BAB II
TINJAUAN PUSTAKA

2.1 Kajian Pustaka

Tinjauan pustaka dalam penelitian ini berisi review jurnal penelitian terdahulu yang mengangkat tema serupa yaitu analisis sentimen terhadap produk *skincare*. *Review* jurnal pada penelitian ini terdiri dari tujuh jurnal nasional dan tiga jurnal internasional. Jurnal tersebut kemudian dirangkum dalam tabel tinjauan pustaka yang terdiri dari beberapa kolom seperti judul, *comparing, constrasting, criticize, synthesize dan summarize. Comparing* berisi kesamaan pada penelitian, *contrasting* berisi perbedaan penelitian, *criticize* berisi kritik terhadap penelitian terdahulu, *synthesize* berisi ide pada penelitian terdahulu dan *summarize* berisi ringkasan dari penelitian terdahulu. Tinjauan pustaka dilakukan bertujuan untuk mendapatkan gambaran terkait proses dan hasil penelitian. *Review* penelitian terdahulu dapat dilihat pada Tabel 2.1.

Penelitian terdahulu menjadi unsur yang penting dalam penelitian dikarenakan menjadi acuan oleh peneliti dalam melakukan dan menyusun penelitian. Dengan adanya penelitian terdahulu, maka peneliti dapat mengetahui keterkaitan dengan penelitian yang akan dilakukan nantinya. Selain itu, juga membantu untuk menghindari adanya duplikasi dari penelitian yang akan dilakukan.

Tabel 2. 1 Penelitian Terdahulu

No	Judul	*Compare*	*Contrast*	*Criticize*	*Synthesize*	*Summarize*
1.	Analisis Sentimen Aplikasi Identitas Kependudukan Digital (IKD) Menggunakan Metode *Naïve Bayes* [13]	Penelitian [13] melakukan analisis sentimen terhadap komentar masyarakat sejalan dengan penelitian yang sedang dilakukan.	Penelitian [13] hanya menggunakan metode *Naïve Bayes* untuk menganalisis sentimen, sedangkan pada penelitian yang sedang dilakukan menggunakan metode *Naïve Bayes* dan *Support Vector Machine* (SVM).	Penelitian [13] dapat membandingkan kinerja metode-metode klasifikasi yang lain untuk mendapatkan pemahaman yang lebih lengkap.	Penelitian [13] berfokus pada memahami pandangan publik terhadap aplikasi Identitas Kependudukan Digital (IKD) di sosial media YouTube, sedangkan pada penelitian yang akan dilakukan berfokus pada komentar pengguna YouTube terhadap konten promosi brand *Scarlett* yang melakukan kolaborasi dengan EXO.	Hasil penelitian [13] menunjukkan nilai akurasi, presisi, recall, dan F1 di atas 90%, yang menunjukkan keefektifan metode dalam melakukan analisis sentimen terhadap IKD.
2.	Analisis Sentimen Kebijakan Kampus Merdeka Menggunakan *Naive Bayes* dan Pembobotan Tf-Idf Berdasarkan Komentar Pada YouTube [10]	Penelitian [10] melakukan analisis sentimen terhadap komentar masyarakat pada YouTube sejalan dengan penelitian yang sedang dilakukan.	Penelitian [10] hanya menggunakan metode *Naïve Bayes* untuk menganalisis sentimen, sedangkan pada penelitian yang sedang dilakukan menggunakan metode *Naïve Bayes* dan *Support Vector Machine* (SVM).	Penelitian [10] selain hanya mengklasifikasikan sentimen menjadi positif dan negatif, penelitian selanjutnya dapat memperluas cakupan analisis sentimen dengan mencakup sentimen netral atau bahkan mengidentifikasi sentimen-sentimen spesifik seperti kepuasan, kekecewaan, kekhawatiran, atau harapan terkait	Penelitian [10] berfokus pada menganalisis sentimen komentar kebijakan kampus merdeka, sedangkan pada penelitian yang akan dilakukan berfokus pada komentar pengguna YouTube terhadap konten promosi brand *Scarlett* yang melakukan kolaborasi dengan EXO.	Hasil penelitian [10] menunjukkan bahwa penggunaan proses *text preprocessing*, pembobotan TF-IDF, dan klasifikasi menggunakan algoritme *Naive Bayes Classifier* dapat menghasilkan akurasi yang signifikan dalam klasifikasi sentimen positif

No	Judul	*Compare*	*Contrast*	*Criticize*	*Synthesize*	*Summarize*
				Kebijakan Kampus Merdeka.		dan negatif berdasarkan komentar-komentar di *platform* YouTube.
3.	*Implementation Of The Naive Bayes Classifier Algorithm For Classification Of Community Sentiment About Depression On YouTube* [14]	Penelitian [14] melakukan analisis sentimen terhadap komentar masyarakat pada YouTube sejalan dengan penelitian yang sedang dilakukan.	Penelitian [14] hanya menggunakan metode *Naïve Bayes* untuk menganalisis sentimen, sedangkan pada penelitian yang sedang dilakukan menggunakan metode *Naïve Bayes* dan *Support Vector Machine* (SVM).	Penelitian [14] disarankan untuk memperluas sumber data dengan mempertimbangkan *platform* media sosial lainnya seperti Twitter, Facebook, atau Instagram.	Penelitian [14] berfokus pada mengklasifikasikan sentimen masyarakat terhadap depresi berdasarkan komentar yang ditemukan di *platform* YouTube, sedangkan pada penelitian yang akan dilakukan berfokus pada komentar pengguna YouTube terhadap konten promosi brand *Scarlett* yang melakukan kolaborasi dengan EXO.	Hasil penelitian [14] menunjukkan bahwa dominasi sentimen positif sebesar 93,31%. Komentar negatif menyumbang sebanyak 6,68%, dan tingkat keakuratan yang dicapai sebesar 84,11%.
4.	Analisis Sentimen Penilaian Masyarakat Terhadap *Childfree* Berdasarkan Komentar di YouTube Menggunakan Algoritma *Naïve Bayes* [15]	Penelitian [15] melakukan analisis sentimen terhadap komentar masyarakat pada YouTube sejalan dengan penelitian yang sedang dilakukan	Penelitian [15] hanya menggunakan metode *Naïve Bayes* untuk menganalisis sentimen, sedangkan pada penelitian yang sedang dilakukan menggunakan metode *Naïve Bayes* dan *Support Vector Machine* (SVM).	Penelitian [15] dapat mengambil komentar YouTube dari luar negeri juga tidak hanya dari Indonesia saja dan dapat juga untuk menggunakan algoritma lain, seperti algoritma SVM (*Support Vector Machine*) dan algoritma *K-Nearest Neighbor*, dalam	Penelitian [15] berfokus pada menentukan respon *audiens* terhadap masalah *childfree* pada komentar di YouTube, sedangkan pada penelitian yang akan dilakukan berfokus pada komentar pengguna YouTube terhadap konten promosi brand *Scarlett* yang melakukan kolaborasi dengan EXO.	Hasil penelitian [15] menunjukkan bahwa sentimen penilaian masyarakat terhadap *childfree* berdasarkan komentar di YouTube cenderung negatif. Dalam analisis sentimen, ditemukan bahwa

No	Judul	*Compare*	*Contrast*	*Criticize*	*Synthesize*	*Summarize*
				menganalisis sentimen terkait masalah *childfree*.		persentase akurasi (*accuracy*) sebesar 97%, nilai presisi (*precision*) sebesar 98%, dan tingkat keberhasilan (*recall*) sebesar 96%. Hal ini menunjukkan bahwa mayoritas komentar yang dikumpulkan dari YouTube memiliki sentimen negatif terhadap *childfree.*
5.	Analisis Sentiment Masyarakat terhadap Kasus Covid-19 pada Media Sosial YouTube dengan Metode *Naive bayes* [16]	Penelitian [16] melakukan analisis sentimen terhadap komentar masyarakat pada YouTube sejalan dengan penelitian yang sedang dilakukan.	Penelitian [16] hanya menggunakan metode *Naïve Bayes* untuk menganalisis sentimen, sedangkan pada penelitian yang sedang dilakukan menggunakan metode *Naïve Bayes* dan *Support Vector Machine* (SVM).	Penelitian [16] kurangnya adanya perbandingan dengan algoritma analisis sentimen lainnya.	Penelitian [16] membahas mengenai menganalisis sentimen terhadap komentar masyarakat mengenai pemberitaan perkembangan kasus Covid-19 di kanal YouTube KompasTV, sedangkan yang sedang dikerjakan membahas mengenai komentar pengguna YouTube terhadap konten promosi brand *Scarlett* yang melakukan kolaborasi dengan EXO.	Hasil dari penelitian [16] yaitu memperoleh tingkat akurasi 74% dengan 361 komentar baik, 800 komentar negatif, dan 490 komentar netral.

No	Judul	*Compare*	*Contrast*	*Criticize*	*Synthesize*	*Summarize*
6.	Analisis Sentimen Pengguna YouTube Terhadap Tayangan #Matanajwamenantiterawan Dengan *Metode Naïve Bayes Classifier* [17]	Penelitian [17] melakukan analisis sentimen terhadap tayangan konten YouTube sejalan dengan penelitian yang sedang dilakukan.	Penelitian [17] hanya menggunakan metode *Naïve Bayes* Classifier untuk menganalisis sentimen, sedangkan pada penelitian yang sedang dilakukan menggunakan metode *Naïve Bayes* dan *Support Vector Machine* (SVM).	Penelitian [17] kurangnya adanya perbandingan dengan algoritma analisis sentimen lainnya.	Penelitian [17] membahas mengenai menganalisis sentimen pada komentar pengguna YouTube terhadap tayangan #MataNajwaMenantiTerawan sedangkan pada penelitian yang sedang dikerjakan membahas mengenai komentar pengguna YouTube terhadap konten promosi brand *Scarlett* yang melakukan kolaborasi dengan EXO.	Hasil penelitian [17] menunjukan bahwa klasifikasi sentimen komentar YouTube menggunakan *Naive Bayes* menghasilkan akurasi tinggi yaitu sebesar 90,36%.
7.	Analisis Sentimen Kenaikan Harga Kebutuhan Pokok dimedia Sosial YouTube Menggunakan *Algoritma Support Vector Machine* [12]	Penelitian [12] melakukan analisis sentimen terhadap tayangan konten YouTube sejalan dengan penelitian yang sedang dilakukan.	Penelitian [12] menggunakan metode *Support Vector Machine* (SVM) dalam melakukan klasifikasi komentar dan menggunakan metode SMOTE (*Synthetic Minority Over-sampling Technique*) untuk mengatasi ketidakseimbangan jumlah data antara label positif dan negatif., sedangkan pada penelitian yang sedang dilakukan	Penelitian [12] kurangnya adanya perbandingan dengan algoritma analisis sentimen lainnya.	Penelitian [12] berfokus pada menganalisis sentimen mengenai kenaikan harga kebutuhan pokok di media sosial YouTube, sedangkan pada penelitian yang sedang dikerjakan menganalisis sentimen komentar pengguna YouTube terhadap konten promosi brand *Scarlett* yang melakukan kolaborasi dengan EXO.	Hasil penelitian [12] menunjukkan bahwa tingkat akurasi yang diperoleh yaitu 86.33%, nilai presisi sebesar 75%, nilai recall sebesar 66.67%, dan nilai f1-score sebesar 70.59%.

No	Judul	*Compare*	*Contrast*	*Criticize*	*Synthesize*	*Summarize*
			menggunakan metode *Naïve Bayes* dan *Support Vector Machine* (SVM).			
8.	*An Evaluation of Preprocessing Steps and Tree-based Ensemble Machine Learning for Analysing Sentiment on Indonesian YouTube Comments* [18]	Penelitian [18] melakukan analisis sentimen terhadap tayangan konten YouTube sejalan dengan penelitian yang sedang dilakukan.	Penelitian [18] menggunakan metode *Naive Bayes* (NB), *Support Vector Machine* (SVM), *Decision Tree, Random Forest,* dan *Extra Tree Classifier* dalam melakukan klasifikasi, sedangkan pada penelitian yang sedang dilakukan menggunakan metode *Naïve Bayes* dan *Support Vector Machine* (SVM).	Penelitian [18] hanya menggunakan metrik akurasi sebagai metode evaluasi kinerja model.	Penelitian [18] menggunakan dataset dari komentar-komentar video YouTube tentang layanan pemerintah terkait pandemi COVID-19 di Indonesia, sedangkan pada penelitian yang akan dilakukan menggunakan dataset dari komentar pengguna YouTube mengenai konten promosi brand *Scarlett* yang melakukan kolaborasi dengan EXO.	Hasil penelitian [18] menunjukan bahwa model dengan akurasi maksimum 89,68% berhasil dicapai dengan menggunakan kombinasi pra-pemrosesan standar (termasuk penghapusan kata-kata penghubung, konversi kata slang dan emotikon, serta *stemming*), ekstraksi fitur menggunakan *count vectorizer*, dan model klasifikasi *Extra Tree.*

No	Judul	*Compare*	*Contrast*	*Criticize*	*Synthesize*	*Summarize*
9.	*Sentiment Analysis of Positive and Negative of YouTube Comments Using Naïve Bayes – Support Vector Machine (NBSVM) Classifier* [19]	Penelitian [19] melakukan analisis sentimen terhadap tayangan konten YouTube sejalan dengan penelitian yang sedang dilakukan.	Penelitian [19] menggunakan *Python* dengan *library* seperti NLTK (*Natural Language Toolkit*) untuk melakukan analisis sentimen, sedangkan pada penelitian yang akan dilakukan menggunakan *software Orange Data Mining.*	Penelitian [19] hanya menggunakan metode pendekatan pemrosesan bahasa alami (*natural language processing*) untuk memproses data	Penelitian [19] tidak mencantupkan fokus utama analisis sentimen yang dilakukan, sedangkan pada penelitian yang akan dilakukan berfokus pada komentar pengguna YouTube terhadap konten promosi brand *Scarlett* yang melakukan kolaborasi dengan EXO.	Hasil penelitian [19] menunjukan bahwa kombinasi *Naïve Bayes* dan *Support Vector Machine* menghasilkan tingkat akurasi yang lebih baik dan kuat dengan menghasilkan yang nilai uji kinerja yaitu presisi sebesar 91%, recall sebesar83% dan skor f1 87%
10.	*Sentiment Analysis of Reviews in Natural Language: Roman Urdu as a Case Study* [20]	Penelitian [20] melakukan analisis sentimen terhadap tayangan konten YouTube sejalan dengan penelitian yang sedang dilakukan.	Penelitian [20] menggunakan sembilan algoritma pembelajaran mesin untuk klasifikasi sentimen, yaitu *Naïve Bayes, Support Vector Machine, Regresi Logistik, K-Nearest Neighbors, Artificial Neural Networks, Convolutional Neural Network, Recurrent Neural Networks, ID3, dan Gradient Boost*	Penelitian [20] dapat memfokus pada penggunaan algoritma deep learning seperti LSTM (*Long Short-Term Memory*) atau *Transformer* untuk meningkatkan kinerja dalam analisis sentimen teks *Roman Urdu.*	Penelitian [20] berfokus pada ulasan 20 lagu dari industri musik Indo-Pakistan dalam teks *Roman Urdu,* sedangkan pada penelitian yang sedang dikerjakan menganalisis sentimen komentar pengguna YouTube terhadap konten promosi brand *Scarlett* yang melakukan kolaborasi dengan EXO.	Hasil penelitian [20] menunjukkan bahwa model klasifikasi menggunakan *Regresi Logistik* memiliki kinerja terbaik dalam analisis sentimen terhadap ulasan dalam teks *Roman Urdu* dengan nilai akurasi sebesar 92,25% dan akurasi validasi

No	Judul	*Compare*	*Contrast*	*Criticize*	*Synthesize*	*Summarize*
			Tree, sedangkan pada penelitian yang sedang dilakukan menggunakan metode *Naïve Bayes* dan *Support Vector Machine* (SVM).			silang sebesar 91,47%.

Berdasarkan tabel 2.1 terdapat kajian pustaka yang digunakan dalam penelitian ini, dapat disimpulkan bahwa persamaan pada penelitian terdahulu dengan penelitian yang akan dilakukan penulis yaitu untuk mengetahui komentar masyarakat Indonesia terhadap adanya penggunaan *Brand Ambassador* Korea pada produk *skincare* lokal khususnya pada produk *Scarlett Whitening*. Serta terdapat banyak perbedaan antara penelitian terdahulu dan penelitian yang akan dilakukan penulis seperti perbedaan lingkup permasalahan, perbedaan metode yang digunakan, dan perbedaan hasil penelitian tentunya. Adanya perbedaan tersebut maka berbeda pula fokus utama dari setiap penelitian – penelitian yang dilakukan, oleh karena itu penelitian ini memiliki kebaharuan dibandingkan penelitian sebelumnya dimana penelitian ini akan melakukan analisis sentimen pada ulasan komentar produk *Scarlett Whitening* dengan EXO sebagai *Brand Ambassador*.

2.2 Dasar Teori

Bagian ini mencakup pembahasan terkait dasar teori yang relevan dengan topik penelitian yang dilakukan. Dasar teori yang digunakan sebagai acuan dalam penelitian yang dilakukan yaitu:

2.2.1 Analisis Sentimen

Analisis sentimen adalah metode yang digunakan untuk merubah data opini,memahami serta mengolah tekstual data secara otomatis untuk melihat sentimen yang terkandung dalam sebuah opini [21]. Dengan menggunakan analisis sentimen, informasi yang awalnya tidak teratur dapat diubah menjadi data yang lebih terstruktur [22]. Tujuan penggunaan analisis sentimen yaitu untuk memetakan opini pengguna berdasarkan topik yang telah ditentukan. Selain itu juga untuk mendapatkan informasi yang lebih akurat dan detail dari ulasan pengguna dari berbagai perspektif dan aspek [23]. Analisis sentimen berfokus pada opini-opini yang mengekspresikan atau mengungkapkan sentimen positif atau negatif [24].

2.2.2 Skincare

Skincare, adalah suatu metode khusus untuk merawat kulit wajah dengan menggunakan beragam produk. *Skincare* memiliki peran penting dalam menjaga kesehatan dan memberikan nutrisi pada kulit. Hal ini karena upaya untuk meningkatkan penampilan tidak hanya tergantung pada penggunaan *makeup*, tetapi juga memerlukan perbaikan dan pencegahan masalah kulit yang umumnya dialami oleh wanita. Produk *skincare* dirancang sebagai solusi kecantikan untuk mengatasi berbagai masalah kulit, seperti jerawat, bekas jerawat, flek wajah, pemutihan kulit, perbaikan kulit kusam, dan penundaan penuaan dini. *Skincare* mencakup berbagai jenis produk, termasuk sabun wajah(*facial wash*), serum, *moisturizer*, *sunscreen*, dan masker wajah [25].

2.2.3 Scarlett Whitening

Scarlett Whitening adalah salah satu brand lokal Indonesia yang didirikan pada akhir 2017 oleh selebriti terkenal Indonesia, yaitu Felicya Angelista. Fokus utama produk ini, yang telah mendapatkan izin dari BPOM, adalah perawatan kulit tubuh dan wajah yang aman digunakan dalam kehidupan sehari-hari. *Scarlett*

Whitening memiliki tiga kategori produk utama, meliputi perawatan wajah, perawatan tubuh, dan perawatan rambut. Produk perawatan wajah mencakup *facial wash* dan *serum*, sedangkan perawatan tubuh terdiri dari *shower scrub*, *body lotion*, dan *body scrub*. Sementara itu, produk perawatan rambut terdiri dari *sea salt shampoo* and *conditioner* [26].

2.2.4 Brand Ambassador

Brand Ambassador adalah seseorang yang memiliki ketertarikan dan pemahaman mendalam terhadap suatu produk atau merek. Kehadiran *Brand Ambassador* dapat mempengaruhi sikap dan perilaku konsumen untuk membeli atau menggunakan produk tersebut. Perusahaan harus memilih *Brand Ambassador* yang tepat, yaitu sosok yang memiliki pengaruh positif dan sesuai dengan target konsumen. *Brand Ambassador* yang tepat dapat mempengaruhi konsumen untuk membeli produk perusahaan dan menjadi *trendsetter* [27]. Selebriti sering dipilih sebagai *Brand Ambassador* karena dipercaya memiliki pengaruh psikologis yang kuat terhadap konsumen. Selebriti dapat bertindak sebagai penyalur, pembicara, dan penghubung dalam sebuah iklan, sehingga dapat memperkenalkan produk atau jasa tersebut kepada konsumen secara lebih efektif. *Personality* dari *Brand Ambassador* akan berpengaruh terhadap *personality* dari merek tersebut dan nantinya akan mempengaruhi persepsi masyarakat akan Citra Merek dan dapat menarik konsumen untuk membeli [28].

2.2.5 Brand Image

Brand image adalah gambaran dari pandangan konsumen terhadap suatu produk yang mencerminkan kenangan pelanggan terhadap produk tersebut. Manajemen *brand image* yang efektif akan memberikan dampak positif dengan meningkatkan pemahaman tentang bagaimana perilaku pelanggan memengaruhi proses pengambilan keputusan [29]. *Brand image* memiliki pengaruh pada cara konsumen melihat dan merespons produk pada berbagai kesempatan. Cara untuk membedakan satu produk dari yang lain, terutama jika produk tersebut sejenis, diperlukan keunikan dalam produk tersebut, yang tercermin dalam *brand image* dengan produk yang dihasilkan harus memiliki kualitas yang baik. Keberhasilan

brand image memberikan keuntungan bagi konsumen karena membantu mereka mengenali manfaat dan kualitas produk. *Brand image* juga berperan dalam pengembangan suatu produk, karena reputasi dan kredibilitas yang terkandung dalam *brand image* menjadi faktor pertimbangan konsumen dalam memilih produk atau jasa [30].

2.2.6 YouTube

YouTube adalah sebuah *platform* situs web video sharing yang sangat populer di seluruh dunia dan menawarkan berbagai macam konten video, mulai dari film pendek, klip film, klip musik, hingga konten amatir seperti vlog, video pendek original, dan video pendidikan [31]. YouTube adalah layanan berbagi video dari Google yang memungkinkan penggunanya mengunggah dan menonton video secara gratis. YouTube memiliki koleksi video yang sangat besar dan beragam, sehingga dapat dikatakan sebagai database video paling populer di dunia internet [32].

2.2.7 Orange Data Mining

Orange adalah salah satu *software data mining* bersifat *open source* yang dapat digunakan untuk menganalisis dan memvisualisasikan data. *Sofware* ini memiliki fitur-fitur yang lengkap dan mudah dipahami. *Widget* yang disediakan dapat digunakan untuk berbagai macam tugas data mining, mulai dari eksplorasi data hingga prediksi. Hasil analisis juga ditampilkan secara jelas, sehingga pengguna dapat dengan mudah memahaminya [33].

2.2.8 Klasifikasi

Klasifikasi merupakan metode pengolahan data yang mengelompokkan objek ke dalam beberapa kelas sesuai dengan jumlah kelas yang diinginkan. Metode ini bertujuan untuk menemukan pola yang dapat memisahkan kelas data, memungkinkan identifikasi objek dalam kategori tertentu berdasarkan perilaku dan atribut dari kelompok yang telah ditentukan. Hasil dari klasifikasi dapat digunakan untuk membuat aturan dan memproses data baru. Dalam konteks klasifikasi, terdapat konsep taksonomi, yang awalnya muncul sebagai ilmu mengelompokkan makhluk hidup dan berkembang menjadi ilmu klasifikasi umum, termasuk prinsip-

prinsip klasifikasi. Dengan demikian, klasifikasi (taksonomi) merupakan proses menempatkan objek dalam kategori berdasarkan karakteristik masing-masing. Berbagai algoritma dapat digunakan untuk menangani masalah klasifikasi data, seperti *Decision Tree, K-Nearest Neighbor (KNN), Artificial Neural Network (ANN), Naive Bayes,* dan *Support Vector Machine* (SVM). Dalam penelitian ini, *Naive Bayes* dan *Support Vector Machine* (SVM) dipilih sebagai algoritma untuk melakukan klasifikasi [34].

2.2.9 *Naïve Bayes*

Naïve Bayes merupakan salah satu algoritma *machine learning* yang digunakan untuk melakukan klasifikasi. Algoritma ini banyak digunakan karena kemudahaan dalam menggunakannya dan sering menghasilkan tingkat akurasi yang sebanding daripada algoritma yang lain. Selain itu, *Naïve Bayes* juga dipandang sebagai algoritma yang efektif dan efisien. Algoritma *Naïve Bayes* terdiri dari dua tahapan yaitu pelatihan dan klasifikasi. Pada tahap pelatihan, kata dalam dataset akan diuji dataset pelatihan untuk menghitung probabilitas masing-masing kategori sentimen (positif dan negatif). Perlu mempelajari pola dalam dataset pelatihan terlebih dahulu sebelum melakukan klasifikasi. Selanjutnya, pada langkah klasifikasi, probabilitas setiap kategori sentimen dihitung terhadap data yang dimasukkan dan hasilnya adalah probabilitas dari setiap kategori sentimen yang diberikan pada dataset [35]. Secara umum, rumus *teorema Bayes* [36] adalah sebagai berikut:

$$P(A|B) = \frac{P(B|A)\ x\ P(A)}{P(B)} \quad (2.1)$$

Keterangan :

A : Hipotesis data pada kelas yang spesifik.

B : Data kelas yang belum diketahui.

$P(A|B)$: Probabilitas A terjadi dengan bukti bahwa B telah terjadi (probabilitas superior)

$P(B|A)$: Probabilitas B terjadi dengan bukti bahwa A telah terjadi

$P(A)$: Peluang terjadinya A

$P(B)$: Peluang terjadinya B

2.2.10 *Support Vector Machine* (SVM)

Support Vector Machine adalah algoritma pembelajaran mesin yang menggunakan fungsi *hyperplane* untuk memisahkan data menjadi daerah-daerah kelas. *Hyperplane* adalah fungsi yang digunakan sebagai pembatas antar kelas. Untuk memprediksi kelas suatu data, SVM akan melabeli data tersebut berdasarkan daerah kelas mana yang menjadi tempatnya. SVM biasanya digunakan pada kumpulan data besar yang diambil dari situs online dan menjadi populer karena penerapannya dalam klasifikasi teks. Prinsip SVM adalah membangun *hyperplane* yang memiliki ukuran margin yang sama dan tidak cenderung mendekati daerah dari salah satu kelas. Hal ini dapat dilakukan dengan mengukur margin dan kemudian mencari titik maksimumnya. Usaha pencarian *hyperplane* terbaik sebagai pembatas antar kelas merupakan inti dari metode SVM [37]. Dalam menangani kasus *nonlinier*, SVM dimodikasi untuk menyertakan fungsi *kernel* agar dapat menemukan hasil dengan cepat. Berdasarkan fungsi rumus yang biasa digunakan dalam SVM[38][39] adalah :

a) Polynomial Kernel

Polynomial adalah metode yang digunakan untuk mengklasifikasi dataset training yang sudah normal. Proses ini dapat diimplementasikan dengan persamaan berikut:

$$K\ (\ x.x' = (x.x' + c)' \qquad (2.2)$$

b) *Radial Bias Function* (RBF)

RBF adalah metode klasifikasi yang dapat digunakan untuk memisahkan data yang tidak dapat dipisahkan secara linear. Metode ini memiliki keunggulan akurasi yang tinggi, baik untuk data training maupun data prediksi. RBF dapat diimplementasikan menggunakan persamaan berikut:

$$K\ (\ x,x' = \exp(1 - \gamma||x - x'||^2) \qquad (2.3)$$

c) Signoid Kermel

Sigmoid adalah fungsi aktivasi yang digunakan dalam pengembangan jaringan saraf tiruan. Fungsi ini dapat diimplementasikan menggunakan persamaan berikut:

$$K\ (\ x,x') = \tanh\ (\alpha x.x' + \beta) \qquad (2.4)$$

2.2.11 Text Mining

Text Mining adalah penerapan konsep dan teknik data mining untuk mengidentifikasi pola dalam teks untuk menemukan informasi yang bermanfaat untuk tujuan tertentu. Penambahan teks dapat dianggap sebagai proses dua tahap yang pertama adalah penerapan struktur pada sumber data teks, dan tahap kedua melibatkan ekstraksi informasi dan pengetahuan yang relevan dari data yang telah terstruktur ini dengan menggunakan teknik dan alat yang sama dengan penambahan data [40]. *Text mining* mencakup seperti *kategorisasi*, pengelompokan teks, ekstraksi konsep/entitas, analisis sentimen, *document summarization*, dan *entity-relation modeling* [24].

2.2.12 Pre-processing Text

Pre-processing Text adalah tahap awal untuk mengolah data teks dengan cara menghilangkan data-data yang tidak relevan agar dapat dilakukan analisis sentimen[41]. Dalam penelitian ini, tahapan yang digunakan [42] antara lain :

1. *Transformation*

Transformation adalah proses mengubah data input untuk *transformasi* huruf kecil secara *default.* Beberapa proses *transformasi* yang dilakukan pada penelitian ini, yaitu :

a. *Lowercase,* berfungsi untuk mengubah semua huruf kapital menjadi huruf kecil
b. *Remove url,* berfungsi untuk menghapus *urls* yang ada pada teks

2. *Tokenization*

Tokenization adalah proses memecah teks menjadi komponen yang lebih kecil. Penelitian ini dilakukan dengan menerapkan *Regexp* atau *Regular Expression*, yaitu untuk memisahkan data menjadi kata tanpa menggunakan tanda baca seperti titik (.) dan koma (,).

3. *Normalization*

Normalization adalah proses memisahkan teks menjadi kata per kata yang dapat berdiri sendiri dalam sebuah kalimat. Dengan menggunakan normalisasi, makna teks tersebut dapat diketahui dengan menggunakan *Porter Stemmer,* yaitu algoritma yang menghilangkan akhiran *morfologis* dan *infleksional* yang lebih umum dari kata-kata dalam bahasa Inggris.

4. *Filtering*

Filtering adalah Proses menghapus data yang tidak penting dan mempertahankan data yang dianggap penting untuk mempermudah proses analisis. Fitur yang akan digunakan pada penelitian ini adalah :

a. *Stopword*, berfungsi untuk menghilangkan kata yang tidak diperlukan
b. *Numbers*, berfungsi untuk menghilangkan data angka
c. *Regexp*, berfungsi untuk menghapus tanda baca yang tidak penting dalam proses analisis seperti \.|,|:|;|!|\?|\(|\)|\||\+|'|"|‘|’|“|”|'|\’|…|\-|–|—|\$|&|*|>|<|\/|\[|\]

2.2.13 *Term Frequency Inverse Document Frequency* (TF-IDF)

TF-IDF adalah metode pembobotan yang sangat populer dan sering digunakan karena memiliki akurasi dan recall yang cukup tinggi. Metode pembobotan ini biasanya digunakan dalam *information retrieval*. Pada pembobotan TF-IDF, pentingnya sebuah kata dinilai berdasarkan frekuensi kemunculannya dalam sebuah dokumen. Jika kata tersebut sering muncul dalam sebuah dokumen, maka kontribusinya akan lebih besar, tetapi jika kata tersebut sering muncul dalam beberapa dokumen, kontribusinya akan lebih kecil. TF-IDF terdiri dari dua komponen, yaitu *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF) [43].

Term Frequency (TF) mengindikasikan seberapa sering suatu kata muncul dalam sebuah dokumen. Semakin banyak kemunculan kata tersebut dalam dokumen, semakin besar bobotnya atau nilai kesesuaiannya. $f(t_k, d_j)$ menunjukkan jumlah kemunculan kata k dalam dokumen j.

Rumus TF sebagai berikut :

$$TF(t_k, d_j) = f(t_k, d_j) \tag{2.5}$$

Dalam sebuah koleksi dokumen, term dapat didistribusikan secara acak dengan menggunakan IDF (*Inverse Document Frequency*). IDF mengukur seberapa sering suatu term muncul di banyak dokumen. Jika suatu term sering muncul dalam banyak dokumen, nilai IDF-nya akan kecil. Jumlah dokumen dalam dataset (D) akan dibagi dengan jumlah dokumen yang mengandung term tersebut, dinyatakan sebagai $df(t)$. Rumus untuk menghitung IDF adalah sebagai berikut:

$$IDF(t_k) = \log \frac{D}{df(t)} \quad (2.6)$$

TF-IDF dapat dirumuskan sebagai berikut:

$$TF\ IDF(t_k, d_j) = TF(t_k, d_j) * IDF(t_k) \quad (2.7)$$

2.2.14 K Fold Cross Validation

K-fold cross validation adalah teknik untuk memperkirakan kesalahan prediksi model pembelajaran mesin. Data dibagi menjadi beberapa bagian yang sama besar, kemudian model dilatih dan diuji sebanyak bagian tersebut. Di setiap pengulangan, satu bagian data digunakan sebagai data uji dan sisanya digunakan sebagai data latih [44]. Langkah-langkah *k-fold cross validation* adalah sebagai berikut:

1. Bagi data menjadi k bagian.
2. Untuk fold ke-1, gunakan bagian ke-1 sebagai data uji dan sisanya sebagai data latih. Hitung akurasi model berdasarkan data uji tersebut.

$$akurasi = \frac{\sum \text{data uji benar klasifikasi}}{\sum total\ data\ uji} \text{ X } 100 \quad (2.8)$$

3. Untuk fold ke-2, gunakan bagian ke-2 sebagai data uji dan sisanya sebagai data latih. Hitung akurasi model berdasarkan data uji tersebut.
4. Ulangi langkah 2 dan 3 hingga *fold* ke-k.
5. Hitung rata-rata akurasi dari k akurasi tersebut. Rata-rata akurasi ini adalah akurasi final model.

2.2.15 Confusion matrix

Confusion matrix adalah suatu metode yang diterapkan dalam konsep data mining untuk mengukur tingkat akurasi. Ada empat istilah yang digunakan untuk merepresentasikan hasil klasifikasi. Keempat istilah ini mencakup *True Positif* (TP), yang merujuk pada nilai positif yang berhasil dideteksi secara benar. *True Negatif* (TN) menyatakan jumlah data negatif yang berhasil dideteksi dengan benar. *False Positif* (FP) merujuk pada situasi di mana data yang sebenarnya negatif terdeteksi sebagai positif. Sementara itu, *False Negatif* (FN) mencerminkan kondisi ketika data yang sebenarnya negatif terdeteksi sebagai negatif [45]. Berikut adalah empat kemungkinan hasil di peroleh dari matrix:

Tabel 2. 2 Model *Confusion matrix* [46]

Correct Classification	***Classified as***		
	Predict +	*Predict -*	*Predict Netral*
Actual +	*True Positive(TP)*	*False Negative(FN)*	*False Netral1(FNt1)*
Actual -	*False Positive1(FP1)*	*False Negative1(FN1)*	*False Netral2(FNt2)*
Actual Netral	*False Positive2(FP2)*	*False Negative2(FNg2)*	*True Netral(TNt)*

Dalam *confusion matrix*, terdapat beberapa metrik evaluasi, yaitu *accuracy, precision, recall, dan F1-Score*. Berikut adalah kemungkinan *confusion matrix* [46]:

1. *Accuracy*, mengukur rasio prediksi yang benar terhadap keseluruhan data. Rumus *Accuracy* adalah sebagai berikut :

$$\frac{TP+TN}{TP+TN+FP+FN} \tag{2.9}$$

2. *Precision*, menghitung perbandingan nilai prediksi yang benar positif dengan total hasil yang diprediksi sebagai positif. Rumus *Precision* adalah sebagai berikut :

$$Positive = \frac{TP}{FP1+FP2+TP}$$
$$Negatif = \frac{TP}{FNg1+FNg2+TNg} \tag{2.10}$$
$$Netral = \frac{TP}{FNt1+FNt2+TNt}$$

3. *Recall*, mengukur perbandingan nilai prediksi benar positif dengan keseluruhan data yang sebenarnya positif. Rumus *Recall* adalah sebagai berikut :

$$Positive = \frac{TP}{FNg1+FNt1+TP}$$

$$Negatif = \frac{TNg}{FP1+FNt2+TNg} \quad (2.11)$$

$$Netral = \frac{TNt}{FP2+FNg2+TNt}$$

4. *F1-Score*, perbandingan rata-rata antara presisi dan recall yang dibobotkan. Rumus *F1-Score* adalah sebagai berikut :

$$2\ x\ \frac{Precision\ X\ Recall}{Precision+\ Recall} \quad (2.12)$$