BAB II
LANDASAN TEORI

2.1 Tinjauan Pustaka

Penilaian suatu *brand* atau produk menjadi *point* penting untuk tetap diperhatikan sedetail mungkin, penilaian suatu *brand* atau produk menjadi titik keberhasilan dari sebuah perusahaan tersebut. Salah satu cara untuk mengetahui dan memproses hasil analisis dari setiap penilaian suatu *brand* atau produk dapat dilakukan dengan menggunakan sentiment analisis. Sentimen analisis bertujuan untuk dapat mengekstrak opini secara terstuktur dan untuk menentukan polaritas sentimennya. Kontribusi dari beberapa penelitia sebagai titik awal kajian ilmu pengetahuan dalam membangun sebuah program sistem perlu untuk diketahui dari setiap aspek bidang ilmu yang akan dikembangkan, oleh karena itu pada paragraph selajutnya akan beisikan beberapa penelitian terdahulu yang telah dilakukan oleh para peneliti dunia.

Tabel 2.1 Penelitian Terkait

Referensi	Metode	Keterangan
(Chen dan Huang, 2019)	*Neural Network*	Analisis sentimen untuk ulasan mobil di China dengan menggunakan *Neural Network* dan membandingkan analisa bersifat konvensional dan modern.
(Alqaryouti dkk, 2019)	*SVM dan leksikon SVM*	Menganalisis pendapat publik pada penggunaan aplikasi *mobile smart government*. Analisis dilakukan, untuk meningkatkan pelayanan publik.
(Diekson dkk., 2023)	SVM, Regresi Logistik dan Naïve Bayes	Sentimen analisis terhadap aplikasi Traveloka dengan membandingkan beberapa metode yaitu SVM, Regresi Logistik dan Naïve Bayes.
(Al-Smadi dkk., 2018)	SVM dan RNN	Menganalisi Sentimen berbasis aspek dengan menggunakan data ulasan Hotel

Referensi	Metode	Keterangan
		Arab, pemodelan dilakukan dengan mengimplementasikan SVM dan RNN.
(Moraes dkk., 2013)	*Support Vector Machine (SVM) dan Artificial Neural Network*	Mengklasifikasikan ulasan tekstual, terhadap sebuah topik yang diberikan terhadap sistem, untuk mengetahui apakah ekspresi sentimen positif atau negatif dengan implementasi SVM dan ANN.
(Hossain dkk., 2020)	CNN dan LSTM	CNN digunakan untuk mempelajari representasi kata dan LSTM digunakan untuk mempelajari representasi kata yang lebih tinggi yang digunakan untuk klasifikasi
(Al-Natour dan Turetken, 2020)	*VADER SENTIMEN ANALISIS*	Sentimen analisis mampu mengukur dan menyimpulkan ulasan pengguna terhadap sebuah produk. Hasil analisisnya terhadap 900 ulasan menunjukan bahwa yang sentiment analisis dapat menampilkan tingkat akurasi, sehingga sangat efektif.
(Borg dan Boldt, 2020)	*VADER SENTIMEN* DAN *SVM*	Sentiment analisis untuk mengetahui respon dan tingkat ketidak puasan pelanggan pada perusahaan telekomunikasi Swedia, data yang digunakan yaitu menggunakan data dari email pelanggannya
(Sadiq dkk., 2021)	*VADER SENTIMEN* DAN *CNN*	Sentimen analisis digunakan untuk membuktikan ketidaksesuain antara hasil ulasan dan jumlah rating yang diberikan pengguna. Data diambil dari ulasan

Referensi	Metode	Keterangan
		aplikasi *Yelp* yang terdapat di halaman *Google Play Store*.
(Mahmud dkk., 2022)	*CNN, LSTM dan DistilBert*	Sentimen analisis dilakukan dengan menggunakan dataset dari ulasan beberapa aplikasi ojek online yang ada di Google Play Store. Analisa dilakukan dengan membandingkan algoritma yang diteliti.
(Wen dkk., 2021)	*MLSTM* dan *CMOS*	Sentimen analisis dilakukan dengan melakukan pengujian terhadap IMDb dataset dan SemEval dataset yang diuji dengan melakukan implementasi dan perbandingan antara LSTM berbasis CMOS dan MLSTM.

Penelitian lain yang berkaitan dengan penelitian ini, dilakukan oleh Chen & Huang, (2019). Penelitiannya bertujuan untuk membandingkan sentimen secara kovensional dan modern. Peneliti memodelkan identifikasi pasangan aspek-opini dan klasifikasi sentimen tingkat aspek sebagai tugas klasifikasi teks. Selain itu, peneliti menggabungkan pengetahuan eksternal ke dalam jaringan saraf untuk mengimbangi kurangnya data pelatihan. Hasil eksperimen menunjukkan bahwa jaringan saraf yang ditingkatkan pengetahuannya secara konsisten mengungguli model konvensional terhadap data pada ulasan mobil di China (Chen dan Huang, 2019).

Penelitian sentimen analisis, yang dilakukan oleh (Alqaryouti dkk., 2019), melakukan analisa terhadap aplikasi *smart government*, yang berguna untuk meningkatkan kinerja aplikasi dan layanan pemerintah. Evaluasi dilakukan dengan menganalisa ribuan data, kemudian diklasifikasi menggunakan SVM dan leksikon SVM. Hasil yang didapat leksikon lebih unggul dibandingkan SVM (Alqaryouti dkk., 2019).

Penelitian (Diekson dkk., 2023), membahas kepuasan pelanggan terhadap layanan Traveloka dengan menganalisis berapa banyak orang yang puas dan tidak puas dengan layanan yang ditawarkan Traveloka. Dataset dikumpulkan dari platform media sosial yang terdiri dari 1200 *tweet* terkait Traveloka. Penelitian ini menggunakan tiga metode klasifikasi: Support Model Vektor (SVM), Regresi Logistik, dan Näıve Bayes. Hasil menunjukkan, dari 1200 *tweets* kebenyakan memberikan nilai postif terhadap traveloka. Hasilnya menunjukkan bahwa SVM lebih baik akurasi dalam menentukan sentimen tweet tentang Traveloka (Diekson dkk., 2023).

Penelitian (Al-Smadi dkk., 2018) melakukan analisis sentimen berbasis aspek (ABSA) dari ulasan Hotel Arab. Penelitian dilakukan dengan mengimplementasikan deep recurrent neural network(RNN) dan Support Vector Machine (SVM). Evaluasi dilakukan menggunakan kumpulan data berupa referensi ulasan Hotel Arab. Hasil evaluasi, menunjukkan bahwa metode SVM lebih baik daripada metode RNN pada kategori sentimen aspek based. Akan tetapi jika dilihat waktu eksekusinya, RNN lebih cepat (Al-Smadi dkk., 2018).

Proses klasifikasi sentimen bertujuan untuk mengklasifikasikan ulasan tekstual, yang diberikan pada satu topik, sebagai ekspresi sentimen positif atau negatif. *Support Vector Machine* (SVM) telah digunakan secara ekstensif dan berhasil sebagai pendekatan pembelajaran sentimen sementara Jaringan Syaraf Tiruan (JST) jarang dipertimbangkan dalam studi komparatif dalam literatur analisis sentiment. Pada penelitian yang berkaitan dengan proses klasifikasi sentiment dilakukan oleh Moraes dkk, (2013), peneliti menggunakan objek sebuah dokumen untuk menganalisa menggunakan teknik sentimen analisis. Peneliti melakukan perbandingan empiris antara SVM dan ANN mengenai analisis sentimen tingkat dokumen. Percobaan penelitian menunjukkan bahwa JST menghasilkan hasil yang lebih unggul atau setidaknya sebanding dengan SVM. Khususnya pada kumpulan data tolok ukur ulasan sebuah produksi film, ANN mengungguli algoritma SVM dengan perbedaan yang signifikan secara statistik, bahkan pada konteks data yang tidak seimbang (Moraes dkk., 2013).

Pada penelitian (Hossain dkk., 2020) tentang ulasan pelanggan restoran yang di ambil dari kumpulan data yang di miliki oleh Bangladesh dalam Bahasa Bangali telah terkumpul data ulasan 1000 ulasan, Langkah pertama yaitu Menyusun data agar terstruktur dan mengatur label data menjadi positif dan negatif, setelah pra-pemrosesan kemudian data dimasukkan ke dalam model yang digunakan, seperti CNN digunakan untuk mempelajari representasi kata dan LSTM digunakan untuk mempelajari representasi kata yang lebih tinggi yang digunakan untuk klasifikasi, dan telah menetapkan beberapa *hyperparameter* sebelum melatih model, dengan word2vec yang dilatih sebelumnya dengan 300 dimensi, dengan akurasi yang mendekati memuaskan pada ulasan restoran Bengali (Hossain dkk., 2020).

Penelitian yang berkaitan dengan sentiment analisis lainnya telah dilakukan oleh peneliti Al-Natour & Turerken, (2020). Pada penelitiannya mampu mengeksplorasi kelayakan analisis secara otomatis. Menurut penelitiannya kemampuan dari sentimen analisis mampu untuk mengukur dan menyimpulkan ulasan pengguna terhadap sebuah produk. Hasil analisisnya terhadap 900 ulasan menunjukan bahwa yang sentiment analisis dapat menampilkan tingkat akurasi, sehingga sangat efektif dalam mendeteksi nada yang mendasari konten yang dianalisis dan dapat digunakan sebagai alternatif untuk meningkatkan penilaian (Al-Natour dan Turetken, 2020),

Penelitian lainnya yaitu dilakukan oleh Borg & Boldt, (2020). Tujuan dari penelitiannya yaitu melakukan sentiment analisis untuk mengetahui respon dan tingkat ketidak puasan pelanggan pada perusahaan telekomunikasi Swedia, data yang digunakan yaitu menggunakan data dari email pelanggannya. Hasil penelitiannya menunjukkan bahwa model Linier *Support Vector Machine* (SVM) mampu mengekstraksi sentiment dengan mean skor F1 0.834 dan mean AUC 0.896, serta algoritma Linier SVM mampu memprediksi sentiment pada email selangkah lebih baik dan terstruktur yang didasarkan pada teks email. *Support Vector Machine* dipilih karena tingkat keakurasian yang cukup tinggi, sekitar 80 hingga 90 persen, serta pengimplementasian algoritma yang cukup mudah dan fleksibilitas algoritma yang dapat digabungkan dengan metode lainnya (Borg dan Boldt, 2020).

Saat ini, ulasan memainkan peran penting dalam mempengaruhi konsumen. Konsumen menyampaikan pengalaman dan informasi mengenai sebuah produk yang disampaikan melalui halaman *Google Play Store*. Hal tersebut dikemukakan dalam penelitian (Sadiq dkk., 2021), peneliti mengambil sampel dari ulasan produk aplikasi *Yelp* yang berkaitan dengan restauran. Tujuan penelitian dilakukan untuk membuktikan adanya ulasan palsu dan rating palsu yang terdapat di *google Play Store*. Hal ini secara tidak langsung dapat mempengaruhi konsumen dan kesuksesan aplikasi. Penelitian dilakukan dengan menggunakan kerangka *deep learning*. Polaritas ulasan dideteksi dengan menggunakan sentimen analisis dan rating dideteksi dengan deteksi hasil ulasan yang diperoleh. Sentimen analisis terhadap ulasan aplikasi *Yelp* diuji menggunakan metode VADER dan CNN. Peneliti menghitung ketidaksesuaian antara ulasan pengguna dan jumlah rating yang diberikan, hasilnya menunjukkan adanya bias antara ulasan dan rating dengan metode VADER sebesar 25,03%. Hasil menunjukkan bahwa metode VADER dan CNN memberikan nilai lebih baik terhadap ulasan, dibandingkan dengan jumlah rating yang diberikan. Sehingga, ditemukan adanya ketidak cocokan antara ulasan dan jumlah rating(Sadiq dkk., 2021).

Penilitian (Mahmud dkk., 2022) melakukan penelitian sentimen analisis dengan menggunakan dataset dari *Google Play Store*. Penelitian dilakukan dengan membandingkan beberapa aplikasi ojek online seperti Uber, Lyft, GO CNG dan Pathao. Data dikumpulkan dari bagian ulasan pengguna yang terdapat di *Google Play Store*. Pemodelan dilakukan dengan menggunakan algoritma CNN, LSTM dan DistilBert. Hasilnya DistilBert terbaik dan mencapai akurasi tertinggi sebesar 98,84%(Mahmud dkk., 2022).

Penelitian (Wen dkk., 2021), melakukan sentimen analisis dengan mengunakan dataset yang bersumber dari *Internet Movie Database* (IMDb) dan *Semantc Evaluation* (SemEval). Penelitian dilakukan untuk menguji dan membandingkan algoritma LSTM berbasis CMOS dan MLSTM. Hasil eksperimen menunjukkan bahwa sistem LTSM berbasis memristor yang diusulkan mencapai akurasi yang lebih tinggi dibandingkan metode berbasis CMOS, serta kinerja yang lebih baik dalam hal konsumsi daya(Wen dkk., 2021).

2.2 Dasar Teori

2.2.1 Aplikasi Belajar Online

Bidang pendidikan memiliki peranan penting dalam kehidupan manusia karena dengan pendidikan manusia dapat menjalani kehidupan dengan lebih baik dan menjadi modal dasar sebuah pengetahuan dalam kehidupan sehari-hari. Pendidikan merupakan salah satu faktor yang sangat fundamental dalam upaya meningkatkan kualitas kehidupan manusia, di samping itu juga merupakan faktor penentu dalam perkembangan dibidang sosial dan ekonomi untuk ke arah kondisi yang lebih baik. Teknologi dan pendidikan, pada saat ini berkembang berdampingan dan sangat maju sangat cepat, berbagai bentuk inovasi teknologi unutk pendidikan sangat bermanfaat bagi kehidupan manusia. Teknologi dan pendidikan tidak dapat dipisahkan sesuai dengan perkembangan zaman karena teknologi dapat membantu proses perkembangan pada bidang pendidikan (Shoumi, 2019).

Seiring perkembangan zaman, bimbel mulai berubah menjadi bimbel online yang dapat diakses dimana saja. Siswa tidak perlu datang ke lokasi bimbel karena mereka bisa belajar kapanpun, dimanapun mereka mau dengan hanya membawa laptop ataupun gadget dengan jaringan internet untuk dapat mengakses aplikasi bimbel tersebut. Bimbingan belajar atau les *private* selain bisa dilakukan secara langsung juga bisa dilakukan secara *online/daring*. Bimbingan belajar secara *online* dirancang untuk memungkinkan proses pembelajaran jarak jauh melalui internet tanpa harus bertatap muka dengan pengajarnya. Bimbingan belajar *online* dapat memberikan alternatif pilihan bagi siswa yang memiliki akses jaringan internet untuk memperoleh layanan bantuan belajar yang efektif, efisien, dan interaktif secara optimal (Ramadhayanti, 2018).

Salah satu perusahan yang berkembang dalam menyediakan bimbingan belajar online saat ini sangatlah banyak. Perusahaan penyedia aplikasi bimbingan *online* dan belajar *online* diantaranya adalah seperti aplikasi Quipper, Zenius Education, Ruang Guru, Prime Mobile dan lainnya. Berdasarkan halaman blog resmi Ruangguru mengatakan bahwa telah ada 6 juta pengguna semenjak didirikan pada tahun 2014, dan berdasarkan situs halaman website resmi *Cable News*

Network Indonesia (CNN Indonesia) memuat berita bahwa aplikasi Ruangguru mencatat pengingkatan pengguna hingga 10 kali lipat sejak tahun 2016. Aplikasi Ruangguru hadir sebagai bimbel *online* yang digunakan dengan jaringan internet di berbagai komputer, gadget ataupun *smartphone* (Hayati, 2020).

2.2.2 Media Sosial

Media sosial berkembang pesat karena banyak digunakan oleh orang-orang dan berguna untuk menganalisis berbagai postingan pengguna untuk mengidentifikasi tingkat emosi mereka. Penggunaan situs jejaring sosial terus meningkat terutama oleh generasi muda. Orang-orang di media sosial mengungkapkan perasaan mereka, kegiatan sehari-hari, pendapat tentang berbagai topik, dan lain sebagainya (Sharma dan Sharma, 2020). Media sosial kini ada berbagai macam dan hampir dimiliki oleh semua orang pada ponsel pintarnya, dengan contohnya, aplikasi *Facebook*, *Twiiter*, *Instagram*, dan lainnya.

2.2.3 Sentimen Analisis

Sentimen analisis merupakan proses pengolahan data yang diperoleh dari berbagai macam data untuk mendapatkan sebuah informasi. Analisis sentimen digunakan untuk menemukan atau menganalisa umpan balik pengguna mengenai sebuah produk dan layanan(Li dkk., 2017). Sentimen analisis dianggap sebagai salah teknik untuk mendefinisikan dan mengekstraksi perasaan manusia melalui teks tidak terstruktur dan dilakukan melalui pemrosesan bahasa alami dan pembelajaran mesin (*Machine Learning*). Pembelajaran mesin (*Machine Learning*) adalah sistem pembelajaran yang membantu dalam mempelajari dan melatih kumpulan data yang diperoleh dari media sosial. Analisis sentimen adalah instrumen terbaik untuk menentukan apakah evaluasi itu positif atau negatif. Hasil dari sentimen analisis didapatkan dari keterkaitan hubungan saling mempengaruhi aspek kognitif, psikologis dan social (Jindal dan Aron, 2021).

Sentimen Analisis memiliki ciri khas sebagai teknik manipulasi kontekstual teks yang mengenali informasi subjektif dalam materi sumber dan kemudian mengekstrak sebuah data. Sentimen analisis telah banyak membantu perusahaan memahami sentimen sosial produk atau layanan setiap perusahaan. Hampir semua perusahaan profesional dan non-teknis menggunakan teknik sentimen analisis

untuk mengartikan maksud dari masukkan dan keinginan pelanggan mengenai bagaimana kualitas dari produk atau layanannya (Chitra dkk., 2021).

2.2.4 Google Play Store

Google Play Store merupakan pasar terbuka untuk berbagai macam aplikasi seperti toko online, permainan, buku elektronik, kelas belajar, dan lain-lain, yang tersedia di *platform* tersebut. Sebagian besar aplikasi tersedia secara gratis untuk diunduh, sehingga memudahkan pengguna untuk mendapatkan aplikasi yang diinginkan. *Google Play Store* memiliki jumlah pengikut yang sangat besar sehingga banyak digunakan perusahaan aplikasi untuk memasarkan dan mengenal produk yang dibuat. Setiap halaman aplikasi di Google Play Store memiliki bagian komentar di mana pengguna menyampaikan kritik membangun mereka terhadap aplikasi yang telah diunduh dan digunakan(Jawad Soumik dkk., 2019).

Teknologi telah mengubah cara manusia untuk melakukan aktifitasnya, sehingga teknologi mempermudah manusia dalam memenuhi kebutuhannya melalui penggunaan gadget dan internet. Teknologi mengakibatkan informasi menjadi sangat mudah diakses, seperti informasi mengenai berbagai macam aplikasi yang dapat membantu dalam proses belajar mengajar yang tersedia di *Google Play Store.* Indonesia memiliki banyak perusahaan pembuat aplikasi belajar, seperti, Ruangguru, Zenius, Quipper, dan lain-lain. Setiap aplikasi memiliki kelemahan dan kelebihannya masing-masing. Sebagaian besar pengguna menyampaikan hal tersebut melalui halaman keluhan dan saran yang tersedia di halaman *Google Play Store*(Ahmadi dkk., 2020).

Halaman *Google Play Store,* merupakan tempat yang menarik untuk dibaca pengguna. Pada halaman tersebut terdapat sebuah ulasan mengenai pendapat pengguna yang telah menggunakan aplikasi tersebut. Hal ini, membuat pengguna baru aplikasi tersebut dapat mempertimbangkan dan mengambil sebuah keputusan untuk meng-unduh aplikasi tersebut atau tidak (Wahyudi dan Kusumawardana, 2021). Ulasan mengenai aplikasi merupakan ruang yang digunakan pengembang aplikasi untuk memperbaiki kinerja aplikasi. sehingga pengembang aplikasi dapat mengambil sebuah keputusan berdasarkan rekomendasi yang diberikan oleh pelanggan(Venkatakrishnan dkk., 2020).

2.2.5 Aspect-Based Sentimen Analysis (ABSA)

Pemodelan analisis sentimen berbasis aspek (ABSA) diperlukan untuk mengekstraksi sentimen secara detail, terutama dalam ulasan produk seperti kamera dan ponsel pintar, serta merek tertentu seperti Apple, Samsung, Google, dan lain-lain (Zainuddin dkk., 2018). *Aspect-based sentiment analysis* digunakan untuk mengetahui aspek apa yang mendapat penilaian positif, netral, atau negatif dari pelanggan. Misalnya dalam sebuah *review* produk, ada pelanggan yang memberikan komentar, "Bahan celana ini halus dan nyaman digunakan." Maka, dapat disimpulkan bahwa aspek yang mendapat penilaian positif dari pelanggan adalah bahan celana tersebut. Perbedaan utama antara analisis sentimen dan ABSA adalah analisis sentimen cenderung mendeteksi sentimen melalui teks yang diberikan. Sedangkan ABSA adalah teknik yang menentukan berbagai aspek dalam sebuah teks dan sentimen (positif, negatif atau netral) dari setiap aspek yang diidentifikasi pada teks itu, seperti pada gambar 2.1 (AL-Smadi dkk., 2023).

Gambar 2.1 Aspect-based sentiment analysis

2.2.6 Teknik Scraping

Teknik *Scraping* merupakan teknik untuk mengubah data web yang tidak terstruktur menjadi data terstruktur yang dapat disimpan dan dianalisis dalam database atau spreadsheet pusat. *Scraping* banyak digunakan ilmuwan untuk melakukan ekstraksi data dalam jumlah besar dan terus menerus, yang dihasilkan secara *online* dengan biaya relatif rendah. Proses *web Scraping,* dibagi menjadi 3 tahap yaitu *Fetching*, *Extraction*, dan *Transformation*. Tahap *Fetching* dilakukan dengan akses melalui protokol HTTP, untuk mengirim dan menerima permintaan dari *server web*. Kemudian dilakukan *Extraction*, melalui halaman HTML dengan

melakukan parsing data. Setelah data terkumpul, dilakukan *Transformation* data, untuk mendapatkan data yang terstruktur (Khder, 2021).

2.2.7 Text Preprocessing

Pada sebuah dokumen atau sebuah teks pada umumnya mempunyai struktur kata yang sembarang atau tidak terstrukrur, sehingga dapat dikatakan bahwa setiap dokumen atau teks dalam sebuah kalimat perlu ada tahapan untuk membuat sebuah kata yang dapat dimengerti oleh sebuah program sistem. Oleh karena itu, text preprocessing merupakan proses pengubah teks yang tidak terstruktur menjadi terstruktur, dimana sudah adanya pengurangan volume kata, menjadikan kata menjadi dimensi kata yang dapat diolah oleh program sistem (Analisis-data.com, 2017).

Sebuah dokumen mengadung banyak makna dan variasi kata dari setiap bentuk huruf sampai dengan tanda baca. Variasi huruf harus diseragamkan dikarenakan harus dijadikan bentuk dari huruf yang seragam (dimana huruf besar menjadi huruf kecil). Selain itu, proses penghilangan tanda baca dilakukan untuk menghilangkan noise pada saat pengambilan informasi. Proses *text preprocessing* dilakukan dengan tujuan agar data yang digunakan bersih dari noise, memiliki dimensi yang lebih kecil, serta lebih terstruktur. Adapun berikut ini merupakan tahapan dari *text preprcessing*, di antaranya (Jumeilah, 2017).

1. Normalisasi, sebuah data teks diperlukan pengecekan apakah sebuah kata termasuk dalam kamus bahasa Indonesia. Oleh karena itu diperlukannya normalisasi atau leksikon bahasa Indonesia yang didapatkan dari berbagai macam data teks yang diambil dari media sosial ataupun internet, sehingga didapatkan data yang seragam untuk mempermudah kerja dari sistem yang akan dibangun.
2. *Case Folding*, merupakan proses dalam *text preprocessing* yang dilakukan untuk menyamakan atau menyeragamkan karakter pada data teks. Proses *case folding* adalah proses mengubah seluruh huruf menjadi huruf kecl. Pada proses ini, karakter huruf “A sampai dengan Z” yang terdapat pada data teks diubah menjadi karakter huruf kecil “a sampai dengan z”, (tanda baca atau angka akan dihilangkan dari data dan dianggap sebagai *delimeter*. *Dilimeter* adalah urutan

satu atau lebih karakter yang digunakan untuk menentukan batas pemisah suatu data (Jumeilah, 2017).

3. *Cleaning, cleaning* merupakan proses membersihkan sebuah *review* atau penilaian dari kata-kata yang tidak diperlukan untuk mengurangi proses noise pada proses klasifikasi. Kata-kata yang dihilangkan merupakan karakter yang tidak berarti pada proses pengolahan data (Analisis-data.com, 2017).
4. *Tokenizing*, proses *tokenizing* adalah tahap pemotongan string masukan bedasarkan kata-kata yang menyusun atau dengan kata lain pemecahan kalimat menjadi kata-kata. Pada umumnya menggunkan strategi yang dilakukan yaitu *white space* atau spasi dan membuang karakter tanda baca. *White space* adalah proses pengoreksian tanda baca, contohnya kesalahan sebuah kata dari sebuah kalimat atau dokumen. Tahap *tokenizing* membagi urutan karakter menjadi kalimat dan kalimat menjadi sebuah token (Jumeilah, 2017).

2.2.8 N-Gram

Metode N-gram adalah sebuah metode pemotongan atau pemisahan string di dalam kalimat atau kata. N-gram diaplikasikan untuk mengambil potongan *string* atau karakter dengan jumlah (n) tertentu dari sebuah kata yang ada secara berkelanjutan dari awal hingga akhir. Metode N-gram dibedakan berdasarkan jumlah pemotongan *string* atau karakter yang diproses. Terdapat banyak jenis N-gram yaitu Uni-gram untuk pemotongan tiap satu karakter *string*, Bi-gram untuk pemotongan tiap dua karakter *string*, Tri-gram untuk pemotongan tiap tiga karakter *string*, Quad-gram untuk pemotongan tiap empat karakter *string* dan seterusnya (Nadya , 2015).

Sebagai contoh dari penerapan metode N-gram, seperti berikut. Bentuk teks yang akan diolah yaitu “STEMMINGWORDS”, maka proses dari N-gram yang akan dilakukan yaitu;

Jenis Uni-gram : S, T, E, M, M, I, N, G, W, O, R, D, S.

Jenis B-gram : S, ST, TE, EM, MM, MI, IN, NG, GW, WO, OR, RD, DS, S

Jenis Tri-gram : ST, STE, TEM, EMM, MMI, MIN, ING, NGW, GWO, WOR, ORD, RDS, DS.

Jenis Quad-gram : STE, STEM, TEMM, EMMI, MMIN, MING, INGW, NGWO, GWOR, WORD, ORDS, RDS.

Salah satu keunggulan dari penggunaan N-gram adalah metode ini tidak terlalu sensitif dengan kesalahan penulisan yang terdapat dalam suatu dokumen atau teks. Metode N-gram juga dinilai lebih efektif untuk digunakan dalam stemming words terhadap bahasa aglutinatif karena proses pemotongan karakter *string* yang dilakukan tidak bergantung pada jenis huruf dan bentuk huruf yang digunakan dalam bahasa tersebut(Laippala dkk., 2015).

2.2.9 TF-IDF (*Term Frequency-Inverse Document Frequency*)

Analisis teks merupakan proses mengekstrak informasi, makna, dan pola dari teks yang tersedia. Beberapa metode yang biasa digunakan dalam bidang analisis teks antara lain TF-IDF, Word2Vec, dan N-gram. Penelitian sentimen analisis terhadap aplikasi belajar online menggunakan TF-IDF. Pemilihan TF-IDF didasari pada penelitian sebelumnya yang mengatakan bahwa TF-IDF mencapai akurasi tertinggi jika dibandingkan N-Gram(Ahuja dkk., 2019) dan kemudian diperkuat kembali terhadap penelitian (Cahyani dan Patasik, 2021), yang menyimpulkan bahwa TF-IDF mencapai akurasi terbaik jika dibandingkan Word2Vec.

TF-IDF (*Term Frequency-Inverse Document Frequency*) memiliki karakteristik yang bermanfaat dalam penelitian ulasan teks terhadap ulasan aplikasi belajar *online*. Hal ini didasarkan pada sifat TF-IDF dapat digunakan untuk memilih fitur-fitur yang paling informatif dalam analisis ulasan. Dengan menggunakan skor TF-IDF, peneliti dapat mengidentifikasi kata-kata yang paling penting dalam ulasan dan menggunakannya sebagai fitur-fitur dalam model analisis sentimen atau klasifikasi ulasan. TF-IDF mampu menginditifikasi kata-kata ulasan yang penting, yang berkaitan dengan aplikasi belajar *online*, Seperti kata belajar, aplikasi, bagus, sangat, dan kata-kata yang bermakna terkait ulasan aplikasi belajar *online*. Penerapan TF-IDF dalam penelitian dapat membantu dalam pemrosesan dan analisis ulasan dengan cara yang lebih informatif dan relevan.

TF-IDF (*Term Frequency-Inverse Document Frequency*) adalah metode yang digunakan dalam pemrosesan bahasa alami untuk mengevaluasi kepentingan

relatif suatu kata dalam dokumen dalam korpus teks. Penilaian TF-IDF mencerminkan pentingnya suatu ekspresi dalam suatu dokumen relatif terhadap kemunculannya dalam dokumen lain dalam korpus(Mee dkk., 2021). Cara kerja TF_IDF didasarkan pada rumus untuk menghitung suatu dokumen pada persamaan;

$$TF(t) = \frac{F_{t,d}}{\sum t, d} \tag{2.1}$$

$$IDFt = log\left(\frac{N}{df_t}\right) \tag{2.2}$$

$$\text{TF. IDF}_{t,d} = TF(t) * IDFt \tag{2.3}$$

Keterangan:

t : Suatu Kata

d : Suatu Dokumen

TFt.d : Frekuensi *term* t pada Dokumen d

IDFt : Pendistribusian *term* t pada seluruh Dokumen

Ft.d : Jumlah *term* t yang ada pada Dokumen d

$\Sigma\ t$,d : Jumlah kata yang ada pada Dokumen d

N : Jumlah Dokumen

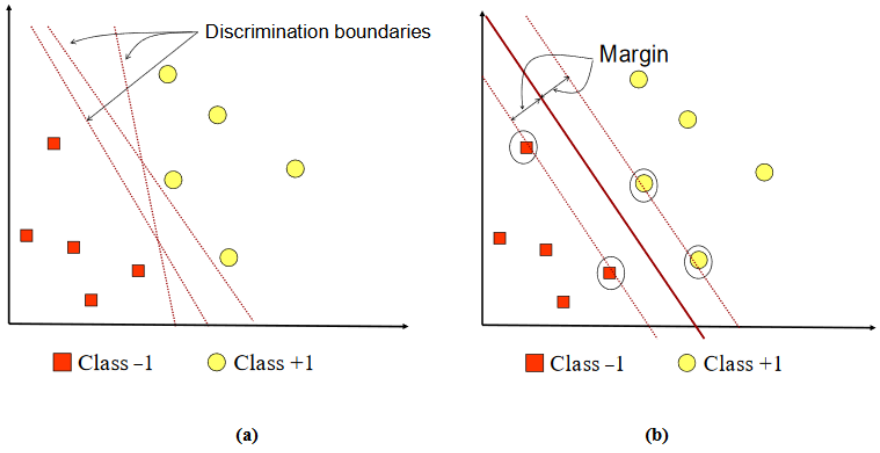
dft : Jumlah dokumen yang mengandung *term* t

Analisa teks menggunakan TF-IDF, meskipun baik, tetapa saja memiliki kelebihan dan kekurangan. Kelebihan TF-IDF membantu dalam mempertimbangkan konteks lokal dokumen dan mengidentifikasi kata-kata yang unik atau khas dalam representasi vektor. Mengurangi pengaruh kata-kata yang umum dan cenderung tidak memberikan makna khusus dalam analisis teks. TF-IDF dapat mempertahankan informasi yang relevan dan penting dalam representasi vektor, sehingga dapat membedakan dokumen-dokumen berdasarkan konten dan mengindentifikasi kata-kata kunci yang berkontribusi pada makna keselrurhan. Konsep dasar TF-IDF sederhana dan mudah diimplementasikan dalam proses penelitian. Kekurangan TF-IDF dalam melakukan analisa teks, yaitu tidak memperhatikan urutan kata dan mengabaikan makna semantik yang luas.

2.2.10 Support Vector Machine

Support Vector Machine adalah salah satu algoritma yang sangat berguna dalam klasifikasi data yang besar. Ide Algoritma SVM diciptakan oleh Vladimir Vapnik dan sebagai algoritma yang efisien dan bisa diterapkan di berbagai bidang(Wang dan Zhao, 2020). Standar dari SVM yaitu mengambil himpunan data masukan dan memprediksi untuk setiap masukkan yang diberikan, hasil kemungkinan masukkan adalah anggota dari salah satu kelas dari dua kelas yang ada, yang mana membuat algoritma SVM sebagai penggolong *non probabilistic linier biner*, dikarenakan algoritma SVM sebuah pengklasifikasi, dengan kemudian diberi himpuanan pelatihan yang mana masing-masing ditandai sebagai milik salah satu dari dua kategori. SVM dikembangkan oleh Boser, Guyon, dan Vapnik, pertama kali diperkenalkan pada tahun 1992 di *Annual Workshop on Computational Learning Theory*. SVM merupakan suatu teknik untuk melakukan prediksi, baik prediksi dalam kasus regresi maupun klasifikasi (Octaviani dkk., 2014).

Berbeda dengan strategi *neural network* yang berusaha mencari nilai *hyperplane* sebagai pemisah antar kelasnya, algoritma SVM bekerja atas prinsip *Structural Risk Minimization* (SRM) yang bertujuan untuk menentukan dan menemukan *hyperplane* yang terbaik yang dapat memisahkan dua buah kelas pada masukan pemisah. Prinsip dasar algoritma SVM adalah klasifikasi linier, dan selanjutnya dikembangkan agar dapat bekerja pada masalah *non-linear*, dengan memasukkan konsep *kernel trick* pada ruang kerja berdimensi tinggi.



Gambar 2.2 Kerja algoritma SVM dalam menentukan nilai *hyperplane*

Pada Gambar 2.2. menunjukan dua kondisi dan memperlihatkan konsep dasar dari algoritma SVM, dengan penyebaran data yang ditunjujan dengan warna kotak merah dan warna lingkaran kuning. Data dengan kotak berwarna merah merupakan anggota dari kelas -1 dan data dengan lingkaran berwarna kuning merupakan anggota dari kelas +1. Permasalah utama dari konsep klasifikasi adalah mencari *hyperplane* sebagai pemisah antara dua kelas atau lebih, dari kedua gambar di atas terdapat ada banyak alternatid sebagai garis pemisah (*discrimination boundaries*) antara dua kelas (Nur , 2015).

Hyperplane merupakan salah satu pemisah, dan merupakan pemisah data yang terbaik antara dua kelas, diperoleh dengan cara mengukur margin dari *hyperplane* dan mencari margin terbesar. *Margin* adalah jarak antara *hyperplane* tersebut dengan data yang terdekat dari tiap masing-masing kelas. Data yang paling dekat dengan *hyperplane* disebut sebagai *support vector*. Garis solid pada Gambar 2.2. (b), menunjukkan *hyperplane* yang terbaik, yaitu terletak tepat pada pertengahan kedua kelas yang terbentuk, sedangkan titik merah dan kuning yang berada dalam lingkaran hitam adalah *support vector*. Usaha untuk mencari titik dari lokasi *hyperplane* ini merupakan inti dari proses pemebelajaran pada SVM (Nur , 2015).

Menurut Santosa (2007) *hyperplane* klasifikasi linier SVM dinotasikan sebagai formula perumusan beriku (Bo dkk., 2009):

$$f(x) = W^T X + b \qquad (2.4)$$

sehingga menurut Vapnik dan Cortes (1995) diperoleh persamaan formula perumusan sebagai berikut:

$$[(W^T . x_i) + b] > 1 \; untuk \; y_i = +1 \qquad (2.5)$$

$$[(W^T . x_i) + b] < 1 \; untuk \; y_i = -1 \qquad (2.6)$$

Dimana:

X = himpunan data *training*,

i = 1, 2, ..., n dan,

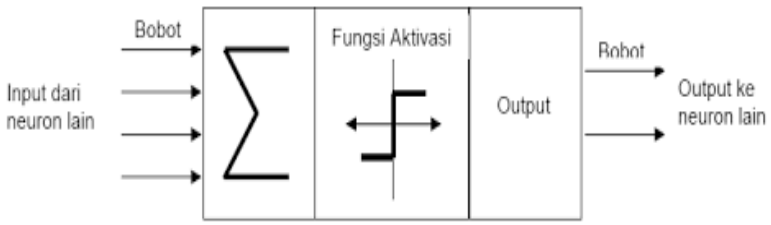
y_i = label kelas dari $\boldsymbol{x^I}$.

2.2.11 Jaringan Syaraf Tiruan

Jaringan Syaraf Tiruan (JST) merupakan suatu sistem pemrosesan informasi yang mempunyai karakteristik menyerupai jaringan syaraf biologis (JSB). Jaringan Syaraf Tiruan tercipta sebagai suatu generalisasi model matematis dari pemahaman manusia (*human cognition*) (Sudarsono, 2016). JST merupakan salah satu representasi buatan dari otak manusia yang selalu mencoba untuk mensimulasikan proses pembelajaran pada otak manusia tersebut. Istilah buatan digunakan karena jaringan syaraf ini diimplementasikan dengan menggunakan program komputer yang mampu menyelesaikan sejumlah proses perhitungan selama proses pembelajaran (Kusumaningtyas dan Asmara, 2016).

Pada proses pembelajaran, ke dalam jaringan saraf tiruan dimasukkan pola-pola masukan dan keluaran lalu jaringan akan diajari untuk memberikan jawaban yang bisa diterima. Seperti halnya otak manusia, jaringan syaraf juga terdiri dari beberapa *neuron*, dan terdapat hubungan antara *neuron-neuron* tersebut. Menunjukkan struktur *neuron* yang mana pada *Neuron-neuron* tersebut akan mentransformasikan informasi yang diterima melalui sambungan keluarannya menuju ke *neuron-neuron* yang lain. Pada jaringan syaraf hubungan ini dikenal dengan nama bobot. Informasi tersebut tersimpan pada suatu nilai tertentu pada bobot tersebut (Sudarsono, 2016).

Keduanya atau mungkin lebih untuk mendapatkan redundansi data. Pada proses ini oleh suatu fungsi perambatan yangakan menjumlahkan nilai-nilai semua bobot yang akan datang. Hasil penjumlahan ini kemudian dibandingkan dengan suatu Informasi yang disebut dengan masukkan dikirim ke *neuron* dengan bobot kedatangan tertentu. Masukkan nilai ambang (*threshold*) tertentu melalui fungsi aktivasi setiap *neuron.*

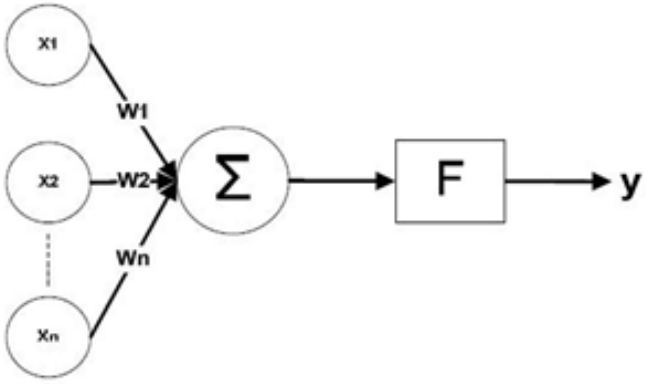


Gambar 2.3 Struktur Neuron Jaringan Saraf Tiruan

Pada jaringan syaraf, *neuron-neuron* akan dikumpulkan dalam lapisan-lapisan yang disebut dengan lapisan *neuron*. Biasanya *neuron* pada satu lapisan akan dihubungkan dengan lapisan sebelum atau sesudahnya terkecuali lapisan masukkan dan lapisan keluaran. Informasi yang diberikan pada jaringan syaraf akan dirambatkan dari lapisan ke lapisan, melalui dari lapisan masukkan sampai lapisan keluaran melalui lapisan tersembunyi. Algoritma pembelajaran menentukan informasi akan dirambatkan kearah mana, gambar kotak/bagian 3 menunjukkan *neuron* jaringan syaraf sederhana dengan fungsi aktivasi F (Sudarsono, 2016).

Pada gambar 2.4 sebuah *neuron* akan mengolah N masukkan (X_1, X_2, X_3,…, X_n) yang masing-masing memiliki bobot W_1, W_2, W_3,..., W_n dengan formula perumusan sebagai berikut.

$$Net = X_1W_1 + X_2W_2 + X_3W_3 + \cdots + X_nW_n \quad (2.7)$$



. Gambar 2.4 Model Neuron Sederhana

2.2.12 *Long Short-Term Memory* (LSTM)

LSTM adalah singkatan dari *Long Short-Term Memory* dan merupakan jenis arsitektur model jaringan saraf berulang (RNN) yang dikembangkan untuk memecahkan masalah gradien hilang dalam pelatihan jaringan saraf berulang tradisional. LSTM memiliki kemampuan untuk memeriksa hubungan jangka panjang dalam kumpulan data, menjadikannya berguna untuk tugas yang melibatkan urutan atau konteks temporal. LSTM mengatur aliran informasi selama proses pelatihan dan prediksi. Contoh dari penerapan jaringan LSTM yaitu dalam pemodelan bahasa, transkripsi ucapan-ke-teks, mesin terjemah, dan aplikasi lainnya(Sherstinsky, 2020).

Sel LSTM terdiri dari tiga gerbang utama, yaitu:

a) Gerbang Input (*Input Gate*): Mengontrol berapa banyak informasi baru yang akan masuk ke sel memori.
b) Gerbang Penghapusan (*Forget Gate*): Mengontrol sejauh mana informasi lama akan dihapus dari sel memori.
c) Gerbang Keluaran (*Output Gate*): Mengontrol berapa banyak informasi dalam sel memori yang akan diberikan sebagai output.

Berikut Perhitungan mengenai LSTM, yang ditunjukkan pada Persamaan (2.8) sampai persamaan (13);

$$f_t = \sigma\left(W_f.[h_{t-1}.x_t] + b_f\right) \quad (2.8)$$

Dalam rumus tersebut:

f_t = Vektor *gate* lupa pada langkah waktu t.

σ = Fungsi sigmoid yang mengaktifkan nilai pada rentang antara 0 dan 1.

W_f = Matriks bobot yang menghubungkan vektor input h_{t-1} dan x_t dengan vektor gate lupa f_t.

h_{t-1} = Vektor representasi state tersembunyi pada langkah waktu sebelumnya (t-1).

x_t = Vektor input pada langkah waktu t.

b_f = vektor bias.

Rumus tersebut menggabungkan vektor input h_{t-1} dan x_t melalui perkalian matriks dan menjalankannya melalui fungsi sigmoid untuk menghasilkan vektor gate lupa f_t. Nilai f_t yang mendekati 1 menunjukkan bahwa informasi dari langkah waktu sebelumnya harus diingat sepenuhnya, sedangkan nilai f_t yang mendekati 0 menunjukkan bahwa informasi tersebut harus dilupakan. Fungsi, Rumus ini merupakan salah satu dari beberapa rumus yang digunakan dalam LSTM untuk mengontrol aliran informasi dan mengatur penggunaan memorinya.

Gerbang masukan *(input gate)* adalah untuk menentukan seberapa besar arus informasi yang ditambahkan ke arus informasi tersebut. Perhitungannya untuk menghitung gate input (*input gate*) dan *state cell* (*sel state*) dalam jaringan LSTM, ditunjukkan pada Persamaan (2.9) dan (2.10):

$$i_t = \sigma(W_i.[h_{t-1}.x_t] + b_i) \quad (2.9)$$

Dalam rumus tersebut:

i_t = Vektor *gate input* pada langkah waktu t.

Σ = Fungsi sigmoid yang mengaktifkan nilai pada rentang antara 0 dan 1.

W_i = Matriks bobot yang menghubungkan vektor input h_{t-1} dan x_t) dengan vektor gate input i_t.

h_{t-1}= Vektor representasi state tersembunyi pada langkah waktu sebelumnya (t-1).

x_t = Vektor input pada langkah waktu t.

b_i = Vektor bias.

Rumus tersebut menggabungkan vektor input (h_(t-1) dan x_t) melalui perkalian matriks dan menjalankannya melalui fungsi sigmoid untuk menghasilkan vektor gate input i_t. Nilai i_t yang mendekati 1 menunjukkan bahwa seberapa banyak informasi baru harus diintegrasikan ke dalam sel LSTM pada langkah waktu t.

Selanjutnya, rumus menghitung *vektor state cell* (sel state) pada langkah waktu t. Pada rumus ini:

$$C_t = \tanh\left(W_c.\left[h_{t-1,}x_t\right] + b_c\right) \quad (2.10)$$

Dimana :

C_t = Vektor *state cell* pada langkah waktu t.

tanh = Fungsi tangen hiperbolik yang menghasilkan nilai pada rentang antara -1 dan 1.

W_c = Matriks bobot yang menghubungkan vektor input $h_{t-1,}$ dan x_t dengan vektor state cell C_t .

b_c = vektor bias.

Rumus tersebut menggabungkan vektor input $h_{t-1,}$dan x_t melalui perkalian matriks dan menjalankannya melalui fungsi tangen hiperbolik (tanh) untuk menghasilkan vektor state cell C_t. Nilai $C_t$menggambarkan informasi yang disimpan dalam sel LSTM pada langkah waktu t.

setelah informasi melewati gerbang *input (input gate)* dan gerbang lupa *(forget gate)*, LSTM memperbarui status sel untuk menghitung keluaran sel LSTM saat ini dan mentransfernya ke sel LSTM berikutnya. Rumus ini memperhitungkan

gate forget (f_t), *state cell* sebelumnya (C_{t-1}), *gate input* (i_t), dan *state cell* saat ini (C_t). Perhitungannya ditunjukkan pada Persamaan (2.11):

$$C_t = f_t * C_{t-1} + i_t * C_t \quad (2.11)$$

Dalam rumus tersebut:

C_t = vektor *state cell* pada langkah waktu t.

f_t = *vektor gate forget* pada langkah waktu t.

C_{t-1} = Vektor state cell pada langkah waktu sebelumnya (t-1).

i_t = Vektor gate input pada langkah waktu t.

Gerbang keluaran *(output gate)* menggabungkan masukan saat ini dan status sel untuk menentukan keluaran dari sel LSTM saat ini. Perhitungannya ditunjukkan pada Persamaan (2.12) dan (2.13):

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (2.12)$$

Rumus ini menghitung *vektor gate output* (o_t) pada langkah waktu t. Vektor ini mengontrol seberapa banyak informasi dalam state cell (C_t) akan dipaparkan ke output. Dalam rumus ini, $[h_{t-1}, x_t]$ adalah konkatenasi vektor state hidden sebelumnya (h_{t-1}) dan vektor input saat ini (x_t). W_o dan b_o adalah parameter yang harus dipelajari, dan σ adalah fungsi aktivasi sigmoid yang menghasilkan nilai antara 0 dan 1.

$$h_t = o_t * tanh(C_t) \quad (2.13)$$

Rumus ini menghitung vektor state hidden (h_t) pada langkah waktu t. Vektor ini merupakan *output* aktual dari jaringan LSTM pada langkah waktu tersebut. Dalam rumus ini, o_t adalah vektor gate output yang mengendalikan seberapa banyak informasi dari state cell akan dipaparkan ke output. tanh(C_t) adalah fungsi tangen hiperbolik yang memperoleh nilai antara -1 dan 1 dari vektor state cell (C_t).

Dengan mengalikan *gate output* (o_t) dengan fungsi tangen hiperbolik dari *state cell* (C_t).), rumus ini menghasilkan output yang terkendali dan terkait dengan informasi dalam *state cell*, dan itulah mengapa *state cell* digunakan sebagai input ke fungsi tangen hiperbolik dalam langkah ini(B dkk., 2018).

2.2.13 Evaluasi Model Sistem

Tahap evaluasi menjadi salah satu tahap penting dalam membangun sebuah program sistem, dengan adanya tahap evaluasi maka pengguna akan mengetahui setiap point akurasi, presisi bahkan *error* dari sebuah program sistem yang dibuatnya. Salah satu teknik untuk mengetahui hasil evaluasi dari program sentimen analisis yaitu bisa menggunakan *confusion matrix*. *Confusion matrix* merupakan sebuah metode yang dapat mempermudah program untuk dapat menentukan nilai penting dalam menimbang nilai akurasi, presisi, recall dan error dari sebuah program sistem, yang diilustrasikan pada Tabel 2.2 (Xu dkk., 2020).

Tabel 2.2 Bentuk *Confusion Matrix* 2 X 2

		Kelas Prediksi	
		Kelas Positif	Kelas Negatif
Kelas Aktual	Kelas Positif	**TP** (True Positive)	**FP** (False Positive)
	Kelas Negatif	**FN** (False Negative)	**TN** (True Negative)

Berdasarkan bentuk dari *confusion matrix* diatas, maka dengan mudah untuk dapat mengetahui nilai-nilai penting yang memudahkan peneliti untuk menghitung akurasi, recall dan presisi dari sebuah program sistem(Heydarian dkk., 2022), maka berikut formula perumusannya:

$$Akurasi = \frac{T_p+T_n}{T_p+T_n+F_p+F_n} \times 100\% \quad (2.14)$$

$$Recall = \frac{T_p}{T_p+F_p} \times 100\% \quad (2.15)$$

$$Akurasi = \frac{T_p}{T_p+F_n} \times 100\% \quad (2.16)$$

Dengan :

Tp : Jumlah kelas positif yang diklasifikasi positif.

Tn : Jumlah kelas positif yang diklasifikasi negatif.

Fp : Jumlah kelas negatif yang diklasifikasi positif.

Fn : Jumlah kelas negatif yang diklasifikasi negatif