BAB II
LANDASAN TEORI

2.1 Tinjauan Pustaka

Metode klasifikasi dapat digunakan untuk analisis komentar positif dan negatif pada trailer film (Alkaff dkk, 2020). Metode ini dilakukan dengan menggunakan klasifikasi SVM dan pembobotan Delta TF-IDF. Hasil penelitian tersebut memberikan gambaran atau informasi umum kepada calon penonton tentang film yang ditonton. Diketahui pula bahwa hasil akurasi menggunakan metode tersebut dapat mencapai 85.76%. Kelebihan dari metode tersebut terbukti memberikan kinerja yang baik dalam melakukan analisis komentar positif dan negatif untuk semua genre film dari video trailer film Indonesia. Kekurangan metode tersebut yaitu belum adanya klasifikasi komentar netral, sarkasme dan ironi.

Metode klasifikasi dapat digunakan untuk analisis *review* produk online (Fauzi, 2019). Metode ini dilakukan dengan menggunakan model Word2Vec dan klasifikasi SVM. Hasil penelitian tersebut yaitu mendapatkan informasi produk yang banyak diminati konsumen pada pembelian online. Diketahui bahwa hasil akurasi metode tersebut dapat mencapai 81%. Kelebihan dari metode tersebut yaitu memiliki kinerja yang baik dalam merepresentasikan fitur klasifikasi positif dan negatif dari sentimen *review* produk dalam bahasa Indonesia. Kekurangan metode tersebut yaitu tidak dapat menggunakan dataset yang kecil dan harus menggunakan data yang lebih besar agar dapat mempelajari representasi kata dan menempatkan kata-kata serupa ke posisi yang lebih dekat.

Metode reparasi teks dapat digunakan untuk analisis sentimen *review* berbagai macam film yang dikategorikan menjadi komentar positif atau negatif (Jing dkk, 2018). Metode ini dilakukan dengan menggunakan LDA dan pembobotan TF-IDF. Hasil Penelitian tersebut membantu pengguna untuk mendapatkan informasi dari film tersebut dan memandu pengguna membuat pilihan yang lebih baik. Diketahui bahwa hasil akurasi metode tersebut dapat mencapai 86%. Kelebihan metode tersebut memiliki struktur yang jelas dalam menggunakan beberapa kata yang sangat terkait dengan topik untuk menggambarkannya.

Kekurangan metode tersebut yaitu belum dapat melakukan multi klasifikasi sentimen.

Metode klasifikasi dapat digunakan untuk analisis berita yang diambil dari twitter secara otomatis (Negara dan Triadi, 2021). Metode ini dilakukan dengan menggunakan reparasi teks LDA. Hasil penelitian tersebut dapat mengklasifikasi teks berdasarkan topik yang digunakan untuk meringkas, mengelompokkan, dan menghubungkan atau mengolah data yang besar serta menghasilkan daftar topik yang berbobot pada setiap dokumen. Kelebihan metode tersebut dapat menganalisis dokumen yang sangat besar dan LDA ini dapat memastikan bahwa model topik yang dihasilkan pada dokumen sudah benar, baik berupa topik maupun kata-kata dalam topik. Kekurangan metode tersebut yaitu dokumen yang belum diberi label tidak terdeteksi dalam kumpulan dokumen yang sudah ada.

Metode klasifikasi dapat digunakan untuk analisis opini publik pada masa covid-19 (Nurmawiya dan Harvian, 2022). Metode ini dilakukan dengan menggunakan LDA. Hasil penelitian tersebut dapat mengidentifikasi topik-topik yang menjadi perhatian masyarakat terkait kegiatan tatap muka selama pandemi COVID-19 dan menghasilkan enam topik antara lain vaksinasi, preferensi publik, pembukaan kembali sekolah, sentimen publik, minat siswa. Kelebihan metode tersebut fokus pada representasi dokumen untuk menemukan struktur laten dari suatu topik atau konsep dalam teks. Kekurangan metode tersebut yaitu belum ada analisis yang lebih mendalam khususnya pada topik yang opininya paling banyak muncul pada masa pandemi. [2]

2.2 Layanan *Streaming*

Dalam dunia internet, layanan *streaming* merupakan sebuah teknologi yang mampu mengkompresi atau menyusutkan ukuran file audio dan video agar mudah ditransfer melalui jaringan internet (Setyawan dan Marzuki, 2018). Pentransferan file audio dan video tersebut dilakukan secara terus menerus (Humaizi & Nasution, 2021). Dari sudut pandang prosesnya, *streaming* berarti sebuah teknologi pengiriman file dari server ke *client* melalui jaringan *packet based* (Jumino & Sahnassari, 2019). *Streaming* merupakan sebuah metode untuk membuat audio, video, dan multimedia yang lain yang tersedia untuk *real-time*

pada tipe jaringan yang berbeda (Rizki dkk, 2019). Data pada file *streaming* di bagi-bagi ke dalam beberapa paket kecil yang dikirim ke sebuah aliran secara terus menerus ke perangkat end-user atau mobile phone (Diwi dkk, 2015). Hingga saat ini *streaming* masih sangat diminati di berbagai kalangan masyarakat.

2.3 Analisis Sentimen

Analisis sentimen atau biasa disebut opinion mining merupakan teknik atau cara yang dilakukan untuk mengidentifikasi opini, emosi, dan sikap (Falahah & Nur, 2015). *Opinion mining* bidang yang bertugas untuk mengekstraction opini dari teks yang tidak terstruktur menggunakan teknik *natural language processing* (NLP) dan ilmu komputer (Rozi dkk, 2012). Sebuah studi dalam *natural language processing* yang berhubungan dengan mengidentifikasi opini dari sebuah teks. Berdasarkan sumber datanya analisis sentimen dapat dibagi menjadi sentimen analisis pada level kalimat dan sentimen analisis pada level dokumen. *Opinion mining* merupakan cabang ilmu dari data mining yang biasanya digunakan untuk menganalisis data tekstual berupa sebuah opini yang mengandung polaritas sehingga nantinya menghasilkan sebuah informasi yang memiliki nilai positif dan negatif (Balazs dan Velazquez, 2016).

2.4 Text Preprocessing

Beberapa tahapan dalam melakukan data *preprocessing* yaitu *case folding*, *Tokenizing*, *filtering*, dan *stemming*. *Case folding* merupakan dokumen teks yang tidak semua dokumen teks konsisten dalam penggunaan huruf kapital. Peran *Case Folding* dibutuhkan dalam mengkonversi keseluruhan teks dalam dokumen menjadi suatu bentuk standar (biasanya huruf kecil atau lowercase) (Hafidz & Liliana, 2021). Sebagai contoh, *user* yang ingin mendapatkan informasi "Komputer" dan mengetik "Kompoter", "KomPUter", atau "komputer", tetap diberikan hasil retrieval yang sama yakni "komputer". *Case folding* adalah mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf 'a' sampai dengan 'z' yang diterima. Karakter selain huruf dihilangkan dan dianggap delimiter. *Tokenizing* merupakan pemotongan *string* input berdasarkan tiap kata yang menyusunnya (Amrizal, 2018). Secara garis besar memecah sekumpulan

karakter dalam suatu teks ke dalam satuan kata, dan bagaimana membedakan karakter-karakter tertentu yang dapat diperlakukan sebagai pemisah kata atau bukan. *Filtering* yaitu mengambil kata-kata penting dari hasil *tokenizing*. Dapat menggunakan *stoplist* (membuang kata yang kurang penting) atau *wordlist* (menyimpan kata yang penting). *Stoplist* atau *stopword* merupakan kata-kata yang seperti "yang", "dan", "di", "dari" (Riyani dkk, 2019). Teknik *Stemming* diperlukan selain untuk memperkecil jumlah *indeks* yang berbeda dari suatu dokumen, dan untuk melakukan pengelompokan kata-kata lain yang memiliki kata dasar dan arti yang serupa namun memiliki bentuk atau *form* yang berbeda karena mendapatkan imbuhan yang berbeda. Sebagai contoh kata bersama, kebersamaan, menyamai, kata dasar yang sama yaitu "sama". Teknik *Stemming* untuk mengurangi variasi dari kata yang mempunyai kata dasar yang sama (Guterres, 2019).

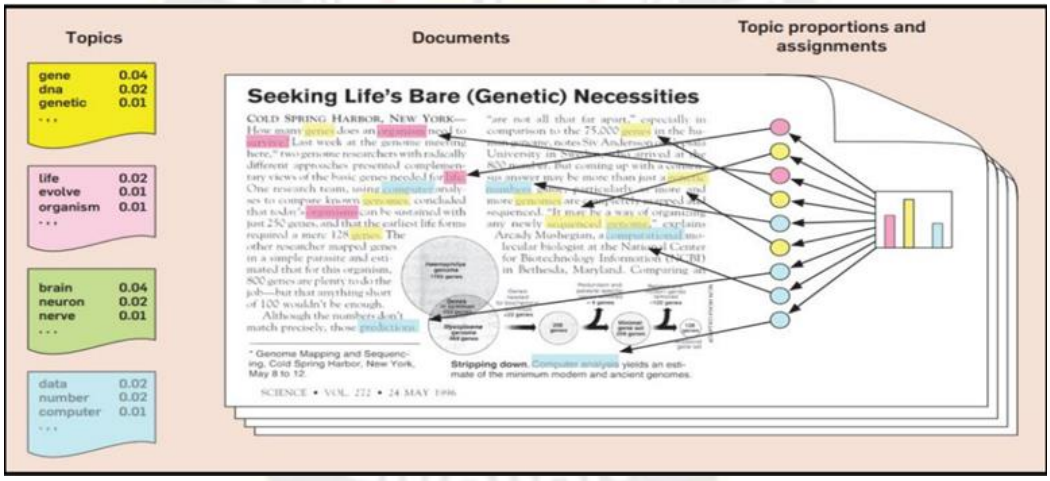
Preprocessing	Input	Output
Case Folding	Menpora Imam Nahrawi memberikan penghargaan kepada Timnas Indonesia yang telah berjuang di final Sea Games tahun ini	menpora imam nahrawi memberikan penghargaan kepada timnas indonesia yang telah berjuang di final sea games tahun ini
Tokenization	menpora imam nahrawi memberikan penghargaan kepada timnas indonesia yang telah berjuang di final sea games tahun ini	menpora, imam, nahrawi, memberikan, penghargaan, kepada, timnas, indonesia, yang, telah, berjuang, di, final, sea, games, tahun, ini
Stemming	menpora, imam, nahrawi, memberikan, penghargaan, kepada, timnas, Indonesia, yang, telah, berjuang, di, final, sea, games, tahun, ini	menpora, imam, nahrawi, beri, harga, kepada, timnas, indonesia, yang, telah, juang, di, final, sea, games, tahun, ini
Stopword Removal	menpora, imam, nahrawi, beri, harga, kepada, timnas, indonesia, yang, telah, juang, di, final, sea, games, tahun, ini	menpora, imam, nahrawi, beri, harga, timnas, indonesia, juang, final, sea, games, tahun

Gambar 1. *Text Preprocessing*

2.5 *Topic Modelling*

Pemodelan topik merupakan teknik untuk menemukan representasi dokumen berupa kata-kata kunci dari dokumen. Kata-kata kunci yang ditemukan selanjutnya digunakan untuk pengindeksan dan pencarian dokumen kembali sesuai kebutuhan dari pengguna (Suhartono, 2014). Menurut Blei pemodelan topik merupakan rangkaian algoritma yang tujuannya untuk menemukan dan memberikan keterangan pada suatu dokumen dengan informasi tematik untuk mengaitkan beberapa tema dengan entitas yang dikaitkan berdasarkan pembelajaran terpadu menggunakan tema (Blei, 2012). *Topic modeling* merupakan

metode yang digunakan untuk menemukan topik utama yang tersembunyi dari rangkaian kata dalam kumpulan dokumen yang besar dan tidak terstruktur. Metode *topic modeling* menganalisis data berdasarkan teks asli, mengenai hubungan antar topik satu sama lain, hubungan antar tema yang dapat berubah sewaktu-waktu, sehingga metode ini dapat dikembangkan untuk pencarian atau meringkas teks yang terdapat dalam dokumen. Secara umum, setiap dokumen dalam korpus memiliki proporsi topik yang dibahas. Hasil analisis pemodelan tematik digunakan untuk merangkum, memvisualisasikan, mengeksplorasi dan berteori tentang korpus. Sehingga, *topic modeling* bertujuan untuk menemukan kumpulan topik dan kata yang terkandung dalam berbagai suatu dokumen (Putri dan Kusumawardani, 2017). Adapun gambaran *topic modeling* menurut Blei (2012) dapat dilihat pada gambar 2.



Gambar 2. *Topic Modelling*

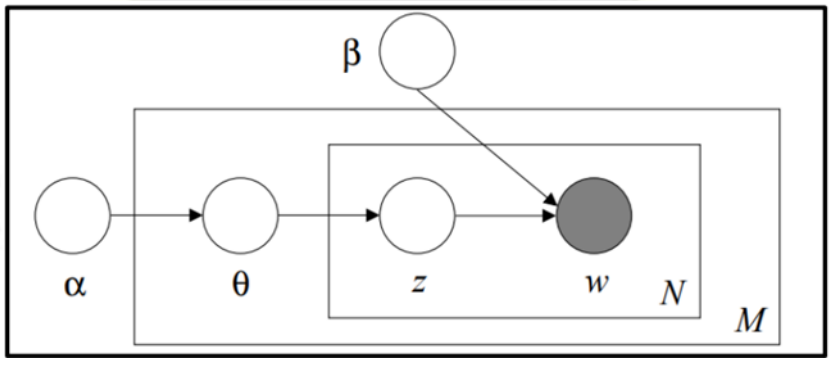
Gambar 2 dari Blei (2012) mengilustrasikan topic modeling seperti pada gambar diatas. Salah satu cara untuk memikirkan tentang bagaimana proses pemodelan topik bekerja adalah dengan membayangkan mengerjakan sebuah artikel dengan sekumpulan highlighters. Untuk menemukan kata kunci dapat menggunakan beberapa warna yang membedakan beberapa topik. Sehingga didapatkan beberapa kata yang dikelompokkan berdasarkan warna yang ditetapkan sehingga menjadi sebuah topik.

2.6 Coherence Score

Coherence score merupakan tahapan untuk menentukan jumlah model topik dalam *topic modelling.* Semakin tinggi nilai *coherence score* menunjukan bahwa model yang dihasilkan akan semakin baik. Selanjutnya jumlah topik dengan nilai *coherence score* tertinggi akan digunakan sebagai *topic modelling* (Matira dkk, 2023). Dengan cara ini, LDA yang dihasilkan dapat memberikan representasi topik yang akurat dan optimal untuk dokumen yang diberikan (Renaldi dkk, 2023).

2.7 Latent Dirichlet Allocation

LDA merupakan metode *topic modeling* dan topik analisis yang paling populer saat ini. Menurut Blei (2003), LDA merupakan model probabilistik generatif dari kumpulan tulisan yang disebut *corpus*. Ide dasar yang diusulkan metode LDA adalah setiap dokumen direpresentasikan sebagai campuran acak atas topik yang tersembunyi, yang mana setiap topik memiliki karakter yang ditentukan berdasarkan distribusi kata-kata yang terdapat di dalamnya LDA model probabilistik generative dari koleksi data diskrit seperti korpus teks. Dalam LDA, dokumen merupakan objek yang diamati, sedangkan topik, distribusi topik, penggolongan setiap kata pada topik merupakan struktur yang tersembunyi, sehingga metode ini disebut LDA (Blei, 2012). Adapun gambaran ilustrasi metode LDA secara visual menurut Blei (2003) dapat dilihat pada gambar 3.



Gambar 3. Ilustrasi Metode LDA

Gambar 3 menunjukan parameter α merupakan parameter dirichlet distribution yang berguna untuk mengontrol distribusi topik pada setiap dokumen, dimana apabila nilai α yang semakin tinggi menyatakan bahwa dokumen memuat

topik campuran atau memuat beberapa topik (Nurid dan Candra, 2021). Sedangkan nilai α yang semakin rendah menyatakan bahwa tidak terdapat topik yang tercampur satu sama lain dalam dokumen. Oleh karena itu, untuk melihat seberapa baik sebaran topik yang dihasilkan dapat dilihat dari nilai α yang semakin rendah. Parameter β merupakan parameter dirichlet distribution yang digunakan untuk mengontrol distribusi kata pada setiap topik, dimana apabila nilai β yang semakin tinggi menyatakan bahwa topik memuat kata-kata yang terdapat pada topik lainnya. Sedangkan nilai β yang rendah menyatakan bahwa sebuah topik terdiri dari kata-kata yang lebih spesifik. Variabel θ merupakan distribusi multinomial atau distribusi topik untuk dokumen tertentu, yang menyatakan probabilitas dokumen tertentu. Apabila semakin tinggi nilai θ maka menyatakan bahwa semakin banyak topik yang termuat dalam dokumen, sedangkan semakin rendah nilai θ maka menyatakan bahwa dokumen semakin spesifik pada topik tertentu. Variabel Z merepresentasikan topik untuk kata tertentu dari suatu dokumen. Variabel W merepresentasikan kata spesifik yang berkaitan dengan topik tertentu dalam suatu dokumen. Bentuk lingkaran menyatakan bahwa kata bersifat individual. Variabel W digambarkan dengan lingkaran yang berwarna abu-abu merepresentasikan variabel yang diobservasi, sedangkan lingkaran yang berwarna putih merepresentasikan variabel laten atau hidden variabel yang secara tidak langsung diobservasi (Blei, 2003).

LDA mempresentasikan topik dengan probabilitas kata, dimana kata yang memiliki probabilitas yang paling tinggi di setiap topiknya mempresentasikan topik tersebut. Metode ini banyak digunakan secara luas pada pembelajaran mesin, pengolahan data, maupun pengolahan citra. LDA bekerja berdasarkan analisis matriks penyebaran yang bertujuan menemukan suatu proyeksi optimal sehingga dapat memproyeksikan data input pada ruang dengan dimensi yang lebih kecil yang semua pola dapat dipisahkan semaksimal mungkin. Tujuan pemisahan tersebut, maka LDA mencoba memaksimalkan penyebaran data input di antara kelas-kelas yang berbeda dan meminimalkan penyebaran input pada kelas yang sama. Adapun algoritma inferensi menggunakan *variational bayes* dari metode LDA sebagai berikut:

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{d,n}} p(z_{d,n}|\theta_d) p(w_{d,n}|z_{d,n},\beta) \right) d\theta_d \quad (1)$$

Dengan:

p : Nilai probabilitas

M : Banyaknya dokumen

n : Banyaknya kata dalam suatu dokumen

α : Nilai distribusi topik pada setiap dokumen

β : Nilai distribusi kata pada setiap topik

θ_d : Probabilitas distribusi topik dalam suatu dokumen

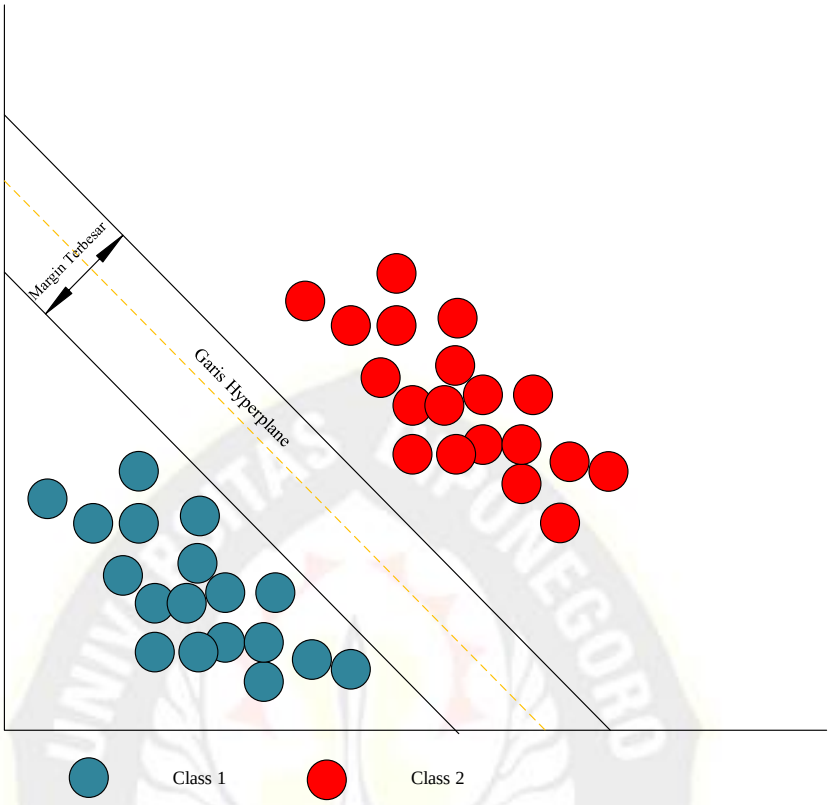
$z_{d,n}$: Identitas topik untuk kata tertentu dari suatu dokumen

$w_{d,n}$: Identitas kata tertentu dalam suatu dokumen

Metode LDA memaksimalkan diskriminasi antar kelas dan meminimalkan persebaran dalam kelas. Ekstraksi *feature* digunakan dalam LDA yang kemudian diekstrak datanya untuk diolah dalam perhitungan *training* maupun *testing*.

2.8 *Support Vector Machine*

SVM adalah metode *supervised* yang mencoba untuk mencari fungsi optimal yang digunakan untuk pemisah antar kelas (*hyperplane)* dua kelas data yang berbeda dan menghasilkan *model* yang terbaik untuk data di masa depan (Ritonga & Purwaningsih, 2018). SVM mencoba mencari nilai maksimum margin antara dua kelas data dari sampel data yang diujikan dan SVM sudah banyak diterapkan dalam studi kasus pengenalan pola dan pengkategorian teks dan menghasilkan akurasi yang tinggi. Algoritma SVM ini bekerja menggunakan pemetaan *nonlinier* untuk mengubah data pelatihan asli menjadi dimensi yang lebih tinggi. Dalam konsep ini, SVM berusaha untuk mencari *hyperplane* terbaik diantara fungsi yang tidak terbatas jumlahnya. SVM mempunyai kemampuan generalisasi yang tinggi serta memiliki konsep formulasi yang jelas dan sedikit parameter yang harus diatur (Hermanto dkk, 2020). Adapun gambar permisah antar kelas dapat dilihat pada gambar 4.



Gambar 4. Garis pemisah antar kelas atau hyperplane di metode SVM

Pada awalnya SVM digunakan untuk memisahkan data linear kemudian SVM berkembang dan dapat bekerja pada data *nonlinear*. Untuk mengklasifikasikan data *nonlinear* SVM menggunakan model kernel *linear* yang menghasilkan akurasi yang baik dalam melakukan klasifikasi teks. Adapun persamaan dari kernel *linear* sebagai berikut:

$$K(x_i, x_j) = x_i \,.\, x_j \qquad (4)$$

Dengan:

$K(x_i, x_j)$	:	Fungsi kernel
x_i	:	Data ke-*i*
x_j	:	Data ke-*j*

Pada klasifikasi SVM memiliki parameter yaitu parameter C yang berfungsi untuk mengontrol optimasi antara margin dan kesalahan klasifikasi (Fitriyah dkk, 2020). Dalam SVM, tujuan utama adalah untuk menemukan

hyperplane yang memaksimalkan margin antara kelas yang berbeda. Semakin besar nilai C maka memberikan penalti yang besar terhadap kesalahan dalam klasifikasi. Klasifikasi menggunakan nilai yang dihasilkan dari ekstraksi *feature* metode LDA sehingga data dapat diklasifikasi. Pada tahap ini klasifikasi menggunakan *library* SVM pada *scikit-learn*. *Scikit-learn* adalah *library* bahasa pemrograman *python* yang umum digunakan untuk membuat model *machine learning* baik itu *supervised* maupun *unsupervised learning*. *Scikit-learn* mudah digunakan dan cukup efisien untuk melakukan klasifikasi SVM. Selain itu, *scikit-learn* saling terintegrasi dengan *library scikit* lainnya seperti *pandas* yang digunakan pada data handling ataupun matplotlib untuk melakukan visualisasi data.

2.9 Evaluasi Performa

Pengujian performa metode dapat dilakukan dengan menggunakan *confusion matrix* untuk mengevaluasi performa dengan hasil akurasi berdasarkan dataset yang diuji. Berikut adalah tabel *confusion matrix* yang dapat dilihat pada tabel 1.

Tabel 1. *Confusion matrix*

		Prediksi	
		Positif	Negatif
Aktual	Positif	TP	FN
	Negatif	FP	TN

Dengan:

True Positive (TP) : Kelas yang digolongkan positif dan faktanya positif

True Negative (TN) : Kelas yang digolongkan negatif dan faktanya negatif

False Positive (FP) : Kelas yang digolongkan positif dan faktanya negatif

False Negative (FN) : Kelas yang digolongkan negatif dan faktanya positif

Accuracy yaitu total prediksi data yang benar dengan jumlah data keseluruhan dapat didefinisikan dengan persamaan berikut:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (5)$$

Precision yaitu prediksi positif dibanding dengan total data yang diklasifikasi positif, dengan persamaan berikut:

$$Precision = \frac{TP}{(TP + FP)} \quad (6)$$

Recall yaitu prediksi positif dibanding dengan total data yang bernilai benar, dengan persamaan berikut:

$$Recall = \frac{TP}{(TP + FN)} \quad (7)$$

F1 Score yaitu kombinasi dari *recall* dan *precision*, dengan persamaan berikut:

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (8)$$