BAB II

TINJAUAN PUSTAKA

2.1 Landasan Teori

2.1.1 Analisis Sentimen

Analisis sentimen merupakan metode komputasi terhadap opini, sentimen dan subjektivitas teks. Analisis sentimen akan mengidentifikasi sentimen yang diungkapkan pada sebuah teks lalu menganalisisnya dengan target menemukan opini, mengidentifikasi sentimen dan mengklasifikasikan polaritasnya (Medhat et al., 2014). Analisis sentimen mengacu pada bidang NLP (*natural languange process*), *computational linguistics*, dan *text mining* yang bertujuan untuk menganalisis pendapat individu terhadap suatu layanan, produk atau aktivitas lainnya (Mustopa et al., 2021). Dalam proses klasifikasi sentimen terdapat tiga teknik yang dapat digunakan yaitu *machine learning, lexicon based* dan *hybrid approach*. Saat ini prosedur klasifikasi yang banyak digunakan adalah metode *machine learning* karena metode ini lebih mendekati prediksi polaritas sentimen berdasarkan data yang disiapkan (Hashfi et al., 2022).

2.1.2 Google Play

Google Play adalah sebuah *platform* distribusi digital yang dikembangkan oleh Google. *Platform* ini menawarkan berbagai produk seperti aplikasi, game, film, musik, dan buku yang dapat diakses melalui perangkat Android, web, dan Google TV. Di dalam Google Play, terdapat fitur ulasan yang memungkinkan pengguna memberikan pendapat tentang aplikasi yang diunduh, yang dikenal

sebagai "*review*". Selain itu, terdapat juga fitur "*rating*" yang mencerminkan tingkat kepuasan pengguna terhadap produk yang mereka gunakan (Asri et al., 2022).

2.1.3 Live.On

Live.On merupakan *digital provider* yang berasal dari perusahaan PT. XL Axiata Tbk yang diluncurkan pertama kali pada tanggal 5 Oktober 2020 dalam rangka perayaan hari jadi ke-24 XL Axiata. Melalui Live.On, pengguna dapat mengatur nomor, melakukan pembayaran, mengaktifkan, dan mengontrol paket, serta menghubungi layanan pelanggan melalui satu aplikasi, dengan memperhatikan kondisi pada saat itu yang sedang berada pada kondisi pandemi COVID-19. Aplikasi Live.On sendiri terdapat pada *platform* Google Play dengan *user interface* yang cukup mudah dipahami bagi pengguna (Idli et al., 2022).

2.1.1 *Net Reputation Score* (NRS)

Net Reputation Score merupakan metrik yang digunakan untuk memberikan gambaran mengenai reputasi *online* suatu merek atau perusahaan. Untuk menghitung NRS dapat dihitung dengan beberapa data, seperti ulasan pelanggan, media sosial dan hasil pencarian. Tujuannya adalah untuk membantu perusahaan meningkatkan reputasi *online*, menjadi alat untuk mengukur efektivitas strategi manajemen dan memperbaiki reputasi online untuk membangun citra perusahaan yang positif. *Range* nilai NRS berada diantara -100 dan 100, dimana semakin tinggi angkanya maka semakin banyak komentar positif

dari konsumen terhadap perusahaan. Dengan perhitungan pada persamaan (2.1) (Marsden, 2010).

$$\text{NRS} = \frac{\text{Positif} - (\text{Netral} + \text{Negatif})}{\text{Total Sentimen}} \text{x } 100\% \qquad (2.1)$$

2.1.4 Naïve Bayes

Metode Naïve Bayes merupakan sebuah metode dalam *machine learning* yang mengandalkan perhitungan probabilitas berdasarkan teorema Bayes. Metode ini memiliki kemampuan yang sebanding dengan Decision Tree dan Neural Network (Kusrini & Luthfi, 2009). Teorema Bayes mengasumsikan adanya indepedensi antara prediktor yang artinya kehadiran suatu fitur dalam suatu kelas tidak bergantung pada fitur lainnya (Kim et al., 2006). Metode Naïve Bayes berguna untuk kumpulan data yang cukup besar yang berdasarkan pada teorema Bayes dimana penentuan hubungan antara probabilitas P dari dua peristiwa A dan B direpresentasikan sebagai P(A) dan P(B) serta probabilitas bersyarat dari kejadian A ketika dikondisikan oleh kejadian B (Ahmad et al., 2017). Persamaan (2.2) merupakan persamaan dari Teorema Bayes:

$$\text{P(A|B)} = \frac{\text{P(B|A)}.\text{P(A)}}{\text{P(B)}} \qquad (2.2)$$

Keterangan:

P(A|B) = Peluang kejadian A jika diketahui B

P(B|A) = Peluang kejadian B jika diketahui A

P(A) = Peluang kejadian A

P(B) = Peluang kejadian B

2.1.5 Random Forest

Random Forest merupakan metode klasifikasi dari sebuah *ensemble* (kumpulan) pohon keputusan. Metode ini melibatkan penggunaan data latih dan fitur acak yang berbeda untuk menghasilkan suara (*vote*) yang akan menentukan hasil akhir. Setiap pohon keputusan dalam *ensemble* akan menentukan node akar dan node akhir dengan beberapa node daun untuk menghasilkan prediksi akhir (Klyueva, 2019). Random Forest memiliki 3 aspek penting yaitu: (1) *bootstrap sampling* untuk membangun pohon prediksi, (2) penggunaan prediktor acak dalam setiap pohon keputusan, dan (3) penggabungan hasil dari setiap pohon keputusan menggunakan metode *majority vote* untuk klasifikasi atau rata-rata untuk regresi (Nugraha et al., 2022). Keuntungan dari metode ini adalah dapat melakukan klasifikasi data yang tidak memiliki atribut yang tidak lengkap (Virra et al., 2019). Persamaan (2.3) dan (2.4) merupakan persamaan dari proses klasifikasi Random Forest (Trivusi, 2022):

$$Entropy\ (Y) = -\sum i\, p(c|Y) Log^2 p(c|Y) \tag{2.3}$$

Keterangan :

Y : Himpunan kasus

P(c|Y) : Proporsi nilai Y terhadap kelas c

$$Information\ Gain\ (Y, a) = Entropy(Y) - \sum v \varepsilon Values(a) \frac{|Yv|}{|Ya|} Entropy(Yv) \tag{2.4}$$

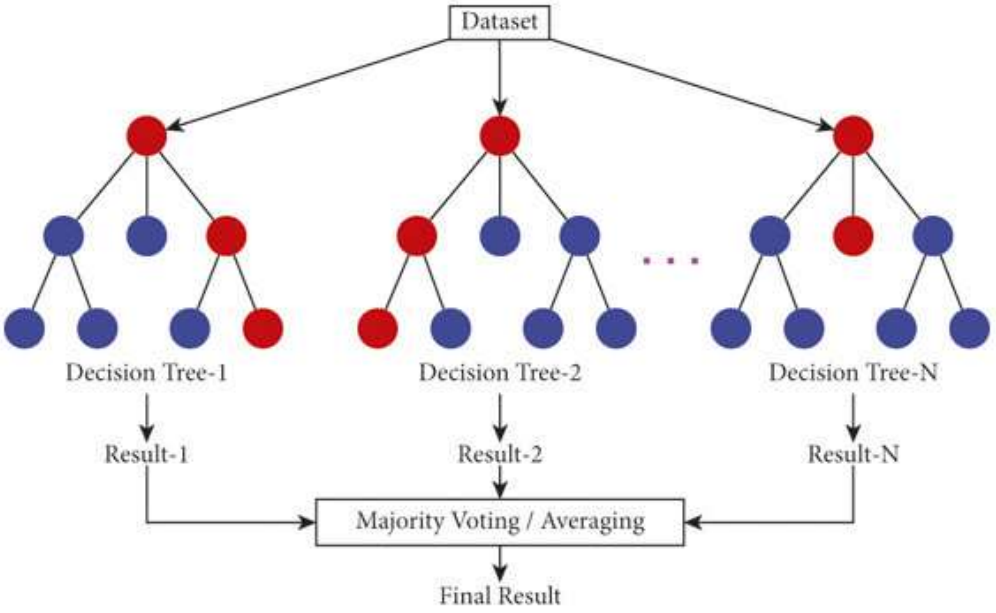
Keterangan :

Values (a) : Nilai yang mungkin dalam himpunan kasus a

Yv : Subkelas dari Y dengan kelas v yang berhubungan dengan kelas a

Ya : Semua nilai yang sesuai dengan a

Gambar 2.1 merupakan gambaran sederhana dari proses klasifikasi metode Random Forest.



Gambar 2.1 Metode Random Forest

2.1.6 *Web Scraper*

Web Scraper merupakan sebuah proses yang dilakukan untuk mengekstrak data dari halaman web secara otomatis menggunakan *crawling* dan bot. *Web Scraper* terhubung melalui *Hypertext Transfer Protocol* (HTTP) yang akan megambil halaman situs web untuk mengekstrak data, lalu kemudian di ekspor dengan format sesuai dengan kebutuhan (Ullah et al., 2018). Terdapat beberapa metode yang dapat digunakan dalam preoses *web scraping* pada halaman web,

seperti HTML *parsing*, DOM *parsing*, *vertical aggregation*, dan *XPath*. Selain itu, terdapat juga salah satu *library* di *python* yaitu *google-play-scrapper*, yang digunakan untuk mengunduh data terstruktur dari Google Play, seperti *review* pengguna terhadap layanan yang digunakan.

2.1.7 *Data Preprocessing*

Data Preprocessing adalah proses yang dilakukan untuk mengubah data menjadi format yang sesuai dengan kebutuhan. Pada tahap ini, dilakukan eksplorasi, pengolahan, dan analisis data yang terstruktur maupun tidak terstruktur guna menghilangkan *noise* sebelum memasuki tahap selanjutnya (Series, 2020). Beberapa proses yang dilakukan pada tahap *data preprocessing* diantaranya:

a. *Filtering*

Filtering merupakan proses pemilihan data tertentu berdasarkan dengan kriteria yang telah ditetapkan.

b. *Cleansing*

Cleansing merupakan proses pembersihan data dari tanda baca, nomor, URL, *white space* di awal dan akhir.

c. *Case Folding*

Case Folding merupakan proses yang digunakan untuk mengubah huruf-huruf besar menjadi huruf-huruf kecil, sehingga semua huruf memiliki format yang seragam.

d. *Slangwords*

Slangwords merupakan proses untuk mengubah atau menggantikan kata atau bahasa informal menjadi bahasa formal atau standar.

e. *Stopwords*

Stopwords adalah proses untuk menghapus kata-kata yang tidak menambah makna atau kurang berkaitan pada sebuah kalimat yang biasanya adalah kata sambung.

f. *Tokenization*

Tokenization merupakan proses untuk mengubah kalimat menjadi potongan *string* yang disebut dengan token.

g. *Stemming*

Stemming merupakan proses untuk mengurangi kata ke bentuk dasarnya dengan menghilangkan semua imbuhan yang menyatu pada kata.

2.1.8 *Confusion Matrix*

Confusion Matrix adalah metode yang digunakan untuk mengukur tingkat keberhasilan klasifikasi dengan mengidentifikasi tupel dari kelas yang berbeda. Pengunaan *confusion matrix* dapat menghitung parameter-parameter evaluasi seperti *accuracy*, *precision, recall,* dan *f1-score*, yang bergantung pada nilai *True Positive* (TF), *True Negative* (TN), *False Positive* (FP), *False Negative* (FN) (Han et al., 2011). Beberapa persamaan pada parameter evaluasi *confusion matrix* yaitu *accuracy, precission, recall,* dan *f1-score*:

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{2.5}$$

$$Precission = \frac{TP}{TP + FP} \tag{2.6}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.7}$$

$$F1 - Score = 2.\frac{precission\,.recall}{precission + recall} \tag{2.8}$$

2.2 Penelitian Terkait

2.1.2 *State of The Art* (SOTA) Penelitian

State of The Art merupakan matriks yang berisi penelitian yang sudah dilakukan mengenai analisis sentimen dengan metode, data yang berbeda-beda. *State of The Art* dari penelitian ini disajikan pada tabel 2.1.

Tabel 2.1 *State of The Art* Penelitian

No	Peneliti	Judul Penelitian	Hasil Penelitian
1.	Cendana, Idli Mulia Asriguna et all (2022)	*Sentiment Analysis of Live.On Digital Provider Application Using Naïve Bayes Classifier Method*	Penelitian ini bertujuan untuk mengevaluasi akurasi algoritma Naïve Bayes dalam menganalisis sentimen aplikasi Live.On dengan menggunakan metode *lexicon* untuk melabeli data kelas positif dan negatif. Hasil penelitian menunjukkan bahwa algoritma Naïve Bayes memiliki tingkat akurasi sebesar 87% dengan menggunakan 1000 data, dengan nilai *Precision* sebesar 61%, *f-measurement* 69%, dan *Recall* 81%.
2.	Hashfi, Farhan et all (2022)	*Sentiment Analysis of An Internet Provider Company Based on Twitter Using Support Vector Machine and Naïve Bayes Method*	Dalam penelitian ini, dilakukan analisis sentimen terhadap komentar Twitter mengenai Indihome yang dibagi menjadi dua kelas, yaitu kelas positif dan negatif. Metode yang digunakan dalam penelitian ini adalah Support Vector Machine dan Naïve Bayes. Meskipun nilai akurasi Support Vector Machine lebih tinggi yaitu 84%, sedangkan Naïve Bayes sebesar 82%, namun penelitian ini masih memiliki kekurangan yaitu pelabelan data dilakukan secara manual dan hanya menggunakan 1000 data.
3.	Rozaq, Abdul et all	Analisis Sentimen Terhadap	Penelitian ini dilakukan untuk melakukan analisis

No	Peneliti	Judul Penelitian	Hasil Penelitian
	(2022	Implementasi Program Merdeka Belajar Kampus Merdeka Menggunakan Naïve Bayes, K-Nearest Neighboars dan Decision Tree	sentimen terhadap implementasi program Merdeka Belajar - Kampus Merdeka di Twitter. Penelitian ini menggunakan metode Naïve Bayes, K-Nearest Neighbors, dan Decision Tree. Hasil penelitian menunjukkan bahwa algoritma Naïve Bayes memiliki tingkat akurasi yang paling besar yaitu 99,22%, diikuti oleh K-Nearest Neighbors sebesar 96,90%, dan Decision Tree sebesar 37,21%. Namun, penelitian ini memiliki keterbatasan yaitu pelabelan data dilakukan secara manual dan hanya menggunakan data dari Twitter.
4.	Dangin, Renyta Kristianti et all (2022)	Perbandingan Naïve Bayes dan Support Vector Machine pada Sentimen Analisis Reputasi Brand Twitter DQLab.id	Penelitian ini melakukan analisis sentimen terhadap perusahaan DQLab.id di Twitter dengan menghitung *Net Brand Reputation* (NBR) menggunakan metode Naïve Bayes dan Support Vector Machine. Tujuan penelitian ini adalah untuk membandingkan performa kedua metode tersebut. Hasil penelitian menunjukkan bahwa Support Vector Machine lebih unggul dalam performanya dibandingkan dengan Naïve Bayes. Selain itu, dari perhitungan NBR, DQLab.id memiliki nilai reputasi sebesar 56%. Namun, pada penelitian ini, pelabelan data masih dilakukan secara manual.
5.	Mustopa, Ali et all (2021)	*Analysis of User Reviews for the PeduliLindungi Application on Google Play Using the Support Vector Machine and Naïve Bayes Algorithm Based on Particle Swarm Optimization*	Tujuan dari penelitian ini adalah untuk menganalisis sentimen dari *review* komentar pengguna aplikasi PeduliLindungi di Google Play menggunakan algoritma Naïve Bayes dan Support Vector Machine (SVM) berbasis *Particle Swarm Optimization* (PSO). Hasil dari penelitian ini menunjukkan bahwa algoritma SVM berbasis PSO menghasilkan nilai akurasi yang lebih

No	Peneliti	Judul Penelitian	Hasil Penelitian
			tinggi dibandingkan dengan algoritma Naïve Bayes berbasis PSO.
6.	Rahmatulloh, Alam et all (2021)	*Sentiment Analysis of Ojek Online User Satisfaction Based on the Naïve Bayes and Net Brand Reputation Method*	Penelitian ini bertujuan untuk mengevaluasi persepsi publik terhadap kualitas layanan dari perusahaan Gojek dan Grab Indonesia di Twitter menggunakan analisis sentimen dan perhitungan *Net Brand Reputation* (NBR) dengan metode klasifikasi Naïve Bayes. Hasil penelitian menunjukkan bahwa perusahaan Grab memiliki skor kepuasan pengguna yang lebih baik daripada Gojek berdasarkan perhitungan NBR, meskipun hasil klasifikasi Naïve Bayes menunjukkan akurasi yang tinggi untuk kedua perusahaan.
7.	Fitri, Evita et all (2020)	Analisis Sentimen Terhadap Aplikasi Ruangguru Menggunakan Algoritma Naïve Bayes, Random Forest dan Support Vector Machine	Tujuan dari penelitian ini adalah untuk melakukan analisis sentimen terhadap aplikasi Ruangguru menggunakan tiga model klasifikasi yaitu Naïve Bayes, Random Forest, dan Support Vector Machine. Hasil dari penelitian ini menunjukkan bahwa model klasifikasi Random Forest memiliki akurasi tertinggi yaitu 97,16% dengan AUC *score* 0,99.
8.	Virra, Khalisa et all (2019)	*Sentiment Analysis of Social Media Users Using Naïve Bayes, Decision Tree, Random Forest Algorith: A Case Study of Draft Law on the Elimination of Sexual Violence (RUU PKS)*	Penelitian ini melakukan analisis sentimen dari pengguna Twitter menggunakan algoritma Naïve Bayes dengan kasus RUU PKS. Hasil dari penelitian ini menunjukkan bahwa analisis sentimen dari kasus ini mendapatkan hasil terbaik menggunakan algoritma Naïve Bayes dengan akurasi sebesar 83,54%, Decision Tree 75,31% dan Random Forest 75,72%. Berdasarkan hasil yang didapatkan bahwa kampanye RUU PKS banyak menghasilkan sentimen positif. Akan tetapi pada penelitian ini sumber data yang diambil hanya dari

No	Peneliti	Judul Penelitian	Hasil Penelitian
			Twitter.
9.	Guia, Marcio et all (2019)	*Comparison of Naïve Bayes, Support Vector Machine, Decision Trees and Random Forest on Sentiment Analysis*	Penelitian ini bertujuan untuk membandingkan kinerja algoritma Naïve Bayes, Support Vector Machine, Decision Trees, dan Random Forest pada analisis sentimen ulasan ponsel. Selain itu, penelitian ini juga bertujuan untuk menganalisis dampak merek dan harga pada ulasan polaritas. Hasil penelitian menunjukkan bahwa Random Forest dan Support Vector Machine memiliki akurasi, presisi, *Recall*, dan skor F1 yang lebih baik dibandingkan dengan Naïve Bayes dan Decision Tree. BlackBerry memiliki nilai polaritas 74,3% ulasan positif dan ZTE dengan 82,9% ulasan positif. Namun, penelitian ini memiliki keterbatasan yaitu penggunaan hanya satu set data dan kurangnya analisis tentang dampak dari atribut lain seperti fitur dan spesifikasi ponsel.
10.	Vidya, Nur Aziza et all (2015)	*Twitter Sentiment to Analyze Net Brand Reputation of Mobile Phone Providers*	Penelitian ini bertujuan untuk mengklasifikasikan analisis sentimen dari Twitter menggunakan Naïve Bayes, Support Vector Machine, dan Decision Tree dan mengusulkan metode baru untuk mengukur reputasi merek menggunakan *Net Brand Reputation*. Hasil penelitian menunjukkan bahwa algoritma Support Vector Machine memiliki skor tertinggi yaitu sebesar 82,40%. Merek yang memiliki nilai NBR tertinggi adalah XL Axiata dengan skor rata-rata NBR sebesar 32,3%.

2.2.2 Matriks Penelitian

Tabel 2.1 merupakan tabel matriks penelitian terkait analisis sentimen yang bersumber dari jurnal nasional, jurnal internasional maupun sumber lain.

Tabel 2.2 Matriks Penelitian Terkait

No.	Penulis (Tahun)	Judul	Analisis Sentimen		Metode					Preprocessing													Validasi Model			Evaluasi						Nilai Akurasi
			Sentiment	Tingkat Kepuasan	Support Vector Machine	Naïve Bayes	Random Forest	K-Nearest Neighboars	Decision Tree	*Cleansing*	*Filtering*	Converting Emoticon	Remove Duplicate	*Case Folding*	Tokenization	Normalization	Stopword Removal	*Stemming*	Manual Labelling	Remove Url	TF-IDF	*Slangwords*	Cross-validation	Hold-out validation	K-fold cross-validation	*Accuracy*	*Precision*	*Recall*	Area Under Curve	Balanced *Accuracy* Score	*F1-Score*	
1	Hashfi, Farhan et al (2022)	Sentiment Analysis of An Internet *Provider* Company Based on Twitter Using Support Vector Machine and Naïve Bayes Method	✓		✓	✓				✓				✓	✓	✓	✓		✓		✓	□	□	□	□	✓	✓	✓			✓	SVM = 84% NB = 82%
2	Rozaq, Abdul et all (2022)	Analisis Sentimen Terhadap Implementasi Program Merdeka Belajar Kampus Merdeka Menggunakan Naïve Bayes, K-Nearest Neighboars, dan Decision Tree	✓			✓	□	✓	✓		□	□		✓	✓		✓				✓	□	□	□	□	✓	□	□			□	NB = 99,22% KNN = 96,90% DT = 37,21%
3	Dangin, Renyta Kristianti; Febriyanto, Ferdy; Sari, Renny Puspita	Perbandingan Naïve Bayes dan Support Vector Machine pada Sentimen Analisis Reputasi Brand Twitter DQLab.id	✓	✓	✓	✓					✓	□	✓		✓	✓	✓	✓	□	□	□	□	□	□	□	✓	✓	✓	□	✓	✓	SVM = 88% NB = 75%

No.	Penulis (Tahun)	Judul	Analisis Sentimen		Metode					Preprocessing												Validasi Model			Evaluasi						Nilai Akurasi
	(2022)																														
4	Mustopa, Ali et all (2021)	Analysis of User Reviews for the PeduliLindungi Application on Google Play Using the Support Vector Machine and Naïve Bayes Algorithm Based on Particle Swarm Optimization	✓		✓	✓				✓			✓	✓	✓		✓								✓	✓	✓	✓			NB = 69% SVM = 93%
5	Rahmatulloh, Alam et all(2021)	Sentiment Analysis of Ojek Online User Satisfaction Based on the Naïve Bayes and Net Brand Reputation Method	✓	✓		✓					✓	✓	✓	✓	✓		✓	✓	✓	✓					✓	✓	✓		✓		NB Gojek = 99,80%, NBR = -58% NB Grab = 99,90%, NBR = -52%
6	Fitri, Evita et all (2020)	Analisis Sentimen Terhadap Aplikasi Ruangguru Menggunakan Algoritma Naïve Bayes, Random Forest dan Support Vector Machine	✓		✓	✓	✓						✓	✓		✓	✓	□	□	□	□		□	□	✓	✓	✓	✓	✓			NB = 94,16% RF = 97,16% SVM = 96,01%
7	Giovani, Angelina Puput et all (2020)	Analisis Sentimen Aplikasi Ruang Guru di Twitter Menggunakan Algoritma Klasifikasi	✓		✓	✓	□	✓	□		✓	□		✓	✓		✓				✓	□	□	□	✓	✓	✓	✓	✓		□	SVM = 78,55% NB 67,32% K-NN = 77,21%

No.	Penulis (Tahun)	Judul	Analisis Sentimen		Metode					Preprocessing												Validasi Model			Evaluasi						Nilai Akurasi
8	Fitri, Veny Amilia et al (2019)	Sentiment Analysis of Social Media Twitter with Case of LGBT Campaign in Indonesia using Naïve Bayes, Decision Tree and Random Forest Algorithm	✓		□	✓	✓		✓	✓				✓	✓		✓	✓	✓	□						✓	✓	✓			✓	NB = 83,43% DT = 82,91% RF = 82,91%
9	Virra, Khalisa et all (2019)	Sentiment Analysis of Social Media Users Using Naïve Bayes, Decision Tree, Random Forest Algorith: A Case Study of Draft Law on the Elimination of Sexual Violence (RUU PKS)	✓				✓		✓		✓	□			✓			✓								✓	✓	✓			✓	NB = 83,94% RF = 75,72% DT = 75,32%
10	Vidya, Nur Aziza; Fanany, Mohamad Ivan; Budi, Indra (2015)	Twitter Sentiment to Analyze Net Brand Reputation of Mobile Phone *Provider*s	✓	✓	✓	✓			✓		✓	□	✓	✓				✓			✓				✓	✓	✓	✓	✓		✓	SVM = 82,40% NB = 78,90% DT = 72,90%
11	Tugas Akhir	Perbandingan Naïve Bayes dan Random Forest pada Analisis Sentimen Reputasi Brand di Google Play *Provider* Live.On	✓	✓		✓	✓			✓	✓			✓	✓		✓	✓	✓	✓		✓			✓	✓	✓	✓			✓	NB = ? DT = ?

Berdasarkan matriks penelitian pada tabel 2.2, penelitian ini dilakukan untuk perbandingan metode Naïve Bayes dan Random Forest dengan data dan jumlah yang berbeda dari penelitian sebelumnya. Selain itu, pada tahap *preprocessing* diterapkan penggunaan *Slangwords* yang mengubah bahasa informal menjadi bahasa formal.