

BAB II

TINJAUAN PUSTAKA

2.1. Studi Sebelumnya

Prachyachuwong & Vateekul [22] dalam penelitian berjudul "*Stock Trend Prediction Using Deep Learning Approach on Technical Indicator and Industrial Specific Information*" melakukan pengembangan model *deep learning* yang dapat memprediksi arah tren pergerakan harga *Thailand Futures Exchange* (TFEX) dengan menganalisis data numerik dari indikator teknikal dan data historis *OHLCV* indeks saham SET 50 yang diintegrasikan dengan data tekstual dari judul berita ekonomi Thailand 2014-2020. Tidak hanya itu, peneliti juga melakukan pemecahan terhadap sektor-sektor industri terhadap data-data yang digunakan. Model *deep learning* yang digunakan, menggabungkan *Long Short-Term Memory Network (LSTM)* untuk data numerik dan *Bidirectional Encoder Representations from Transformers (BERT)*, khususnya *Multilingual BERT*, untuk data tekstual. Penelitian ini menghasilkan temuan bahwa pengintegrasian data sentimen dan pemecahan terhadap sektor-sektor industri dapat meningkatkan kinerja prediksi pada model. Performa model yang diciptakan menghasilkan akurasi rata-rata sebesar 61,28%, rata-rata *F1-score* sebesar 59,58%, dan menghasilkan rata-rata *return* tahunan sebesar 8,47% dari simulasi *trading*.

Gu, et al. [23] dalam penelitian berjudul "*Predicting Stock Prices with FinBERT-LSTM: Integrating News Sentiment Analysis*" melakukan peningkatan terhadap presisi dalam prediksi pergerakan harga indeks NASDAQ-100 dengan mengintegrasikan data numerik dari data historis harga NASDAQ-100, dengan data tekstual dari artikel berita. Peningkatan ini dilakukan dengan memanfaatkan *multimodal model*, yaitu model *FinBERT* untuk analisis sentimen terhadap artikel berita dan *LSTM* yang digunakan untuk prediksi pergerakan harga NASDAQ-100. Hasilnya menunjukkan bahwa model *FinBERT-LSTM* mencapai kinerja terbaik dengan *testing loss* 0.00083, *MAE* 173.67, *MAPE* 0.045, dan Akurasi 0.955. Performa model ini mengungguli model *LSTM* yang hanya menggunakan data numerik tanpa diintegrasikan dengan data tekstual dan model *DNN* yang juga hanya menggunakan data numerik.

Kim, et al. [24] melalui penelitian "*CBITS: Crypto BERT Incorporated Trading System*" mengembangkan *trading model Bitcoin* yang disebut CBITS. CBITS merupakan *trading model* yang bertujuan untuk meningkatkan akurasi prediksi pergerakan harga *Bitcoin*, dengan mengintegrasikan skor sentimen dari data tekstual dengan data numerik dari data historis *Bitcoin*. Penelitian ini menggunakan data tekstual berupa berita *cryptocurrency* Korea dari *Coinness Korea* dan data numerik dari data historis *BTC/USDT* dengan *timeframe trading* 4 jam. Pengintegrasian skor sentimen menunjukkan peningkatan performa semua model *time-series analysis*. Akurasi model tertinggi didapat oleh model *TabNet* yang telah dikombinasikan dengan *CryptoRoBERTa* dengan mencapai *F1-score* sebesar 69.83% dan akurasi sebesar 73.35%, dan menghasilkan profit *backtest* sebesar 304.65%. Model *TabNet-CryptoRoBERTa* mengungguli model-model lain seperti *LSTM-CryptoRoBERTa* dan *XGBoost-CryptoRoBERTa*.

Kim, et al. [25] dalam penelitian berjudul "*Forecasting the S&P 500 Index Using Mathematical-Based Sentiment Analysis and Deep Learning Models: A FinBERT Transformer Model and LSTM*" melakukan analisis performa model *LSTM* untuk prediksi pergerakan harga indeks *S&P 500*. Performa model ini dianalisis dengan setelah dan sebelum diintegrasikan dengan data tekstual berupa ringkasan artikel berita. Data tekstual tersebut dimanfaatkan dengan menggunakan skor sentimennya yang didapat dengan menggunakan *FinBERT*. Penelitian ini menunjukkan bahwa penggabungan data tekstual dengan skor sentimen meningkatkan performa model dengan *MSE* yang berubah dari 0.0026 menjadi 0.0016, *RMSE* dari 0.0513 menjadi 0.0407, *MAE* dari 0.0394 menjadi 0.0313 dan R^2 dari 0.8339 menjadi 0.8950. Peningkatan performa ini juga terjadi bahkan selama periode krisis seperti COVID-19 dan perang Ukraina Rusia.

Ko & Chang [26] melalui penelitian berjudul "*LSTM-based sentiment analysis for stock price forecast*" berupaya meningkatkan akurasi prediksi harga saham *Taiwan Stock Exchange* dengan menggabungkan data tekstual dan data numerik. Penelitian ini menggunakan artikel berita dari situs berita dan forum diskusi Taiwan sebagai data tekstual dan data historis enam saham sebagai data numerik. Hasil penelitian menunjukkan bahwa penggabungan informasi sentimen dari berita dan forum

menghasilkan prediksi paling akurat, dengan peningkatan rata-rata *RMSE* sebesar 12.05% dibandingkan model yang hanya menggunakan data numerik.

Tabel 2. 1. Tabel Penelitian Terdahulu

No	Peneliti	Judul	Tujuan	Metode	Hasil
1	Prachyachuwong, K. & Vateekul, P. (2021)	<i>Stock Trend Prediction Using Deep Learning Approach on Technical Indicator and Industrial Specific Information</i>	Mengembangkan model <i>deep learning</i> untuk memprediksi tren <i>Thailand Futures Exchange</i> (TFEX) dengan menganalisis data numerik (indikator teknikal dan data historis) dan data tekstual (judul berita ekonomi Thailand, dikelompokkan berdasarkan sektor industri) untuk meningkatkan kinerja prediksi dan <i>return</i> dari simulasi perdagangan.	Data: Judul berita ekonomi Thailand (2014-2020), data indeks saham SET50. Model: <i>LSTM</i> untuk data numerik, <i>BERT</i> (<i>Multilingual BERT</i>) untuk data tekstual. Evaluasi: Akurasi prediktif, <i>return</i> simulasi perdagangan.	Kerangka kerja berbasis sektor yang diusulkan (<i>model BERT_SEC + NUM</i>) mengungguli <i>baseline</i> dalam kinerja prediksi dengan akurasi rata-rata 61,28% dan <i>F1-score</i> 59,58%. <i>Return</i> tahunan dari simulasi perdagangan dengan model ini mencapai 8,47%.
2	Gu, W., Zhong, Y., Li, S., Wei, C., Dong, L., Wang, Z., & Yan, C. (2024)	<i>Predicting Stock Prices with FinBERT-LSTM: Integrating News Sentiment Analysis</i>	Meningkatkan presisi prediksi harga saham dengan mengintegrasikan kategori berita yang dibobot (terkait pasar, industri, dan saham) beserta sentimennya dengan harga saham historis, menggunakan model <i>FinBERT-LSTM</i> .	Data: Indeks NASDAQ-100, artikel berita <i>Benzinga</i> (15 Februari 2009 - 12 Juni 2020). Model: <i>FinBERT</i> untuk analisis sentimen berita (positif, netral, negatif), <i>LSTM</i> untuk harga saham historis (8 hari sebelumnya) dan skor sentimen. Dibandingkan dengan <i>LSTM</i> (hanya	Model <i>FinBERT-LSTM</i> berkinerja terbaik dengan <i>Testing loss</i> 0,00083, <i>MAE</i> 173,67, <i>MAPE</i> 0,045, dan Akurasi 0,955. Model <i>LSTM</i> (hanya harga) berada di peringkat kedua, dan model <i>DNN</i> (hanya harga) berkinerja terburuk.

				harga) dan <i>DNN</i> (hanya harga). Evaluasi: <i>MAE</i> , <i>MAPE</i> , Akurasi.	
3	Kim, G., Kim, M., Kim, B., & Lim, H. (2023)	<i>CBITS: Crypto BERT Incorporated Trading System</i>	Mengembangkan sistem perdagangan Bitcoin (CBITS) dengan menggabungkan skor sentimen dari <i>pre-trained Language Model</i> (LM) khusus <i>cryptocurrency</i> berbahasa Korea dengan data grafik <i>BTC</i> untuk meningkatkan kinerja model time-series analysis dan menciptakan strategi pasar.	Data: Berita <i>cryptocurrency</i> Korea dari <i>Coinness Korea</i> (CK) (19 Januari 2018 - 16 April 2022), data grafik <i>BTC/USDT</i> interval 4 jam dari <i>Binance</i> . Model: <i>fine-tuned pre-trained LM</i> khusus <i>cryptocurrency</i> (<i>CryptoBERT</i> , <i>CryptoRoBERTa</i> , <i>CryptoDeBERTa</i>) untuk klasifikasi sentimen <i>BTC</i> . Pencarian semantik (<i>KR-SBERT</i>) untuk artikel panjang. Model time-series analysis (<i>TabNet</i> , <i>LSTM</i> , <i>XGBoost</i>). Evaluasi: Uji KS, akurasi model <i>time-series analysis</i> , <i>F1-score</i> , profit <i>backtest</i> , volatilitas, <i>MDD</i> .	<i>CryptoRoBERTa</i> adalah <i>LM</i> optimal untuk analisis sentimen (<i>F1-score</i> 83,01%). Penambahan skor sentimen meningkatkan semua model perdagangan. <i>TabNet</i> + <i>CryptoRoBERTa</i> (<i>top 10</i> pencarian semantik) mencapai <i>F1-score</i> tertinggi (69,83%) dan akurasi (73,35%). <i>Backtesting TabNet RoBERTa Top 10</i> menghasilkan profit 304,65%.
4	Kim, J., Kim, H.-S., & Choi, S.-Y. (2023)	<i>Forecasting the S&P 500 Index Using Mathematical-Based Sentiment Analysis and</i>	Menyelidiki pengaruh sentimen dari data teks berita (ringkasan) terhadap pergerakan indeks <i>S&P 500</i> menggunakan	Data: Teks ringkasan dari <i>The New York Times</i> (1 Januari 2018 - 31 Desember 2022), data	Penambahan skor sentimen pada model <i>LSTM</i> (<i>LSTM-NYT</i>) meningkatkan akurasi prediksi dibandingkan

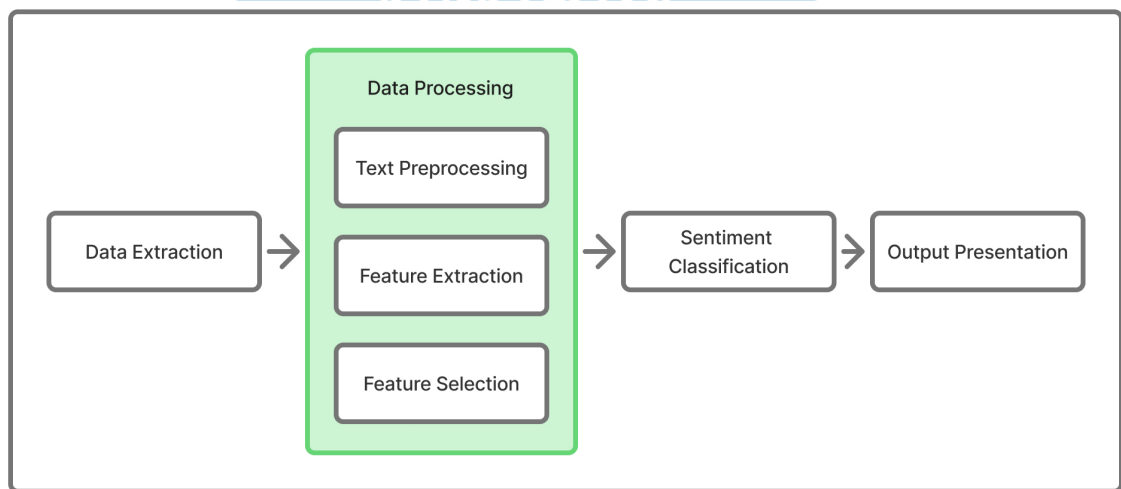
		<i>Deep Learning Models: A FinBERT Transformer Model and LSTM on line text data</i>	<i>FinBERT</i> untuk analisis sentimen dan <i>LSTM</i> untuk prediksi harga saham.	indeks <i>S&P 500</i> dari <i>Yahoo Finance</i> . Model: <i>FinBERT</i> untuk analisis sentimen pada ringkasan berita. Model time-series analysis (<i>LSTM</i> dengan fitur <i>S&P 500</i> saja vs. dengan fitur <i>S&P 500</i> + Skor Sentimen). Evaluasi: <i>MSE</i> , <i>RMSE</i> , <i>MAE</i> , <i>R</i> ² .	<i>LSTM-Only</i> : <i>MSE</i> (0,0016 vs 0,0026), <i>RMSE</i> (0,0407 vs 0,0513), <i>MAE</i> (0,0313 vs 0,0394), dan <i>R</i> ² (0,8950 vs 0,8339). Peningkatan ini juga teramati selama periode krisis (COVID-19 dan Perang Rusia-Ukraina).
5	Ko, C.-R. & Chang, H.-T. (2021)	<i>LSTM-based sentiment analysis for stock price forecast</i>	Meningkatkan akurasi prediksi harga saham <i>Taiwan Stock Exchange</i> dengan menggabungkan analisis fundamental (data tekstual sentimen dari artikel berita dan diskusi forum <i>PTT</i> menggunakan <i>BERT</i>) dan analisis teknikal (data numerik historis saham menggunakan <i>LSTM</i>).	Data: Artikel berita Taiwan, diskusi forum <i>PTT</i> (papan saham), data numerik historis saham (1 Januari 2015 - 31 Maret 2020) untuk 6 saham. Model: <i>BERT</i> (<i>pre-trained Mandarin</i>) untuk analisis sentimen teks (probabilitas positif/negatif). Model time-series analysis (<i>LSTM</i> dengan data historis saham 20 hari + sentimen berita + sentimen forum vs. <i>LSTM</i> (hanya harga) vs. <i>LSTM</i> (harga + berita), <i>LSTM</i>	Model <i>LSTMNF</i> (kombinasi data historis, sentimen berita, dan sentimen <i>PTT</i>) menghasilkan prediksi paling akurat (nilai <i>RMSE</i> terkecil untuk keenam saham). Rata-rata peningkatan <i>RMSE</i> sebesar 12,05% dibandingkan model <i>LSTM</i> (hanya harga historis).

				(harga + forum)). Evaluasi: <i>RMSE</i> .	
--	--	--	--	--	--

2.2. Dasar Teori

2.2.1. Analisis Sentimen

Analisis sentimen merupakan proses pemahaman dan pengolahan data tekstual untuk memahami opini, sikap, dan emosi terhadap suatu subjek, objek, maupun fenomena tertentu [27].



Gambar 2. 1. Kerangka Kerja Umum Analisis Sentimen

Birjali, et al. [28] telah melakukan kajian komprehensif terhadap lebih dari 500 model analisis sentimen. Seperti pada Gambar 2. Kerangka kerja umum untuk proses analisis sentimen mencakup empat tahapan utama. Tahapan pertama adalah *Data Extraction*, tahapan ini merupakan tahapan pengumpulan dan standardisasi data dari beragam sumber seperti platform media sosial, situs *e-commerce*, dan lainnya. Tahapan kedua merupakan tahapan *Data Processing*. *Data processing* merupakan tahapan eliminasi *noise* yang ada dalam data tekstual yang dapat mengganggu tahapan selanjutnya, tahapan ini berfungsi untuk mempersiapkan data agar menjadi bersih dan terorganisir dengan baik sehingga dapat memfasilitasi tahapan selanjutnya. Tahapan ini terdiri dari beberapa bagian di dalamnya, yaitu *text preprocessing* yang berfungsi untuk membersihkan kualitas data, *Feature Extraction* untuk mengekstraksi fitur atau kata kunci representatif dari keseluruhan data, dan *Feature Selection* untuk menghilangkan

fitur yang tidak relevan dan redundan. Tahapan ketiga merupakan tahapan paling penting dalam proses analisis sentimen, yaitu tahapan *Sentiment Classification*. Tahapan ini penting karena berfungsi untuk mengklasifikasikan sentimen dengan menggunakan beberapa pilihan pendekatan yang sesuai dengan kebutuhan dan karakteristik penelitian seperti leksikon atau *machine learning*. Tahapan terakhir adalah *Output Presentation* yang bertujuan menyajikan visualisasi atau rangkuman informasi dari keseluruhan sebaran sentimen hasil *sentiment classification* pada data.

2.2.2. Pendekatan dalam Tahapan Prediksi dan Klasifikasi Sentimen

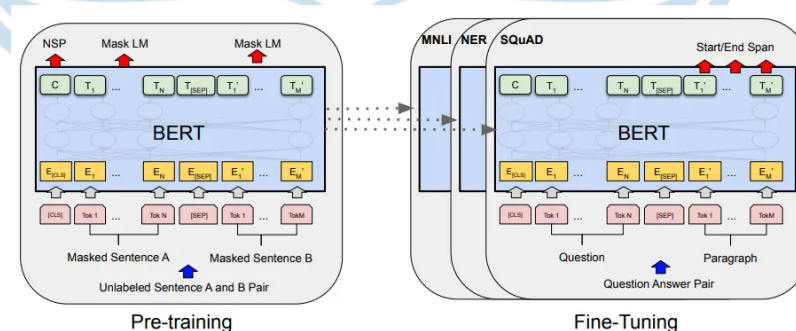
Tahapan *sentiment classification* merupakan inti dari proses analisis sentimen yang bertujuan untuk mengklasifikasikan sentimen dalam kalimat atau dokumen secara menyeluruh guna menunjang kebutuhan analisis atau penelitian. Berdasarkan Bordoloi & Biswas [29], terdapat dua pendekatan utama dalam tahapan ini, yaitu *machine learning approach* dan *lexicon-based approach*. Pendekatan *Lexicon-based* merupakan pendekatan yang menggunakan kamus atau daftar kata *lexicon* berisi kata-kata yang telah diberi label sentimen. Pendekatan *lexicon-based* memiliki kesulitan dalam menangani orientasi kata yang sensitif terhadap domain dan konteks [30].

Berdasarkan penggunaan data dan metode pembelajarannya, pendekatan *machine learning* dalam tahapan *sentiment classification* terbagi menjadi empat kategori. *Supervised learning* merupakan kategori klasifikasi sentiment dengan melakukan pelatihan dengan menggunakan *dataset* berlabel untuk memberikan label sentimen terhadap data baru yang memiliki karakteristik yang sama. Model analisis sentimen yang berbasis *supervised learning*, seperti *Support Vector Machine (SVM)*, *Naive Bayes*, *Maximum Entropy*, *Artificial Neural Network (ANN)*, *Random Forest*, dan *Gradient Boosting*. *Unsupervised learning* merupakan kategori teknik pengklasifikasian sentimen yang bekerja dengan *dataset* tanpa label. Kategori ini menggunakan metode *clustering* untuk mengklasifikasikan sentimen pada data dengan mengelompokkan berdasarkan kesamaan karakteristik, seperti *K-Means*. *Self-supervised learning* merupakan kategori teknik pengklasifikasian sentimen yang memanfaatkan *Pretrained Language Models (PLM)* seperti *BERT*, *NLTK*, *Stanford NLP*, *ULMFiT*, dan *ELMo*. *Self-Supervised learning* telah dilatih sebelumnya pada korpus berkapasitas besar untuk penggunaan klasifikasi sentimen. Setelah tahapan pelatihan selesai, model kemudian

disebarluaskan untuk pemanfaatan analisis sentimen pengguna lain tanpa harus melatih ulang model kecuali dengan tujuan penyesuaian. Kategori yang terakhir adalah *semi-supervised learning*, kategori yang mengombinasikan data berlabel dan tidak berlabel. *Semi-supervised learning* merupakan kategori dimana melakukan pelatihan dengan data berlabel yang kemudian diimplementasikan kepada data tidak berlabel. Contoh penggunaan *semi-supervised learning* ini adalah dengan *Graph Convolution Networks (GCN)*.

2.2.3. Bidirectional Encoder Representations from Transformers (BERT)

Pemanfaatan *self-supervised learning* memberikan berbagai macam keuntungan dalam analisis sentimen, terlebih dalam efisiensi komputasi dan efisiensi waktu [31]. Terdapat beberapa model untuk teknik pengklasifikasian sentimen yang menggunakan *self-supervised learning*, salah satu yang memiliki performa superior yaitu *BERT* [32]. *BERT* merupakan *pre-trained language model* berbasis *Transformer* yang terdiri dari serangkaian *encoder Transformer* yang ditumpuk satu sama lain [33]. Dengan adanya mekanisme *attention* pada *Transformer* dan kemampuannya yang dilatih untuk memahami teks secara dua arah, *BERT* mampu mengungguli model-model lain dalam memahami hubungan antar kata dan konteks secara lebih mendalam, sehingga mampu dalam menganalisis teks yang kompleks, panjang, dan memiliki sarkasme [34].



Gambar 2. 2. Arsitektur BERT

BERT berjalan dengan dua tugas utama, yaitu *Masked Language Model (MLM)* dan *Next Sentence Prediction (NSP)*. *Masked language model* merupakan sebuah tugas untuk melatih model dengan menutup beberapa kata atau token dan membuat model menebak token tersebut. Dalam *MLM*, sekitar 15% token dalam sebuah urutan ditutup secara acak, dan model dilatih untuk memprediksi token-token yang ditutup tersebut

berdasarkan konteks sekitarnya [35]. *NSP* merupakan tugas dimana model diberikan dua kalimat dan dilatih untuk memprediksi apakah kedua kalimat berurutan atau tidak berhubungan sama sekali [36].

Sejak kemunculannya, *BERT* telah melahirkan beberapa model turunan yang dikembangkan untuk meningkatkan kinerja atau menyesuaikan dengan kebutuhan spesifik, seperti *RoBERTa*, *BERTweet* yang dioptimalkan untuk teks *Twitter*, *DistilBERT* yang merupakan versi lebih kecil dan cepat dari *BERT*, serta model-model lain.

2.2.4. Kapitalisasi Pasar

Kapitalisasi pasar merupakan metrik finansial untuk mengukur nilai pasar keseluruhan dari suatu entitas atau instrumen investasi berdasarkan harga pasar saat ini. Metrik ini diaplikasikan secara luas pada berbagai aset investasi, termasuk saham, komoditas, hingga *cryptocurrency* [37]. Bagi pelaku pasar, kapitalisasi pasar memberikan gambaran mengenai ukuran, stabilitas, dan profil risiko dari suatu aset atau entitas, serta menjadi acuan dalam menentukan strategi investasi yang sesuai. Nilai kapitalisasi pasar dapat dihitung dengan melakukan perkalian antara harga saham saat ini dengan jumlah saham yang beredar. Nilai kapitalisasi pasar bersifat dinamis, berubah mengikuti fluktuasi harga, jumlah unit yang beredar, ataupun sentimen pasar [38].

Dalam konteks *cryptocurrency*, kapitalisasi pasar dihitung dengan melakukan perkalian antara harga koin saat ini dengan jumlah koin yang beredar. Kapitalisasi pasar *cryptocurrency* tidak hanya dapat dihitung secara individual atau terkhusus pada satu koin, tetapi juga dapat dihitung pada kapitalisasi pasar keseluruhan *cryptocurrency*, atau yang biasa dikenal dengan nama *TOTAL*. *TOTAL* merupakan sebuah indeks yang diciptakan oleh *Tradingview*, sebuah platform trading terbesar [39]. Tidak hanya *TOTAL*, *Tradingview* juga menciptakan *TOTAL2*, sebuah indeks kapitalisasi pasar *Altcoin* atau keseluruhan *cryptocurrency* diluar *Bitcoin* untuk menganalisis pergerakan keseluruhan *Altcoin*. Dengan perkembangannya, *Tradingview* juga menciptakan *TOTAL2ES*, indeks yang sama seperti *TOTAL2ES* tetapi dengan mengecualikan *Bitcoin* dan *Stablecoin*. *Stablecoin* dikecualikan karena memiliki karakteristik yang berbeda secara fundamental dengan *cryptocurrency* lainnya. *Stablecoin* dirancang untuk memberikan stabilitas harga

dengan nilai yang dipatok pada aset tertentu seperti *USD*, sehingga tidak mengalami volatilitas ekstrem seperti *Bitcoin* dan *Altcoin* lainnya [40], [41].

2.3. Tabel Perbandingan

Tabel 2. 2. Tabel Perbandingan Teknik Pendekatan Machine Learning

Aspek Perbandingan	Supervised Learning	Unsupervised Learning	Self-Supervised Learning	Semi-Supervised Learning
Jenis Data Utama	Data Berlabel	Data Tidak Berlabel	Data Tidak Berlabel	Campuran Data Berlabel dan Tidak Berlabel
Kebutuhan Pelabelan Manusia	Tinggi	Tidak Ada	Tidak Ada	Rendah hingga Sedang
Akurasi	Sangat Tinggi	Rendah hingga Sedang	Tinggi	Sedang hingga Tinggi