BAB II
TINJAUAN PUSTAKA

2.1. Kajian Pustaka

Dalam melakukan sebuah penelitian tentu akan lebih efektif jika penelitian tersebut didukung oleh hasil penelitian sebelumnya. Berikut merupakan beberapa penelitian sebelumnya yang berguna sebagai pendukung untuk penelitian ini:

Tabel 2.1 Kajian Pustaka

No.	Judul	Comparing	Constracting	Critize	Synthesize	Summarize
1	Analisis Sentimen Masyarakat Terhadap Vaksin Covid-19 di Twitter Menggunakan Metode *Random Forest Classifier* (Studi Kasus: Vaksin *Sinovac*)[11]	Persamaan dari penelitian ini adalah analisis sentimen menggunakan metode *Random Forest Classifier*	Perbedaan terletak pada objek dan subjek penelitian, model evaluasi, metode pengambilan data, dan metode transformasi teks	Penelitian ini dibuat sebuah metode pembelajaran mesin untuk menganalisis opini publik pada program Vaksin Sinovac. Penelitian ini menggunakan dua klasifikasi dan pemrosesan	Penelitian ini menggunakan 1500 data yang dibagi menjadi dua kategori yaitu *sentiment* positif dan negatif.	Analisis sensitivitas vaksin Sinovac mendapatkan prediksi sebanyak 13% klasifikasi komentar negatif dan 87% klasifikasi komentar positif.

No.	Judul	Comparing	Constracting	Critize	Synthesize	Summarize
				data dilakukan secara manual		
2	Analisis Sentimen berbasis Aspek terhadap Ulasan Hotel Tentrem Yogkarta menggunakan Algoritma *Random Forest Classifier*[12]	Persamaan dari penelitian ini adalah analisis sentimen memakai metode *Random Forest Classifier*	Perbedaan terletak dalam objek penelitian, metode transformasi teks dan metode pengambilan data.	Pengklasifikasian dilakukan dengan menggunakan algoritma klasifikasi *Random Forest Classifier* dan dengan pembobotan kata (TF-IDF)	Pengujian dilakukan berdasarkan skenario dari parameter jumlah *Tree* dan dalam Tree. Skenario jumlah *Tree* yang dilakukan pada penelitian ini yaitu sebanyak 100, 200 dan 300 *Tree* dengan kedalaman *Tree* sebanyak 5, 7, dan 10	Penerapan algoritma *Random Forest Classifier* Sistem prediksi mendapatkan hasil nilai yang terbaik dengan nilai rata-rata *accruracy* dan skor f1 yang sama yaitu sebanyak 90%
3	Analisis Sentimen Objek Wisata Di Provinsi Sulawesi Selatan Berdasarkan Ulasan Pengunjung Menggunakan Metode *Random*	Persamaan dari penelitian ini yaitu analisis sentimen dengan menggunakan metode *Random Forest Classifier*	Perbedaan terletak pada objek penelitian, metode transformasi teks, dan metode	Penggunaan *Trip Advisor* sebuah situs perjalanan yang membuat pengunjung membuat ulasan tempat, ulasan dari pengunjung ini digunakan	Tujuan dari penelitian ini adalah untuk mendapatkan informasi dari analisis sentimen melalui *rating* pengunjung tempat	Kajian yang mengklasifikasikan ulasan wisatawan Sulawesi Selatan menghasilkan akurasi sebesar 82%, akurasi sebesar 86%, dan nilai pengenalan

No.	Judul	Comparing	Constracting	Critize	Synthesize	Summarize
	Forest Classifier[13]		pengambilan data.	untuk sumber data pada penelitian.	wisata di Sulawesi Selatan.	sebesar 86% yang cukup untuk digunakan dalam sistem.
4	Analisis Sentimen Pengguna Twitter Terhadap Pembayaran *Cash* Menggunakan Shopeepay Dengan Algoritma *Random Forest*[9]	Persamaan dari penelitian ini yaitu analisis sentimen dengan menggunakan metode *Random Forest Classifier*	Perbedaan terletak pada objek penelitian, metode transformasi teks, dan metode pengambilan data.	Banyak pengguna *Shopeepay* di Indonesia membawa banyak opini di platform tersebut, termasuk situs *microblogging Twitter.platform shopeepay*	Dalam penelitian ini, algoritma klasifikasi *Random Forest* digunakan untuk mengklasifikasikan pendapat pengguna *Twitter* di *platform shopeepay* dengan pembobotan kata TF-IDF	Dari kedalaman 300 *Tree* dan 55 *Tree* mendapatkan skenario parameter terbaik *precision* 95%, *recall* 94%, *F1-Score* 95% dan *accuracy* 95%.
5	Analisis Sentimen Pengguna Twitter terhadap Vaksinasi COVID-19 di Indonesia menggunakan Algoritma *Random Forest* dan BERT[14]	Persamaan dari penelitian ini yaitu analisis sentimen dengan menggunakan metode *Random Forest Classifier*	Perbedaan terletak dalam objek penelitian, metode pengambilan data dan metode transformasi teks	Analisis suasana hati diperlukan untuk mengetahui animo orang terhadap program vaksinasi Covid-19 yang telah di keluarkan oleh pemerintah.	Tujuan penelitian ini yaitu menganalisis sentimen pada *user Twitter* di Indonesia tentang vaksinasi Covid-19 menggunakan *Random Forest Algorithm* dan *Transformer*	Klasifikasi *Random Forest* menggunakan data asli memberikan akurasi 81%, akurasi 82%, *recall* 70%, skor F1 70%, dan skor 3761. Memori 74%, hasil F1

No.	Judul	Comparing	Constracting	Critize	Synthesize	Summarize
				Analisis sentimen biasanya dilakukan untuk mendapatkan informasi terbaru dari korpus besar.	*Bidirectional Encoder Representation* (BERT).	74%, dukungan 3760.
6	Analisis Sentimen Terhadap Aplikasi Ruangguru Menggunakan Algoritma Naive Bayes, *Random Forest* Dan *Support Vector Machine*[15]	Persamaan dari penelitian ini yaitu analisis sentimen dengan menggunakan metode *Random Forest Classifier*	Perbedaan terletak pada objek dan subjek penelitian, dan metode pengujian klasifikasi.	Peringkat pengguna aplikasi dapat membantu dalam mengembangkan peningkatan kualitas pada aplikasi dan dapat menjadi cara untuk menilai apakah pengguna mendapatkan kepuasan dalam penggunaannya	Di dalam penelitian ini, dilakukan *sentiment analysis* pada aplikasi RuangGuru dengan melakukan pengujian pada model klasifikasi yang dipakai seperti *Naive Bayes*, *Random Forest*, dan *Support Vector Machines.*	Studi ini menemukan bahwa 97,16% model klasifikasi hutan acak dan model klasifikasi *Random Forest* adalah model klasifikasi hutan acak yang berkinerja terbaik.
7	Analisis Sentimen Ulasan Aplikasi Dana dengan	Persamaan dari penelitian ini yaitu analisis sentimen	Perbedaan terletak dalam objek	Karena banyaknya pengguna di	Dengan membagi data menjadi 250 titik data per kelas,	Berdasarkan hasil pengujian yang dilakukan dan

No.	Judul	Comparing	Constracting	Critize	Synthesize	Summarize
	Metode *Random Forest*[8]	dengan menggunakan metode *Random Forest Classifier*	penelitian, metode pengambilan data, metode pengujian klasifkasi, dan metode transformasi teks.	aplikasi Dana, sering kali ada ulasan positif, negatif, dan netral ini tidak ada hubungannya dengan *rating* pengguna di *Play Store*.	pengujian dijalankan pada 1354 titik data berdasarkan jumlah pohon dan kedalaman pohon.	analisis perbandingan data 80% data pelatihan dan pengujian 20% mencapai akurasi 84%, ingatan 84%, skor dan akurasi F1 84% Peningkatan 84% untuk kedalaman pohon 65, total 400 pohon.
8	Analisis Sentimen Ulasan Aplikasi Peduli Lindungi dengan Metode Random Forest[16]	Persamaan dari penelitian ini yaitu melakukan analisis sentimen menggunakan metode *Random Forest Classifier*.	Perbedaan terletak pada objek penelitian, metode pengambilan data dan metode transformasi teks.	Data tinjauan aplikasi Peduli Lindungi menyediakan sumber data yang dapat dianalisis dan juga digunakan untuk klasifikasi suasana hati.	Data verifikasi aplikasi Peduli Lindungi diambil menggunakan teknik pengikisan yang memungkinkannya mengekstrak data dari halaman *web* dalam proses ini.	Hasil penelitian dengan kedalaman 65 pohon dan 400 pohon mendapatkan nilai terbaik. Yaitu, 71% presisi, 71% recall, dan 71% skor F1. Akurasi 72% pada data latih 90% dan rasio data uji 10%.
9	Pendeteksian Sarkasme pada	Persamaan dari penelitian ini yaitu	Perbedaan terletak	Penelitian ini bertujuan untuk	Pada proses pengolahan	Hasil penelitian ini menunjukkan

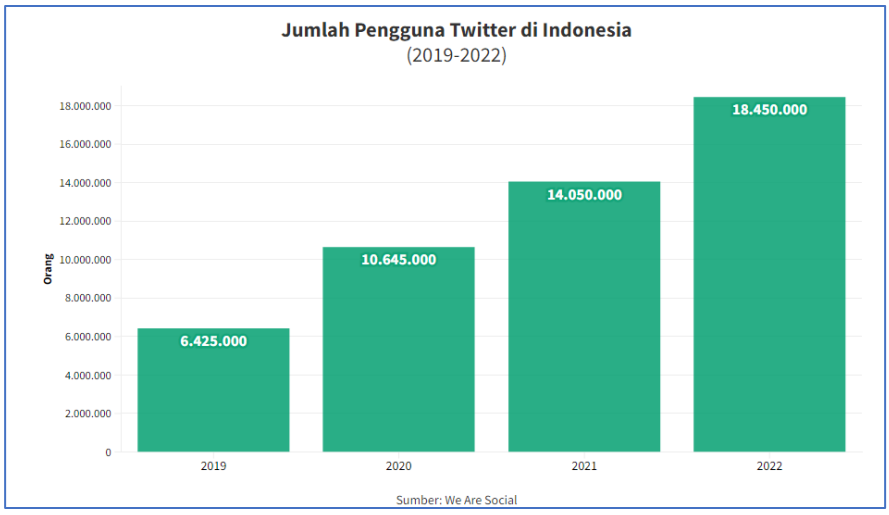
No.	Judul	Comparing	Constracting	Critize	Synthesize	Summarize
	Proses Analisis Sentimen Menggunakan *Random Forest Classifier*[17]	melakukan analisis sentimen menggunakan metode *Random Forest Classifier*	dalam objek dan subjek penelitian, dan metode pengujian klasifkasi.	melakukan penggabungan *sentiment analysis* dan deteksi sarkasme untuk menentukan peringkat opini pada pengguna media sosial *Twitter*.	sentimen dilakukan melalui tahap *preprocessing text* dan ekstraksi fitur, lalu diklasifikasikan dengan menggunakan metode algoritma metode *random forest classifier* dan *support vector machine.*	peningkatan rata-rata skor presisi 16,61%, skor presisi 5,45%, skor *recall* 9,64%, dan skor skor F1 11,27% dengan total 2.027 data. Dari total 587 data didapatkan 462 data berlabel netral.
10	*Sentiment Analysis Using Random Forest Algorithm-Online Social Media Based*[18]	Persamaan dari penelitian ini yaitu melakukan analisis sentimen menggunakan metode *Random Forest Classifier*	Perbedaan terletak pada objek penelitian, metode pengambilan data, metode transformasi teks.	Analisis sentimen area *Natural Language Processing* (NLP) yang membangun sistem untuk mengenali dan mengekstrak opini dalam teks.	Penelitian ini menerapkan metode klasifikasi algoritma *random forest* untuk melakukan *sentiment analysis* pada sumber data *Twitter* untuk mengukur hasil skoring dari algoritma yang digunakan	Akurasi pengukuran pada penelitian ini sekitar 75%. Modelnya cukup bagus. Namun, menyarankan untuk mencoba algoritma lain dengan penyelidikan lebih lanjut.

Seperti yang terlihat dari tabel 2.1 terdapat persamaan metode dari kesepuluh penelitian terdahulu. Dari sepuluh penelitian diatas bisa dilihat bahwa algoritma klasifikasi *Random Forest* sebagai metode penelitian yang sering digunakan untuk melakukan sentimen analisis dengan persentase akurasi yang besar. Hal tersebut membuat penulis untuk melakukan penelitian analisis sentimen dengan menggunakan metode klasifikasi *Random Forest*, perbedaan dari penelitian yang dilakukan sebelumnya dengan yang dilakukan oleh penulis terletak pada subjek penelitian, objek penelitian dan metode transformasi teks.

2.2. Dasar Teori

2.2.1. *Twitter*

Twitter pertama kali digagas oleh seorang mahasiswa sarjana dari Universitas New York yang bernama Jack Dorsey saat melakukan diskusi di sebuah acara yang dibuat oleh perusahaan *podcast* yang bernama Odeo. Kicauan (*tweet)* merupakan fitur yang ada di *Twitter* yang dapat dipakai oleh pengguna untuk membagikan tulisan, foto, video, dan gif. Pada umumnya, *tweet* yang di unggah dapat dilihat oleh semua pengguna, namun pengguna juga dapat mengatur kiriman *tweet* hanya bagi pengikut tertentu saja[19]. Berdasarkan hasil laporan *We Are Social*, terdapat jumlah pengguna *Twitter* di Indonesia mencapai 18, 45 juta di tahun 2022.



Gambar 2.1 Jumlah Pengguna *Twitter* Di Indonesia Tahun 2019-2022

Dengan jumlah tersebut Indonesia menempati peringkat kelima dengan salah satu negara pengguna *Twitter* terbanyak di dunia[20]. Melalui fitur *tweet*, para pengguna *Twitter* dapat berinteraksi dengan pengguna *Twitter* lainnya dengan mengirimkan berbagai hal yang mereka pikirkan, berita terbaru, kejadian yang sedang berlangsung, dan topik-topik lainnya. Adapun beberapa fitur yang dapat digunakan oleh pengguna *Twitter* sebagai berikut[21]:

1. *Followers dan Following*

Pengikut (*followers*) merupakan akun atau individu yang mengikuti akun lainnya, sementara yang diikuti (*following*) merupakan akun atau individu yang diikuti oleh akun tersebut. *Twitter* juga dapat digunakan sebagai tempat berbagi informasi dengan pengikutnya. Pengguna *Twitter* yang mengikuti akun tersebut dapat menerima pembaruan dari akun tersebut, yang kemudian ditampilkan di halaman utama *Twitter* mereka.

2. *Direct Message*

   Direct Message juga dapat dikirim ke pengikut akun di *Twitter*. Ini pada dasarnya adalah program email terintegrasi *Twitter*. Bahkan jika pengguna Twitter tidak mengikuti akun tersebut, mereka tetap dapat berkomunikasi secara pribadi.

3. *Twitter Search*

   Salah satu fitur paling menarik dari *Twitter* adalah kemampuan pengguna untuk mencari orang tertentu, kata kunci, topik, dan lokasi.

4. *Trending Topics*

   Fitur menarik lainnya dari *Twitter* adalah tren topik. Tren topik terdiri dari sepuluh topik teratas yang sering dibicarakan oleh pengguna di *Twitter* untuk waktu tertentu. Topik-topik tersebut dapat berupa berita, olahraga, hingga hiburan.

5. *Tweets*

   Fitur ini dapat digunakan untuk mengirim pesan, gambar, gif, *polling*, atau lokasi yang ingin di *publish* atau dibagikan kepada orang lain.

6. *Profile* dan *Settings*

   Fitur ini memungkinkan pengguna untuk mengubah informasi pribadi dan melakukan perubahan terkait keamanan dan privasi akun.

2.2.2. *Twitter API (Application Proramming Interface)*

Twitter API atau disebut dengan *Application Programming Interface* adalah sebuah fungsi yang digunakan untuk mengakses aplikasi perangkat lunak berbasis *web* atau *web tool*. *Twitter* menyediakan *Twitter* API untuk memungkinkan pengembang dari pihak ketiga untuk membuat program yang

terintegrasi dengan layanan *Twitter*. Pengembang dapat menggunakan API *Twitter* untuk membuat perangkat lunak, situs web, dan konten lain yang dapat berinteraksi dengan *Twitter*[22]. Dikutip dari pusat bantuan *Twitter*, *Platform* API menyediakan akses yang luas terhadap data *Twitter* publik yang telah dipilih oleh pengguna untuk dibagikan ke publik. *Twitter* juga mendukung API yang memungkinkan para pengguna untuk mengelola informasi yang ada pada *Twitter* mereka yang non-publik dan memberikan informasi tersebut pada pengembang yang telah diizinkan pengguna untuk melakukannya.

2.2.3. *Sentiment Analysis*

Sentiment Analysis atau yang juga dikenal sebagai istilah opini *mining* adalah sebuah proses otomatis untuk memahami suatu data, mengekstrak, dan mengolah data yang berbentuk *text* dengan tujuan untuk mendapatkan informasi mengenai *sentiment* atau pendapat yang terkandung pada suatu kalimat opini. Sentimen analisis dilakukan untuk melihat kecenderungan pendapat seseorang terhadap suatu objek atau masalah, baik itu bersifat pendapat negatif atau pendapat positif[23].

Analisis sentimen bisa dibedakan dari sumber datanya, beberapa level yang sering digunakan dalam penelitian analisis sentimen adalah analisis sentimen dengan level dokumen dan analisis sentimen dengan level kalimat. Berdasarkan level sumber datanya sentimen analisis terbagi menjadi dua kelompok yaitu[24]:

1. *Sentiment Analysis Coarse-grained*

   Sentimen analisis yang dilakukan pada level dokumen melihat seluruh isi dokumen sebagai sebuah kesatuan, dan mencoba untuk menentukan apakah sentimen keseluruhan dalam dokumen tersebut adalah positif, negatif, atau netral. Analisis sentimen pada level dokumen ini berguna untuk memahami sentimen umum yang terkandung dalam sebuah dokumen, seperti artikel berita, laporan penelitian, atau *review* produk. Dalam analisis sentimen level dokumen, dokumen diperlakukan sebagai

satu kesatuan, sehingga tidak memperhatikan sentimen pada kalimat atau kata-kata yang mungkin berbeda dalam dokumen tersebut.

2. *Fined-grained Sentiment Analysis*

   Fined-grained Sentiment Analysis adalah analisis sentimen yang terkandung pada level kalimat, yang di mana fokus utamanya adalah menentukan sentimen yang terdapat pada setiap kalimat secara terpisah. Dalam analisis ini, setiap kalimat dianggap sebagai unit analisis yang terpisah, dan sentimen yang ditemukan dapat berbeda antara satu kalimat dengan kalimat lain dalam dokumen yang sama. Tujuan dari analisis ini yaitu untuk menghasilkan pemahaman yang lebih rinci dan terperinci tentang sentimen yang terkandung dalam sebuah dokumen.

2.2.4. *Text Mining*

Dalam buku yang berjudul "*Text Mining Handbook*: *Advanced Approaches in Analyzing Unstructured Data*[22], Text *mining* dapat diartikan sebagai proses pengambilan informasi dari sekelompok dokumen secara berkala yang memungkinkan pengguna untuk berinteraksi dengan menggunakan berbagai jenis alat analisis. Tujuan dari *text mining* yaitu untuk menemukan kata-kata yang merepresentasikan isi dokumen agar dapat dilakukan analisis hubungan antara dokumen-dokumen lainnya.

Secara prinsip proses dari *text mining* banyak mengadopsi penelitian tentang *data mining*, tetapi perbedaannya terletak pada pola yang digunakan. Pola yang digunakan dalam *text mining* diperoleh dari kumpulan bahasa alami yang tidak terstruktur, sedangkan dalam *data mining* pola diperoleh dari *database* yang terstruktur[25].

2.2.5. *Pre-Processing*

Pre-processing adalah tahap awal dalam analisis data di mana data mentah atau tidak terstruktur dipersiapkan dan diolah agar dapat dijalani proses analisis lebih lanjut dengan lebih baik. Kegunaan dari *pre-processing* yaitu untuk melakukan proses pembersihan data yang memiliki karakter yang tidak diperlukan untuk melakukan suatu proses pengolahan data. Ada

beberapa tahapan dalam melakukan *pre-prosesing* seperti *case folding, filtering, stopword removal, stemming* dan lainnya[26]. *Preprocessing* merupakan *text mining* yang diharapkan dapat untuk menghilangkan atau mengurangi kata-kata yang tidak memiliki arti di dalam suatu dokumen. *Preprocessing tex*t pada umumnya adalah untuk mengubah informasi dari masing-masing sumber data yang asli ke dalam suatu format kata dasar dengan menerapkan berbagai jenis metode-metode ekstraksi[27].

2.2.6. *Random Forest*

Random Forest merupakan salah satu metode klasifikasi yang berbasis komputasi yang diperkenalkan oleh Leo Breiman tahun 2001, *Random Forest* juga merupakan pengembangan dari metode CART (*Classification And Regression Tree*) sehingga *Random Forest* memiliki beberapa kelebihan yang tidak dimiliki metode CART tersebut[28]. Proses *bagging* dan *resampling bootstrap* digunakan untuk membuat beberapa versi pohon klasifikasi dengan menggunakan sampel data yang berbeda dan menggabungkannya untuk menghasilkan prediksi akhir. Namun pada algoritma *Random Forest*, proses pengacakan tidak hanya digunakan pada sampel data, tetapi juga pada variabel prediktor yang digunakan untuk membentuk pola pohon klasifikasi. Hal tersebut menghasilkan kumpulan pohon klasifikasi dengan ukuran dan bentuk yang berbeda-beda. Tujuannya adalah untuk menghasilkan kumpulan pohon klasifikasi yang saling memiliki korelasi yang kecil, karena hal ini dapat mengurangi kesalahan prediksi pada hasil akhir dari *Random Forest*[28]. Ada tiga poin utama pada *Random Forest*, (1) melakukan *bootstrap sampling* untuk membangun pohon prediksi; (2) masing-masing pohon keputusan memprediksi dengan prediktor acak; (3) lalu mengombinasikan hasil dari setiap pohon keputusan dengan cara *majoriy vote* untuk klasifikasi atau rata-rata untuk regresi[6].

2.2.7. *RapidMiner*

RapidMiner adalah *software* komputer *opensource* dan digunakan untuk melakukan analisa *data mining, text mining*, dan dapat melakukan

prediksi. *RapidMiner* menyediakan berbagai teknik deskriptif dan prediksi yang dapat memberikan wawasan untuk pengguna dan membantu dalam membuat keputusan terbaik. Perangkat lunak ini dilengkapi dengan GUI (*Graphical User Interface*) yang memudahkan pengguna merancang *pipeline* analitis sesuai kebutuhan. Setelah merancang *pipeline* analitis, *RapidMiner* akan menghasilkan file XML yang dapat diterapkan ke data untuk melakukan analisis yang diinginkan[29].

2.2.8. *Orange Data Mining*

Orange Data Mining adalah perangkat lunak komputer *opensource* untuk *data analytics* dan *data mining*. Yang membedakan *Orange* dengan perangkat lunak *data minin*g lainnya adalah fokusnya pada visualisasi atau *visual programming*. *Orange data mining* menyediakan berbagai macam *widget* yang dapat ditempatkan di dalam *canvas board* dan dihubungkan dengan *widget* lainnya. Dengan penggunaan *canvas board* ini, pengguna dapat dengan mudah memproses data dan melakukan analisis data secara intuitif.

2.2.9. *Stemming*

Stemming merupkan suatu proses untuk mengubah kata-kata (non-baku) menjadi kata dasarnya (baku) dengan cara menghilangkan imbuhan-imbuhan pada kata dalam dokumen atau mengubah kata kerja menjadi kata benda. Proses *stemming* dilakukan dengan menghapus awalan dan akhiran dari suatu kata sehingga tersisa kata intinya atau akar kata. Sebagai contoh, kata "dihilangkan" setelah imbuhan "di" dan "kan" dihilangkan akan menjadi "hilang"[30].

2.2.10. *Confusion Matrix*

Confusion Matrix merupakan suatu teknik yang digunakan untuk menghitung akurasi pada metode *data mining*. Teknik ini ditunjukkan dalam bentuk tabel yang menunjukkan jumlah data uji yang diklasifikasikan dengan

benar dan jumlah data uji yang salah diklasifikasikan. Komponen dari *Confusion Matrix* adalah tabel yang terdiri dari:[31]

Tabel 2.2 *Confusion Matrix*

Actual Label	*Predicted Label*	
	Positive (+)	*Negative* (-)
Positive (+)	*True Positive* (TP)	*False Negative* (FN)
Negative (-)	*False Postive* (FP)	*True Negative* (TP)

1. True Positives: merupakan jumlah *record* data positif yang di klasifikasikan sebagai nilai benar positif
2. False Positive: merupakan jumlah *record* data negatif yang diklasifikasikan sebagai nilai salah positif
3. False Negative: merupakan jumlah *record* data positif yang diklasifikasikan sebagai nilai salah negatif
4. True Negative: Merupakan jumlah *record* data negatif yang diklasifikasikan sebagai nilai benar negatif

Berikut rumus untuk menghitung model evaluasinya:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \quad (2.1)$$

$$\text{Precission} = \frac{\text{TP}}{\text{TP}+\text{FP}} \quad (2.2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \quad (2.3)$$

$$\text{F1} - \text{Score} = 2 * \frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \quad (2.4)$$

2.2.11. *Bag Of Words*

Bag Of Words adalah salah satu teknik dalam *Natural Language Processing* untuk mengonversi kata ke dalam bentuk vektor dan digunakan pada beberapa aplikasi NLP. *Bag of Words* juga merupakan metode ekstraksi fitur yang disederhanakan untuk data teks yang mudah diimplementasikan. Hal tersebut mempertahankan kosakata dan menghitung frekuensi kata, mengabaikan berbagai abstraksi bahasa alami seperti tata bahasa dan urutan

kata. Pendekatan *Bag of Words* mengambil dokumen sebagai *input* dan memecahnya menjadi kata-kata. Kata-kata ini juga dikenal sebagai token dan prosesnya disebut sebagai tokenisasi[32].