BAB II

LANDASAN TEORI

2.1 Tinjauan Teori

2.1.1 Analisis Sentimen

Analisis sentimen adalah sebuah metode yang digunakan untuk menilai perasaan dari para pengguna terhadap suatu subjek yang digunakannya [21]. Melalui analisis sentimen, perasaan pengguna akan dibagi pada umumnya ke dalam dua tipe yaitu positif dan negatif saja atau positif, negatif, dan netral seperti pada gambar 2.1. Akan tetapi, dalam beberapa kasus terdapat beberapa penelitian yang membaginya ke dalam bentuk yang lebih spesifik lagi seperti positif kuat, positif lemah, negatif kuat, dan negatif lemah. Proses analisis sentimen tidak terlepas dari bantuan *natural language processing* (NLP) yang mampu mengidentifikasi makna emosional di balik teks ulasan atau *review* yang ditulis.

Gambar 2.1 Kategori analisis sentimen

Sumber: MonkeyLearn [22]

Tujuan akhir dari dilakukannya analisis sentimen adalah untuk mengetahui bagaimana kepuasan konsumen atau pengguna terhadap objek yang dikonsumsi atau digunakannya sehingga dapat dilakukan tindakan lebih

lanjut misalnya melakukan revisi atau perbaikan dalam beberapa bagian yang dinilai masih kurang memuaskan agar diperoleh objek yang lebih baik dari sebelumnya. Analisis sentimen sendiri dapat dibagi ke dalam dua metode yang berbeda antara lain:

- Analisis sentimen berbasis *lexicon*

  Lexicon atau leksikon merupakan salah satu metode analisis sentimen yang dilakukan dengan cara menggunakan kumpulan diksi yang sudah dimiliki sebelumnya oleh leksikon yang digunakan. Masing-masing leksikon tersebut sudah mengkategorikan positif, negatif, dan netral. Contoh beberapa leksikon yang saat ini dapat digunakan yaitu *AFINN Lexicon*, *SentiWordNet*, *VADER*, dan lainnya.
- Analisis sentimen berbasis *machine learning*

  Analisis sentimen dengan metode *machine learning* memerlukan penggunaan algoritma untuk mengklasifikasi kata-kata ke dalam positif, negatif, dan netral. Contoh dari beberapa algoritma yang digunakan adalah *NB*, *DT*, *SVM*, *RF*, dan lainnya.

2.1.2 *Text Mining*

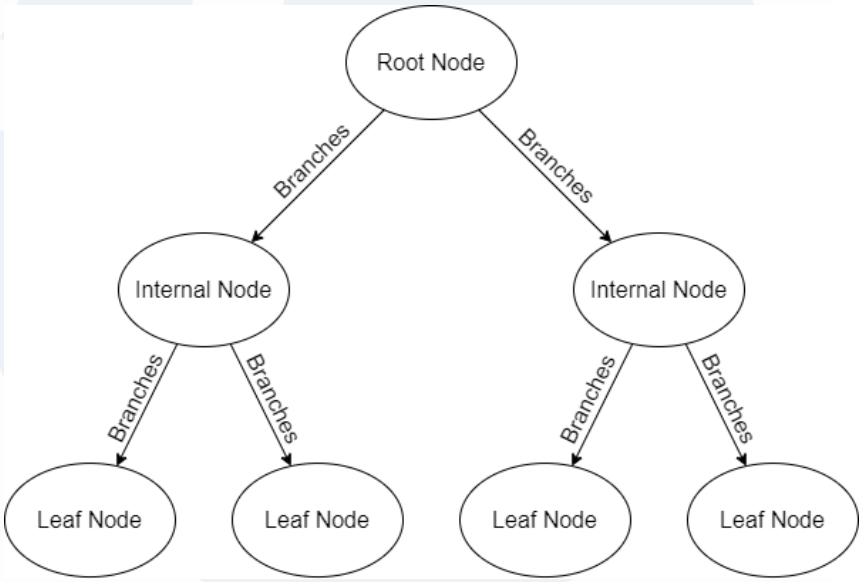
Text mining merupakan sebuah proses untuk mengekstraksi pola yang menarik dan signifikan untuk mendapatkan pengetahuan dari data tekstual yang sangat besar [23]. Terdapat beberapa teknik *text mining* yang berbeda-beda dan masing-masing memiliki tujuannya tersendiri. Penelitian kali ini akan melakukan salah satu teknik *text mining* yaitu melakukan analisis sentimen. Pada saat melakukan analisis sentimen, terdapat beberapa contoh proses text mining yang nanti akan dilakukan seperti *tokenizing*, *case folding*, *stop words*, dan *stemming*.

2.2 *Framework*, Algoritma, dan Teknik dalam Penelitian

2.2.1 *Decision Tree*

Decision Tree (*DT*) adalah sebuah algoritma *machine learning* yang dapat digunakan untuk menyelesaikan berbagai macam masalah khususnya dalam masalah pengklasifikasian. Salah satu contoh pemecahan masalah pengklasifikasian adalah analisis sentimen karena nantinya algoritma tersebut

akan digunakan untuk mengklasifikasi sentimen pengguna ke dalam label positif atau negatif. Algoritma ini memiliki konsep yang mudah dipahami akibat kesederhanaan cara kerja algoritmanya sendiri yaitu seperti akar pohon selain itu, algoritma ini juga memiliki kesamaan dengan cara manusia menentukan pilihan [24].



Gambar 2.2 Bentuk Decision Tree (DT)

Gambar 2.2 adalah bentuk paling sederhana dari algoritma *DT* terdiri dari beberapa istilah yaitu *root node, branches, internal node,* dan *leaf node*. *Root node* adalah kepala dari seluruh percabangan sehingga *root node* akan selalu hanya ada di paling atas dari pohon keputusan. *Internal node* adalah pengujian terhadap sebuah atribut. Setiap cabang atau *branches* akan menghubungkan atribut dengan hasilnya yang dinamakan dengan *leaf node*.

2.2.2 *Naïve Bayes*

Naïve Bayes (*NB*) adalah sebuah algoritma klasifikasi probabilitas sederhana yang mengkalkulasi atas probabilitas dengan cara menghitung frekuensi dan kombinasi dari nilai dalam data yang dimiliki [25]. Algoritma ini menganut asumsi naif bahwa hubungan antar seluruh variabel yang ada adalah independen atau saling tidak berkaitan. Aplikasi dari algoritma ini tidak hanya terbatas pada analisis sentimen saja melainkan juga untuk

menyelesaikan permasalahan klasifikasi lainnya seperti penyaringan spam dan pembuatan sistem rekomendasi.

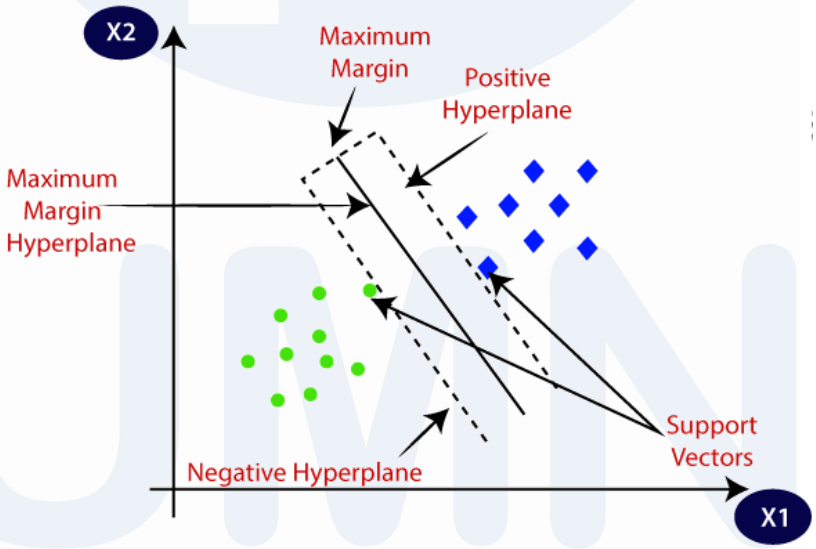
$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$$

Rumus 2.1 Rumus algoritma Naïve Bayes (NB)

Keterangan rumus 2.1:

a. P(H|E) menunjukkan bagaimana peristiwa H terjadi ketika peristiwa E terjadi.
b. P(E|H) menyatakan seberapa sering kejadian E terjadi ketika kejadian H terjadi terlebih dahulu.
c. P(H) merupakan peluang terjadinya H
d. P(E) merupakan peluang terjadinya E

2.2.3 Support Vector Machine



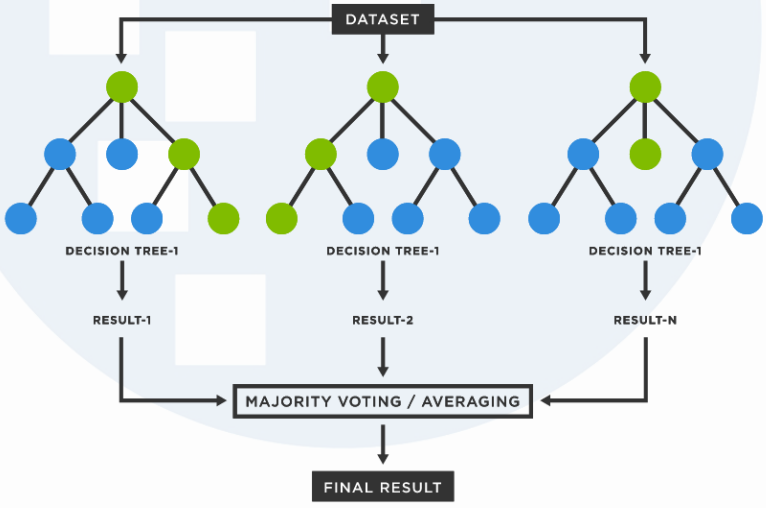
Gambar 2.3 Cara kerja algoritma Support Vector Machine (SVM)

Sumber: Javatpoint [26]

Support Vector Machine (*SVM*) adalah sebuah algoritma *machine learning* yang dapat digunakan untuk melakukan klasifikasi biner yaitu membagi data menjadi 2 kelas misalnya positif dan negatif. Algoritma ini bekerja dengan cara mempelajari data *training* terlebih dahulu yaitu dengan

melihat distribusi datanya dan mencari data-data yang berada pada titik ekstrim atau disebut juga *outlier*. Data-data dari titik ekstrim tersebut akan dijadikan sebagai vektor pendukung yang nanti akan digunakan sebagai patokan untuk membentuk garis *hyperplane* positif dan *hyperplane* negatif sehingga diperoleh garis *hyperplane* maksimumnya seperti pada gambar 2.3 [27]. Oleh karena itu, dapat disimpulkan bahwa algoritma *SVM* dapat bekerja dengan baik apabila distribusi kelas data yang digunakan jelas terlihat.

2.2.4 Random Forest



Gambar 2.4 Algoritma Decision Tree (DT)

Sumber: Tibco [28]

Random Forest (*RF*) adalah algoritma *machine learning* yang dapat digunakan untuk melakukan klasifikasi dan regresi. Algoritma ini terbentuk dari kombinasi beberapa *DT* yang membentuk beberapa pohon keputusan sehingga mampu menghasilkan keputusan seperti pada gambar 2.4 yang lebih baik daripada satu *DT* saja. Berdasarkan beberapa pohon keputusan tersebut, masing-masing hasilnya akan dikombinasikan atau dalam algoritma ini disebut *voting* untuk memperoleh hasil mayoritasnya [29]. Dukungan dari beberapa *DT* sekaligus tersebut membuat *RF* menjadi sebuah model yang memiliki akurasi klasifikasi yang tinggi, toleransi terhadap *outlier* dan *noise* yang baik, dan dapat menangani sebagian besar kasus *overfitting* yaitu kasus di mana akurasi *training* tinggi tetapi akurasi prediksi pada data lainnya rendah [29].

2.2.5 *Bootstrapping*

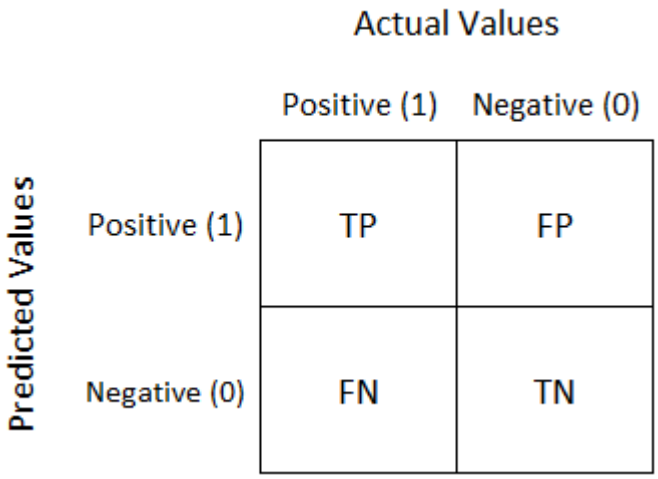
Bootstrapping adalah sebuah teknik *resampling* yang digunakan untuk mengimbangi data pada kelas-kelas yang jumlahnya tidak seimbang. *Bootstrapping* bekerja dengan cara menduplikasi data secara acak pada kelas yang dipilih hingga mencapai jumlah sampel data yang diinginkan. Teknik *bootstrapping* ini hanya diterapkan pada data *training* saja karena tujuannya hanya untuk meningkatkan kemampuan algoritma untuk mengidentifikasi data dengan tepat. Data *training* dengan kelas yang tidak seimbang dapat menyebabkan algoritma terlalu bias kepada kelompok kelas yang lebih banyak. Oleh karena itu, dengan bantuan *bootstrapping* data-data pada kelas minoritas mendapat penekanan yang sama dengan kelas mayoritas sehingga model dapat memprediksi dengan lebih akurat [30].

2.2.6 *TF-IDF*

TF-IDF merupakan singkatan dari *term frequency-inverse document frequency* yang menjadi salah satu *feature extraction* yang dapat digunakan pada tahapan persiapan data. *TF-IDF* bekerja dengan cara mengevaluasi frekuensi dari kata-kata yang dianggap berguna atau memiliki bobot lebih. Hasil akhirnya dapat digunakan untuk mengidentifikasi sentimen [31]. Sebagai contoh, *TF-IDF* dapat menemukan kata-kata yang paling sering dicantumkan dalam sebuah ulasan seperti “baik”, “bagus”, “buruk”, “jelek”, dan lainnya. Pada umumnya penggunaan *feature extraction* seperti *TF-IDF* ini dapat meningkatkan performa akurasi algoritma yang digunakan dalam melakukan analisis sentimen.

2.2.7 *Confusion Matrix*

Confusion matrix adalah sebuah alat yang dapat digunakan untuk mengukur performa dari hasil klasifikasi sebuah *machine learning* [32]. Bentuk dari *confusion matrix* adalah tabel ringkasan yang dinamakan matriks 2 dimensi dan berisikan jumlah prediksi yang benar dan salah dari sebuah model klasifikasi.



Gambar 2.5 Tabel confusion matrix

Sumber: Towards Data Science [33]

Gambar 2.5 adalah bentuk tabel *confusion matrix* paling sederhana yang terdiri dari 4 kolom. 4 kolom tersebut akan terbentuk apabila *output* dari prediksi model hanya 2 buah sedangkan, apabila hasil keluaran dari prediksi model lebih dari 2 buah maka kolom dan baris pada tabel juga akan bertambah banyak. Pada dasarnya, tabel *confusion matrix* memiliki makna sebagai berikut:

a. *True Positive* (TP): jumlah prediksi positif dan tepat
b. *False Positive* (FP): jumlah prediksi yang positif dan salah
c. *False Negative* (FN): jumlah prediksi yang negatif dan salah
d. *True Negative* (TN): jumlah prediksi negatif yang tepat

Confusion matrix memiliki 4 rumus antara lain:

1. *Precision*

   Precision menyatakan rasio prediksi positif yang tepat dari seluruh prediksi positif yang ada. Rumus 2.1 merupakan rumus *precision*.

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

Rumus 2.2 Precision

2. *Recall*

Recall menyatakan rasio jumlah prediksi positif dari kelas positif yang dapat diprediksi dengan tepat oleh model. Rumus 2.2 merupakan rumus recall.

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

Rumus 2.3 Recall

3. *F1-score*

   F1-score memberikan rasio mengenai gabungan *precision* dan *recall* yang biasanya memiliki hubungan terbalik. Rumus 2.3 merupakan rumus F1-score.

$$F1score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Rumus 2.4 F1-score

4. *Accuracy*

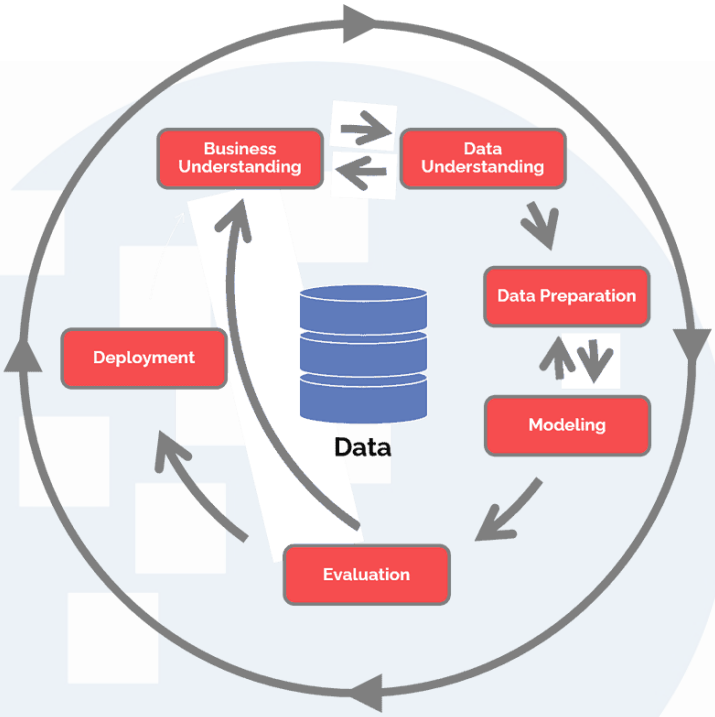
   Accuracy atau akurasi akan memberikan informasi terkait jumlah prediksi yang tepat dari total seluruh prediksi yang dilakukan. Rumus 2.4 merupakan rumus *accuracy*.

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

Rumus 2.5 Accuracy

2.2.8 CRISP-DM

CRISP-DM adalah kepanjangan dari *Cross-Industry Standard Process for Data Mining* yang merupakan alur kerja atau metode yang dijadikan standar bagi para praktisi data khususnya dalam melakukan *data mining* sekaligus juga *data science* [34]. Keberadaan metode CRISP-DM bagi para peneliti membuat proses jalannya penelitian lebih terarah kepada tujuan utamanya. CRISP-DM memiliki 6 tahapan yaitu *business understanding, data understanding, data preparation, modeling, evaluation,* dan *deployment*. Gambaran alur metode CRISP-DM dapat dilihat pada gambar 2.6.



Gambar 2.6 Alur metode CRISP-DM

Sumber: Data Science Process Alliance [35]

Berikut ini adalah 6 tahapan dari metode CRISP-DM [36]:

1. *Business Understanding*

   Business understanding merupakan pemahaman terhadap situasi bisnis yang menjadi permasalahan. Tahapan pertama ini penting sekali untuk dipahami secara menyeluruh agar proses *data mining* memiliki tujuan yang jelas.

2. *Data Understanding*

   Data Understanding merupakan tahapan mengumpulkan data, memahami, dan melakukan pengecekan terhadap kualitas data. Pada tahap ini dilakukan penelitian menyeluruh terhadap sumber-sumber data yang ada.

3. *Data Preparation*

   Data Preparation adalah tahapan mempersiapkan data seperti *data cleaning* dan *data labeling* sehingga data nantinya siap untuk digunakan. Proses *data preparation* ini juga bisa disebut sebagai *data pre-processing*.

4. *Modeling*

Modeling adalah tahapan memilih teknik pemodelan yang tepat. Tidak ada satu teknik pemodelan yang paling tepat karena semuanya tergantung kepada masalah bisnis dan data yang dimiliki seperti pada tahapan *business understanding* dan *data understanding*. Oleh karena itu, akan lebih baik jika model yang digunakan lebih dari satu sehingga diperoleh model yang paling optimal.

5. *Evaluation*

   Evaluation adalah tahapan memeriksa kembali hasil akhir dengan tujuan bisnis yang sudah didefinisikan sebelumnya di tahapan *business understanding*.

6. *Deployment*

   Deployment adalah tahapan implementasi dari hasil yang sudah didapatkan sebelumnya. Implementasi dalam hal ini dapat berupa laporan akhir bentuk visualisasi hasil penelitian ataupun perangkat lunak seperti aplikasi atau *website*.

2.3 *Tools* yang digunakan

2.3.1 Bibit

Gambar 2.7 Logo Bibit

Sumber: Bibit [37]

Bibit adalah sebuah aplikasi jual beli reksa dana secara online yang dapat dikategorikan ke dalam aplikasi *marketplace* seperti Tokopedia dan Shopee. Melalui Bibit, pengguna dapat membeli berbagai macam reksa dana yang ditawarkan dalam aplikasi. Pengguna hanya perlu memasukkan dana yang ingin diinvestasikan selanjutnya, dana tersebut akan dikelola oleh Robo Advisor yang akan membagi secara otomatis dana tersebut ke dalam berbagai macam reksa dana yang ditawarkan sesuai dengan profil risiko pengguna. Alokasi dana juga akan secara otomatis berubah apabila profil risiko pengguna mengalami perubahan.

Terdapat 3 tipe reksa dana yang dapat dibeli di Bibit yaitu pasar uang, obligasi, dan saham. Reksa dana jenis pasar uang ini cocok bagi pengguna yang memiliki profil risiko yang rendah karena reksa dana jenis ini cenderung stabil kenaikannya meskipun pada umumnya imbal hasil yang diperoleh juga paling sedikit. Contoh instrumen pasar uang adalah deposito. Reksa dana jenis kedua adalah obligasi. Reksa dana jenis obligasi ini cocok bagi pengguna yang memiliki profil risiko menengah karena nilai reksa dana ini pada umumnya cukup fluktuatif namun masih dalam batas yang wajar. Contoh instrumen obligasi adalah surat utang pemerintah dan korporasi. Reksa dana yang ketiga adalah reksa dana saham. Reksa dana ini cocok bagi pengguna yang memiliki profil risiko tinggi karena nilai dari reksa dana ini sangat fluktuatif. Contoh instrumen reksa dana saham adalah saham itu sendiri.

2.3.2 Google Play Store

Gambar 2.8 Logo Google Play

Sumber: 1000Logos [38]

Google Play Store adalah *marketplace* aplikasi yang dapat digunakan oleh para pengguna *smartphone* dengan sistem operasi Android. Melalui Google Play Store pengguna dapat mengunduh berbagai macam aplikasi termasuk aplikasi Bibit. Pada Google Play Store, pengguna dapat melihat keterangan dari setiap aplikasi yang tersedia seperti kapan aplikasi tersebut dimasukkan ke dalam Google Play Store, kapan aplikasi tersebut mendapatkan pembaharuan, versi aplikasi, besaran aplikasi, ulasan pengguna, *rating* atau nilai aplikasi, dan lain sebagainya. Google Play Store memiliki keunggulan dalam hal sinkronisasi sehingga apabila pengguna memiliki perangkat lain maka data dari aplikasi di perangkat sebelumnya dapat dihubungkan.

Tentunya hal ini terbatas kepada dukungan masing-masing aplikasi akan tetapi, sebagian besar aplikasi sudah mendukung kemampuan sinkronisasi tersebut.

2.3.3 Google Colaboratory

Gambar 2.9 Logo Google Colaboratory

Sumber: Google Cloud Console [39]

Google Colaboratory atau biasa juga disebut Google Colab adalah sebuah produk dari Google Research yang memungkinkan pengguna untuk menuliskan program dalam bahasa Python. Google Colab dibuat sedemikian rupa agar dapat menunjang kegiatan *machine learning*, analisis data, dan pendidikan dengan baik. Secara teknis, Google Colab adalah bentuk Jupyter Notebook yang tidak memerlukan persiapan terlebih dahulu sehingga pengguna dapat langsung menuliskan kodenya. Google Colab memberikan layanan secara gratis terhadap penggunaan *Central Processing Unit* (CPU) dan *Graphics Processing Unit* (GPU).

2.3.4 Visual Studio Code

Gambar 2.10 Logo Visual Studio Code

Sumber: Wikimedia Commons [40]

Visual Studio Code merupakan sebuah program *code editor* yang sering digunakan dalam menunjang kegiatan koding untuk keperluan apapun. Pada penelitian kali ini, Visual Studio Code digunakan sebagai sarana menulis code untuk membuat *website* yang dapat melakukan prediksi terhadap sentimen dalam bentuk kalimat atau kata yang ditulis oleh pengguna. Penggunaan Visual Studio Code ini juga memungkinkan *website* dijalankan secara lokal yaitu melalui *localhost* pada komputer/laptop terlebih dahulu. Setelah pembuatan *website* selesai maka *website* tersebut dapat di *deploy* di berbagai macam layanan *hosting* online seperti Heroku, Render, Replit, Railway, Glitch, dan lain-lain.

2.3.5 Excel

Gambar 2.11 Logo Microsoft Excel

Sumber: Microsoft Tech Community [41]

Microsoft Excel atau Excel adalah sebuah produk Microsoft dalam bentuk perangkat lunak atau aplikasi yang dapat digunakan untuk mengolah dan menghitung data. Penggunaan Excel juga memungkinkan penggunanya untuk membuat visualisasi data hingga analisis statistik [42]. Pada penelitian ini, Microsoft excel akan digunakan sebagai penyimpanan data sementara dalam bentuk format CSV (*Comma-Separated Values*) selama *data pre-processing* berlangsung. Selain itu, Excel juga akan digunakan untuk membuat beberapa visualisasi data ulasan aplikasi Bibit untuk melihat distribusi kelasnya.

2.4 Penelitian Terdahulu

Tabel 2.1 Penelitian terdahulu

No	Jurnal	Judul	Penulis	Metode	Hasil
1	International Journal of Environmental Research and Public Health, Vol. 18, No. 11, 2021	Sentiment Analysis on COVID-19-Related Social Distancing in Canada Using Twitter Data	Carol Shofiya dan Samina Abidi	*TF-IDF* dan *SVM*	Akurasi algoritma *SVM* adalah 87% dengan meningkatkan rasio data *training*
2	IJNMT (International Journal of New Media Technology), Vol. 7, No. 2, 2020	The Right Sentiment Analysis Method of Indonesian Tourism in Social Media Twitter	Cristian Steven, Wella	*SVM*, *KNN*, *DT*, dan *NB*	Akurasi terbaik diperoleh algoritma *KNN* sebesar 74,26% persen. Namun, terdapat pertimbangan hasil evaluasi lain yaitu *AUC* (*Area Under the ROC Curve*) yang menyatakan Algoritma *SVM* sebagai algoritma terbaik dengan nilai *AUC* 0,805
3	Journal of Information Systems Engineering and Business Intelligence, Vol. 4, No. 1, 2018	Sentiment Analysis in the Sales Review of Indonesian Marketplace by Utilizing Support Vector Machine	Anang Anggono Lutfi, Adhistya Erna Permanasari, Silmi Fauziati	*TF-IDF*, *SVM*, dan *NB*	Akurasi tertinggi diperoleh algoritma *SVM* sebesar 93,65% dengan menggunakan parameter max_df 25%
4	IJNMT (International Journal of	Sentiment Analysis about	Nico Nathanael Wilim,	*KNN*, *NB*, dan *DT*	Akurasi sebesar 76,94%

	New Media Technology) , Vol. 8, No. 1, 2021	Indonesian Lawyers Club Television Program Using K-Nearest Neighbor, Naïve Bayes Classifier, and Decision Tree	Raymond Sunardi Oetama		diperoleh algoritma *KNN*. Namun, penggunaan *dataset* pada tahun yang berbeda menunjukkan bahwa akurasi tertinggi diperoleh algoritma *NB*
5	(IJACSA) International Journal of Advanced Computer Science and Applications , Vol. 11, No. 4, 2020	Sentiment Analysis for Assessment of Hotel Services Review Using Feature Selection Approach based-on Decision Tree	Dyah Apriliani, Taufiq Abidin, Edhy Sutanta, Amir Hamzah, Oman Somantri	*Feature selection*, *NB*, *DT*, dan *SVM*	Kombinasi antara *DT* dan *feature selection* memperoleh hasil akurasi terbaik yaitu 88,54%
6	International Journal of Information System & Technology, Vol. 6, No. 4, 2022	Comparison of the Multinomial Naive Bayes Algorithm and Decision Tree with the Application of AdaBoost in Sentiment Analysis Reviews PeduliLindungi Application	Cecep Muhamad Sidik Ramdani, Andi Nur Rachman, Rizki Setiawan	*Multinomial NB*, *DT*, *AdaBoost*	Penggunaan aplikasi *AdaBoost* efektif meningkatkan akurasi dari algoritma *NB* dan *DT* menjadi 88,8% dan 84,1%
7	International Journal of Information Management Data	Sentiment Analysis and Classification of Indian Farmers'	Ashwin Sanjay Neogi, Kirti Anilkumar Garg, Ram	*NB*, *SVM*, *DT*, *RF*, *Bag of Words*, *TF-IDF*	Algoritma *RF* mendapatkan akurasi yang paling tinggi diantara 4

	Insights, Vol. 1, No. 2, 2021	Protest Using Twitter Data	Krishn Mishra, Yogesh K Dwivedi		algoritma yang digunakan
8	Journal of Physics: Conference Series, Vol. 1641, 2020	Improved Accuracy of Sentiment Analysis Movie Review Using Support Vector Machine Based Information Gain	Reza Maulana, Panny Agustia Rahayuningsih, Windi Irmayani, Dedi Saputra, dan Wanty Eka Jayanti	*Information Gain*, *NB*, *SVM*, *KNN*	*Information Gain* dapat meningkatkan akurasi algoritma *SVM* menjadi 85,65% untuk *dataset* Cornell dan 86,62% untuk *dataset* Stanford
9	IEEE Systems Journal, Vol. 13, No. 1, 2019	Forecasting Stock Market Movement Direction Using Sentiment Analysis and Support Vector Machine	Rui Ren, Desheng Dash Wu, Tianxiang Liu (2019)	*SVM*	Menemukan bahwa mengkombinasikan fitur sentimen dengan data pasar saham akan memberikan akurasi yang baik yaitu 89,93%, lebih tinggi 18,6% dibandingkan menggunakan data pasar saham saja
10	Sinkron: Jurnal dan Penelitian Teknik Informatika, Vol. 8, No. 1, 2023	Sentiment Analysis on App Reviews Using Support Vector Machine and Naïve Bayes Classification	Marchenda Fayza Madjid, Dian Eka Ratnawati, Bayu Rahayudi	*TF-IDF*, *SVM*, dan *NB*	*SVM* memperoleh akurasi 94,29% dan *NB* memperoleh akurasi 93,97%

Penelitian yang pertama dilakukan untuk mengetahui analisis sentimen masyarakat terhadap kebijakan *social distancing* di Kanada menggunakan data dari Twitter. Algoritma yang digunakan adalah *SVM* dengan *feature extraction TF-IDF*. Sebanyak 40% masyarakat Kanada bersikap netral terhadap kebijakan *social distancing* sementara 35% masyarakat Kanada menyatakan sentimen negatif. Algoritma *SVM* memperoleh akurasi 87% dengan pembagian data *training* dan *testing* sebesar 90% dan 10%. Sementara itu, pembagian data *training* dan *testing* sebesar 80% dan 20% membuat akurasi model menurun menjadi 81% [11].

Penelitian kedua membahas mengenai analisis sentimen masyarakat terhadap kota Bali. Total terdapat 4000 twit yang diambil dari media sosial Twitter mulai pada tanggal 1 Februari 2020 sampai 21 Maret 2020. Algoritma yang digunakan pada penelitian kedua ini adalah *SVM, KNN, Naïve Bayes*, dan *Decision Tree*. Keseluruhan proses penelitian dilakukan dengan menggunakan *tools* Rapidminer. Hasil dari penelitian ini diukur dengan menggunakan *confusion matrix* dan *AUC* pada *ROC curve*. Melalui *confusion matrix*, algoritma *KNN* memiliki akurasi tertinggi sebesar 73,91%. Sementara itu, hasil *AUC* menyatakan algoritma terbaik adalah *SVM* dengan nilai 0,805. Berdasarkan hasil *AUC* tersebut, algoritma *SVM* mendapatkan peringkat terbaik, yaitu *good classification* [12].

Penelitian ketiga menganalisis analisis sentimen pada ulasan aplikasi Bukalapak dengan menggunakan algoritma *SVM*. Penelitian ini menggunakan *feature extraction TF-IDF* dan dipadukan dengan algoritma *SVM* dan *NB*. Terdapat beberapa *max feature* pada *TF-IDF* yang digunakan pada penelitian ini yaitu 25%, 50%, 75%, dan 100%. Hasil akhir penelitian menyatakan bahwa akurasi tertinggi terdapat pada algoritma *SVM* sebesar 93,65% dengan parameter *max feature* 25% [13].

Penelitian keempat membahas mengenai perbandingan sentimen Twitter pada acara televisi Indonesian Lawyers Club dan Mata Najwa di tahun 2018 dan 2019. Terdapat 3 algoritma berbeda yang akan digunakan untuk penelitian tersebut antara lain *NB, KNN*, dan *DT*. Hasil yang diperoleh menyatakan bahwa tidak ditemukan salah satu algoritma yang memiliki performa paling tinggi secara tetap karena untuk

tahun 2018, algoritma yang memiliki performa paling akurat adalah *NB* sedangkan, pada tahun 2019 algoritma yang memiliki performa terbaik adalah *KNN* [14].

Penelitian kelima membahas mengenai analisis sentimen pada ulasan layanan hotel menggunakan pendekatan seleksi fitur dengan menggunakan 3 algoritma yaitu *SVM*, *NB*, dan *DT*. Pertama-tama, peneliti menggunakan model standar yaitu ketiga algoritma tersebut. Selanjutnya, peneliti membandingkan hasil model standar tersebut dengan model yang sudah dimodifikasi dengan penambahan seleksi fitur *Information Gain* (IG). Hasil akurasi tertinggi adalah 88,54% yang diperoleh algoritma *DT* dengan tambahan seleksi fitur [15].

Penelitian keenam membahas mengenai analisis sentimen pada aplikasi PeduliLindungi dengan menggunakan algoritma *NB* dan *DT*. Kedua algoritma tersebut masing-masing akan dikombinasikan dengan AdaBoost untuk menguji apakah penggunaan *AdaBoost* mampu meningkatkan performa algoritma. Hasil rata-rata akurasi sebelum menggunakan *AdaBoost* pada algoritma *NB* adalah 83,7% sedangkan, algoritma *DT* memperoleh 82,8%. Setelah penggunaan *AdaBoost*, rata-rata algoritma *NB* adalah 88,8% sedangkan, algoritma *DT* memperoleh 84,1%. Hasil akhir tersebut membuktikan bahwa penggunaan *AdaBoost* pada kedua algoritma mampu meningkatkan akurasinya [16].

Penelitian ketujuh meneliti sentimen analisis protes para petani di India melalui media sosial Twitter. Total data pada penelitian tersebut berjumlah 20.000 twit. Twit diolah menggunakan *feature extraction Bag of Words* dan *TF-IDF* kemudian dilanjutkan dengan proses dari 4 algoritma yaitu *NB*, *DT*, *RF*, dan *SVM*. Hasil penelitian menyatakan bahwa penggunaan *feature extraction Bag of Words* memberikan dampak terhadap performa algoritma yang lebih baik jika dibandingkan dengan *TF-IDF*. Akurasi tertinggi didapatkan oleh algoritma *RF* dengan tingkat akurasi mencapai 96,6% sedangkan, tingkat akurasi terendah diperoleh algoritma *NB* sebesar 72% [17].

Penelitian kedelapan dilakukan untuk mengetahui apakah *Information Gain* sebagai *feature selection* dapat digunakan untuk meningkatkan akurasi model. Terdapat 3 algoritma yang digunakan pada penelitian kedelapan ini yaitu *SVM*, *NB*,

dan *KNN*. Ketiga algoritma diuji coba dengan menggunakan *dataset* ulasan film Cornell dan Stanford. Melalui penelitian ini dibuktikan bahwa penggunaan *Information Gain* dapat meningkatkan algoritma *SVM* menjadi 85,65% pada *dataset* Cornell dan 86,62% pada *dataset* Stanford [18].

Penelitian kesembilan membuat sebuah model yang dapat digunakan untuk memprediksi pergerakan pasar saham yaitu indeks saham SSE 50. Model dibuat dengan terlebih dahulu melakukan analisis sentimen menggunakan algoritma *SVM*. Pada umumnya, model yang dibuat untuk memprediksi pergerakan saham hanya menggunakan data pasar saham itu sendiri namun, penelitian kali ini ingin melihat akurasi prediksi model apabila model tersebut digabungkan dengan fitur sentimen dari para investor terhadap pergerakan saham. Hasil akhir model dengan tambahan fitur sentimen mampu memprediksi pergerakan indeks saham SSE 50 dengan akurasi mencapai 89,93% yaitu 18,6% lebih tinggi jika dibandingkan dengan menggunakan data pasar saham saja [19].

Penelitian kesepuluh dilakukan untuk mengetahui bagaimana sentimen pengguna terhadap aplikasi Allo Bank. Algoritma yang digunakan adalah *SVM* dan *NB* dengan tambahan *feature extraction TF-IDF*. Hasil akhir yang diperoleh pada penelitian tersebut adalah meskipun kedua algoritma memiliki performa sangat baik yaitu di atas 90%, tetapi algoritma yang memiliki akurasi tertinggi adalah *SVM* sebesar 94,29% [20].