2. LANDASAN TEORI

Pada bab ini akan dijelaskan mengenai teori-teori yang digunakan dalam pembuatan program analisis sentimen pada *review* atau komentar terhadap merek *skincare* lokal yang ditulis di Twitter. Teori-teori yang dijelaskan adalah mengenai *skincare*, *review*, sarkasme, analisis sentimen, *lexicon-based*, *Support Vector Machine*, dan *Random Forest*. Pada bab ini juga akan dijelaskan tinjauan studi dari penelitian sebelumnya yang berhubungan dengan penelitian ini.

2.1 Tinjauan Pustaka

2.1.1 *Skincare*

Skincare adalah produk perawatan kulit yang memiliki berbagai macam kegunaan, mulai dari menenangkan, memperbaiki dan memulihkan kulit, hingga melindungi kulit. Produk *skincare* digunakan terhadap kulit secara teratur. Namun, karena setiap orang memiliki permasalahan kulit yang berbeda-beda, maka rutinitas dan cara penggunaan *skincare* juga akan berbeda pada setiap orang. Tiga tahap dasar dalam penggunaan *skincare* atau perawatan kulit, sebagaimana dilansir oleh laman *Self dan New York Times*, melibatkan pembersihan (membersihkan wajah), pencerahan (menyamakan warna kulit), dan pelembaban (menghidrasi dan membuat kulit lebih lembut). Berdasarkan tiga dasar dari perawatan kulit, terdapat berbagai macam produk *skincare* antara lain pembersih wajah atau *facial cleanser*, *toner*, *serum*, *sunscreen,* dan pelembab atau *moisturizer* (*Mengenal Apa Itu Skincare Hingga Urutan Penggunaannya*, 2021).

2.1.2 *Review*

Review adalah suatu ulasan singkat yang dibuat berdasarkan analisis dan fakta, memberikan manfaat baik bagi penulisnya sendiri, pemilik barang atau jasa, maupun masyarakat umum. Dengan pesatnya perkembangan media, *review* kini dapat diberikan oleh siapapun melalui postingan pada media sosial miliknya. Tujuan utama dari melakukan *review* adalah memberikan gambaran dan informasi tentang suatu produk atau karya. Ulasan tersebut dapat mencakup kelebihan dan kekurangan produk yang disampaikan dalam bentuk kritikan dan komentar. *Review* tidak hanya bermanfaat bagi masyarakat atau calon pembeli, tetapi juga bagi para produsen produk atau penyedia jasa. Fokus utama dari

review adalah pada nilai guna suatu produk atau jasa dan bidang terkait lainnya (Badriyah, 2023).

2.1.3 Sarkasme

Sarkasme adalah nama lain dari sindiran, sehingga dapat diartikan bahwa sarkasme adalah sebuah ungkapan yang memiliki tendensi untuk mencemooh seseorang. Sarkasme juga merupakan suatu majas sindiran yang digunakan untuk mengungkapkan kebalikan dari makna sebenarnya. Umumnya, sarkasme bernada negatif dan sarkasme diungkapkan dengan menggunakan ungkapan positif untuk mengungkapkan sesuatu yang negatif. Sarkasme sangat bergantung pada nada bicara. Oleh karena itu, sarkasme biasanya lebih mudah dipahami ketika diucapkan. Namun, bukan berarti tidak bisa digunakan dalam tulisan (Gischa, 2023).

2.1.4 Analisis Sentimen

Analisis sentimen atau yang sering disebut sebagai *opinion mining*, merupakan suatu teknik dalam *Natural Language Processing* (NLP) yang bertujuan untuk menganalisis emosi yang terdapat dalam suatu teks. Proses analisis sentimen melibatkan pengidentifikasian kata dan frasa yang memiliki sifat positif, negatif, atau netral, serta pengamatan pola penggunaan kata-kata tersebut. Tujuan utama dari analisis sentimen adalah memahami opini, sikap, dan emosi yang diungkapkan oleh individu dalam teks, seperti ulasan, postingan media sosial, umpan balik pelanggan, survei, dan berbagai bentuk teks lainnya (*Apa Itu Analisis Sentimen? Pengertian Dan Contoh 2023 | RevoU*, n.d.).

Analisis sentimen dapat membantu perusahaan dalam memberikan evaluasi cepat mengenai data transkrip obrolan masukan pelanggan, komentar media sosial, dan ulasan produk guna untuk meningkatkan mutu layanan pelanggan dan meningkatkan reputasi merek. Cara kerja analisis sentimen dalam memperoleh data dapat dibagi menjadi tiga langkah yaitu klasifikasi, evaluasi, dan visualisasi hasil.

1. Klasifikasi

   Pertama, klasifikasi data yang dinilai sebagai opini dari sebuah teks. Ada tiga tipe klasifikasi dalam metode analisis sentimen yang dapat dilakukan, yaitu:

   - *Machine learning*: Fitur yang ada dapat mengenali sentimen atau sudut pandang seseorang dalam sebuah teks. Metode *machine learning* saat ini semakin populer dalam konteks pengenalan sentimen.
   - *Lexicon-based*: Memanfaatkan berbagai kata yang dinilai dengan skor polaritas untuk memahami tanggapan terhadap suatu topik. Kelebihan

dari metode ini adalah tidak memerlukan data latih, namun kelemahannya terletak pada ketidakmampuannya mengatasi kata-kata yang belum terdaftar dalam leksikon atau kamus.

- *Campuran*: Merupakan kombinasi antara metode *Machine Learning* dan *Lexicon-based*. Meskipun metode ini tidak sering digunakan, namun cenderung memberikan hasil yang lebih menjanjikan.

2. Evaluasi

   Setelah proses klasifikasi data, langkah selanjutnya dalam metode analisis sentimen adalah menggunakan matriks evaluasi seperti *Precision*, *Recall*, *F-score*, dan *Accuracy*. Proses ini juga mencakup pengukuran rata-rata seperti makro, mikro, dan *f1-score*. Pemilihan matriks evaluasi didasarkan pada keseimbangan klasifikasi *dataset*. Secara umum, prosedurnya melibatkan peninjauan *dataset*, *pre-processing*, *tokenizer*, penghapusan *stopwords*, transformasi, klasifikasi, dan evaluasi.

3. Visualisasi Data

   Tahap berikutnya dalam metode analisis sentimen adalah visualisasi data. Visualisasi data dilakukan dengan memanfaatkan model yang sesuai dengan kebutuhan perusahaan atau pihak yang menggunakan data tersebut. Banyak orang cenderung menggunakan teknik yang sudah umum dikenal, seperti grafik, histogram, atau matriks. Meskipun begitu, hasil akhir dari analisis sentimen dapat sangat bervariasi, terutama karena data yang tersedia seringkali terkait dengan domain lain. Oleh karena itu, teknik visualisasi data berupa *word cloud,* peta interaktif, dan gaya *sparkline* juga cukup efektif untuk menggambarkan hasil analisis dengan baik (*Analisis Sentimen (Sentiment Analysis) : Definisi, Tipe Dan Cara Kerjanya*, 2022).

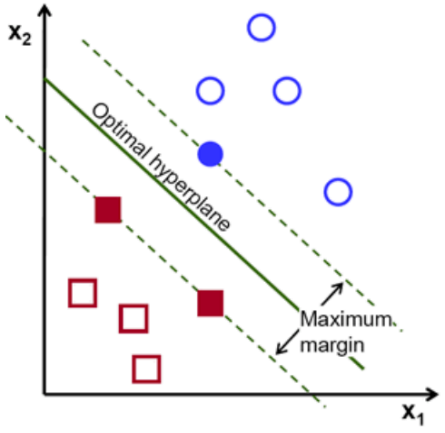
2.1.5 *Lexicon-based*

Lexicon adalah sebuah daftar kata atau kamus yang dipakai dalam suatu bahasa atau bidang khusus. Istilah ini umumnya diterapkan dalam *Natural Language Processing* (NLP), bagian dari kecerdasan buatan. *Lexicon* menggambarkan sekumpulan kata beserta nilai-nilai bobotnya yang menjadi elemen penting dalam berbagai tugas NLP, seperti analisis sentimen, klasifikasi teks, dan penerjemahan bahasa. (Mulyawan, 2023). Bobot pada kamus *Lexicon* memiliki nilai dengan rentang skor dari -5 sampai dengan +5. Data yang telah melalui tahap

preprocessing akan diklasifikasikan menggunakan kamus dari *Lexicon* dan akan diperoleh hasil klasifikasi berdasarkan hasil perhitungan skor dari data tersebut (Francesco, 2023).

2.1.6 *Support Vector Machine* (SVM)

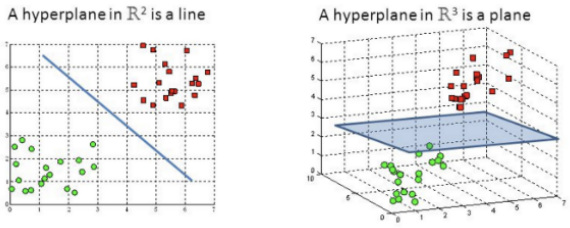
Support Vector Machine (SVM) adalah algoritma *machine learning* yang bergantung pada *supervised learning* dan dapat digunakan untuk menyelesaikan masalah regresi maupun klasifikasi. Cara kerjanya adalah dengan mencari batas keputusan atau *hyperplane* terbaik untuk memisahkan dua atau lebih kelas. Pada dua dimensi, *hyperplane* dapat berupa *line* atau garis, sedangkan pada *multiple plane*, dapat berupa *flat plane*.



Gambar 2.1 *Cara Kerja SVM*

Sumber: Trivusi. (2022, July 3). *Penjelasan Lengkap Algoritma Support Vector Machine (SVM)*. Trivusi. https://www.trivusi.web.id/2022/04/algoritma-svm.html

Hyperplane merupakan batas keputusan dalam SVM yang memisahkan dua kelas. Titik data di kedua sisi *hyperplane* dapat dikaitkan ke kelas yang berbeda. Dimensi *hyperplane* ditentukan oleh fitur-fitur dalam kumpulan data. Artinya, *hyperplane* adalah garis lurus dengan dua fitur, dan bidang dua dimensi dengan tiga fitur.



Gambar 2.2 *Ilustrasi Hyperplane*

Sumber: Trivusi. (2022, July 3). *Penjelasan Lengkap Algoritma Support Vector Machine (SVM)*. Trivusi. https://www.trivusi.web.id/2022/04/algoritma-svm.html

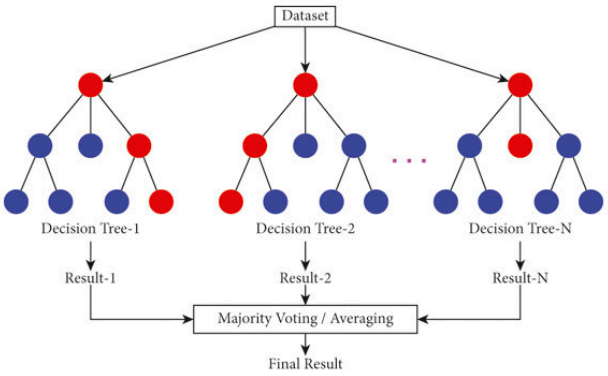
Algoritma SVM beroperasi dengan mengubah data ke dalam ruang fitur yang memiliki dimensi tinggi sehingga bisa dilakukan pengkategorian titik data, bahkan ketika pemisahan data tidak dapat dilakukan secara linier. Jika berhasil menemukan pemisah antara kategori-kategori tersebut, data dapat diubah sedemikian rupa. Selanjutnya, fitur-fitur dari data dapat digunakan untuk memprediksi ke dalam kelompok mana *record* baru sebaiknya ditempatkan (Trivusi, 2022).

2.1.7 *Random Forest*

Random Forest juga merupakan algoritma *machine learning* yang bergantung pada *supervised learning* dan dapat digunakan untuk menyelesaikan masalah regresi maupun klasifikasi. Algoritma ini menggabungkan sejumlah *decision tree* untuk mencapai hasil tertentu. Dibentuk dari banyak pohon (*tree*), algoritma ini dinamakan *'Forest'* atau *'Hutan'*. Setiap pohon dalam *Random Forest* memberikan prediksi kelas, dan prediksi kelas yang mendapat suara terbanyak dijadikan sebagai kandidat prediksi dalam model. Semakin banyak jumlah *tree*, tingkat akurasi akan semakin tinggi dan akan mencegah terjadinya *overfitting*.

Random Forest bekerja dalam dua fase yaitu menggabungkan sejumlah N *decision tree* untuk membuat *Random Forest* dan kemudian dilanjutkan dengan fase berikutnya yaitu

membuat prediksi untuk setiap *tree* yang pada fase pertama.



Gambar 2.3 *Cara Kerja Random Forest*

Sumber: Trivusi. (2022, September 17). *Algoritma Random Forest: Pengertian dan Kegunaannya*. Trivusi. https://www.trivusi.web.id/2022/08/algoritma-random-forest.html

Cara kerja algoritma Random Forest adalah dengan langkah-langkah sebagai berikut:

1. Algoritma memilih sampel acak dari *dataset* yang disediakan.
2. *Decision tree* akan disusun untuk setiap contoh yang terpilih, dan hasil prediksi akan diperoleh dari masing-masing *decision tree* yang telah dibuat.
3. *Voting* akan diterapkan pada setiap hasil prediksi. Pada permasalahan klasifikasi, metode modus (nilai yang paling sering muncul) akan digunakan, sementara pada permasalahan regresi, metode *mean* (nilai rata-rata) akan diadopsi.
4. Algoritma akan memilih hasil prediksi yang mendapatkan suara terbanyak (*vote* terbanyak) sebagai prediksi akhir. (Trivusi, 2022).

2.2 Tinjauan Studi

Bagian ini akan menjelaskan mengenai penelitian-penelitian berhubungan yang telah dilakukan sebelumnya. Beberapa penelitian tersebut akan dijadikan referensi dalam penelitian ini.

a. Analisis Sentimen terhadap *Brand Skincare* Lokal menggunakan *Naïve Bayes Classifier* (Putri et al., 2023)
    - Masalah yang diangkat adalah dimana penulis ingin mengetahui persepsi masyarakat dengan analisis sentimen terhadap data *tweet* dari *brand skincare*

lokal. Hal ini dikarenakan Twitter menjadi *platform* diskusi yang seringkali digunakan oleh masyarakat untuk mengekspresikan opini dan salah satu topik yang secara konsisten dibahas adalah mengenai perawatan kulit atau *skincare*. Dalam penelitian ini, *brand skincare* lokal yang digunakan sebagai sumber data adalah Avoskin, Azarine, dan Somethinc.

- Metode yang digunakan dalam penelitian ini adalah pengambilan data menggunakan *crawling dataset* dari Twitter dan hasil dari *dataset* dibagi menjadi 3 berdasarkan *brand skincare* dan masing-masing *dataset* terbagi lagi menjadi data latih (80%) dan data uji (20%). Dilakukan juga proses *preprocessing* yang meliputi *data cleaning*, *case folding*, *tokenizing*, penghapusan *stopword* dan *slang word*, serta *stemming*. Algoritma yang digunakan untuk analisis sentimen adalah *Naïve Bayes Classifier* untuk klasifikasi sentimen dengan pembobotan kata menggunakan metode TF-IDF.
- Perbedaan dengan skripsi ini adalah penelitian ini tidak terdapat deteksi sarkasme dan perbedaan metode yang digunakan untuk analisis sentimen.
- Hasil dari penelitian ini menunjukkan bahwa ketiga *brand* dalam sumber data memiliki hasil perhitungan klasifikasi mendapatkan nilai akurasi sebesar 79% untuk *dataset* Avoskin, 78% untuk *dataset* Azarine, dan 75% untuk *dataset* Somethinc. Pengujian dengan *k-fold cross validation* menghasilkan nilai sebesar 79% untuk *dataset* Avoskin dan Somethinc, serta 78% untuk *dataset* Azarine.

b. Analisis Sentimen terhadap Perkuliahan Daring di Indonesia dari Twitter *Dataset* menggunakan *InSet Lexicon* (Musfiroh et al., 2021)

- Masalah yang diangkat adalah penulis ingin melakukan klasifikasi sentimen terhadap respon dan suara masyarakat terhadap penerapan kebijakan baru yang muncul karena adanya pandemi COVID-19, yaitu kebijakan mengenai perkuliahan di universitas yang dilakukan secara daring. Penulis memilih Twitter sebagai *platform* yang digunakan untuk dijadikan fokus penelitian karena Twitter merupakan *platform* dimana masyarakat banyak menyuarakan pendapatnya.
- Metode yang digunakan adalah metode pengambilan data menggunakan API dari Twitter dan menggunakan kata kunci “kuliah daring”, dari hasil pengambilan data ini, didapatkan jumlah data sebanyak 6000 *tweet*. Namun, jumlah data ini dianggap kurang ideal karena dapat memperlambat proses

preprocessing, sehingga penulis membatasi jumlah data awal. Dalam klasifikasi sentimen, penelitian ini menerapkan metode *Lexicon-based Approach* yang mencocokkan masing-masing kata pada kalimat *tweet* dengan suatu *lexicon* untuk menghitung keseluruhan bobot kata untuk mendapatkan *polarity score*. Kalimat tweet diklasifikasikan sebagai kelas positif jika bobot *polarity score*-nya lebih besar dari 0, tergolong kelas negatif apabila *polarity score*-nya lebih kecil dari 0, dan tergolong netral jika *polarity score*-nya sama dengan 0.

- Perbedaan dengan skripsi ini adalah penelitian ini tidak terdapat deteksi sarkasme dan domain yang diteliti berbeda sehingga perlakuan saat proses analisis sentimen juga berbeda walaupun sama menggunakan metode *Lexicon-based*. Selain itu, pada penelitian ini juga hanya menggunakan InSet *Lexicon*, berbeda dengan skripsi ini yang mengkombinasikannya dengan SentistrengthID *Lexicon*.
- Hasil penelitian ini didapatkan kelas positif sebanyak 27.6%, kelas negatif sebanyak 63.6% dan kelas netral sebanyak 8.9%. Perhitungan tingkat akurasi diperoleh sebesar 79.2%, untuk *precision* 72.9%, *recall* 62.8% dan *f-measure* 67.4% dengan komposisi data latih 80% dan data uji 20%. Penulis juga memberikan saran-saran untuk penelitian selanjutnya, yaitu untuk memperkaya kamus kata, menangani singkatan dan bahasa tidak formal, mengecualikan *tweet* yang bukan kalimat opini misalnya iklan dan berita, dan mempertimbangkan penggunaan atau penghilangan emotikon sebagai indikator kecenderungan sentimen.

c. Pendeteksian Sarkasme pada Proses Analisis Sentimen menggunakan *Random Forest Classifier* (Alita & Isnain, 2020)

- Masalah yang diangkat dalam penelitian ini adalah kalimat sindiran atau sarkasme yang seringkali membuat opini salah dalam pengklasifikasiannya. Sehingga dalam klasifikasi analisis sentimen, dibutuhkan deteksi sarkasme. Dalam penelitian ini, digunakan sumber data dari opini-opini berbahasa Indonesia pada Twitter yang membahas mengenai pelayanan publik, khususnya terkait dengan jaringan telekomunikasi seluler, layanan BPJS dan layanan PLN.
- Metode yang digunakan adalah dalam pengambilan data penelitian ini menggunakan API yang disediakan oleh Twitter. Untuk *preprocessing* data dilakukan *case folding*, *tokenizing*, *stopword removal*, *convert emoticon*, dan

convert slang word. Selanjutnya dilakukan proses *Feature Extraction* dengan empat fitur yang digunakan yaitu *Unigram*, POS *Tagging*, TF dan TF-IDF. Kemudian, dilakukan proses analisis sentimen dengan deteksi sarkasme. Deteksi sarkasme dilakukan untuk data dengan polaritas positif dan pengujian deteksi sarkasme menggunakan metode *random forest classifier* dan diekstraksi dengan 4 fitur yaitu *sentiment relate feature*, *punctuation relate feature*, *lexical and syntactic feature*, dan *pattern relate feature*.

- Perbedaan dengan skripsi ini adalah untuk skripsi ini juga mencoba deteksi sarkasme menggunakan metode SVM dan juga mencoba vektorisasi dengan CountVectorizer, TF-IDF dan Word2Vec. Selain itu, untuk percobaan ekstraksi fitur pada skripsi ini juga berbeda dengan ekstraksi fitur yang ada pada penelitian ini.
- Hasil dari penelitian ini didapatkan bahwa pendeteksian sarkasme memberikan kenaikan nilai rata-rata akurasi dari analisis sentimen sebesar 16.61%, kenaikan nilai presisi sebesar 5.45%, kenaikan nilai *recall* sebesar 9.64% dan kenaikan nilai *f1-score* sebesar 11.27%.