BAB 2
DASAR TEORI

2.1 KAJIAN PUSTAKA

Telah ada beberapa penelitian sebelumnya terkait deteksi berita palsu seperti yang dilakukan oleh Faisal Rahutomo, Inggrid Yanuar Risca Pratiwi, dan Diana Mayangsari Ramadhan mengenai berita *hoax* menggunakan *naïve bayes* yang tertuang dalam Jurnal Penelitian Komunikasi dan Opini Publik Vol.23 No.1, Juni 2019: 1 - 15. Proses yang dilalui mulai dari mengumpulkan dataset, melakukan manual tagging pada setiap data berita yang telah dikumpulkan dengan *voting* diantara koresponden, hingga merancang sistem bangun yang dibuat, penelitian ini berfokus pada berita *hoax* yang berbahasa Indonesia saja. Penelitian ini menghasilkan nilai rata-rata *precision hoax* adalah 81%, 80% dan 72,5% dan untuk nilai rata-rata *precision* alid adalah 83,7%, 84,6% dan 89% untuk pengujian statis, dan menghasilkan luaran bahwa berita tersebut adalah valid pada pengujian dinamis [7].

Hery Mustofa, dan Adzhal Arwani Mahfudh dalam jurnal WJIT: Walisongo *Journal of Information Technology* - Vol 1 No. 1, melakukan penelitian pada tahun 2019. Algoritma *naïve bayes* digunakan dalam penelitian ini, dengan *input* data berupa dokumen teks. Dokumen teks tersebut kemudian diproses terlebih dahulu tanpa menggunakan *stemming*. Setelah itu, proses pembobotan kata diterapkan pada data pelatihan (*training* data). Selanjutnya, algoritma *naïve bayes* digunakan untuk mengklasifikasikan teks berita. Metode 10-*Ford Cross Validation* digunakan untuk melakukan pengukuran. Proses klasifikasi akan menghasilkan *output* berupa berita dan fakta *hoax* sebagai hasil dari proses klasifikasi *naïve bayes.* Berdasarkan temuan penelitian, nilai *fold* 6 memberikan nilai akurasi terbaik dengan hasil terbaik dengan nilai akurasi sebesar [8].

Alvin Pratama, Dwi Marisa Midyanti, dan Syamsul Bahri dalam penelitiannya yang berjudul Penggunaan nave bayes classifier dalam hubungannya dengan algoritma *stemming nazief* dan adriani untuk aplikasi deteksi ujaran kebencian berbasis *web.* Tujuan dari penelitian ini adalah untuk mengetahui nilai akurasi, presisi, dan *recall*, serta kata-kata yang biasa digunakan dalam ujaran kebencian. Aplikasi pendeteksi ujaran kebencian memperoleh nilai akurasi 68%, nilai presisi 93%, dan nilai *recall* 69% dari pengujian 183 data uji menggunakan algoritma *stemming Nazief* dan Adriani, serta classifier *naïve bayes*. Selain nilai akurasi,

presisi, dan *recall*, diketahui juga bahwa kata-kata yang sering digunakan dalam 104 data pelatihan kategori ujaran kebencian adalah kata-kata kebencian, dengan total 128 kata [9].

M.Ibrahim, Efori Bu'ulolo, dan Ikwan Lubis melakukan penelitian yang terbit pada tanggal 1 September 2020 yang dituangkan dalam jurnal Rekayasa Teknik Informatika dan Informasi dengan judul penerapan algoritma *naïve bayes classifier* untuk mendeteksi tingkat kredibilitas *hoax news/fake news* pada sosial media di Indonesia berbasis android (studi kasus: kantor tribun medan). Dalam jurnal ini dituliskan bahwa berita bohong atau disebut juga berita *hoax* menjadi masalah yang paling serius di Indonesia. Pasalnya, tersebarnya berita *hoax* membuat masyarakat sulit membedakan antara berita asli dan *hoax*. Berita *hoax* hanya merugikan satu pihak dan merusak reputasi seseorang. Polri atas nama pemerintah mengancam akan menindak para penyebar berita bohong. Sayangnya, tindakan mengancam dianggap sebagai kebebasan berbicara. Investigasi ini menghasilkan perbandingan antara berita palsu dan fakta [10].

Candra Surya Sriyano dan Erwin Budi Setiawan melakukan penelitian yang diberi judul pendeteksian berita *hoax* menggunakan *naïve bayes multinominal* pada *twitter* dengan pembobotan TF-IDF. Penelitian ini bertujuan untuk mengetahui bagaimana cara untuk mendeteksi berita *hoax* dengan menggunakan *naïve bayes multinominal* dan fitus pembobotan TF-IDF, serta bagaimana hasil akurasi dari pendeteksian berita hoax dengan menggunakan *naïve bayes multinominal* dan fitur pembobotan TF-IDF. Dari hasil pengujian yang penulis lakukan dapat dilihat bahwa pembagian data yang diuji dengan data train 80% menggunakan fitur pembobotan TF-IDF memperoleh akurasi tertinggi sebesar 72.06%, dan tanpa fitur pembobotan TF-IDF dengan data train 80% mendapatkan akurasi sebesar 71.65%. Dengan ini dapat kita lihat bahwa pembagian rasio data berpengaruh dengan hasil dari nilai akurasi. Dan dapat kita lihat juga bahwa hasil rata-rata yang diperoleh menggunakan metode pembobotan TF-IDF lebih besar dibandingkan hasil yang diperolah tanpa TF-IDF, maka dapat disimpulkan metode pembobotan TF-IDF berpengaruh pada hasil pengujian yang dilakukan [11].

Esther Irawati Setiawan, Sugiharto Johannes, Arya Tandy Hermawan, dan Yuni Yamasari melakukan penelitian pada tahun 2021 tentang berita *hoax* pada media sosial *twitter* yang dituangkan dalam Vol.3 No.2 *Journal of Intelligent System and Computation*. Data dari *platform* media sosial *twitter* diolah agar dapat digunakan untuk pelatihan dan pengujian. Tahap pelatihan menggunakan metode klasifikasi *naïve bayes* untuk mengolah data *tweet,* dan tahap pengujian bertujuan untuk mengklasifikasikan *tweet* di media sosial palsu atau tidak.

Preprocessing sebelum proses pengklasifikasian berita bohong di media sosial ternyata dapat mempengaruhi hasil akurasi uji coba. Untuk data uji coba sebanyak 309 data, penelitian ini menghasilkan nilai akurasi tertinggi sebesar 92% [12].

Penelitian selanjutnya dilakukan oleh Nova Agustina, Adrian, dan Merry Hernawati dan di tuliskan dalam sebuah *Journal of Intelligent Systems And Computation* Vol. 14 No. 4, December 2021, pp. 206 - 213. Metode penelitian ini yaitu data *maining*, Algoritma *naive bayes classifier*, dan teknik analisis data. Penelitian ini melakukan prediksi Hoax dengan membandingkan metode klasifikasi lain, contohnya *Convolutional Neural Network* (CNN) dan *Pre-training of Deep Bidirectional Transformers for Language Understanding* (BERT). Eksplorasi ini dapat dilakukan pada dataset yang sama dalam studi masa depan, sehingga model terbaik dari hasil perbandingan dapat diukur, penelitian ini menghasilkan total akurasi sebesar 81% [13].

Sayyid Muhammad Habib dalam penelitiannya Klarifikasi Berita Menggunakan Metode *Naïve Bayes Classifier* Jurusan Teknik Informatika UIN Suska Riau pada Juli 2022 menggunakan metode kualitatif. Penelitian ini dilakukan untuk mengklasifikasikan berita menggunakan algoritma *naïve bayes* dengan mengangkat berita dengan tema sosial yang diambil dari Badan Pusat Statistik Provinsi Riau. Proses klasifikasi menggunakan metode *naïve bayes classifier* menghasilkan akurasi tertinggi, 94%, dengan sebaran 90% data latih dan 10% data uji dari dataset berita yang digunakan [14].

Fani Prasetya dan Ferdiansyah melakukan penelitian tentang Pada 1 September 2022 akan dilakukan analisis data *mining* klasifikasi berita *hoax* Covid-19 dengan menggunakan algoritma *naïve bayes.* Para peneliti kemudian mencoba mengklasifikasikan berita *hoax covid*-19 dengan menggunakan algoritma klasifikasi *naïve bayes*. Menurut temuan penelitian, model *naïve bayes* dan validasi silang dapat mengklasifikasikan berita *hoax* dengan tepat; akurasi yang dihasilkan adalah 86,3%, dengan 80-90% masuk dalam kriteria klasifikasi baik. Tidak banyak data yang diprediksi salah; dari total 300 dataset, hanya 41 yang dinyatakan salah dalam pelabelan. Karena model ini menyumbang hingga 2% dari total dataset, dapat disimpulkan bahwa model ini dapat digunakan sebagai referensi [15].

Noor Aliyah Susanti, Miftahul Walid, dan Hoiriyah melakukan penelitian klasifikasi data *tweet* ujaran kebencian di media sosial menggunakan *naive bayer classifier* pada 2 September 2022. Data yang digunakan dalam penelitian ini bersumber dari situs www.kaggle.com. Data

sentimen yang diunduh merupakan sekumpulan *tweet* dari pengguna *twitter* mengenai ujaran kebencian kepada pemerintah. Data telah diberi label 0 dan 1. Nilai 0 disini berarti sentimen negatif dan nilai 1 merupakan sentimen positif. Kemudian data yang berupa sekumpulan teks dari *tweet* pengguna *twitter* diproses pada tahapan *text processing*. Berdasarkan hasil pengujian prediksi dengan model perhitungan algoritma *Naive Bayes Classifier* dengan nilai akurasi mendekati 70% termasuk dalam kategori *good*. Sebelum melakukan perhitungan nilai akurasi, data tweet harus diolah melalui teks *preprocessing* agar kata (*term*) dapat dikonversikan ke dalam bentuk matriks. Untuk kemudian diolah sebagai data numerik dengan perhitungan matriks dan split data *tweet* menggunakan model Multinomial *naïve bayes* dapat disimpulkan bahwa perhitungan pembobotan TF-IDF memiliki nilai akurasi yang sama untuk masing-masing pembobotan n-gram, yakni 69,23076923 [16].

2.2 DASAR TEORI

2.1.1 *Naïve Bayes*

Naïve bayes adalah metode yang cocok untuk klasifikasi biner dan *multiclass*. Metode yang juga dikenal sebagai *naïve bayes classifier* ini menerapkan teknik *supervised* klasifikasi objek di masa depan dengan menetapkan label kelas ke *instance*/catatan menggunakan probabilitas bersyarat. Probabilitas bersyarat adalah ukuran peluang suatu peristiwa yang terjadi berdasarkan peristiwa lain yang telah (dengan asumsi, praduga, pernyataan, atau terbukti) terjadi. *Naïve bayes* merupakan metode pengklasifikasian paling populer digunakan dengan tingkat keakuratan yang baik. Banyak penelitian tentang pengklasifikasian yang telah dilakukan dengan menggunakan algoritma ini. Berbeda dengan metode pengklasifikasian dengan *logistic regression ordinal* maupun nominal, pada algoritma *naïve bayes* pengklasifikasian tidak membutuhkan adanya pemodelan maupun uji statistik.

Naïve bayes merupakan metode pengklasifikasian berdasarkan probabilitas sederhana dan dirancang agar dapat dipergunakan dengan asumsi antar variabel penjelas saling bebas (independen). Pada algoritma ini pembelajaran lebih ditekankan pada pengestimasian probabilitas. Keuntungan algoritma *naïve bayes* adalah tingkat nilai *error* yang didapat lebih rendah ketika dataset berjumlah besar, selain itu akurasi *naïve bayes* dan kecepatannya lebih tinggi pada saat diaplikasikan ke dalam dataset yang jumlahnya lebih besar. Cara kerja *naïve bayes classifier* melalui dua tahapan, yaitu :

Pertama, *Learning* (pembelajaran) *naïve bayes* adalah suatu metode yang termasuk ke dalam *supervised learning*, maka akan dibutuhkan pengetahuan awal untuk dapat mengambil keputusan. Langkah-langkah :

Langkah 1: Bentuk *vocabulary* pada setiap dokumen data *training*

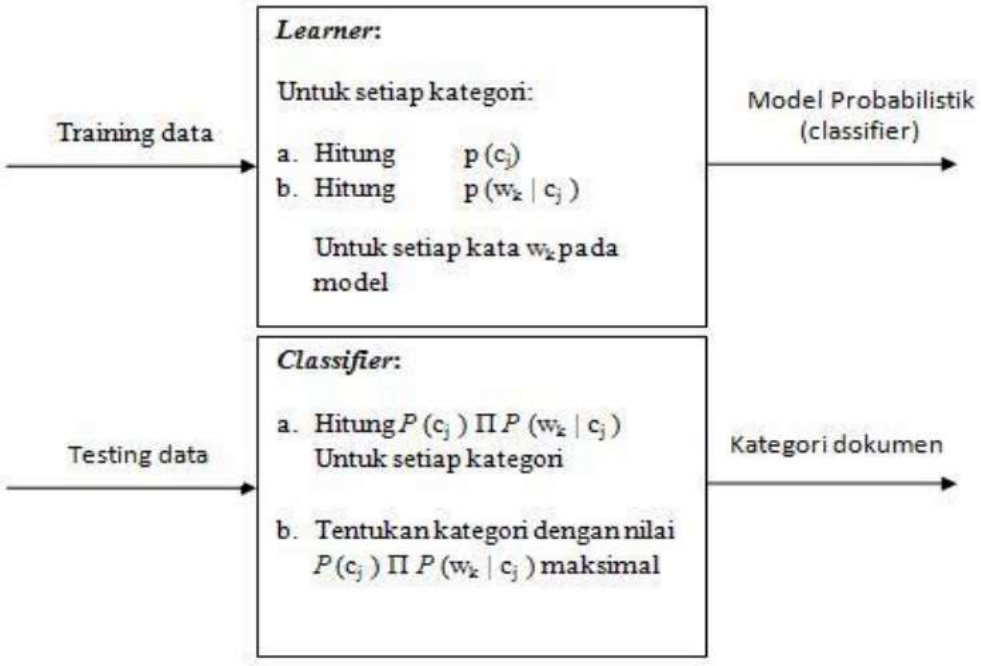
Langkah 2: Hitung probabilitas pada setiap kategori P(vj)

Langkah 3: Tentukan frekuensi setiap kata wk pada setiap kategori P(wk | vj)

Kedua, *Classify* (pengklasifikasian). Langkah-langkahnya adalah :

Langkah 1: Hitung P(vj) II P(wk | vj) untuk setiap kategori

Langkah 2: Tentukan kategori dengan nilai P P(vj) II P(wk | vj) maksimal



Gambar 2. 1 Tahapan Proses Klasifikasi Dokumen *Naïve Bayes Classifier* [26]

Untuk mengklasifikasikan data biner dan multiclass digunakan metode naive bayes. Metode ini, juga dikenal sebagai Guileless Bayes Classifier, menggunakan probabilitas bersyarat untuk menetapkan kelas tanda ke contoh/rekaman. Probabilitas bersyarat adalah ukuran kemungkinan suatu peristiwa terjadi berdasarkan peristiwa sebelumnya yang telah terjadi (dengan asumsi, anggapan, pernyataan, atau terbukti).

Pendekatan yang diawasi didasarkan pada pengelompokkan data pelatihan yang telah diberikan identitas sebagai kelas. Selain itu, penawaran terpisah akan diperlakukan sebagai

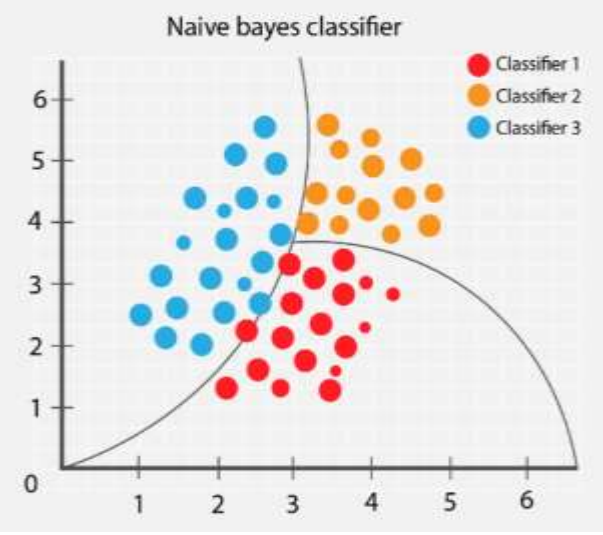
data transaksional. Jika kita akan memverifikasi transaksi di tengah hari sebagai penipuan atau non-penipuan, proses verifikasi dapat diawasi [12].

Metode klasifikasi *naïve bayes* dapat diterapkan pada berbagai situasi di dunia nyata seperti kesalahan dalam bentuk sistem klasifikasi untuk dokumen atau spam Hal ini disebabkan fakta bahwa *naïve bayes* mengandalkan berbagai data untuk menentukan parameter yang sesuai. Penerapannya sangat luas, contohnya dalam penelitian ini *naïve bayes* digunakan dalam pengklasifikasian berita *hoax* dan benar, sehingga dengan *naive bayes* ini kita dapat dengan mudah mengklasifikasikan hal yang banyak dan rumit [12].

Rumus *naïve bayes*:

$$P(A|B) = \frac{P(B|A)\,.\,P(A)}{P(B)} \qquad (\mathbf{2.1})$$

Pada dasarnya, ketika mencoba mencari peluang kejadian A, dan apabila kejadian B bernilai benar, kejadian B juga disebut sebagai bukti. P(A) adalah apriori dari A (probabilitas sebelumnya, yaitu probabilitas peristiwa sebelum bukti terlihat). Bukti adalah nilai atribut dari *instance* yang tidak diketahui (peristiwa B). P(A|B) adalah probabilitas posteriori dari B, yaitu probabilitas kejadian setelah bukti terlihat [13].



Gambar 2. 2 *Naïve Bayes Classifier* [12]

2.1.2 Dataset

A. Pengertian Dataset

Menurut IBM, *dataset* digambarkan sebagai kumpulan dokumen dengan fokus (*record)* tertentu yang biasanya terdiri dari satu atau beberapa fokus. Dalam setiap kelas *record* dalam contoh saat ini dikenal dengan *dataset* dan mempunyai kemampuan dalam mencadangkan informasi contohnya *medical record*, asuransi, program, dan sistem data institusional. *Dataset* diperuntukkan dalam menyimpan informasi yang diperlukan aplikasi atau sistem operasi lainnya, misalnya ringkasan program, pola berskala besar, atau variabel dan parameter sistem.

Selain itu, *dataset* juga bisa diartikan sebagai gabungan data bisa juga sebagai bahan untuk ditampilkan di kolom tabel. Dalam masing-masing kolom di tabel data yang dimaksud menunjukkan variabel yang relevan, jadi ada beberapa variabel dalam satu *dataset*. Secara teknis, *dataset* adalah potongan dari sistem manajemen data. Namun, angka-angka dari kumpulan data ini disebut sebagai data.

Secara teknik, *dataset* adalah gabungan *item* terkait yang bisa dijangkau secara individual atau digabungkan dengan *item* terkait lainnya sebagai satu kesatuan. Kumpulan data dapat diubah menjadi berbagai jenis struktur data. Contoh *dataset* dalam dunia bisnis dapat dilihat dari nama, gaji, informasi tentang karyawan, hingga jumlah penilaian, dan lain-lain.

Dapat ditarik kesimpulan, *dataset* adalah gabungan data yang terstruktur dan diambil dari gabungan informasi. Informasi secara keseluruhan didapatkan dengan studi, analisis, atau penelitian sampai menjadi data. Data dapat berupa kenyataan, angka, nama, atau bahkan deskripsi. Oleh karena itu, *dataset* memiliki hubungan yang kuat dengan kegiatan *data mining*, yang memungkinkan ilmuwan data mengubah data menjadi informasi yang koheren [6].

Dataset juga didefinisikan sebagai gabungan atau kombinasi data yang ditampilkan dengan bentuk pola tabel. Setiap kolom pada tabel data menggambarkan variabel yang berbeda, jadi terdapat beberapa variabel dalam satu dataset. Secara teknis, kumpulan data adalah bagian dari manajemen data.

B. Jenis-jenis Dataset

1) *Private* dataset

   Private dataset yaitu kumpulan data yang dapat diambil oleh suatu institusi untuk digunakan sebagai sasaran penelitian, seperti data universitas, rumah sakit, kantor kelurahan, BMKG, dan lain-lain.

2) *Public* dataset

   Yaitu kumpulan data yang dapat diambil dari arsip publik yang telah disetujui oleh para ahli dalam penelitian data *mining.* Tujuan kumpulan data ini adalah untuk menguji metode penelitian peneliti berpengalaman menggunakan materi publik dan pribadi. Saat ini, bahan terbanyak yang digunakan pada penelitian data *mining* yaitu proses pengujian yang dikembangkan oleh peneliti berpengalaman menggunakan bahan yang tersedia untuk umum agar penelitian dapat dibandingkan, diulang, dan diverifikasi.

2.1.3 Teks Berita

Teks berita merupakan paragraf yang memuat kabar terbaru dan akurat berlandaskan kebenaran suatu peristiwa yang terjadi. Dalam teks berita juga terdapat informasi penting yang perlu diketahui atau hendak dikenali oleh masyarakat luas. Menurut Kamus Besar Bahasa Indonesia (KBBI), berita berisi komentar atau informasi yang gamblang tentang isu-isu mendesak. Maka daripada itu, dapat disimpulkan bahwa berita merupakan tulisan yang memuat informasi terkini atau sedang berlangsung.

Berita pengelompokannya berita terbagi menjadi dua jenis, yaitu berdasarkan cara penulisannya lisan dan tulisan (*writing*). Di televisi, kita sering mendengar dan melihat berita yang disampaikan dengan bahasa yang sederhana. Selain itu, berita yang dikomunikasikan secara jelas dan ringkas sering dimuat di media cetak dan *online.* Agar pesan yang dimaksud dapat tersampaikan kepada khalayak, diperlukan keterampilan dan penguasaan dasar penulisan dalam menyunting artikel berita [7].

A. Ciri-ciri Teks Berita

Untuk penyusunan teks berita harus mengenali ciri- cirinya terlebih dahulu. Umumnya, berita biasanya terdiri atas elemen yang disatukan menjadi suatu karakteristik dari teks berita itu sendiri. Karakteristik teks berita yang wajib diketahui yaitu:

1) Faktual

   Berisi peristiwa yang bersifat nyata serta betul- betul terjadi tanpa adanya rekaan dan tidak terbelenggu waktu misalnya peristiwa di masa sebelumnya. Tetapi berita wajib berbentuk peristiwa terbaru, sedang terjadi, terbaru, konkret, dan baru saja berlangsung.

2) Aktual

   Memuat peristiwa yang bersifat valid, sedang berlangsung dan lagi ramai-ramainya serta jadi bahan pembicaraan masyarakat luas.

3) Unik

   Berita wajib menunjukkan kabar yang bisa menarik minat dan kalimat yang di pakai mengenakan kata yang khas sehingga membuat pembaca tertarik untuk membaca berita yang kita buat. Unik serta menarik artinya bisa memunculkan rasa ketertarikan pembaca untuk membaca. Peristiwa yang menarik umumnya berciri menghibur, memiliki nilai kemanusiaan, kejahatan, peristiwa yang sedang hangat-hangatnya, konflik, serta yang lainnya.

4) Berdampak untuk warga luas

   Bacaan kabar yang bisa pengaruhi seorang tercantum kabar yang baik sebab bila warga luas tertarik hingga hendak dipercayai oleh banyak serta mempengaruhi pada warga selaku pendengar.

5) Ada waktu serta runtutan peristiwa

   Teks untuk dimuat dalam berita umumnya tetap mencantumkan runtutan waktu peristiwa serta urutannya. Kapan serta di mana peristiwa itu terjadi senantiasa dimasukkan dalam sebuah berita, gunanya agar orang yang membaca bisa menguasai waktu serta tempat terjadinya peristiwa itu.

6) Sesuai dengan keadaan yang sebenarnya

   Kabar yang di informasikan wajib cocok keadaannya tanpa mengaitkan pemikiran ataupun opini individu yang bisa pengaruhi orang yang membaca.

7) Bahasa formal, mudah, dan informatif

   Teks dalam berita biasanya memakai bahasa formal, lumrah, dan informatif dengan harapan pembaca tidak akan mengerti jika kalimat yang disampaikan tidak

menggunakan bahasa baku. Akibatnya, bahasa baku harus digunakan karena sepadan dengan aturan bahasa Indonesia.

8) Ejaan Yang Disempurnakan (EYD)

Penerapannya juga ringkas dan informatif sehingga mampu berdampak terhadap orang yang membaca berdasarkan suatu peristiwa yang sedang berlangsung [17].

B. Unsur-unsur Berita

Unsur- unsur berita dapat diketahui dengan 5W+1H, seperti dijelaskan dibawah ini:

1) *What* (apa yang terjadi?)

   What merupakan faktor sangat awal yang ada dalam suatu berita. Ini adalah salah satu aspek terpenting, karena what hendak menerangkan menimpa peristiwa apa yang hendak dinaikan buat diberitakan.

2) *Where* (di mana itu terjadi?)

   Where merupakan unsur berita yang menyatakan lokasi peristiwa itu terjadi sehingga data berisi tempat peristiwa dalam berita, untuk memperjelas informasi yang diberikan kepada pembaca.

3) *When* (kapan peristiwa itu terjadi?)

   When merupakan unsur berita yang berisi kapan terjadinya peristiwa yang di informasikan dalam berita.

   1) *Who* (siapa yang ikut serta dalam peristiwa itu?)

      Who merupakan unsur berita yang berisi siapa dalam peristiwa di berita itu ataupun siapa yang ikut serta dalam peristiwa itu. Orang- orang yang ikut serta wajib dipaparkan supaya tidak memunculkan kesalahpahaman.

   2) *Why* (mengapa perihal itu terjadi?)

      Why merupakan unsur berita yang berisi mengapa peristiwa dalam berita tersebut dapat terjadi, umumnya yang melatarbelakangi peristiwa tersebut.

   3) *How* (Bagaimana peristiwa itu terjadi?)

      How merupakan unsur berita yang berisi bagaimana peristiwa itu terjadi, umumnya informasi yang diberikan dipaparkan secara kronologis [17].

C. Struktur Berita

Struktur teks berita adalah cerminan metode suatu bacaan tersebut dibentuk. Teks berita mempunyai struktur yang jelas, dan disusun berdasarkan sumber yang ada pada susunan teks berita yang memuat peristiwa dalam berita, diiringi dengan konteks kejadian, serta diiringi sumber kabar. Secara universal susunan berita terdiri dari judul, kepala berita, badan berita, serta ekor berita.

1) Judul berita

   Judul sangat berarti dalam berita, sebab berfungsi selaku penarik pembaca untuk membaca isi berita tersebut. Maka dari itu judul hendaknya diciptakan dengan sangat menarik agar dapat menarik minat orang yang akan membaca.

2) Kepala berita

   Kepala berita mempunyai cakupan pembahasan yang lebih luas, dan terdapat beragam data yang ditampilkan. Biasanya pada bagian awal berita penulis memulai informasi dalam berita dengan memperlihatkan 4 faktor, yaitu apa, di mana, kapan, dan siapa.

3) Badan berita

   Bagian badan berita berisi uraian ataupun data yang di informasikan pada bagian kepala berita. Bagian ini merupakan jawaban atas persoalan kenapa dan bagaimana". Biasanya, mencantumkan latar belakang ataupun bukti sesuatu peristiwa dapat terjadi.

4) Ekor berita

   Bagian ini mencakup lebih banyak kabar. Cerita utama tidak akan terpengaruh jika bagian ini dihilangkan [18].

D. Jenis-jenis Berita

Dalam jurnalistik, pembagian tersebut didasarkan pada isi berita dan dapat dilihat pada proses penataan dan penyajian berita. Dalam publisistik juga ada berbagai jenis berita, seperti:

1) Berita langsung

   Berita langsung (*straight news*) merupakan penjelasan kejadian yang ditulis secara singkat, padat, dan sederhana, tanpa tambahan keterangan, terutama penafsiran. Berita ini terbagi menjadi dua kategori yaitu berita keras atau hangat (*hard news*) dan berita lunak atau ringan (*soft news*).

2) Berita opini

Berita opini adalah mengenai komentar, penjelasan, atau dari buah pikiran seseorang, biasanya dari cendekiawan, pakar, atau pejabat pemerintah tentang suatu kejadian.

3) Berita interpretatif

Berita interpretatif didefinisikan sebagai berita yang dipaparkan berdasarkan pendapat atau penilaian wartawan atau narasumber yang berkompeten atas berita yang muncul lebih dulu, sehingga menghasilkan kombinasi realitas dan interpretasi, diawali dengan data yang dianggap cacat arti dan maknanya.

4) Berita mendalam

In-depth news adalah berita yang tercipta dari berita yang telah ada sebelumnya dan memperluas informasi yang terkandung dalam basis berita. Diawali dengan cerita yang belum berakhir dan dapat diteruskan lagi (*follow up system*). Pengkajian diupayakan agar memperoleh tambahan fakta dari sumber atau berita yang relevan.

5) Berita penjelasan

Berita deskripsi (*explanatory news*) adalah informasi yang bersifat penjelas digambarkan secara lengkap tentang suatu peristiwa. Kenyataan yang didapatkan dirinci dalam pendapat atau komentar penulis berita. Karena berita ini lumrahnya panjang, maka harus disediakan secara berangkai serta urut.

6) Berita penyelidikan

Berita investigasi adalah informasi yang dikumpulkan dan disebarluaskan berlandaskan pada penelitian atau dapat juga investigasi dari berbagai sumber. Disebut juga mencari karena wartawan mencari informasi dari berbagai sumber, apalagi melakukan investigasi langsung di tempat, diawali dengan data mentah atau berita pendek. Biasanya, berita investigasi disajikan dalam bentuk artikel *feature* [18].

E. Pengertian Berita Menurut Para Ahli

1) Nasution

"Berita adalah laporan terkait peristiwa-peristiwa yang terjadi dan ingin diketahui secara umum, bersifat aktual, telah terjadi dalam lingkungan pembaca, berhubungan dengan tokoh terkemuka, dan akibat peristiwa tersebut bisa berpengaruh kepada pembaca".

2) Djuraid

Menurut Djuraid, berita merupakan suatu laporan ataupun pemberitahuan mengenai terjadinya peristiwa atau keadaan bersifat umum dan baru saja terjadi, yang disampaikan oleh wartawan media massa.

3) Yani Josef

Jani Yosef mendefinisikan berita sebagai laporan terkini tentang fakta penting atau menarik bagi khalayak, yang disebarluaskan lewat media massa.

4) Hoeta Soehot

Hoeta Soehoet menyatakan bahwa berita adalah keterangan mengenai peristiwa atau isi pernyataan manusia.

5) Doug Newson dan James A. Wollert

"Berita adalah apa saja yang ingin dan perlu diketahui orang atau lebih luas lagi oleh masyarakat".

6) Paul De Maeseneer

Menurut Paul De Maeseneer, berita adalah informasi mengenai kejadian baru bersifat penting dan bermakna (signifikan), yang berpengaruh pada pendengarnya serta relevan dan layak dinikmati oleh mereka.

7) Sumadiria

Sumadiria menjelaskan bahwa berita adalah laporan tercepat lewat media berkala, mengenai ide atau fakta terbaru yang menarik, benar, dan penting bagi sebagian besar khalayak.

8) Dean M. Lyle Spencer

"Berita ialah kenyataan ide secara benar dan dapat menarik perhatian lebih besar bagi para pembacanya".

9) Adi Negoro

Menurut Adi Negoro, berita adalah suatu pernyataan antara manusia yang saling memberitahukan satu sama lainnya.

10) Freda Morris

berita adalah sesuatu yang baru, penting, dan dapat memberi dampak dalam kehidupan manusia. Menurutnya, berita terdiri dari unsur baru, penting, serta bermanfaat bagi manusia [8].

2.1.4 Berita Hoax

Berita *hoax* terus diperbarui sepanjang hari dan dikonsumsi oleh pengguna *internet*. *Hoax* dibuat khusus untuk digunakan dalam meyakinkan masyarakat menggunakan materi seperti foto atau resep otentik. Tujuan dari berita *hoax* yaitu untuk mengungkap kesalahan terhadap seseorang atau sekelompok orang tertentu. Selain itu, dapat berupa provokasi, propaganda, menghasut opini publik, atau bahkan pernyataan resmi yang dibuat untuk mengatasi masalah tertentu yang sedang dihadapi. Berita-berita yang banyak diberitakan di masyarakat umum seringkali berkaitan dengan isu-isu politik, agama, dan politik. Namun ada juga penipuan seperti informasi lowongan kerja sedangkan menurut KBBI, *hoax* adalah informasi bohong atau informasi tanpa data pasti [9].

Berita bohong atau disebut juga berita palsu merupakan kabar atau informasi yang beredar biasanya bertujuan meresahkan masyarakat sekaligus mengecilkan motif pembuatnya. Atribut ini dapat diklasifikasikan sebagai komunitarian, politik, ideologis, dan lainnya. Dikutip dari pernyataan Pellegrini, 2008 berita palsu dapat menyertakan URL, sumber, atau fakta alternatif yang terbukti. mungkin dianggap tepat. Selain itu, hoaks diartikan sebagai “kebohongan” yang di peruntukkan bagi kepentingan perseorangan baik yang bersifat esensial maupun ekstrinsik, yang dilakukan dengan sedemikian rupa oleh individu guna memalingkan pengamatan dari objek tertentu.

Eko Septiaji, Ketua Gerakan Anti Fitnes Indonesia (Mafindo), menggarisbawahi pemahaman yang jelas bahwa *hoax* adalah informasi yang harus digunakan untuk melengkapi informasi yang lebih terpercaya. Dengan maksud lain *hoax* dapat digambarkan seperti strategi pengecekan kenyataan yang menggunakan penjelasan yang sugestif tetapi tidak memungkinkan diversifikasi asumsi yang mendasarinya. *Hoax* juga bisa digunakan sebagai minuman [10].

Berita *hoax* dapat ditemukan sejak tahun 1600. Informasi mengenai era ini tersedia melalui korespondensi. Pembaca bertanggung jawab untuk memverifikasi informasi yang berkaitan dengan orang, agama dan individu pada saat itu. Kebanyakan berbohong pada waktu tersebut terbentuk karena penembakan. Menurut Pennsylvania Gazette, pada 17 Oktober 1745, Benjamin Franklin menerbitkan sebuah laporan tentang batuan Cina yang dimaksudkan untuk menyebarkan rabies, kanker, dan penyakit lainnya.

Namun, kesaksian pribadi sering digunakan untuk memverifikasi informasi tersebut. Saat tulisan ini dibuat, salah satu bagian Gazette yang dapat diklarifikasi menjelaskan bahwa batuan tersebut terutama dimiliki oleh Amerika Serikat dan tidak mengandung bukti medis apa pun. Pada tahun 1726, Jonathan Swift menyusun strategi tipuan untuk menyebarkan informasi yang salah tentang Perjalanannya ke Beberapa Negara Terpencil. Tahun berikutnya, pada tahun 1708, dia juga melakukan tipuan yang tidak melibatkan prediksi astrologi pada Hari April Mop. Pada tahun 1835, Edgar Allan Poe membuat pengamatan berikut: Petualangan Hans Pfaall yang Tak Tertandingi, tentang seorang anak yang lahir di tengah malam, memiliki balon udara, dan tinggal di sana selama lima tahun [11].

Hoax umumnya menyebar bagai virus, sehingga normal saja banyak berita *hoax* yang jadi populer serta viral, apalagi orang-orang dengan tanpa sadar turut menyebarkan kabar tersebut. Situs hoakses.org melaporkan kalau agar bisa terkategori selaku hoaks, suatu kebohongan wajib mempunyai "nilai lebih" semacam bersifat dramatis ataupun sensasional. Lebih dari itu, juga harus sanggup menyedot atensi public, ublik jadi semacam kata kunci. Karena, tidak ada hoaks yang sifatnya privat. Kian luas capaian sesuatu kabar *hoax*, kian besar tingkat kabar hoaks tersebut. Inilah yang membedakannya dengan tipe kebohongan yang lain semacam penipuan dan olok- olokan. Dalam masyarakat luas, tiap anggota masyarakat memiliki ketergantungan terhadap media komunikasi serta data. Pada kegiatan pertukaran serta mengkonsumsi data yang mendominasi setiap kegiatan masyarakat tersebut, kabar *hoax* sangat deras timbul dan memforsir buat disantap. Jika dicermati dengan seksama, istilah *hoax* sama tidak menyenangkannya dengan maknanya. *Hoax* memiliki sejarah panjang dan dampak yang agak negatif pada masyarakat umum. Di dunia sekarang ini, dimana data mudah tersebar, *hoax* juga mudah tersebar.

A. Sebab Maraknya Hoax

Dalam beberapa tahun terakhir, dengan maraknya teknologi informasi, hoaks dan berita bohong lainnya juga menyebar ke seluruh Indonesia. Kurang lebih terdapat 800.000 website di Indonesia yang sudah ditandai menjadi situs *hoax*, menurut data Kementerian Komunikasi dan Informasi RI. Berita bohong tersebut kemudian terungkap melalui situs

media sosial seperti grup Facebook dan *WhatsApp*. Sebagian besar lelucon yang digunakan menyangkut masalah politik, kesehatan, dan agama.

Menurut hasil survei *Katadata Insight Center* (KIC), antara 30% hingga 60% masyarakat Indonesia tertipu *hoax* saat mengakses dan berkomunikasi melalui internet di berbagai kantor redaksi. Kekuatan *hoax* ini, yang kemudian dibarengi dengan ujaran kebencian dan semburan kebohongan, tidak bisa dipungkiri tentunya ini sangat berbahaya bagi kehidupan sosial, politik, budaya, bangsa, dan kehidupan berbangsa.

Terlepas dari kenyataan bahwa pemerintah telah melakukan banyak penyelidikan terhadap pelaku dan penyebar hoax, hasilnya di bawah standar karena masalah yang mendasarinya tetap tidak konsisten. Perlu dicatat bahwa peningkatan produksi *hoax* tidak dapat dikaitkan hanya dengan jumlah orang yang menggunakan *hoax* tersebut. *Hoax* terus diproduksi karena ada pengguna atau konsumen, sesuai dengan hukum permintaan dan penawaran. Akibatnya, jika akan dihentikan atau setidaknya dikurangi pembuatan dan peredaran *hoax*, kita harus berbicara kepada pengguna atau *user*.

Korban tipuan yang dimaksud adalah politisi yang tidak jujur, jajaran yang tidak puas, dan ideolog transnasional radikal. politisi nakal dan jajaran yang tidak puas berusaha mendelegitimasi kekuatan rezim politik yang berkuasa, sementara pendukung ideologi radikal transnasional berusaha memprovokasi rakyat jelata dengan tujuan mengganti ideologi Pancasila dengan ideologi mereka sendiri. Selain politik yang buruk, kecil hati, dan paham transnasional radikal, mereka juga merupakan orang kaya dengan sedikit pengetahuan dan tingkat literasi digital yang tinggi [19].

B. Jenis-jenis *Hoax*

1) Satire ataupun parodi

   Satire ataupun prodi merupakan suatu konten yang dibanguns. Tipe Konten ini kerap diperuntukkan mencela pihak tertentu. Tidak hanya itu, konten satire ini pula berupa wujud kritik. Kritik yang di informasikan dapat dalam ikatan indvidual, kelompok dalam suatu golongan, ataupun mengomentari hal- hal yang sangat sering terjadi di masyarakat.. Sementara itu konten satire tidaklah konten yang seluruhnya beresiko. Konten satire ini pula umumnya tidak berpotensi mempunyai faktor kejahatan. Hanya saja senantiasa konten- konten semacam ini masih banyak mengecoh warga. Banyak

orang menerima dengan sungguh-sungguh konten tersebut. Yang lebih mengkhawatirkan yaitu bila isi konten yang di informasikan merupakan keadaan yang keasliannya masih diragukan. Masyarakat yang menyaksikannya dengan kontan dapat yakin, ini juga dapat menggambarkan terjadinya pemicu kabar *hoax*.

2) *Misleading content* (konten menyimpang)

Misleading content atau konten yang menyimpang sering direncanakan. Jenis konten ini dirancang untuk mengkritik orang lain atau objek tertentu. Keadaan yang dimunculkan dalam informasi yang sudah dikemas dengan topik tertentu juga bisa melibatkan satu atau bahkan banyak manusia. Ini dibuat guna memengaruhi pendapat orang. Konten menyesatkan atau konten menyesatkan dibuat dengan menggunakan data asli. Informasi tersebut dapat berupa ungkapan formal, foto atau gambar, statistik dan sebagainya. Data itu akan disunting dengan berbagai cara sehingga data dan konten yang hendak diolah tidak memiliki keterikatan.

3) *Imposter content* (konten tiruan)

Konten imitasi merupakan konten palsu. Informasi dalam konten ini berasal dari sumber yang dapat dipercaya. Contohnya antara lain mengutip ungkapan tokoh populer atau berpengaruh. Jenis konten ini tidak hanya untuk individu. Banyak juga konten jenis ini yang dibuat untuk menipu, bukan untuk mempromosikan sesuatu. Penipu ingin membuat konten serupa dengan menggunakan konten yang mirip dengan aslinya. Banyak orang, misalnya, bertindak atas nama aplikasi, seperti layanan aplikasi, untuk menipu masyarakat umum.

4) *Fabricated Content* (konten palsu)

Jenis *hoax* berikut adalah konten yang direkayasa atau palsu. Jenis konten tipuan ini sangat rawan. Konten ini dirancang untuk mencurangi pemirsa. Banyak orang yang terpengaruh oleh konten palsu seperti ini. Tidak ada cara untuk menjelaskan informasi yang terkandung. Informasi dalam data itu salah. Data lowongan pekerjaan adalah contoh umum dalam jenis konten ini. Data lowongan kerja dibuat mirip dengan aslinya atas nama industri atau institusi.

5) *False connection* (konten yang salah)

False connection atau koneksi yang salah, merupakan konten yang juga sangat umum di sosial media. Perbandingan antara konten, judul konten, dan foto konten

adalah contoh umum. Konten ini dimaksudkan untuk memberikan keunggulan kompetitif.

6) *False context* (konten galat)

False context adalah konten yang salah. Disebut *error* lantaran mengandung informasi yang tidak benar. Misalnya mencakup ekspresi, video, atau gambar sebelumnya. Setelah itu, peristiwa itu dimuat kembali dan tidak menyesuaikan kenyataan yang sebenarnya.

7) *Manipulated content* (konten manipulasi)

Konten yang diedit disebut sebagai konten yang dimanipulasi. Konten ini hendak disunting kembali sehingga tidak sesuai lagi dengan konten aslinya. Jenis konten ini dirancang untuk menipu mereka yang akan membaca. Insiden semacam itu sering dilaporkan di media arus utama. Orang yang tidak bertanggung jawab akan menyunting atau menyunting konten yang mereka buat [20].

C. Cara Penyebaran Berita *Hoax*

Komunitas anti penipuan yang tergabung dalam *Turn Back Hoax* menjelaskan beberapa cara penyebaran berita bohong, antara lain:

1. ke arah yang kemudian

   Melegitimasi kebenaran isi berita. Menciptakan kebenaran baru menurut kehendak
2. Melalui akun yang sedang hangat

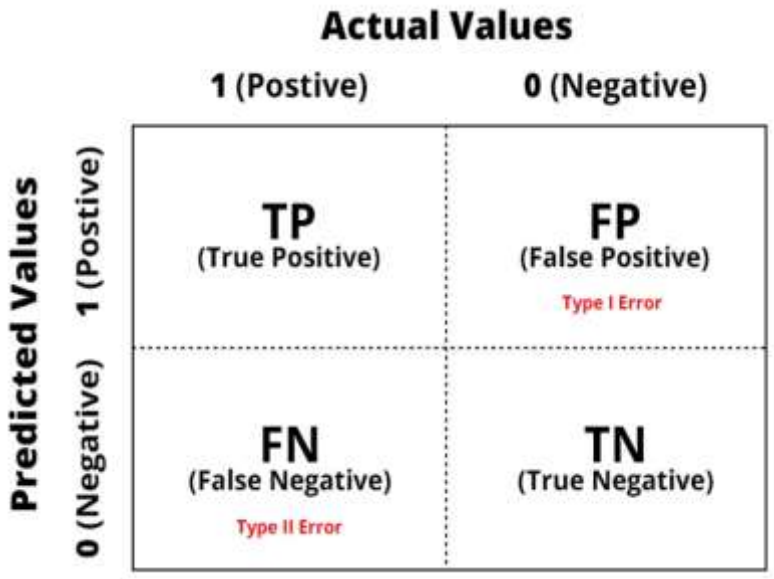
   Menyebarkan provokasi penipuan melalui tagar dan melalui sebuah permainan.
3. Selain itu, konsumen menerima berita yang biasanya mereka sukai secara sukarela dan membagikan berita berdasarkan kepentingan pribadi atau karena membenci pihak lain sehingga seolah-olah bersaing.

2.1.5 *Confusion Matrix*

Confusion matrix merupakan tabel yang sangat umum dipakai dalam *machine learning* yang berfungsi meningkatkan performa model klasifikasinya. Tabel ini berisi lebih banyak informasi tentang jumlah data yang tergolong signifikan atau minor. Salah satu dari beberapa metode analisis prediktif yang mencocokkan nilai yang nyata atau tersirat dengan hasil model prediksi adalah matriks kebingungan. Ini dapat digunakan

untuk memberikan metrik evaluasi seperti *Accuracy, Precision, Recall,* dan *F1-Score atau F-Measure.*

Confusion Matrix juga dikenal sebagai matriks kesalahan. Pada intinya, *Confusion* Matrix membandingkan hasil klasifikasi yang dihasilkan oleh sistem dengan hasil klasifikasi yang sesungguhnya. *Confusion matrix* adalah tabel matriks yang menggambarkan kinerja model klasifikasi pada sekumpulan data uji yang nilai sebenarnya diketahui. Gambar di bawah menggambarkan *Confusion Matrix* dengan berbagai kombinasi nilai prediksi dan actual [20].



Gambar 2. 3 *Confusion Matrix* [20]

Ada beberapa keuntungan dalam penggunaan *confusion matrix* yaitu, menunjukkan bagaimana model memprediksi dan memberikan informasi yang tidak hanya seputar kegalatan yang diciptakan oleh model, tapi juga tentang berbagai kegalatan yang aada. Setiap kolom *confusion matrix* mewakili turunan dari kelas predictor, dan setiap baris *confusion matrix* mewakili turunan dari kelas nyata.

Ada empat nilai dalam tabel *Confusion matrix* yaitu: *True Positive* (TP), *False Positive* (FP), *False Negative* (FN), dan *True Negative* (TN). Seperti terlihat pada tabel di bawah ini:

Tabel 2. 1 *Confusion Matrix* [20]

	Prediksi	
Aktual	TRUE	FALSE
TRUE	*TP*	*FP*
FALSE	*FN*	*TN*

Keterangan:

True Positive (TP) : Jumlah data positif yang telah diprediksi dengan benar sebagai positif.

False Positive (FP) : Jumlah data negatif yang diprediksi positif.

False Negative (FN) : Jumlah data positif yang diprediksi negatif.

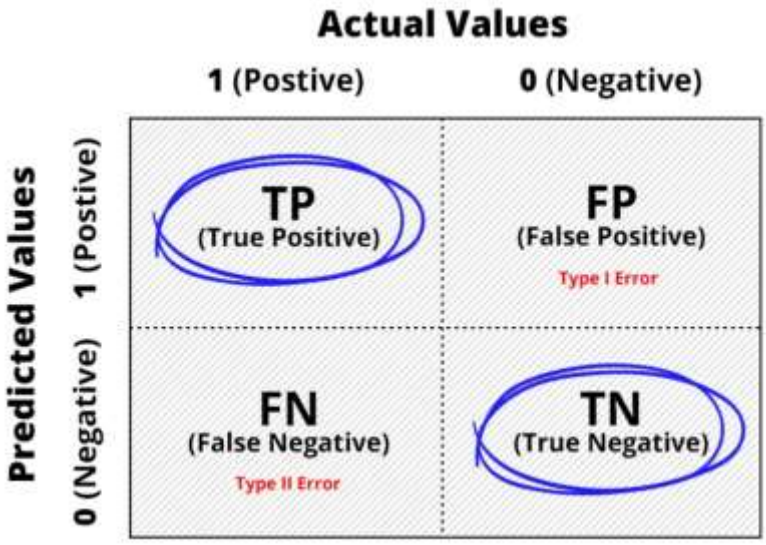
True Negative (TN**)** : Jumlah data negatif yang telah diprediksi dengan benar sebagai negatif.

Confusion matrix digunakan untuk mengatur tingkat akurasi, *Precision, Recall*, dan *F1-Score*. Metode evaluasi yang kuat dapat membantu untuk menilai kinerja pengklasifikasi atau algoritma pembelajaran mesin yang digunakan untuk membuat prediksi. Berikut cara menghitung metode evaluasi dengan menggunakan *confusion matrix*.

a. *Accuracy*

Nilai akurasi diperoleh dari sejumlah kumpulan data dengan nilai positif dan negatif yang ditentukan benar untuk masing-masing dataset. Akurasi menggambarkan seberapa akurat model dapat mengklasifikasikan dengan tepat. Oleh karena itu, akurasi adalah rasio prediksi yang benar (positif dan negatif) terhadap total data. Dengan kata lain, akurasi adalah derajat kedekatan nilai prediksi dengan nilai sebenarnya (sebenarnya). Nilai akurasi dapat diperoleh dengan menggunakan persamaan:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (\mathbf{2.2})$$

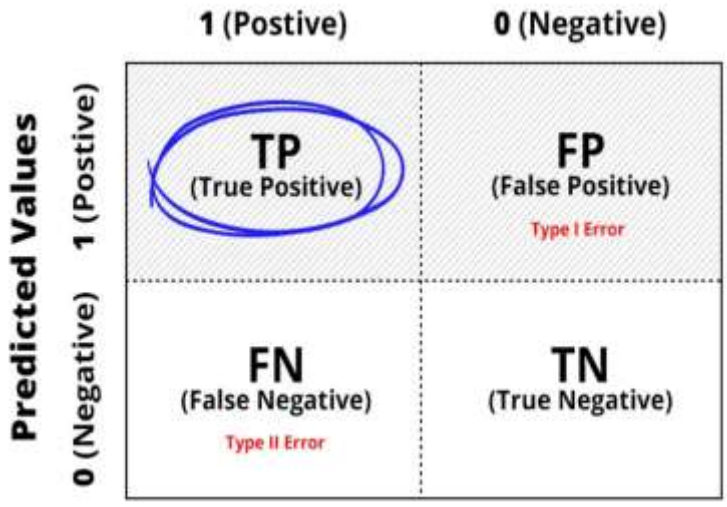


Gambar 2. 4 *Confusion Matrix* yang menggambarkan nilai *Accuracy* [21]

b. *Precision*

Presisi adalah probabilitas kasus yang diperkirakan positif tetapi ternyata positif. Presisi menggambarkan tingkat kesepakatan antara data yang diminta dan hasil prediksi model. Presisi didefinisikan sebagai rasio prediksi positif yang benar terhadap total hasil positif yang diprediksi. Berapa banyak dari kelas positif yang diprediksi dengan benar yang benar-benar positif? Nilai presisi dapat dihitung dengan menggunakan persamaan berikut:

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

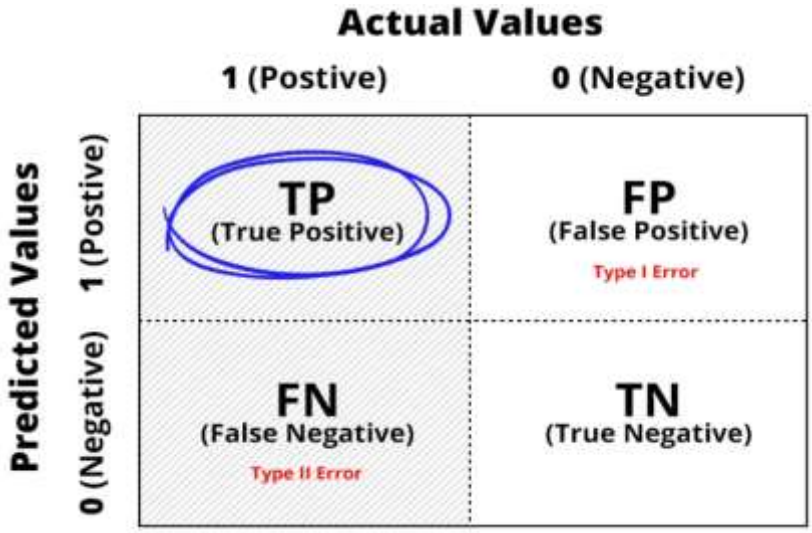


Gambar 2. 5 *Confusion Matrix* yang menggambarkan nilai *Precision* [21]

c. *Recall*

Recall adalah kemungkinan kasus dengan kategori positif akan menjadi positif. Ingat menggambarkan keberhasilan model dalam pencarian informasi. Akibatnya, pemulihan didefinisikan sebagai rasio prediksi positif yang sebenarnya terhadap semua prediksi positif yang sebenarnya. Nilai *recall* dapat dihitung dengan menggunakan persamaan berikut:

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



Gambar 2. 6 *Confusion Matrix* yang menggambarkan nilai *Recall* [14].

2.1.6 *Text Mining*

Teks Mining adalah proses menemukan informasi baru atau sesuatu yang unik dan menghapus sejumlah besar informasi. *Teks mining* menganalisis teks tidak terstruktur yang tidak harus berhubungan dengan satu hal tertentu dan memiliki hubungan dengan prinsip dan hukum lainnya. Hasil yang diantisipasi adalah informasi baru yang belum sepenuhnya dipahami atau belum jelas.

Teks mining membahas banyak sub-tugas, termasuk pencarian informasi, kategorisasi, penandaan POS, pengelompokan, dan lainnya yang dapat dikategorikan dalam kerangka Penemuan Pengetahuan di *Database*. Salah satu metode tersebut adalah metode mengidentifikasi pola dalam data yang otentik, unik, dan dapat dilihat. Penemuan pengetahuan dan penambangan data yaitu prosedur dengan menggunakan komputer yang

berfungsi menganalisis serta memeriksa hamper keseluruhan data, mengumpulkan kabar yang berguna, serta mengungkap pola tersembunyi. Penambangan data (*teks mining)* adalah proses untuk menemukan dan mengumpulkan data yang berguna dari berbagai sumber [15].

Dalam proses text mining, digunakan berbagai metode yang dapat digunakan untuk mengekstraksi makna dari informasi. Berikut tujuh teknik *text mining* yang bisa diterapkan.

1. *Information Extraction* (IE)

   Teknik *text mining* yang pertama adalah information *extracting* atau penggalian informasi dari informasi yang sudah ada. Ini adalah langkah pertama dalam penambangan data teks karena menggambarkan struktur kata yang tidak tepat, mencari kata kunci dan menentukan sentimen yang terkandung dalam teks.

2. *Information Retrieval* (IR)

   Selanjutnya, data mining adalah teknik penambangan data, yaitu. mencari data yang tepat. Jadi setelah mengidentifikasi informasi dengan arti dan kata kuncinya, informasi terkait yang serupa dapat ditemukan. Misalnya, jika kita mengetik *google* di kolom pencarian. Mesin akan menampilkan beberapa hasil lain yang mirip dengan kata kunci yang suda dimasukkan tadi.

3. *Natural Language Processing* (NLP)

   Berikut ini adalah teknik penambangan data teks yang disebut pemrosesan bahasa alami. Tugas teknologi ini adalah memproses informasi yang dihasilkan oleh data tekstual secara otomatis, meskipun dalam bentuk yang tidak terstruktur. Komputer atau mesin mencoba memproses data dengan menganalisis bahasa.

4. *Clustering*

   Teknik selanjutnya dalam *text mining* adalah mengelompokkannya berdasarkan kategori. Contoh sederhananya adalah mengkategorikan kalimat mana yang mengandung kata-kalta/kalimat negatif, seperti pada analisis sentimen status *twitter*. Dengan demikian, terdapat tiga kelompok yaitu teks dengan emosi negatif, netral dan positif.

5. *Categorization*

   Teknik ini digunakan untuk mengklasifikasikan data dalam format teks sesuai dengan kategori yang telah ditentukan. Metode ini melibatkan beberapa teknik untuk

mengidentifikasi data yang diklasifikasikan atau tidak diklasifikasikan, termasuk pengindeksan, reduksi dimensi, dan klasifikasi otomatis.

6. *Visualization*

   Teknik berikutnya yang paling umum digunakan adalah visualisasi atau melihat. Penambangan data tidak dapat diubah menjadi bentuk visual, tetapi penambangan teks dapat diubah menjadi bentuk visual. Teks yang diklasifikasikan diberi warna tertentu sesuai dengan kategorinya. Langkah ini menyederhanakan proses analisis data yang tidak terstruktur.

7. *Text Sumarization*

   Teks adalah data yang tidak terstruktur, jadi bisa satu paragraf panjang atau hanya satu kata. Untuk memudahkan pengolahan data teks yang berisi paragraf-paragraf yang panjang, perlu dilakukan rangkuman atau pemadatan teks. Sekalipun teks dipersingkat, proses ini tidak boleh meninggalkan isi asli teks yang panjang.

   Contoh penerapan *text mining* dapat dilihat di *customer service* sebuah perusahaan. Layanan pelanggan adalah industri yang berhubungan langsung dengan konsumen. Dengan demikian, mereka memahami bagaimana pelanggan bereaksi terhadap perusahaan melalui panggilan, obrolan, ulasan, dan lain sebagainya. Dalam praktiknya, beberapa teknologi juga telah dikembangkan yang dapat secara otomatis menanggapi pesan konsumen. Kita dapat melakukan ini dengan mengidentifikasi pertanyaan umum dan menjawabnya sekaligus, dengan demikian layanan pelanggan perusahaan bisa berjalan dengan lebih cepat [22].

2.1.7 Masyarakat

Society (masyarakat) berasal dari kata latin socius yang berarti persahabatan (persahabatan atau persahabatan). Persahabatan identik dengan sosialisasi. Morris Ginsberg mendefinisikan masyarakat sebagai "kumpulan orang yang dipersatukan oleh ikatan atau sikap tertentu yang membedakan mereka dari orang lain yang tidak termasuk dalam ikatan tersebut atau yang memiliki sikap yang berbeda dari mereka.

Abdul Syani menjelaskan, kata “komunitas” berasal dari bahasa Arab “musyarak” yang berarti “bersama”. Mereka kemudian menjadi warga negara yang berarti “berkumpul bersama, hidup bersama, saling berhubungan, dan saling mempengaruhi”, dan kemudian

terbentuk menjadi masyarakat Indonesia. Dapat disimpulkan bahwa penduduk memiliki sepuluh karakteristik, termasuk wilayah, kolektivitas orang, kelompok yang kuat, hubungan timbal balik orang dan golongan, interaksi timbal balik, interaksi yang dilembagakan, ikatan tertutup dan informal, kesamaan budaya, nilai dan kepercayaan universal, dan hubungan impersonal.

Wilayah adalah ruang dan lokasi tertentu. Dengan kemajuan teknologi internet, daerah tidak lagi harus berbentuk fisik tetapi juga dapat berbentuk ruang publik virtual, dan kolektivitas masyarakat mengacu pada sekelompok orang. Menurut George Simmel, kelompok diklasifikasikan menjadi dua jenis: diat dan triad. Kelompok masyarakat ini memiliki perasaan yang kuat, dibuktikan dengan adanya kesamaan makna dan tujuan, serta adanya hubungan timbal balik antara orang dengan orang dan antara orang dengan kelompok [23].

2.1.8 TF-IDF

TF-IDF (*Term Frequency Inverse Document Frequency*) merupakan metode yang digunakan untuk menentukan nilai frekuensi sebuah kata di dalam sebuah dokumen atau artikel dan juga frekuensi di dalam banyak dokumen. Perhitungan ini menentukan seberapa relevan sebuah kata di dalam sebuah dokumen. TFIDF juga merupakan sebuah algoritma yang umumnya digunakan untuk pengolahan data besar. Algoritma TF-IDF melakukan pemberian bobot pada setiap kata kunci disetiap kategori untuk mencari kemiripan kata kunci dengan kategori yang tersedia. Sebelum melakukan pembobotan maka akan dilakukan lima tahap pencarian text preprocessing yaitu pemecahan kalimat, *case folding, tokenizing, filtering,* dan *stemming,* lalu selanjutnya dilakukan proses menghitung bobot TF-IDF, bobot *query relevance* dan bobot *similarity.*

Nilai TF-IDF meningkat secara proporsional berdasarkan jumlah atau banyaknya kata yang muncul pada dokumen, tetapi diimbangi dengan frekuensi kata dalam *corpus.* Variasi dari skema pembobotan TF-IDF sering digunakan oleh mesin pencari sebagai alat utama dalam mencetak nilai (*scoring*) dan peringkat (*rankng*) sebuah relevansi dokumen yang diberikan *user*. TF-IDF pada dasarnya merupakan hasil dari perhitungan antara (*Term Frequency*) dan IDF (*Inverse Document Frequency*). Banyak cara untuk menentukan nilai yang tepat dari kedua statistik yang ada.

Nilai idf sebuah term (kata) dapat dihitung menggunakan persamaan sebagai berikut: [26]

$$IDF = log10(\frac{D}{dfi}) \quad (\mathbf{2.3})$$

D adalah jumlah dokumen yang berisi t*erm* (t) dan dfi adalah jumlah kemunculan (frekuensi) kata terhadap D. Adapun algoritma yang digunakan untuk menghitung bobot (W) masing-masing dokumen terhadap kata kunci (*query*), yaitu :[26]

$$Wd,t = tf\ d,t \times IDFt \quad (\mathbf{2.4})$$

Dimana d merupakan dokumen ke–d, t merupakan kata ke–t dari kata kunci, W merupakan bobot dokumen ke–d terhadap kata ke–t, dan tf merupakan *term* frekuensi/frekuensi kata. Setelah bobot (W) masing-masing dokumen diketahui, maka dilakukan proses pengurutan (*sorting*) dimana semakin besar nilai W, semakin besar tingkat kesamaan *(similarity*) dokumen tersebut terhadap kata yang dicari, demikian pula sebaliknya [26].

2.1.9 *Stemming*

Stemming adalah proses menghilangkan infleksi kata ke bentuk dasarnya, namun bentuk dasar tersebut tidak berarti sama dengan akar kata (*root word*). Misalnya kata "mendengarkan", "dengarkan", "didengarkan" akan ditransformasi menjadi kata "dengar". Melakukan teknik pemrosesan teks ini seringkali berguna untuk menangani kelangkaan dan/atau standarisasi kosa kata. Tidak hanya membantu mengurangi redundansi, karena sebagian besar kata induk dan kata infleksinya memiliki arti yang sama, ini juga memungkinkan model NLP untuk mempelajari hubungan antara kata infleksi dan kata induknya, yang membantu model memahami penggunaannya dalam konteks serupa.

Algoritma *stemming* berfungsi dengan mengambil daftar prefiks dan sufiks yang sering ditemukan dalam kata-kata infleksi dan memotong akhir atau awal kata. Hal ini terkadang dapat menghasilkan akar kata yang bukan kata sebenarnya dengan demikian, dapat disimpulkan bahwa pendekatan ini pasti memiliki kelebihannya, tetapi bukan tanpa keterbatasannya [27].