BAB II
TINJAUAN PUSTAKA

2.1. Ujaran Kebencian

Ujaran kebencian (*hate speech*) adalah ujaran yang mengandung kebencian, menyerang dan berkobarkobar yang dimaksudkan untuk menimbulkan dampak tertentu, baik secara langsung (aktual) maupun tidak langsung (berhenti pada niat) yaitu menginspirasi orang lain untuk melakukan kekerasan atau menyakiti orang atau kelompok lain, berdasarkan Buku saku penanganan ujaran kebencian (*hate speech*). [3]. Dalam arti hukum, *hate speech* adalah perkataan, perilaku, tulisan, ataupun pertunjukan yang dilarang karena dapat memicu terjadinya tindakan kekerasan dan sikap prasangka entah dari pihak pelaku, atau korban dari tindakan tersebut. Situs internet yang menggunakan atau menerapkan *hate speech* ini disebut *hate site*. Kebanyakan dari situs ini menggunakan forum internet dan berita untuk mempertegas sudut pandang tertentu.

Para kritikus berpendapat bahwa istilah *hate speech* adalah contoh modern dari novel *Newspeak*, ketika *hate speech* digunakan untuk memberikan kritik secara diam-diam kepada kebijakan sosial yang diimplementasikan dengan buruk dan seakan-akan kebijakan tersebut terlihat benar secara politik.

Sementara di Indonesia, R. Susilo menjelaskan bahwa yang dimaksud dari "menghina" adalah "menyerang kehormatan dan nama baik seseorang" dan yang terkena dampak *hate speech* biasanya merasa malu. Menurutnya, penghinaan terhadap satu individu ada 6 macam yaitu;

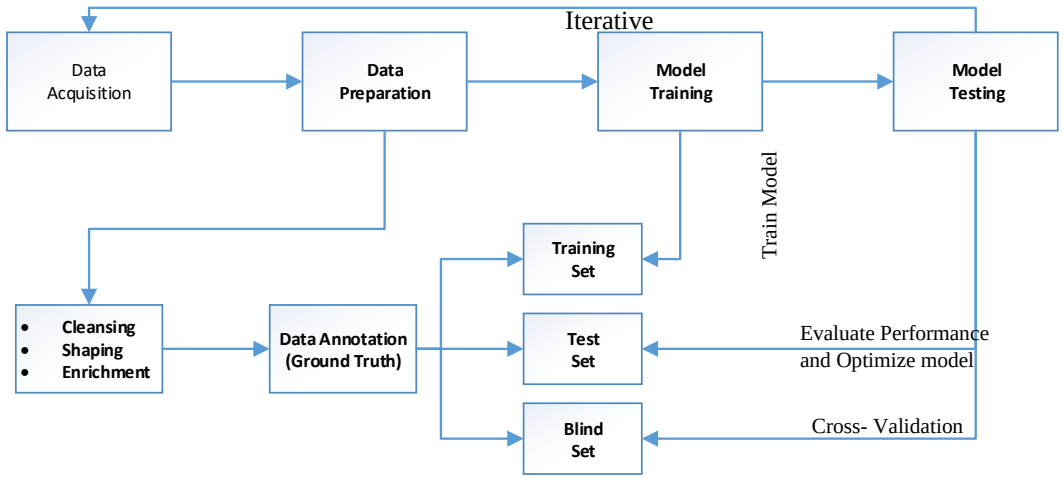
a) Menista secara lisan (smaad)
b) Menista dengan surat/tertulis (smaadschrift)
c) Memfitnah (laster)
d) Penghinaan ringan (eenvoudige belediging)
e) Mengadu secara memfitnah (lasterlijke aanklacht)
f) Tuduhan secara memfitnah (lasterlijke verdachtmaking)

2.2. Machine Learning

Machine learning adalah teknik atau ilmu yang menggunakan algoritma serta statistik komputasi untuk belajar dari data tanpa diprogram secara eksplisit. Teknik machine learning digunakan untuk menemukan pola dalam data yang kompleks. Pola serta pengetahuan tersembunyi dapat digunakan untuk memprediksi peristiwa pada masa depan dan melakukan pengambilan keputusan yang kompleks, menurut D. Elisabeth dkk[5].

Machine learning dikategorikan menjadi supervised leaning, unsupervised learning dan reinforcement learning. Supervised learning mengambil data berlabel dan membuat model untuk memprediksi dari data baru. Kategori ini dapat berupa masalah clasification dan regression. Unsupervised learning mengambil data tidak berlabel untuk menemukan pola dan membuat struktur dalam data.

Machine learning adalah program komputer yang dapat dilatih dengan data. Jenis data dalam machine learning yaitu data terstruktur dan data tidak terstruktur. Data terstruktur misalnya nama, umur dan tempat tanggal lahir. Data tersebut dapat disimpan dalam tabel berbasis baris dan kolom. Data tidak terstruktur misalnya file gambar, video dan suara. Data tersebut tidak dapat disimpan dalam tabel. Gambar 2.1 merupakan gambar *Typical Machine Learning Flow Diagram.*



Gambar 2. 1 *Typical Machine Learning Flow Diagram*

2.3. *Text Mining*

Menurut Mailoa[9], *Text mining* adalah salah satu teknik yang dapat digunakan untuk melakukan klasifikasi dokumen, *clustering, information extraction*, analisis sentimen dan information retrival dimana text mining merupakan variasi dari data mining yang berusaha menemukan pola yang menarik dari sekumpulan data tekstual yang berjumlah besar. Text mining dapat memberikan solusi dari permasalahan seperti pemrosesan, pengorganisasian atau pengelompokkan dan menganalisa unstructured text dalam jumlah besar. Dalam memberikan solusi, text mining mengadopsi dan mengembangkan banyak teknik dari bidang lain, seperti *Data mining, Information Retrieval*, Statistik dan Matematik, *Machine Learning*, *Linguistic*, *Natural Languange Processing* (NLP), dan *Visualization.*

Tujuan dari *Text Mining* adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen, tetapi tujuan utama text mining adalah mendukung proses *knowledge discovery* pada koleksi dokumen yang besar. Adapun tugas khusus dari text mining antara lain yaitu pengkategorisasian teks (*text categorization*) dan pengelompokkan teks *(text clustering*). *text mining* memiliki peran penting dalam bidang data mining. Dengan mengaplikasikan proses-proses dalam *text mining*, maka akan diperoleh pola-pola data, tren, dan ekstraksi dari pengetahuan-pengetahuan yang potensial dari data teks. Diantara proses yang dapat dilakukan dalam *text mining* adalah klasifikasi teks. Klasifikasi teks dapat didefinisikan sebagai proses untuk menentukan suatu dokumen teks ke dalam suatu kelas tertentu. Untuk melakukan proses klasifikasi teks, ada beberapa algoritma yang dapat digunakan diantaranya *Support Vector Machine (SVM), Naive Bayes, k-Nearest Neighbor (KNN), Decision Tree, Logistic Regression,* dan *Artificial Neural Networks (ANN).*

2.4. Analisis Sentimen

Analisis Sentimen (opinion mining) merupakan suatu bidang yang luas dari pengolahan bahasa alami, komputasi linguistik dan text mining dimana memiliki tujuan dalam menganalisa pendapat, sentiment, evaluasi, sikap, penilaian dan emosi seseorang apakah pembicara atau penulis berkenaan dengan suatu topik, produk, layanan, organisasi, individu, ataupun kegiatan tertentu, Kolkata dkk[10]. Analisis

Sentimen atau bisa disebut dengan opinion mining juga dapat memahami proses, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam suatu kalimat opini. Analisis sentimen dilakukan untuk melihat pendapat atau kecenderungan opini terhadap sebuah masalah atau objek oleh seseorang, apakah cenderung berpandangan atau beropini negatif atau positif. Sentiment analysis memiliki tugas dasar dalam mengelompokkan teks yang ada pada sebuah kalimat atau dokumen kemudian menentukan pendapat yang dikemukakan dalam kalimat atau dokumen tersebut apakah bersifat positif, negatif atau netral, menurut Aulia and Amelia[11].

2.5. *Support Vector Machine* (SVM)

Nasution and Hayaty berpendapat[12], *Support Vector Machine* (SVM) adalah salah satu metode klasifikasi yang banyak dikembangkan saat ini. Konsep dasar metode ini adalah memaksimalkan batas *hyperplane* yang memisahkan suatu set data. *Support Vector Machine* (SVM) adalah metode klasifikasi yang menggunakan konsep mencari *hyperplane* (bidang pemisah) yang optimal dalam suatu ruang *feature* untuk memisahkan dua kelas. *Hyperplane* yang dicari adalah yang memberikan jarak paling jauh dari setiap titik data. SVM dapat digunakan untuk klasifikasi *linearly separable* dan *non-linearly separable*.

a) SVM linear

   Mencari hyperplane yang memisahkan dua kelas dengan cara mencari garis yang paling baik memisahkan antara dua kelas. Rumus dasar SVM linear adalah:

   $$f(x) = wx + b \quad (1)$$

   di mana x adalah input (vektor fitur), w adalah vektor bobot, dan b adalah bias. Untuk membuat klasifikasi, maka ditambahkan dengan fungsi kondisional seperti:

   $$y = \{1\ if\ f(x) \geq 0, -1\ otherwise\} \quad (2)$$

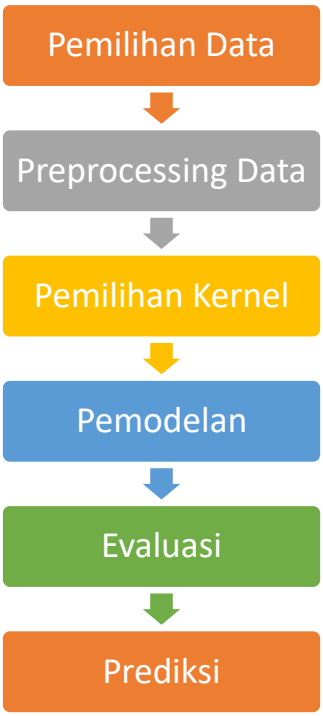
b) SVM non-linear

Jika data tidak linearly separable, maka akan digunakan kernel trick untuk mengubah data ke dalam dimensi yang lebih tinggi sehingga dapat di separasi dengan linear. Rumus dasar SVM non-linear adalah:

$$f(x) = \sum(a\text{a\_iy\_iK(x\_i, x)}) + b \qquad (3)$$

di mana x adalah input (vektor fitur), a adalah vektor bobot, y adalah label, K adalah kernel yang digunakan (misalnya radial basis function, polynomial, atau sigmoid), b adalah bias.

SVM mencari hyperplane yang memberikan jarak paling jauh dari setiap titik data. Untuk menemukan nilai a_i dan b yang optimal digunakan metode Quadratic Programming (QP) yang mencari solusi yang meminimalkan fungsi objective dengan memperhatikan kendala yang ditentukan. Secara umum, SVM mencari solusi optimal dari persamaan quadratic programming (QP) yang di dalamnya termasuk kendala dari persamaan yang akan di optimalkan. Gamabar 2.2 merupakan Alur umum *Support Vector Machine*.



Gambar 2. 2 Alur umum *Support Vector Machine* SVM

Alur umum dalam penerapan Support Vector Machine (SVM) adalah sebagai berikut:

a) Pemilihan data: Pertama-tama, data yang akan digunakan dalam SVM harus dipilih dan disiapkan. Data ini harus terdiri dari fitur (variabel independen) dan label (variabel dependen) yang akan digunakan untuk membuat klasifikasi.
b) Preprocessing data: Data yang telah dipilih harus diolah sebelum digunakan dalam SVM. Langkah-langkah ini dapat termasuk pembersihan data, pengukuran normalisasi, atau pengembangan fitur.
c) Pemilihan kernel: Jika data tidak linearly separable, maka kernel trick harus digunakan untuk mengubah data ke dalam dimensi yang lebih tinggi sehingga dapat di separasi dengan linear. kernel yang digunakan bisa berupa radial basis function (RBF), polynomial, atau sigmoid.
d) Pemodelan: Langkah selanjutnya adalah pemodelan dengan mencari solusi yang optimal dari persamaan quadratic programming (QP) yang di dalamnya termasuk kendala dari persamaan yang akan di optimalkan.
e) Evaluasi model: Melakukan evaluasi model dengan menggunakan metrik seperti Accuracy, Precision, Recall, dan F1-Score.
f) Prediksi: Menggunakan model yang telah dibangun untuk melakukan prediksi pada data baru.

2.6. *Logistic Regression*

Regresi Logistik (Logistic Regression atau LR) adalah sebuah metode statistik yang digunakan untuk melakukan klasifikasi pada data dengan dua kelas target, yaitu kelas positif dan negatif. LR umumnya digunakan dalam tugas klasifikasi biner. LR menggunakan batas ambang untuk memisahkan ulasan positif dari yang negatif. LR menggunakan fungsi logistik untuk memperkirakan probabilitas antara label positif atau negatif y dan fitur data w yang diberikan oleh input x. Dengan demikian, LR, seperti yang ditunjukkan dalam persamaan (3), menggunakan fungsi sigmoid untuk mendapatkan likelihood secara langsung dengan meminimalkan nilai infinitif $+\infty$ dan $-\infty$ menjadi skala antara 0 hingga 1, Omari dkk[13].

Dalam konteks klasifikasi, LR digunakan untuk memprediksi probabilitas kelas target berdasarkan serangkaian fitur atau variabel prediktor.

Rumus Regresi Logistik (LR) adalah sebagai berikut:

$$P(Y=1|X) = 1 / (1 + e^{-(\beta0 + \beta1X1 + \beta2X2 + ... + \betanXn)}) \quad (1)$$

Di mana:

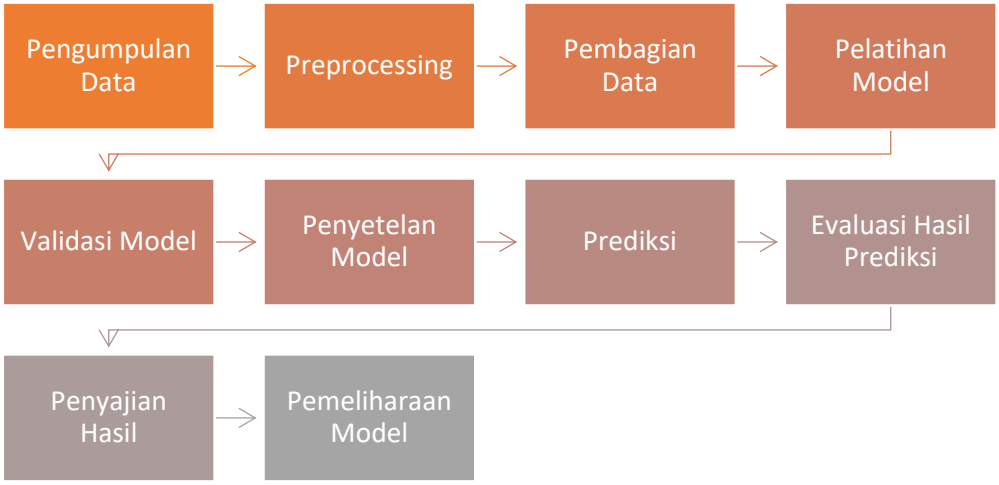
- P(Y=1|X) adalah probabilitas bahwa variabel dependen Y (kelas target) memiliki nilai 1 (positif) berdasarkan nilai-nilai variabel prediktor X.
- e adalah bilangan dasar logaritma.
- β0, β1, β2, ..., βn adalah koefisien yang harus ditentukan selama proses pelatihan model. Koefisien ini menggambarkan pengaruh relatif dari masing-masing variabel prediktor terhadap probabilitas kelas target.

Pada rumus di atas, ekspresi β0 + β1X1 + β2X2 + ... + βnXn merupakan fungsi linier yang menggabungkan nilai-nilai variabel prediktor dengan koefisien terkait. Fungsi ini menghasilkan log-odds atau logit, yang merupakan logaritma dari peluang relatif antara kelas target positif dan negatif.

Untuk mengonversi logit menjadi probabilitas yang dapat diinterpretasikan, digunakan fungsi sigmoid atau logistik. Fungsi sigmoid mengubah logit menjadi angka antara 0 dan 1, yang merepresentasikan probabilitas kelas target positif.

Dalam pelatihan model LR, langkah utama adalah mengestimasi nilai-nilai koefisien (β0, β1, β2, ..., βn) yang optimal melalui proses yang disebut pembelajaran. Hal ini dapat dilakukan dengan menggunakan metode optimasi seperti Maximum Likelihood Estimation (MLE) atau metode lainnya.

Dengan memperoleh koefisien yang sesuai, model Regresi Logistik dapat digunakan untuk memprediksi probabilitas kelas target positif berdasarkan nilai-nilai variabel prediktor baru. Nilai probabilitas tersebut kemudian dapat digunakan sebagai dasar untuk mengklasifikasikan sampel ke dalam kelas positif atau negatif dengan menggunakan ambang batas (threshold) yang ditentukan. Gambar 2.3 merupakan Alur umum *Logistic Regression*



Gambar 2. 3 Alur umum *Logistic Regression*

Alur umum dalam penerapan Logistic Regression adalah sebagai berikut:

a) Pengumpulan Data: Kumpulkan data yang akan digunakan untuk membangun model Regresi Logistik. Data ini terdiri dari variabel dependen (kelas target) yang bersifat biner (misalnya, 0 atau 1) dan variabel independen atau fitur yang digunakan untuk memprediksi kelas target.
b) Preprocessing Data: Lakukan preprocessing data untuk membersihkan dan mempersiapkan data sebelum digunakan dalam model. Langkah-langkah preprocessing ini dapat mencakup mengisi nilai-nilai yang hilang, mengatasi outliers, dan mengubah format data jika diperlukan.
c) Pembagian Data: Bagi data menjadi dua subset: data pelatihan (training data) dan data uji (test data). Data pelatihan akan digunakan untuk melatih model, sementara data uji akan digunakan untuk menguji kinerja model yang telah dilatih.
d) Pelatihan Model: Latih model Regresi Logistik menggunakan data pelatihan. Proses pelatihan ini melibatkan estimasi nilai-nilai koefisien (β0, β1, β2, ..., βn) yang optimal dengan meminimalkan fungsi log-likelihood atau fungsi kesalahan lainnya. Teknik seperti Maximum Likelihood Estimation (MLE) sering digunakan untuk mengestimasi koefisien.
e) Validasi Model: Setelah model dilatih, evaluasi kinerjanya menggunakan data uji. Ukur metrik seperti akurasi, presisi, recall, atau area di bawah kurva

Receiver Operating Characteristic (ROC AUC) untuk mengevaluasi sejauh mana model dapat mengklasifikasikan data dengan benar.

f) Penyetelan Model: Jika diperlukan, lakukan penyetelan model dengan mencoba variasi hyperparameter atau teknik regularisasi untuk meningkatkan kinerja model pada data uji.
g) Prediksi: Setelah model dianggap cukup baik, gunakan model tersebut untuk memprediksi kelas target pada data baru yang belum pernah dilihat sebelumnya.
h) Evaluasi Hasil Prediksi: Evaluasi hasil prediksi model pada data baru untuk mengukur kinerjanya. Jika model memiliki kinerja yang memuaskan, model dapat digunakan untuk klasifikasi lebih lanjut. Jika tidak, Anda mungkin perlu kembali ke langkah sebelumnya untuk melakukan penyetelan lebih lanjut.
i) Penyajian Hasil: Sajikan hasil dari model Regresi Logistik dengan cara yang mudah dimengerti untuk memfasilitasi pengambilan keputusan berdasarkan prediksi model.
j) Pemeliharaan Model: Model Regresi Logistik perlu dipelihara secara berkala untuk memastikan kinerjanya tetap optimal. Hal ini mungkin melibatkan pembaruan data pelatihan, penyetelan ulang hyperparameter, atau penggantian model jika diperlukan.

2.7. *K-Nearest Neighbor*

Menurut Farokhah[14], *K-Nearest Neighbor* (KNN) merupakan langkah pengelompokan yang dapat dikatakan sederhana untuk memisahkan suatu citra dengan melihat kedekatan dengan citra tetangganya.

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:

$$\cos(\theta ij) = \frac{\sum_k (dik.dij)}{\sqrt{\sum_k d^2 ik}\sqrt{\sum_k d^2 jk}} \quad (1)$$

Keterangan:

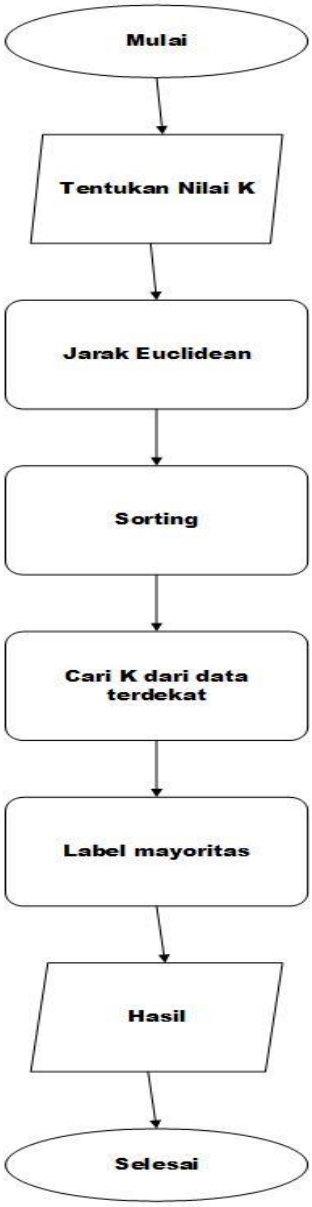
Cos(Øij) : *Similiarity K-Nearest Neighbor*
dik : bobot Data dokumen *Testing*
dij : bobot Data dokumen *Training*
d^2 : Panjang Vektor dokumen

Bobot Label pada Dataset

1 = Positif

2 = Netral

3 = Negatif



Gambar 2. 4 *Flowchart K-Nearest Neighbor*

2.8. Decision Tree

Decision Tree C4.5 merupakan salah satu algoritma klasifikasi yang banyak digunakan untuk memperoleh hasil klasifikasi non biner. Menurut Supangat dkk[15], dibanding algoritma sejenis, *Decision Tree* C4.5 memiliki kelebihan pada kemampuan untuk mengelola data dalam berbagai format. Struktur algoritma ini mirip *flowchart* dimana masing-masing node mewakili nilai atribut dan masing masing cabang merepresentasikan hasil pengujian, dan masing-masing daun merepresentasikan kelas atau distribusi kelas. *Decision Tree* C4.5 merupakan

bentuk pengembangan dari algoritma ID3 yang mengadopsi pendekatan *greedy* dengan pengambilan keputusan berdasarkan *tree* yang terbentuk menggunakan pendekatan *rekursif top down* dan sistem bagi serang.

Algoritma *Decision Tree* C4.5 memiliki kelebihan dibanding ID3 dan CART karena kemampuannya untuk tidak membatasi cabang dalam bentuk biner. Selain itu, C4.5 secara *default* juga memis ahkan cabang untuk masingmasing nilai ke dalam atribut kategori sehingga klasifikasi yang dihasilkan lebih terkelompokkan dibanding ekspektasi.

Berdasarkan kondisi tersebut dapat dilihat bahwa metode *Decision Tree* C4.5 merupakan metode yang dapat memberikan tingkat akurasi tinggi pada kasus-kasus prediktif dengan beberapa atribut kategori.

Secara umum, langkah-langkah pembentukan algoritma C4.5 adalah sebagai berikut:

a) Langkah 1: tentukan atribut root node
b) Langkah 2: tentukan cabang untuk masing-masing nilai atribut
c) Langkah 3: pisahkan kasus sesuai cabang
d) Langkah 4: Ulangi proses hingga semua kasus dalam cabang memiliki kelas yang sama

Untuk menentukan atribut yang menjadi root, dilakukan penghitungan nilai gain untuk masing-masing atribut menggunakan Rumus (1) berikut:

$$Gain\ (S, A) = Entropy\ (s) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i) \quad (1)$$

dimana: {S1, S2, S3,..., Si, ..., Sn} = partisi S sejumlah nilai atribut A n = jumlah atribut A

$|S_i|$ = Jumlah kasus dalam partisi S_i

|S| = Total kasus S

Sementara untuk memperoleh nilai entropi digunakan Rumus (2) berikut.

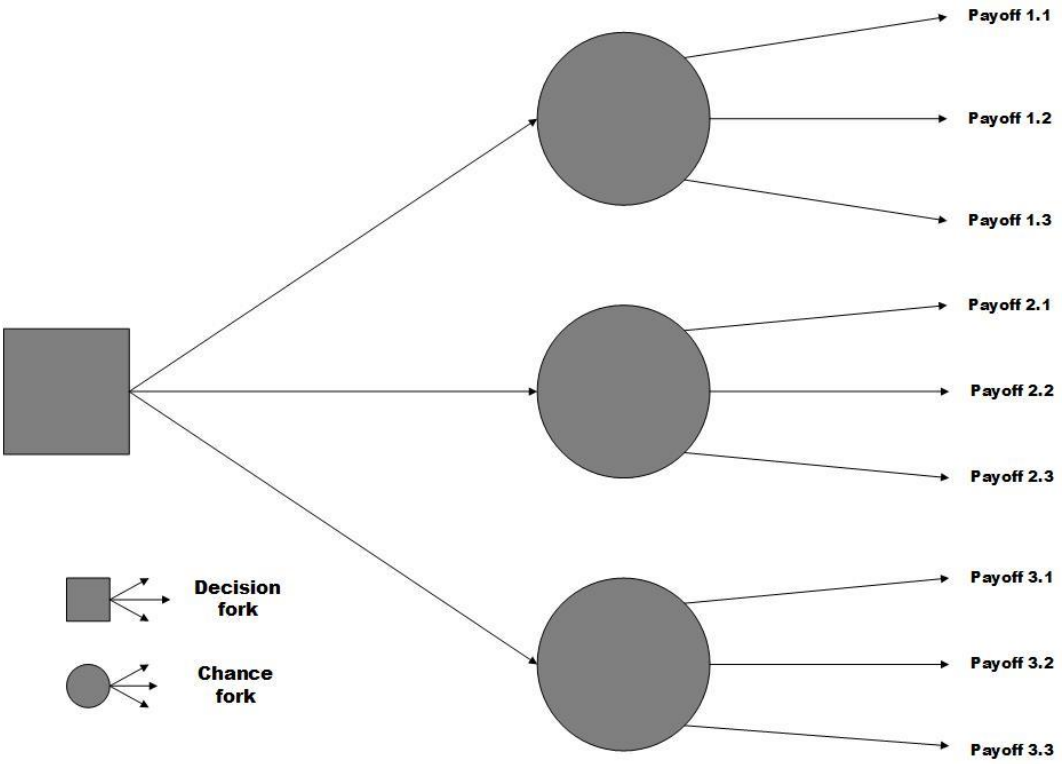
$$Entropy\ (S) = \sum_{i=1}^{n} - p_i * log_2 p_i \quad (2)$$

dimana:

S = jumlah kasus

n = jumlah kasus dalam partisi S

pi = proporsi Si terhadap S



Gambar 2. 5 Bentuk *Decision Tree* Secara umum

2.9. Twitter

Twitter merupakan sosial media masif yang berubah menjadi situs berbagi informasi dan berkomunikasi secara cepat. Kecepatan dan kemudahan *Twitter* dalam hal publikasi, membuat *Twitter* menjadi sebuah medium pilihan bagi pengguna untuk berkomunikasi setiap hari. *Twitter* mempunyai peran dan andil penting dalam pergerakan sosial-politik seperti *Arab Spring* dan *The Occupy Wall Street movement*. *Twitter* juga dapat digunakan untuk melakukan laporan kerusakan dan persiapan informasi terkait bencana pada saat bencana alam akan dan sedang terjadi, Suryono[16].

a) Glosarium Twitter

Glosarium Twitter berisi kosakata dan istilah yang sering digunakan untuk membahas fitur dan aspek dari *Twitter*. Berikut ini merupakan kosakata pada *Twitter* berdasarkan *Support.*

1. @. Simbol @ digunakan untuk memanggil nama pengguna dalam *Tweet*: "Halo @twitter!" Orang lain akan menggunakan

@namapengguna Anda untuk menyebut Anda di *Tweet* dan mengirim *Direct Message* atau tautan ke profil Anda.

2. @username. Anda dikenali di *Twitter* melalui nama pengguna yang selalu diawali simbol @. Misalnya, Bantuan Twitter adalah @BantuanTwitter.
3. #hashtag. *Hashtag* adalah kata atau frasa yang diawali langsung dengan simbol #. Bila Anda melakukan klik atau menyentuh *hashtag*, Anda akan melihat *Tweet* lain yang berisi kata kunci atau topik yang sama.
4. Geolokasi. Dengan menambahkan lokasi pada *Tweet* (geolokasi atau *geotag*), pengguna yang melihat *Tweet* anda akan mengetahui lokasi Anda saat mengirimkan *Tweet*.
5. *Time Stamp*. Tanggal dan waktu ketika *Tweet* dikirim ke *Twitter*. Cap waktu *Tweet* terlihat sebagai teks abu-abu di setiap tampilan rincian

   Tweet.
6. *Following*. Berlangganan ke sebuah akun *Twitter* disebut "mengikuti". Untuk mulai mengikuti, klik atau sentuh tombol ikuti di samping nama akun atau di halaman profil mereka untuk melihat *Tweet* mereka begitu mereka mengirim sesuatu yang baru. Pengguna di *Twitter* dapat mengikuti atau berhenti mengikuti pengguna lain kapan pun, kecuali akun yang diblokir.
7. *Follower*. Mengikuti dihasilkan dari pengguna yang mengikuti akun *Twitter* Anda. Anda dapat mengetahui jumlah mengikuti (atau pengikut) yang Anda miliki dari profil *Twitter* anda.
8. *Retweet*. Tindakan menyebarkan *Tweet* akun lain ke semua pengikut Anda dengan mengeklik atau menyentuh tombol *Retweet.*

b) Streaming API

Streaming API merupakan fitur pada *Twitter* yang membantu *developer* untuk melakukan akses secara langsung ke dalam *stream global Twitter* dengan *latency* yang rendah, sehingga memudahkan *developer* untuk

melakukan pengambilan data. Beberapa tipe *endpoint* dalam *Streaming API* berdasarkan *Support Twitter.*

1. *Public Streams*. Menyediakan *streams* yang berasal dari data publik yang bergabung dengan *Twitter*. Jenis *endpoint* ini berguna untuk mencari *user* tertentu, mencari topik, dan melakukan *data mining.*
2. *User Streams*. *Single-user streams* yang menyediakan seluruh data yang berkesesuaian dengan seluruh informasi mengenai urus pilihan.
3. *Site Streams*. S*treams* untuk melakukan pencarian data yang dikhususkan untuk mecari seluruh informasi pada banyak *user*. *Endpoint* ini mengharuskan *developer* untuk melakukan koneksi ke *Twitter* dengan otentikasi banyak *user.*

2.10. Klasifikasi

Klasifikasi adalah urutan yang sangat penting dalam data komunitas pertambangan. Klasifikasi adalah salah satu prediksi teknik data mining yang membuat prediksi tentang data nilai menggunakan hasil yang diketahui yang ditemukan dari kumpulan data yang berbeda. Masalah akurasi dari banyak algoritma klasifikasi adalah diketahui mengalami penurunan informasi saat dihadapi dengan data yang tidak seimbang, misalnya ketika distribusi sampel lintas kelas sangat miring[17]. Dalam klasifikasi, ada variabel kategoris target, seperti braket pendapatan, yang, misalnya, dapat dipartisi menjaditiga kelas atau kategori: berpenghasilan tinggi, menengah pendapatan, dan pendapatan rendah. Menurut Apostolakis[18], model data mining memeriksa satu set besar catatan, masing-masing catatan yang berisi informasi tentangvariabel target serta satu set input atau prediktor variable. Contoh tugas klasifikasi dalam bisnisdan penelitian meliputi:.

a) Menentukan apakah transaksi kartu kredit tertentu adalah penipuan
b) Menempatkan siswa baru pada jalur tertentu yang berkaitan dengan kebutuhankhusus
c) Menilai apakah aplikasi hipotek adalah risiko kredit yang baik atau buruk
d) Mendiagnosis apakah ada penyakit tertentu

e) Menentukan apakah surat wasiat ditulis oleh almarhum yang sebenarnya, ataucurangoleh orang lain
f) Mengidentifikasi apakah perilaku keuangan atau pribadi tertentu menunjukkankemungkinan ancaman teroris

Klasifikasi yang dilakukan secara manualadalahklasifikasi yang dilakukan oleh manusiatanpa adanya bantuan dari algortima cerdas komputer. Sedangkan klasifikasi yang dilakukan dengan bantuan teknologi, memiliki beberapa algoritma, diantaranya Naïve Bayes, SupportVector Machine, Decission Tree, Fuzzy dan Jaringan Saraf Tiruan, Wibawa dkk[19].

2.11. Lexicon

Menurut Al-Shabi[20], lexicon adalah kumpulan fitur seperti kata-kata dan klasifikasi emosi untuk setiap kata. Metode analisis sentimen ini didasarkan pada perbandingan kata-kata yang digunakan dalam teks dengan salah satu lexicon yang telah dipersiapkan sebelumnya. Lexicon juga dapat berupa kumpulan kata-kata yang diberi label atau skor sentimen, seperti positif, negatif, atau netral. Fungsi utama lexicon adalah sebagai panduan untuk mengidentifikasi sentimen atau emosi yang terkandung dalam teks yang dianalisis. Sentimen analisis teks bertujuan untuk menentukan perasaan atau pendapat yang diungkapkan oleh penulis teks, dan lexicon menjadi salah satu komponen penting dalam memahami dan mengklasifikasikan sentimen.

Lexicon dalam sentimen analisis teks adalah kumpulan kata-kata yang diberi label sentimen (misalnya positif, negatif, atau netral). Lexicon ini berfungsi sebagai panduan untuk mengidentifikasi dan mengekstraksi sentimen atau emosi yang terkandung dalam teks yang dianalisis. Penerapan lexicon dalam sentimen analisis teks bergantung pada bahasa yang digunakan, sehingga untuk bahasa Indonesia, kita menggunakan lexicon bahasa Indonesia.

Penerapan lexicon bahasa Indonesia dalam sentimen analisis teks dapat membantu dalam banyak aplikasi, seperti analisis sentimen media sosial, ulasan produk, atau berita. Namun, perlu diingat bahwa pendekatan ini tidak selalu akurat karena bergantung pada keberadaan kata-kata sentimen dalam lexicon. Oleh karena itu,

kombinasi dengan metode machine learning seperti KNN, Naive Bayes, atau deep learning dapat meningkatkan performa analisis sentimen teks secara keseluruhan.

2.12. Sastrawi

Sastrawi adalah proyek sumber terbuka yang bertujuan untuk mengembangkan alat pemrosesan bahasa alami (NLP) khusus untuk bahasa Indonesia. Sastrawi menyediakan pustaka (library) pemenggalan kata yang berfungsi untuk mengatasi masalah perubahan kata menjadi kata dasar. Pemenggalan kata (stemming) dalam Sastrawi diterapkan menggunakan algoritma berdasarkan Nazief dan Adriani, yang kemudian diperkuat dengan algoritma CS (Confix Stripping), algoritma ECS (Enhanced Confix Stripping), dan lebih lanjut diperbaiki dengan Modified ECS, Rosid dkk[21].

Lebih spesifik, Sastrawi merupakan pustaka untuk bahasa pemrograman PHP yang menyediakan berbagai alat pemrosesan teks dalam bahasa Indonesia, termasuk pemenggalan kata (stemming), analisis sentimen, pengenalan entitas bernama, dan lain-lain. Penggunaan Sastrawi sangat berguna dalam berbagai aplikasi NLP di Indonesia, termasuk analisis sentimen teks. Salah satu fitur utama Sastrawi adalah algoritma stemming untuk bahasa Indonesia. Stemming adalah proses menghilangkan imbuhan atau awalan kata sehingga hanya tersisa kata dasar. Penggunaan stemming ini dapat membantu dalam mengenali pola sentimen dalam kalimat tanpa harus memperhitungkan perbedaan kata berimbuhan. Selain itu, Sastrawi juga dilengkapi dengan kamus kata positif dan negatif (lexicon) yang membantu dalam analisis sentimen.

Penerapan Sastrawi dalam analisis sentimen teks memiliki kegunaan utama pada proses stemming, di mana stemming membantu dalam mengenali pola sentimen tanpa harus mempertimbangkan imbuhan kata. Hal ini memungkinkan sistem untuk lebih efektif mengklasifikasikan sentimen dalam teks bahasa Indonesia.

2.13. Akurasi

Menurut Nasution and Hayaty [12], akurasi adalah salah satu metrik untuk mengevaluasi model klasifikasi. Secara informal,akurasiadalah sebagian kecil dari

prediksi model kami yang benar. Secara formal, akurasi memiliki definisi sebagai berikut:

$$Akurasi = \frac{Number\ of\ Cprreect\ Prediction}{Total\ Number\ of\ Prediction} \quad (1)$$

Untuk klasifikasi biner, akurasi juga dapat dihitung dalam hal positif dan negatif sebagai berikut :

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Dimana

TP = True Positif

TN = True Negatif

FP = False Positif

FN = False Negatif

2.14. Hasil Penelitian yang Relevan

Berikut adalah beberapa hasil penelitian yang menjadi referensi dan memberikan banyak masukan kepada penulis:

Table 2. 1 Hasil Penelitian yang Relevan

No	Judul	Nama & Tahun	Metode	Fokus Deteksi	Jumlah Data	Akurasi
1	Hate Code Detection in Indonesian Tweets using Machine Learning Approach: A Dataset and Preliminary Study	[5]	Logistic Regression (LR), Naive Bayes (NB), dan Random Forest Decision Tree (RFDT)	Bentuk frasa kata	13.169 tweet	Skor f-measure terbaik adalah 94,90%
2	Hate Speech Detection In Indonesian Language Instagram	[6]	Metode word2vec dengan model skip-gram dan TextCNN yang dimodifikasi dan metode random oversampling	Teks ujaran kebencian	13.194 comments	Akurasi terbaik, dalam hal F-score. adalah 93,70%
3	Hate Speech Identification in Text Written in Indonesian with Recurrent Neural Network	[7]	Algoritma GRU dan algoritma LSTM	Data teks Twitter	415.844 tweet	GRU 85,37% F1-score dan LSTM 76,30% F1-score.
4	Hate Speech Detection In Indonesian Language On Instagram Comment Section	[22]	K-Nearest Neighbor	Data teks	83.752 comments	Akurasi 98,13%, dengan presisi sebesar 98%,

	Using K-Nearest Neighbor Classification Method					metode KNearest Neighbor dengan K=3.
5	Analisis Sentimen Ujaran Kebencian Dalam Postingan Twitter Menggunakan Pendekatan Machine Learning	-	*Support Vector Machine, Logistic Regression, K-Nearest Neighbor* dan *Decision Tree*	Data teks	51.250 tweet	?