BAB II

LANDASAN TEORI

2.1 Penelitian Terdahulu

Pada subbab 2.1, akan menampilkan penelitian-penelitian terdahulu yang menjadi acuan utama dalam penelitian ini. Pada Tabel 2.1 menampilkan 10 penelitian terdahulu yang relevan dengan fokus penelitian yang akan dilakukan, pada tabel 2.1 memberikan gambaran yang komprehensif tentang landasan teoritis yang mendukung kajian penelitian.

Tabel 2.1 Penelitian terdahulu

No	Jurnal	Judul	Penulis	Metode	Hasil
1	Jurnal Riset Komputer (JURIKOM), vol. 9, No. 5, 2022	Analisis Sentimen Tokopedia Pada Ulasan di Google Playstore Menggunakan Algoritma Naïve Bayes Classifier dan K-Nearest Neighbor [13].	Muhammad Farid El Firdaus, Nurfaizah, dan Sarmini	Naïve bayes dan KNN	Hasil yang diperoleh pada penelitian tersebut yaitu algoritma KNN memiliki akurasi yang lebih baik dari pada Naïve bayes. Akurasi Naïve bayes sebesar 75,30% sedangkan akurasi algoritma KNN sebesar 86,09%
2	Jurnal Teknologi Informasi: Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika, Vol. 17, No. 2, 2023	Komparasi Algoritma Naive Bayes Dan K-Nearest Neighbor Pada Analisis Sentimen Terhadap Ulasan Pengguna Aplikasi Tokopedia [14].	Ryfan Maulana Putra Hertaryawan, Muhammad Raihan, dan Imam Santoso	Naïve bayes dan KNN dengan Particle Swarm Optimization (PSO)	Hasil yang diperoleh pada penelitian tersebut, algoritma KNN memiliki hasil yang unggul. Algoritma KNN tanpa PSO menghasilkan akurasi 83,1% sedangkan naïve bayes tanpa PSO menghasilkan akurasi sebesar 76,3%.
3	Jurnal Informa, Vol. 5 No. 2, 2019	Analisis Sentiment Twitter Terhadap	Abdul Malik Zuhdi, Ema Utami, dan Suwanto	KNN	Hasil pengujian yang diperoleh pada penelitian tersebut

No	Jurnal	Judul	Penulis	Metode	Hasil
		Capres Indonesia 2019 Dengan Metode K-NN [10].	Raharjo		pembagian sebanyak 70% data latih sebanyak 30% data uji diperoleh hasil akurasi sebesar 81,83%
4	Jurnal Inovasi dan Sains Teknik Elektro (INSANTEK), Vol. 2 No. 1, 2021	Comparison of K-NN Methods, Support Vector Machines, and Random Forests in E-Commerce Shopee [12].	Sri Watmah, Suryanto, Martias	K-Nearest Neighbor, Support Vector Machine, Random Forest	Pada penelitian tersebut diperoleh algoritma KNN dan SVM menunjukan hasil akurasi lebih yang unggul jika dibandingkan dengan algoritma Random Forest dengan nilai precision sebesar 89.7% dan 89.5%
5	Jurnal CoSciTech (Computer Science and Information Technology), Vol. 4 No. 1, 2023	Perbandingan Algoritma K-Nearst Neighbor dan Naïve Bayes pada Aplikasi Shopee [15].	A. Oktian Permana, Sudin Saepudin	K- Nearst Neighbor dan Naïve bayes	Temuan dari riset ini menunjukkan bahwa algoritma naïve bayes menghasilkan tingkat akurasi yang lebih tinggi daripada algoritma KNN. Dengan akurasi algoritma naïve bayes mencapai 80%, sedangkan algoritma KNN hanya mencapai 55%
6	ILKOM Jurnal Ilmiah, Vol. 15 No. 3, 2023	Sentiment Analysis of Shopee App Reviews Using Random Forest and Support Vector Machine [16].	Suswadi, Moh. Erkamim	Random Forest dan Support Vector Macine	Hasil penelitian menunjukkan bahwa algoritma Random Forest memiliki tingkat akurasi sebesar 82.21%, sementara algoritma SVM menunjukkan akurasi yang

No	Jurnal	Judul	Penulis	Metode	Hasil
					lebih tinggi, yakni sebesar 84.71%
7	Indonesian Journal of Computer Science (IJSC), Vol. 12 No.5, 2023	Komparasi Metode KNN dan Naïve BayesTerhaap Analisis Sentimen Pengguna Aplikasi Shopee [14].	Salman Alfaris, Kusnadi	K-Nearest Neigbour dan Naïve bayes	Hasil yang diperoleh berdasarkan penelitian tersebut yaitu algoritma KNN memiliki nilai akurasi yang lebih unggul dibandingkan dengan naïve bayes dengan PSO maupun tanpa optimasi PSO. Hasil algoritma naïve bayes sebesar memperoleh akurasi sebesar 76.30% dan algoritma KNN memperoleh akurasi 83.10%
8	Jambura Journal of Electrical and Electronics Engineering (JJEEE), Vol. 5, No. 1, 2023	Analisis Sentimen Terhadap Penggunaan Aplikasi Shopee Mengunakan Algoritma Support Vector Machine (SVM) [18].	Irma Surya Kumala Idris, Yasin Aril Mustofa, Irvan Abraham Salihi	Support Vector Machine	Hasil penelitian menunjukkan bahwa algoritma SVM menunjukkan tingkat akurasi yang sangat tinggi, mencapai 98%, dengan nilai F1-score yang juga tinggi, yakni 0.98 atau setara dengan 98%.
9	Jurnal Informatika, Jurnal Pengembangan IT (JPIT), Vol. 8, No. 3, 2023	Analisis Sentimen Masyarakat Terhadap Penggunaan E-Commerce Menggunakan Algoritma K-Nearest Neighbor [19].	Ikhsan Habib Kusuma, Nuri Cahyono	KNN dan NLP	Hasil penelitian memperoleh hasil akurasi model sebesar 82%, *presision* sebesar 86%, *recall* sebesar 79%, dan *f1-score* sebesar 82%.
10	Jurnal Informatika Teknologi dan Sains, Vol. 5 No. 1, 2023	Analisis Sentimen Review Wisatawan Pada Objek Wisata	I Wayan Budi Suryawan, Nengah Widya Utami, Ketut Queena Fredlina	SVM	Hasil penelitian memperoleh hasil akurasi sebesar 84%, *recall* sebesar

No	Jurnal	Judul	Penulis	Metode	Hasil
		Ubud Menggunakan Algoritma Support Vector Machine [20].			89.83%, *precision* sebesar 90.4%, dan *f1-score* sebesar 90.11%.

Berdsarkan tabel 2.1 tersebut, menampilkan beberapa jurnal penelitian terdahulu yang telah dilakukan oleh peneliti sebelumnya yang dijadikan acuan pada penelitian ini. Penelitian pertama dilakukan oleh Muhammad Farid El Firdaus, Nurfaizah, dan Sarmini [13]. Pada penelitian tersebut menjadikan *e-commerce* Tokopedia sebagai objek dalam penelian. Hasil yang diperoleh yaitu algoritma KNN memiliki akurasi tertinggi jika dibandingkan dengan algoritma Naïve bayes. Jumlah data penelitan yang digunakan pada penelitian tersebut yaitu sebanyak 992 komentar *review* aplikasi Tokopedia. Akurasi yang diperoleh oleh algoritma naïve bayes yaitu sebanyak 75,30% sedangkan akurasi yang diperoleh oleh algoritma KNN sebesar 86,08% [13].

Selanjutnya merupakan penelitian yang dilakukan oleh Ryfan Maulana Putra Hertaryawan, Muhammad Raihan, dan Imam Santoso dengan judul penelitian "Komparasi Algoritma Naive Bayes Dan K-Nearest Neighbor Pada Analisis Sentimen Terhadap Ulasan Pengguna Aplikasi Tokopedia" menunjukan sentimen masyarakat terhadap aplikasi Tokopedia. Pada penelitian tersebut menggunakan *Particle Swarm Optimization* (PSO) untuk meningkatkan optimasi klasifiaksi pada sentimen. Hasil yang diperoleh pada penelitian tersebut yaitu algoritma KNN memmiliki akurasi yang lebih baik, yaitu memiliki akurasi 83,53% sedangkan algoritma naïve bayes memperoleh akurasi sebesar 74,09% [14].

Selanjutnya, pada penelitian yang berjudul "Analisis Sentiment Twitter Terhadap Capres Indonesia 2019 Dengan Metode K-NN" yang dilakukan oleh Abdul Malik Zuhdi, Ema Utami, dan Suwanto Raharjo. Hasil pengujian yang diperoleh pada penelitian tersebut pembagian sebanyak 70% data latih sebanyak 30% data uji diperoleh hasil akurasi sebesar 81,83% [10].

Berdasarkan penelitian sebelumnya, penelitian ini akan menerapkan algoritma SVM (*Support Vector Machine*) dan KNN (*K-Nearest Neighbors*) dalam melakukan analisis sentimen terhadap aplikasi Shopee berdasarkan ulasan pengguna yang terdapat di Google Play Store. Pemilihan algoritma ini didasarkan pada tingkat akurasi yang tinggi yang telah terbukti pada penelitian sebelumnya. Namun, penelitian ini juga akan mengembangkan aspek-aspek tertentu dari penelitian sebelumnya. Salah satunya adalah memperkaya dataset yang digunakan, dengan menghindari penggunaan dataset yang hanya mencakup data kurang dari 6 bulan, yang dapat dianggap kurang objektif dalam mewakili pandangan pengguna terhadap suatu topik. Selain itu, penelitian ini akan membandingkan performa antara algoritma SVM dan KNN, yang merupakan pendekatan yang belum banyak dieksplorasi dalam penelitian sebelumnya. Terakhir, penelitian ini akan melakukan analisis untuk merangkum kata-kata yang paling sering muncul dalam ulasan pengguna, sehingga dapat memberikan pemahaman yang lebih mendalam terhadap sentimen yang terkandung dalam ulasan tersebut.

2.2 Teori Tentang Topik Skripsi

2.2.1 Sentimen Analisis

Sentimen analisis adalah suatu metode analisis yang digunakan untuk mengevaluasi sentimen di balik teks yang ditulis oleh pengguna. Sentimen yang dianalisis dapat berupa sentimen positif, negatif ataupun netral [21]. Sentimen analisis dapat dilakukan pada berbagai jenis teks, seperti ulasan produk, tweet, artikel berita, atau posting media sosial. Tujuan dari sentimen analisis adalah untuk memahami dan menganalisis perasaan dan pandangan orang terhadap suatu topik, merek, atau produk. Menurut Clayton dalam penelitiannya Sentimental Analysis dapat dibedakan menjadi dua [22]. Yakni *Coarse-grained Sentiment Analysis* dan *Fined-grained Sentiment Analysis*:

1. *Coarse-grained Sentiment Analysis*

   Coarse-grained Sentiment Analysis adalah teknik analisis sentimen yang membagi sentimen ke dalam beberapa kategori atau level yang lebih

umum. Biasanya, ada tiga kategori umum yang digunakan dalam *coarse-grained sentiment analysis*: positif, negatif, dan netral. Dalam melakukan analisis terkait opini publik terhadap suatu *product* biasanya *Coarse grained* biasa di gunakan suatu perusahaan untuk mengetahui positif atau negatifnya respon publik terhadap produk atau layanan perusahaan.

2. *Fined-grained Sentiment Analysis*

   berbeda dengan coarse-grained, Fined-grained Sentiment Analysis adalah teknik analisis sentimen yang lebih mendetail seperti perasaan marah, senang, sedih, kecewa dan lainnya. Analisis dengan menggunakan *Fined-grained Sentiment* relatif digunakan saat melakukan analisis di sosial media survei atau data teks yang mampu direpresentasikan sebagai sebuah respon konsumen terhadap suatu produk yang dikeluarkan oleh sebuah perusahaan.

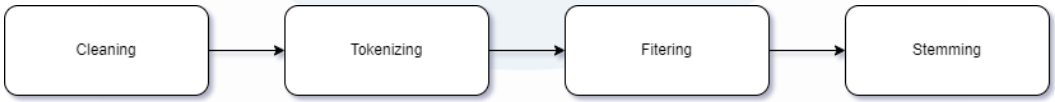
2.2.2 Text Mining

Text mining merupakan suatu proses penting dalam penelitian yang bertujuan untuk mengekstraksi informasi dari teks yang tersimpan dalam data, dengan tujuan utama untuk mengidentifikasi pola-pola yang tersembunyi di dalam data tersebut [23]. *Text mining* dilakukan untuk mengolah teks dari dokumen-dokumen yang umumnya berupa teks *unstructured*, dengan cara mencari korelasi antara kata-kata yang digunakan dalam data [23].

Terdapat perbedaan yang cukup mendasar antara *text mining* dan *data mining* perbedaan tersebut terletak pada sumber data yang digunakan. *Text mining* menggunakan data dalam bentuk teks atau dokumen, sementara pada *data mining* dapat memanfaatkan berbagai jenis sumber data yang tidak terbatas pada teks atau dokumen saja, biasanya data yang digunakan pada *data mining* dapat dalam bentuk data *structured* [24]. Tujuan dari *text mining* tidak hanya sebatas untuk mengidentifikasi keterkaitan antara kata-kata, tetapi juga untuk memberikan alternatif keputusan bagi pengguna.

Hasil analisis *text mining* dapat sangat bermanfaat bagi berbagai pihak, seperti para pemasar dalam menentukan strategi pemasaran yang

akan dilakukan, perusahaan ritel untuk mengidentifikasi pelanggan potensial, dan produsen produk untuk memahami opini publik terhadap produk yang mereka tawarkan. Tahapan dalam proses *text mining* mencakup *cleaning data*, *tokenizing*, *filtering*, dan *stemming*, yang mana tahap-tahap tersebut merupakan langkah-langkah yang cukup penting untuk dilakukan dalam memastikan data yang dihasilkan merupakan data yang berkualitas dan dapat diandalkan untuk analisis lebih lanjut [25]. Melalui serangkaian langkah tersebut, hasilnya adalah data yang telah diolah secara rapi dan terstruktur, yang akan sangat memudahkan proses klasifikasi data yang akan dilakukan pada tahap-tahap selanjutnya. Dengan melakukan tahap *preprocessing*, informasi yang awalnya sulit untuk dimengerti dan diolah dapat diubah menjadi bentuk yang lebih mudah dipahami dan dimanfaatkan dalam analisis dan pengambilan keputusan, sehingga memungkinkan pengguna untuk mendapatkan wawasan yang lebih mendalam dari data yang tersedia [26].



Gambar 2.1 Tahapan preprocesing

Pada Gambar 2.1 diatas mengilustrasikan tahapan dari *proses preprocessing data*, yang merupakan tahapan penting dalam mempersiapkan data untuk analisis lebih lanjut. Tahapan *preprocessing* dimulai dengan *cleaning*, yang merupakan proses untuk membersihkan dataset dari berbagai jenis *noise* dan redundansi sehingga menghasilkan dataset dengan *value* data yang seragam. Pada tahap *cleaning*, akan dilakukan penghapusan angka, emoji, tanda baca, data duplikat, serta normalisasi huruf menjadi huruf kecil [27]. Selain itu, pada tahap *cleaning* juga melakukan pembersihan terhadap baris-baris data yang tidak memiliki nilai atau *value* kosong dan pada tahap *cleaning* juga dilakukan pemilihan beberapa kolom yang digunakan pada penelitian, kolom tersebut antara lain kolom *'userName'*, *'Content'*, *'at'* dan *'score'*. Setelah data melewati tahap *cleaning*, langkah berikutnya adalah tahap *tokenizing*.

Tahap *tokenizing* adalah proses untuk memisahkan teks menjadi unit-unit yang lebih kecil, seperti kata-kata atau frasa, sehingga memfasilitasi analisis lebih lanjut, khususnya dalam analisis sentimen [28].

Selanjutnya, tahap *filtering* dilakukan untuk menghilangkan kata-kata yang termasuk dalam *stopwords* atau kata hubung. *Stopwords* merupakan kata-kata yang umumnya diabaikan dalam pemrosesan teks karena kurangnya makna kontribusi. *Stopwords* biasanya tersimpan dalam daftar stop list [29]. Contoh kata yang termasuk kedalam *stopword* antara lain, ‘yang’, ‘di’, ‘dari’, ‘oleh’, ‘yang’ kata-kata tersebut dihilangkan karena kata tersebut termasuk kedalam kata-kata umum yang sering digunakan dan tidak memiliki makna ataupun informasi konteks dalam sebuah kalimat [30]. Langkah terakhir dalam *preprocessing* adalah *stemming data*, langkah ini bertujuan untuk mengubah kata-kata ke bentuk dasarnya tanpa memperhitungkan imbuhan atau akhiran kata. Proses *stemming* dapat membantu dalam mencegah adanya duplikasi kata-kata dengan makna yang sama tetapi berbeda bentuk cara penulisan [31]. Dengan melewati serangkaian pada *preprocessing*, maka data akan menjadi lebih terstruktur dan siap untuk digunakan dalam analisis serta pengambilan keputusan lebih lanjut. Tahapan *preprocessing* ini menjadi fondasi penting dalam memastikan kualitas dan keakuratan hasil analisis yang akan dilakukan.

2.3 Teori tentang Framework/ Algoritma yang digunakan

2.3.1 CRISP-DM

Gambar 2.2 Diagram CRISP-DM [32]

Metode CRISP-DM (*Cross-Industry Standard Process for Data Mining*) adalah kerangka kerja yang digunakan dalam proses pengembangan solusi data mining. Kerangka kerja ini terdiri dari enam tahap yang saling terkait dan berulang, yaitu pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan *deployment* [33]. Tahap pertama, pemahaman bisnis, memfokuskan pada pemahaman yang mendalam terhadap tujuan bisnis dan kebutuhan pemodelan data yang diinginkan. Tahap kedua, pemahaman data, melibatkan eksplorasi awal terhadap data yang tersedia untuk mengidentifikasi potensi informasi yang relevan. Tahap berikutnya adalah persiapan data, di mana data dipersiapkan untuk proses pemodelan dengan membersihkan, menggabungkan, dan mengintegrasikan data dari berbagai sumber. Selanjutnya, tahap pemodelan melibatkan penggunaan teknik dan algoritma data mining untuk mengembangkan model yang sesuai dengan tujuan analisis. Tahap evaluasi dilakukan untuk menilai kinerja dan validitas model yang dikembangkan, serta untuk memastikan bahwa model tersebut memenuhi kebutuhan bisnis yang telah ditetapkan. Terakhir, tahap penyebaran melibatkan penerapan model ke lingkungan produksi dan pengambilan keputusan berbasis data. Dengan mengikuti metode CRISP-DM, organisasi dapat mengembangkan solusi data mining yang efektif dan efisien, serta memastikan bahwa hasil analisis dapat diaplikasikan secara langsung dalam konteks bisnis yang relevan.

2.3.2 Python

Python adalah sebuah bahasa pemograman yang bertujuan untuk memudahkan dalam pengembangan aplikasi secara cepat. Python merupakan Bahasa pemograman yang mudah dipahami karena memiliki bentuk yang jelas untuk digunakan [34]. Python juga bisa digunakan dalam melaksanakan analisis statistik, pembangunan model *machine learning* maupun tentang yang berhubungan dengan *data science*, pada bahasa pemrograman Python terdapat berbagai macam *library open-source* yang dapat dipakai penggunanya dalam melakuka pemrograman dengan menggunakan Pyhton. Sebagian *package* yang popular di bahasa pemrograman Python antara lain.

1. *Spicy* dan *Scikit*, *library* untuk membuat model *Artificial Intellegence* dan *machine learning*.
2. *OpenCV* Python, *library* untuk membuat aplikasi *Computer vision*.
3. *TensorFlow*, library untuk membuat model dalam implementasi *deep learning*, dan beberapa lainnya.
4. *Matplotlib*, *library* dapat digunakan dalam membuat visualisasi data seperti grafik. Setiap sumbu memiliki sumbu horizontal (x) dan sumbu vertikal (y).

2.3.3 Support Vector Machine

Support Vector Machine (SVM) menyediakan metode klasifikasi, dan merupakan algoritme supervised. Algoritma *supervised* merupakan algoritma yang berlatih menggunakan data yang telah ditandai sebagai dasar, atau dalam bahasa lain telah diajari sebelumnya [27]. SVM dapat juga dibagi dalam model linear dan nonlinear. Beberapa langkah yang harus dilakukan yaitu memetakan data yang telah ada, dan mencari data yang berada di ujung pemisah. Data domain yang dipisahkan kemudian dibagi dengan pemisah antara data yang telah dipisahkan menjadi beberapa *cluster* tergantung dengan *feature* yang dipilih. Pemisah antara data yang menjadi *cluster* ini disebut dengan *hyperplane* [35]. Penentuan dari *hyperplane* tersebut berdasarkan titik yang berada di antara tiap cluster dengan jarak terdekat dengan cluster lainnya, sehingga dapat disebut *support vector*. Penentuan dalam pemilihan *hyperlane* dengan menggunakan

support vector dan mencari margin terbesar. *Margin* merupakan jarak yang ada terhadap support vector terhadap garis hyperlane. Dalam beberapa kasus, bentuk *hyperlane* dalam dua *feature* tidak selalu berbentuk garis lurus yang membagi dua *feature*. Jika menggunakan garis lurus, maka dapat disebut menggunakan kernel linear, sedangkan bila tidak dipisahkan dengan laris lurus maka akan disebut nonlinear. Dikarenakan memiliki lebih dari banyak fitur, maka dilakukan metode ovr (*one versus rest*) yaitu dengan membagi perhitungan sesuai dengan fitur yang dimiliki, jika memiliki tiga fitur maka akan memiliki tiga perhitungan, dan dalam perhitungan satu kelas akan dianggap positif dan kedua kelas lain akan dianggap negatif.

2.3.4 K-Nearest Neighbors

K- Nearest Neighbors (KNN) ialah salah satu tata cara *machine learning* yang mengklasifikasikan objek menurut informasi pembelajaran yang sangat dekat dengan objek tersebut [36]. Tata cara ini sangat simpel, mudah direpresentasikan, mempunyai ketangguhan untuk melatih informasi yang mempunyai banyak *noise*, serta efisien untuk proses pengelompokan. Tujuan dari algoritma ini mengklasifikasikan objek baru, atribut serta pelatihan ilustrasi Nilai k terbaik untuk algoritma ini bergantung pada informasi. Terutama, nilai k yang besar akan mengurangi dampak *noise* pada klasifikasi, namun membuat batas antara tiap klasifikasi terus menjadi kabur. Nilai k yang baik bisa diseleksi berlandaskan parameter optimasi, misalnya dengan memanfaatkan *crossvalidation* [37]. Permasalahan khusus dimana klasifikasi diprediksi bersumber pada informasi pelatihan terdekat (dengan kata lain, k = 1) disebut algoritma tetangga terdekat.

Keakuratan algoritma KNN dipengaruhi oleh ada ataupun tidak terdapatnya fitur yang tidak relevan ataupun apabila nilai fitur tersebut tidak setara dengan relevansinya untuk klasifikasi. Sebagian besar penelitian tentang algoritme ini mangulas metode memilih dan menimbang fitur sehinggabahwa kinerja klasifikasi lebih baik.

2.3.5 TF-IDF

TF-IDF adalah algoritma yang digunakan untuk mengukur signifikansi relatif dari setiap kata dalam suatu dokumen berdasarkan frekuensi kemunculan kata tersebut. *Term Frequency* (TF) mengindikasikan seberapa sering sebuah kata muncul dalam dokumen, di mana semakin sering muncul, semakin tinggi bobotnya. Di sisi lain, *Inverse Document Frequency* (IDF) mengevaluasi seberapa jarang sebuah kata muncul di seluruh kumpulan dokumen. Kata-kata yang jarang muncul cenderung memiliki bobot yang lebih tinggi karena mereka cenderung lebih unik atau spesifik. Dengan demikian, TF-IDF mempertimbangkan kedua aspek ini untuk menentukan relevansi kata-kata terhadap dokumen tertentu dalam konteks kata kunci atau tema yang dicari [38].

2.3.6 Confussion Matrix

Confussion Matrix merupakan sebuah tabel yang digunakan untuk mengukur performa dari sebuah algorima klasifikasi. Pada tabel *confussion matrix* memvisualisasikan dan merangkum performa dari sebuah algoritma. *Confussion matrix* dibagi menjadi empat bagian seperti yang ditampilkan pada tabel 2.2. Pada tabel 2.2 tabel *confussion matrix* dibagi menjadi 4 value yaitu *True Positive*, *False Negative*, *False Negative*, dan *True Negative* [39].

Tabel 2.2 Confussion Matrix Table

	True	*False*
True (Positive)	*TP (True Positive)*	*FP (False Positive)*
False (Negative)	*FN (False Negative)*	*TN (True Negative)*

Berdasarkan tabel *Confussion Matrix* pada tabel 2.2 diatas, terdapat 4 kelompok klasifikasi yang ditampilkan pada bagian dibawah:

- *True Positive* (TP) : data memiliki nilai positif dan terdeteksi benar bahwa data positif.
- *False Positive* (FP) : data memiliki nilai negatif tetapi terdeteksi sebagai data positif
- *False Negative* (FN) : data memiliki nilai positif tetapi terdeteksi sebagai data negatif.

- *True Negative* (TN) : data memiliki nilai negatif dan terdeteksi benar sebagai data negatif.

Hasil dari *Confussion matrix* akan menghasilkan nilai yang diperoleh melalui *accuracy, precision, recall,* dan *F1-Score*.

2.3.6.1 Accuracy

Accuracy memprediksi bahwa seberapa akurat dalam mengklasifikasikan model dengan benar. *Accuracy* merupakan rasio untuk memprediksi positif dan negatif dari seluruh isi data [39]. Berikut adalah perhitungan nilai *accuracy*:

$$\text{Accuracy} = \frac{TN+TP}{TN+FP+FN+TP}$$

Rumus 2.1 Rumus Accuracy

2.3.6.2 Precision

Precision merupakan perbandingan antara metrik yang merepresentasikan rasio dari data yang memiliki nilai prediksi benar positif (*True Positive*) dengan data yang diprediksikan memiliki hasil positif dan atau *precision* merupakan rasio terhadap data yang memiliki nilai negatif dengan data yang memiliki hasil positif. Pada tahap *precision* akan menghasilkan akurasi data yang diminta dan hasil prediksi yang telah disediakan oleh model. *Precision* disebut juga sebagai rasio prediksi positif [39]. Berikut merupakan rumus perhitungan nilai *precision*:

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

Rumus 2.2 Rumus Precision

2.3.6.3 Recall

Recall mengambil keunggulan model dalam mengambil sebuah informasi. *Recall* mempresentasikan presentase prediksi hasil *true positive* dengan jumlah keseluruhan data *positive* [39]. Berikut merupakan rumus perhitungan nilai *recall*:

$$\text{Recall} = \frac{TP}{TP+FP}$$

Rumus 2.3 Rumus Recall

2.3.6.4 F1-Score

F1-Score menggambarkan perbandingan antara rata-rata presisi dan *recall*. *Score* memprediksi hasil positif palsu dan negatif palsu [39]. Berikut merupakan rumus perhitungan *F1-Score*:

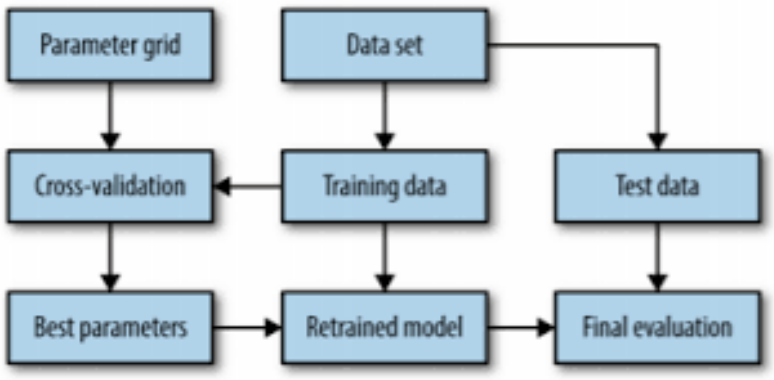
$$\text{FI} - \text{score} = F1 = score = 2 \times \frac{precision \times recall}{precision + recall}$$

Rumus 2.4 Rumus F1-score

2.3.7 GridSearch CV

GridSearch CV merupakan sebuah fungsi yang dapat digunakan untuk mencari parameter yang paling ideal yang ingin dicari untuk dilakukan pengujian pada model untuk memperoleh hasil yang maksimal. Cara kerja dari *GridSearch CV* yaitu dengan membagi data menjadi beberapa lipatan atau *fold* [40]. Lipatan tersebut bertujuan untuk melakukan iterasi berulang sesuai dengan nilai K iterasi. Seluruh nilai parameter yang diuji pada parameter akan diuji untuk mendapatkan parameter paling ideal pada setiap iterasi.

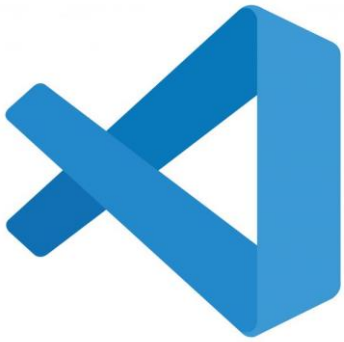
Setelah seluruh iterasi selesai dilakukan, *GridSearch CV* akan memberikan *best parameters* yang dapat digunakan pada model untuk memperoleh hasil yang maksimal, yaitu kombinasi parameter yang memiliki nilai rata-rata tertinggi atas *f-measure* pada setiap iterasi *fold* [40].



Gambar 2.3 Alur GridSearch CV [40]

2.4 Teori tentang Tools/ Software yang digunakan

2.4.1 Visual Studio Code



Gambar 2.4 Logo Visual Studio Code [41]

Visual Studio Code (VS Code) telah menjadi salah satu pilihan utama bagi para peneliti di berbagai bidang untuk pengembangan perangkat lunak dan analisis data [42]. Dikenal karena fleksibilitasnya dan kemampuan adaptasinya yang tinggi, VS Code menyediakan lingkungan pengembangan yang kaya fitur dan terintegrasi dengan berbagai alat yang mendukung penelitian. Para peneliti dapat dengan mudah mengelola proyek-proyek mereka, menulis kode, dan menganalisis data dengan menggunakan berbagai ekstensi yang tersedia, mulai dari pengelola versi seperti Git hingga alat analisis data seperti *Jupyter Notebooks*. Selain itu, fitur *IntelliSense* yang cerdas mempercepat proses penulisan kode dengan memberikan saran dan penyelesaian otomatis, sehingga memungkinkan peneliti untuk fokus pada inti dari penelitian mereka tanpa terganggu oleh detail teknis [41].

Selain fitur-fitur yang kuat, Visual Studio Code juga terkenal karena komunitas yang luas dan dukungan yang aktif. Para peneliti dapat dengan mudah menemukan dukungan dan solusi untuk masalah yang mereka hadapi melalui forum, tutorial, dan sumber daya online lainnya yang disediakan oleh komunitas pengguna VS *Code* [42]. Kemampuan untuk berkolaborasi dengan tim penelitian lainnya juga ditingkatkan melalui integrasi dengan layanan berbagi kode seperti GitHub, memungkinkan para peneliti untuk bekerja secara bersama-sama pada proyek-proyek yang kompleks dan berbagi penemuan mereka dengan lebih efisien. Dengan semua fitur dan dukungan yang ditawarkannya, Visual Studio

Code telah membuktikan diri sebagai alat yang tak tergantikan bagi para peneliti yang berusaha mencapai keberhasilan dalam penelitian mereka.

2.4.2 Google Colaboratory

Gambar 2.5 Logo Google Colaboratory [43]

Google Colaboratory, dikenal juga sebagai Google Colab, adalah sebuah *platform* yang revolusioner dalam dunia pemrograman dan penelitian. *Google Colaboration* dirancang oleh Google, dengan *google Colaboration* memberikan akses ke lingkungan pengembangan Python yang kuat secara langsung melalui browser web, tanpa memerlukan instalasi perangkat lunak tambahan. Salah satu fitur utamanya adalah kemampuan untuk menggunakan GPU dan TPU yang diberdayakan oleh *Google*, memungkinkan para pengguna untuk melakukan komputasi dan pemrosesan data yang intensif dengan cepat dan efisien [43]. Pengguna juga dapat berkolaborasi dengan mudah, dengan fitur berbagi yang memungkinkan tim untuk bekerja bersama pada *notebook* yang sama, membuatnya menjadi pilihan ideal bagi tim pengembang dan peneliti yang membutuhkan kerja sama yang lancar dan efektif.

Selain itu, *Google Colab* terintegrasi secara sempurna dengan *Google Drive*, memungkinkan pengguna untuk menyimpan, mengelola, dan berbagi *notebook* mereka langsung dari penyimpanan awan tersebut. Ini memberikan fleksibilitas yang besar kepada pengguna untuk mengatur dan mengakses proyek-proyek mereka dari berbagai perangkat dengan mudah. Dengan kombinasi kemudahan akses, performa tinggi, dan kemampuan berkolaborasi, Google *Colaboratory*

menjadi pilihan utama bagi para profesional di bidang data *science*, *machine learning*, dan pengembangan perangkat lunak yang mencari solusi yang handal dan efisien untuk memenuhi kebutuhan komputasi [43].

2.4.3 Google Play Store

Gambar 2.6 Logo Play Store [44]

Google Play Store merupakan *platform* distribusi aplikasi yang tersedia bagi pengguna perangkat berbasis sistem operasi Android. Di sini, pengguna memiliki akses untuk mengunduh beragam aplikasi, termasuk aplikasi Shopee. Google Play Store menyediakan informasi lengkap mengenai setiap aplikasi yang tersedia, termasuk tanggal peluncuran aplikasi, pembaruan terbaru, versi aplikasi, ukuran file, ulasan dari pengguna, peringkat atau rating aplikasi, dan informasi-informasi lainnya. Salah satu keunggulan utama dari *Google Play Store* adalah kemampuannya untuk melakukan sinkronisasi antar perangkat, memungkinkan pengguna untuk mengakses data aplikasi dari perangkat yang berbeda. Meskipun fitur ini tergantung pada dukungan dari masing-masing aplikasi, namun sebagian besar aplikasi telah menyediakan kemampuan sinkronisasi ini.

2.4.4 Shopee

Gambar 2.7 Logo Shopee [45]

Shopee, yang berdiri sejak tahun 2015 di Singapura di bawah naungan *Sea Group*, Shopee telah menjadi salah satu perusahaan *e-commerce* terkemuka di Asia Tenggara. Ekspansi Shopee tidak hanya terbatas pada negara Singapura saja, tetapi juga telah merambah ke berbagai negara di kawasan, termasuk Malaysia, Thailand, Taiwan, Vietnam, Filipina, dan Indonesia. Dengan kehadiran Shopee di sejumlah negara, Shopee telah memberikan peran penting dalam mengubah lanskap perdagangan daring di kawasan tersebut [46].

Pada akhir tahun 2023, jumlah kunjungan pengguna Shopee mencapai angka fantastis, yakni 2,35 miliar, hal tersebut menunjukkan betapa besar pengaruh platform ini dalam perekonomian Indonesia [6]. Dengan berbagai fitur yang ditawarkan tersebut, Shopee tidak hanya menjadi tempat transaksi jual beli biasa, tetapi juga menjadi pusat aktivitas ekonomi dan hiburan bagi masyarakat.

Fitur-fitur unggulan yang ditawarkan Shopee seperti *Cash on Delivery* (COD), *live streaming, Shopee Food, Shopee Games, Shopee Mall*, dan *SpayLater* adalah sebagian kecil dari beragam fitur yang disediakan Shopee untuk meningkatkan pengalaman pengguna. Melalui aplikasi Shopee, pengguna tidak hanya dapat membeli produk dengan mudah, tetapi juga memiliki akses ke ulasan dan rating produk yang membantu dalam membuat keputusan pembelian yang lebih terinformasi. Dengan demikian, Shopee tidak hanya menjadi platform e-commerce biasa, tetapi juga menjadi komunitas interaktif tempat pembeli dan penjual saling berinteraksi dan bertransaksi secara aman dan nyaman [47].

2.4.5 Microsoft Excel

Gambar 2.8 Logo Microsoft Excel [48]

Microsoft Excel merupakan perangkat lunak yang disediakan oleh *microsoft corporation* yang dapat dijalankan pada sistem operasi *Windows* dan *Mac OS.* Aplikasi Excel dapat digunakan untuk melakukan manipulasi terhadap data seperti membuat perhitungan, mengolah data, dan membuat visualisasi terhadap data [49]. Aplikasi *Microsoft Excel* dapat diguankan untuk menyimpan berbagai format data beberapa diantaranya, yaitu *.xlsx, .xlsm, .xlsb,* dan *.csv.* Pada penelitian ini, *Microsoft Excel* akan digunakan dalam format CSV (*Comma Separated Values*).