akan dianalisis menggunakan Analisis Sentimen dalam menentukan apakah suatu postingan tersebut termasuk ujaran kebencian atau tidak.

Perbedaan dalam penelitian ini adalah cara dalam menentukan *Labeling Data* yaitu dengan menggunakan *Library* Phyton yaitu *TextBlob* dengan variable menentukan berapa banyak kata positif di dalam kalimatnya dan jumlah data serta keterbaruan data dari penelitian sebelumnya. Dalam jurnal penelitian 1 sampai 5 metode yang digunakan adalah K-NN dengan pendekatan jarak yaitu *Ecludian Distance,* sedangkan untuk penelitian 6 sampai 8 menggunakan metode K-NN *Similiarity* atau kemiripan dengan kasus sebelumnya.