BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Sebelumnya

Berdasarkan penelitian yang dilakukan, terdapat beberapa penelitian sebelumnya yang membahas tentang topik yang berkaitan dengan penelitian yang akan dilakukan. Fokus penelitian ini yaitu bagaimana algoritma *Support Vector Machine* dan *Naïve Bayes* dalam mengklasifikasikan komentar positif dan negatif.

1. **Analisis Sentimen Program Acara Di SCTV pada *Twitter* Menggunakan Metode *Naïve Bayes* Dan *Support Vector Machine* (Dery Anjas Ramadhan, Erwin Budi Setiawan S.Si., M.T : 2019)**

   Penelitian ini memiliki tujuan untuk menentukan kepuasan penonton dengan mengetahui opini pada media sosial *Twitter*. Sehingga perusahaan dapat membuat jadwal tayang yang baik serta memastikan berapa lama episode yang harus dibuat. Tidak hanya itu perusahaan dapat mengurangi kerugian karena jumlah penonton tidak sesuai dengan ekspetasi. Metode *Naïve Bayes* dan *Support Vector Machine* (SVM) merupakan metode yang digunakan dalam penelitian ini.

   Hasil pengujian antara *Naïve bayes* dan algoritma SVM menghasilkan akurasi tertinggi, yaitu *Support Vector Machine* sebesar 88,57%. Berdasarkan kategori didapatkan hasil akurasi sebesar 79,81 untuk kategori berita. 89,80% untuk hasil akurasi entertainment, 73,68% untuk hasil akurasi sinetron dan 87,74% untuk hasil akurasi FTV [11].

2. **Analisis Sentimen Terhadap Produk *The Body Shop Tea Tree Oil* (Abyan Ghiffarie, Ken Dheanis, Rayesha Putra : 2018)**

Penelitian ini bertujuan untuk melakukan klasifikasi terhadap produk The Body Shop Tea Tree Oil ke dalam kelas positif atau negatif pada situs ulasan produk kecantikan Female Daily. Data sampel yang digunakan sebanyak 2820 dengan perbandingan data training dan data testing 7:3.

Pada penelitian ini menggunakan metode *Naïve Bayes Classifier*. *Naïve Bayes* juga didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara kondisional saling bebas jika diberikan nilai output.

Hasil dari penelitian yang telah dilakukan adalah berupa jumlah komentar pengguna, *topic modelling*, dan *word cloud* dengan hasil tingkat akurasi model terhadap komentar poitif dan negatif pada situs Female Daily adalah 61,51%. Berdasarkan data dari analisis sentimen, *topic modelling* dan *word cloud* pada ulasan dari produk Tea Tree Oil didapatkan jumlah konsumen yang memberikan komentar positif lebih banyak dibandingkan dengan komentar negatif. Kata "jerawat" menjadi kata yang sering digunakan dalam komentar baik itu positif maupun negatif [12].

3. ***Analisis Sentimen Berbasis Aspek Pada Review Female Daily Menggunakan TF-IDF Dan Naïve Bayes (Clarisa Hasya, Adiwijaya, Said Al Faraby : 2021)***

   Penelitian ini menggunakan dataset *review* Female Daily untuk mengetahui sentiment berbasis aspek. Dataset *review* yang digunakan untuk menentukan tingkat sentimen komentar apakah bersifat positif, negatif atau netral yang sesuai dengan masing-masing aspek. Kemudian dari *review* yang telah diperoleh bersifat Bahasa yang *multilingual.* Bahasa Indonesia merupakan hasil yang diperoleh dari terjemahan Bahasa *multilingual*.

   Naïve Bayes Classifier merupakan metode yang digunakan penelitian ini dalam model *probabilistic* untuk menentukan proses klasifikasi berdasarkan aturan *bayes*. Pada proses pembobotan yang

paling sering digunakan adalah TF-IDF karena memiliki tingkat akurasi dan *recall* yang cukup tinggi oleh karena itu penelitian ini menggunakan TF-IDF.

Hasil dari penelitian ini yaitu menghasilkan nilai *F1-Score* sebesar 62,81%, untuk data yang diterjemahkan melalui tahapan ke dalam Bahasa Inggris selanjutnya diterjemahkan ke dalam Bahasa Indonesia dan tidak menggunakan *stopword removal* [13].

4. **Analisis Sentimen Transportasi Online Menggunakan *Support Vector Machine* Berbasis *Particle Swarm Optimization* (Valentino Kevin, Ade Iriani, Hindriyanto Dwi Purnomo : 2020)**

   Penelitian ini memiliki tujuan untuk mengetahui opini masyarakat mengenai penilaian terhadap transportasi *online* serta membandingkan tingkat akurasi pada SVM dan SVM-PSO menggunakan nilai parameter *default.* Data *Tweet* digunakan pada metode *scraping* menggunakan *octoparse*.

   Support Vector Machine (SVM) dan SVM *Optimasi Particle Swarm Optimization* (SVM-PSO) merupakan metode yang digunakan dalam penelitian ini. Dimana SVM merupakan metode *supervised learning* untuk melakukan klasifikasi. Sedangkan untuk PSO adalah salah satu Teknik optimasi untuk meningkatkan hasil akurasi.

   Hasil dari penelitian ini yaitu mengenai sentimen masyarakat terhadap transportasi online yang ada di Indonesia dengan hasil akhir nilai opini positif pada metode SVM sebesar 62% dan nilai akhir opini negatif sebesar 38%. Sedangkan nilai akhir pada metode SVM-PSO memperoleh nilai opini positif sebesar 53% dan negatif sebesar 47%. Berdasarkan hasil optimasi pada metode SVM menggunakan PSO lebih baik dibandingkan dengan SVM biasa [14]

5. **Analisis Sentimen Pengguna Gopay Menggunakan Metode *Lexicon Based Dan Support Vector Machine* (Rachmad**

Mahendrajaya, Ghulam Asrofi Buntoro, Moh. Bhanu Setyawan : 2022)

Tujuan dari penelitian ini untuk menganalisis opini masyarakat terkait Gopay pada media sosial Twitter. Metode yang digunakan dalam melakukan pelabelan yaitu menggunakan metode Lexicon Based. Sedangkan untuk mengklasifikasikannya menggunakan metode *Support Vector Machine*.

Hasil dari penelitian ini yaitu menghasilkan 923 opini positif, 287 opini negatif. sedangkan untuk hasil klasifikasinya mendapatkan nilai sebesar 89,17% untuk kernel linear dan 84,38% untuk kernel polynomial [15].

6. **Analisis Sentimen *Zoom Cloud Meetings* Di *Play Store* Menggunakan *Naïve Bayes* Dan *Support Vector Machine* (Nuraeni Herlinawati, Yuri Yuliani, Siti Faizah, Windu Gata, Samudi : 2020)**

Penelitian ini bertujuan untuk menganalisa label sentimen positif atau negatif pada ulasan *Play Store* di aplikasi *zoom.* Jumlah dataset yang digunakan sebanyak 1.007 data.

Metode yang digunakan dalam penelitian ini yaitu *Naïve Bayes* dan *Support Vector Machine* untuk menghasilkan akurasi terbaik dalam melakukan analisis sentimen pada ulasan dari para pengguna aplikasi *zoom cloud meetings* di *Google Play Store*.

Hasil yang didapatkan pada penelitian ini yaitu menghasilkan sentimen positif dan negatif sebanyak 546 dan 461 ulasan. Evaluasi model menggunakan 10 *cross validation* didapatkan hasil akurasi dan nilai AUC dari kedua algoritma yaitu *Naïve Bayes* akurasinya 74,37% dan AUC sebesar 0,659. Sedangkan algoritma SVM nilai akurasinya 81,22% dan AUC sebesar 0,886. Sehingga dapat disimpulkan bahwa tingkat akurasi SVM lebih besar dibandingkan dengan *Naïve Bayes* [16].

7. **Analisa Perbandingan Tingkat Performansi Metode *Support Vector Machine* Dan *Naïve Bayes Classifier* Untuk Klasifikasi Jalur Minat SMA (Oki Arifin, Theopilus Bayu Sasongko : 2018)**

   Penelitian ini bertujuan untuk membentuk model klasifikasi yaitu jalur minat SMA berdasarkan nilai yang dimiliki oleh peserta didik pada saat melakukan pendaftaran di salah satu Sekolah Menengah Atas. Model klasifikasi tersebut menggunakan perbandingan metode *Support Vector Machine* dan *Naïve Bayes* sehingga didapatkan analisis perbandingan terhadap tingkat performansi dengan cara membandingkan dalam dua buah dataset peminatan. Dataset yang digunakan yaitu dataset penjurusan abc yang berjumlah 288 peserta didik dan dataset penjurusan xyz yang berjumlah 280 peserta didik.

   Hasil dari penelitian ini menghasilkan hasil pengujian rata – rata (*mean*) untuk SVM hasil mean akurasi = 97.01, mean presisi = 99.03, *mean recall* = 95.41, AUC = 0.997. sedangkan untuk *Naïve Bayes* mean akurasi = 90.86, *mean presisi* = 91.33, *mean recall* = 91.77, AUC = 0.975. sehingga dapat disimpulkan bahwa algoritma SVM memiliki rata – rata performansi klasifikasi jalur minat yang lebih unggul daripada algoritma *Naïve Bayes* [17].

8. ***Text Mining* Untuk Analisis Sentimen Pelanggan Terhadap Layanan Uang Elektronik Menggunakan Algoritma *Support Vector Machine* (Fajar Romadoni, Yuyun Umaidah, Betha Nurina : 2020)**

   Penelitian ini memiliki tujuan untuk melakukan klasifikasi opini pengguna OVO dalam kategori positif dan negatif. Pengambilan data pada *twitter* dengan keyword @ovo_id dilakukan dengan cara *scraping* dan mendapatkan data sebesar 3852.

Metode algoritma *Support Vector Machine* digunakan dalam penelitian ini untuk proses klasifikasi dengan rasio perbandingan 6:4, 7:3, 8:2 dan 9:1. Dengan menggunakan empat kernel yaitu *kernel linear, RBF, sigmoid, dan polynomial.*

Hasil dari penelitian ini memperoleh nilai akurasi yang terbesar oleh *kernel linear* dengan perbandingan 90:10 yaitu sebesar 98,7% [18].

9. **Analisis Perbandingan Metode *TF-IDF* Dan *Word2Vec* Pada Klasifikasi Teks Sentimen Masyarakat Terhadap Produk Lokal Di Indonesia (Ivan Rifky, Ema Utami, Anggit Dwi : 2022)**

   Penelitian ini bertujuan untuk analisis sentiment guna mengetahui produk yang sedang digemari dan dibutuhkan masyarakat pada hasil ulasan produk *Marketplace*. Sehingga dapat dijadikan sebagai bahan pertimbangan sebelum membeli suatu produk berdasarkan *review* orang sebelumnya yang pernah membeli.

   Metode yang digunakan yaitu TF-IDF dan *Word2vec*. Sedangkan untuk klasifikasi menggunakan algoritma XGBoost. Hasil yang didapatkan pada penelitian ini yaitu menghasilkan nilai F1 score lebih tinggi 0.941% menggunakan kombinasi Word2vec + XGBoost dibandingkan dengan menggunakan TFIDF+XGBoost yaitu 0,940%. Sehingga dapat disimpulkan bahwa Word2vec lebih unggul karna dapat melihat hubungan semantic antar kata [19].

10. **Analisis Sentimen Pada *Review* Kosmetik Bahasa Indonesia Dengan Metode *Naïve Bayes* (Hendy Ardian, Sandy Kosasi : 2019)**

    Penelitian ini dilakukan untuk membangun sistem yang dapat mengklasifikasikan komentar positif atau komentar negatif terhadap suatu produk kosmetik. Metode *Naïve Bayes* digunakan pada penelitian ini kemudian dilakukan ekstrasi fitur TFIDF guna meningkatkan kinerja *Naïve Bayes*. .

Penelitian ini menghasilkan akurasi sebesar 82% yang didapakan menggunakan *Confusion Matrix*. Perhitungan akurasi sistem memiliki performa yang baik dengan perbandingan *data training* dan *data testing* 90:10. Sehingga diperoleh kesimpulan performa sistem yang dinyatakan semakin baik dengan banyaknya jumlah *data training* [20].

Pada Tabel 2.1 merupakan ringkasan yang dilakukan oleh penulis terhadap penelitian yang pernah dilakukan sebelumnya. Dalam ringkasan terdapat penjelasan mengenai judul penelitian, masalah yang diteliti, metode yang digunakan serta hasil yang dilakukan oleh beberapa peneliti terhadap penelitian yang pernah dilakukan.

Tabel 2.1 Penelitian Sebelumnya

No.	Judul	*Comparing*	*Contrasting*	*Criticize*	*Synthesize*	*Summarize*
1.	Analisis Sentimen Program Acara Di SCTV pada Twitter Menggunakan Metode Naïve Bayes Dan Support Vector Machine Oleh Dery dkk, 2019	Penelitian ini melakukan Analisis Sentimen pada twitter menggunakan metode Support Vector Machine dan Naïve Bayes dalam menentukan klasifikasi.	Melakukan klasifikasi menggunakan metode SVM untuk mengetahui kepuasan masyarakat terhadap program acara TV.	Penelitian ini dilakukan untuk memprediksi seberapa bagus program acara TV untuk memikat ketertarikan penonton pada saluran televisi tersebut. Sehingga bisa berinovasi menghadapi masa yang akan datang.	Menggunakan Support Vector Machine dan Naïve Bayes dalam melakukan perbandingan untuk mengetahui tingkat akurasi dari masing-masing metode.	Penelitian ini menghasilkan akurasi terbaik dari Metode SVM dengan seluruh program acara mendapatkan hasil akurasi 88,57%. Untuk hasil akurasi 79,81% diperoleh Berita. 89,80%diperoleh Entertainment, 73,68% diperoleh Sinetron, dan 87,74 diperoleh FTV.
2.	Analisis Sentimen Terhadap Produk *The Body Shop Tea Tree Oil* Oleh Abyan Ghiffarie, Ken Dheanis, Rayesha Putra 2018	Penelitian ini melakukan Analisis Sentimen pada produk The Body Shop Tea Tree Oil berdasarkan ulasan pengguna yang terdapat pada situs femaledaily.com menggunakan metode Naïve Bayes.	Melakukan klasifikasi menggunakan metode Naïve Bayes untuk mengetahui tingkat kepuasan pengguna terhadap produk The Body Shop Tea Tree Oil.	Penelitian ini dilakukan untuk mengetahui tingkat kepuasan pengguna terhadap produk The Body Shop Tea Tree Oil sehingga bisa dijadikan bahan evaluasi untuk produk The Body Shop supaya bisa lebih baik lagi.	Menggunakan Naïve Bayes dalam melakukan klasifikasi untuk mengetahui tingkat kepuasan pengguna terhadap produk The Body Shop Tea Tree Oil.	Penelitian ini menghasilkan tingkat akurasi sebesar 61,51%. Berdasarkan data dari analisis sentiment, *topic modelling* dan *word cloud* pada ulasan dari produk *Tea Tree Oil* didapatkan jumlah konsumen yang memberikan komentar positif lebih banyak dibandingkan dengan komentar negatif. Kata “jerawat” menjadi kata yang sering digunakan dalam komentar baik itu positif maupun negatif.
3.	Analisis Sentimen Berbasis Aspek Pada Review	Penelitian ini melakukan analisis sentimen berbasis aspek pada review Female	Melakukan klasifikasi menggunakan metode Naïve Bayes serta	Penelitian ini dilakukan untuk mengetahui review Female Daily supaya	Menggunakan Naïve Bayes dalam melakukan	Penelitian ini menghasilkan nilai F1-Score sebesar 62,81%, untuk data yang diterjemahkan ke dalam Bahasa Indonesia

No.	Judul	*Comparing*	*Contrasting*	*Criticize*	*Synthesize*	*Summarize*
	Female Daily Menggunakan TF-IDF Dan *Naïve Bayes* Oleh Clarisa Hasya, Adiwijaya, Said Al Faraby 2021	Daily menggunakan metode Naïve Bayes	TFIDF untuk melakukan pembobotan kata	dapat menjadi sebuah informasi yang berharga bagi konsumen.	klasifikasi, serta complement naïve bayes untuk mengatasi data yang tidak seimbang.	dari sebelumnya Bahasa Inggris dan tidak menggunakan *stopword removal.*
4.	Analisis Sentimen Transportasi Online Menggunakan *Support Vector Machine* Berbasis *Particle Swarm Optimization* Oleh Valentino Kevin, Ade Iriani, Hindriyanto Dwi 2020	Penelitian ini bertujuan untuk mengetahui sentimen masyarakat terhadap transportasi online menggunakan metode SVM dan SVM-PSO.	Melakukan klasifikasi menggunakan metode SVM dan SVM-PSO untuk mengetahui sentiment masyarakat terhadap transportasi online.	Penelitian ini dilakukan untuk mengetahui sentimen masyarakat terhadap transportasi online sehingga dapat dijadikan bahan evaluasi bagi perusahaan.	Menggunakan metode SVM dan SVM-PSO dalam melakukan perbandingan untuk mengetahui metode mana yang lebih baik.	Penelitian ini menghasilkan hasil sentimen masyarakat terhadap transportasi online di Indonesia adalah positif dengan hasil opini positif pada metode SVM sebesar 62% dan negatif sebesar 38%, sedangkan pada metode SVM-PSO opini positif sebesar 53% dan negatif sebear 47%. Berdasarkan hasil tersebut didapatkan bahwa metode SVM yang sudah dioptimasi menggunakan PSO lebih baik dibanding SVM biasa
5.	Analisis Sentimen Peengguna Gopay Menggunakan Metode *Lexicon Based* Dan *Support Vector Machine* Oleh Rachmad Mahendrajaya, Ghulam Asrofi, Moh. Bhanu 2022	Penelitian ini bertujuan untuk mengetahui sentimen masyarakat terhadap pengguna layanan Gopay melalui Twitter menggunakan metode *Lexicon Based* dan *Support Vector Machine.*	Melakukan klasifikasi menggunakan metode Lexicon Based dan SVM untuk mengetahui sentimen masyarakat.	Penelitian ini dilakukan untuk mengetahui kelas sentimen masyarakat terhadap layanan Gopay dengan jumlah data yang diambil 1210 dan dibatasi pada 24 juli 2019 sampai 30 juli.	Menggunakan metode Lexicon Based dan Support Vector Machine dalam melakukan klasifikasi.	Penelitian ini menghasilkan pelabelan sebesar 923 opini positif, 287 opini negatif, dan klasifikasi SVM menggunakan kernel *linear* sebesar 89,17%, dan 84,38% menggunakan kernel *polynomial.*
6.	Analisis Sentimen *Zoom Cloud Meetings* Di *Play Store* Menggunakan *Naïve Bayes* Dan *Support Vector Machine* Oleh	Penelitin ini bertujuan untuk mengetahui sentiment masyarakat terhadap aplikasi Zoom Cloud Meetings di Play Store menggunakan metode Naïve Bayes dan Support Vector Machine.	Melakukan klasifikasi menggunakan metode Naïve Bayes dan Support Vector Machine untuk mengetahui sentimen pengguna.	Penelitian ini dilakukan untuk menganalisa label sentimen positif atau negatif pada ulasan para pengguna aplikasi zoom di Google Play Store.	Menggunakan metode Naïve Bayes dan SVM dalam melakukan klasifikasi.	Penelitian ini didapatkan hasil akurasi dan nilai AUC dari kedua algoritma yaitu *Naïve Bayes* akurasinya 74,37% dan AUC sebesar 0,659. Sedangkan algoritma SVM nilai akurasinya 81,22% dan AUC sebesar 0,886. Sehingga dapat disimpulkan bahwa tingkat akurasi SVM lebih besar dibandingkan dengan *Naïve Bayes.*

No.	Judul	*Comparing*	*Contrasting*	*Criticize*	*Synthesize*	*Summarize*
	Nuraeni Herlinawati, Yuri Yuliani, Siti Faizah, Windu Gata, Samudi 2020					
7.	Analisa Perbandingan Tingkat Performansi Metode *Support Vector Machine* Dan *Naïve Bayes Classifier* Untuk Klasifikasi Jalur Minat SMA Oleh Oki Arifin, Theopilus Bayu Sasongko 2018	Penelitian ini bertujuan untuk mengklasifikasi jalur minat SMA dan mengetahui hasil akurasi dari kedua metode yaitu SVM dan Naïve Bayes.	Melakukan klasifikasi menggunakan metode SVM dan Naïve Bayes untuk mengetahui akurasi dari performansi metode SVM dan Naïve Bayes.	Penelitian ini dilakukan untuk mengklasifikasi jalur minat SMA dan untuk mengetahui hasil akurasi dari kedua metode tersebut.	Menggunakan metode SVM dan Naïve Bayes dalam melakukan klasifikasi.	Hasil dari penelitian ini menghasilkan hasil pengujian rata – rata (*mean*) untuk SVM hasil mean akurasi = 97.01, mean presisi = 99.03, mean recall = 95.41, AUC = 0.997. sedangkan untuk *Naïve Bayes mean* akurasi = 90.86, mean presisi = 91.33, *mean recall* = 91.77, AUC = 0.975. sehingga dapat disimpulkan bahwa algoritma SVM memiliki rata – rata performansi lebih unggul daripada algoritma *Naïve Bayes*.
8.	*Text Mining* Untuk Analisis Sentimen Pelanggan Terhadap Layanan Uang Elektronik Menggunakan Algoritma *Support Vector Machine*. Oleh Fajar Romadoni, Yuyun Umaidah, Betha Nurina 2020	Penelitian ini bertujuan untuk mengklasifikasikan sentimen masyarakat terhadap pengguna layanan uang elektronik OVO melalui Twitter menggunakan metode *Support Vector Machine.*	Melakukan klasifikasi menggunakan metode SVM.	Penelitian ini dilakukan untuk mengklasifikasi sentimen para pelanggan terhadap layanan uang elektronik OVO ke dalam kelas positif dan negatif.	Menggunakan metode SVM dalam melakukan klasifikasi.	Hasil dari penelitian ini didapatkan nilai akurasi yang terbesar oleh kernel linear dengan perbandingan 90:10 yaitu sebesar 98,7%
9.	Analisis Perbandingan Metode TFIDF Dan *Word2Vec* Pada Klasifikasi Teks Sentimen Masyarakat Terhadap Produk	Penelitian ini bertujuan untuk mengklasifikasikan review produk dari orang yang melakukan pembelian ke dalam kelas positif dan negatif menggunakan Metode Tf-Idf dan Word2vec	Melakukan klasifikasi *menggunakan TFIDF* Dan Word2Vec	Penelitian ini bertujuan untuk mengklasifikasikan review produk local di Indonesia	Melakukan perbandingan Metode Tf-Idf dan Word2vec dalam melakukan klasifikasi.	Hasil yang didapatkan yaitu nilai F1 score lebih tinggi 0.941% menggunakan kombinasi Word2vec + XGBoost dibandingkan dengan menggunakan TFIDF+XGBoost yaitu 0,940%. Sehingga dapat disimpulkan bahwa Word2vec lebih unggul karna dapat melihat hubungan semantic antar kata.

No.	Judul	*Comparing*	*Contrasting*	*Criticize*	*Synthesize*	*Summarize*
	Lokal Di Indonesia. Oleh Ivan Rifky, Ema Utami, Anggit Dwi 2022					
10.	Analisis Sentimen Pada Review Kosmetik Bahasa Indonesia Dengan Metode *Naïve Bayes.* *Oleh* Hendy Ardian, Sandy Kosasi 2019	Penelitian ini bertujuan untuk mengklasifikasikan review produk ke dalam kelas positif dan negatif menggunakan metode *Naïve Bayes.*	Melakukan klasifikasi menggunakan *Naïve Bayes* untuk mengetahui klasisifikasi dari review produk kosmetik.	Penelitian ini bertujuan untuk mengklasifikasikan review ke dalam kelas positif dan negatif menggunakan metode *Naïve Bayes*.	Menggunakan metode Metode Naive Bayes dalam melakukan klasifikasi	Hasil dari penelitian ini didapatkan akurasi dengan menggunakan confusion matrix sebesar 82%. Dari perhitungan akurasi diketahui sistem memiliki performa yang baik dengan menggunakan perbandingan data training dan data testing 90:10.

Berdasarkan penelitian sebelumnya dapat disimpulkan bahwa beberapa penelitian tidak menggunakan *stopword removal* sehingga mempengaruhi dalam proses pelabelan dan nilai akurasi. Untuk penelitian sebelumnya juga menggunakan pelabelan manual dan menggunakan pelabelan otomatis dengan mentranslate data terlebih dahulu dari Bahasa inggris kedalam Bahasa Indonesia sehingga membutuhkan waktu yang cukup lama. Maka dari itu pada penelitian ini dilakukan dengan menggunakan metode pelabelan otomatis dengan Lexicon Based. Sehingga perbedaan dengan penelitian yang akan dilakukan yaitu menggunakan penggunaan keempat skenario yaitu ada TFIDF tanpa stemming, TFIDF dengan stemming, Word2vec tanpa *stemming*, *Word2vec* dengan stemming. Penggunaan stemming pada penelitian ini dilakukan karena mempengaruhi akurasi pada masing-masing metode.

2.2 Landasan Teori

2.2.1 Lacoco

Lacoco merupakan produk dari Indonesia yang didirikan pada tahun 2017. Arti nama Lacoco sendiri memiliki filosofi "sebatang pohon untuk segala manfaat". Setiap produk mengandung bahan alami berkualitas tinggi, dan dibuat untuk semua orang yang ingin mendapatkan kulit tampak muda, sehat, dan berseri. Berdasarkan penelitian dan eksperimen yang telah dilakukan Lacoco menghasilkan berbagai macam produk kecantikan yang diformulasikan dengan bahan terbaik dari alam. Berikut adalah produk Lacoco Watermelon Glowmask.

Gambar 2.1 Lacoco Watermelon Glowmask [21]

2.2.2 Klasifikasi

Klasifikasi merupakan tahapan pengelompokkan suatu objek dengan karakteristik yang sama pada kategori. Pada dasarnya klasifikasi dilakukan dengan menentukan ciri-ciri dengan kalimat yang penting [22]. *Data training* dan *data testing* yang digunakan sebagai pembelajaran dan pengujian dalam melakukan klasifikasi. Beberapa metode yang digunakan dalam melakukan klasifikasi yaitu *Naïve Bayes, Decision Tree, ANN, K-Nearest Neighbor*, dan *Support Vector Machine* [23].

2.2.3 Analisis Sentimen

Analisis Sentimen adalah proses mencerna atau memahami suatu data, serta mengekstrak data yang berbentuk teks guna memperoleh informasi yang berupa sikap ataupun pendapat seseorang terhadap suatu topik. Hasil dari analisis yaitu berupa kategori positif dan negatif [24]. Hal tersebut dilakukan untuk menganalisis opini atau pendapat seseorang, mengekstrak informasi terhadap suatu entitas seperti layanan, produk, serta topik tertentu [25].

2.2.4 Female Daily

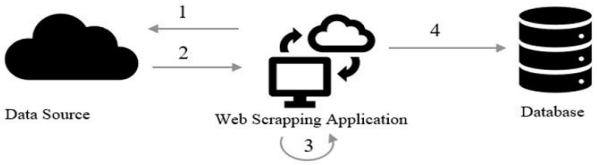
Female Daily merupakan situs No. 1 di Indonesia yang menyediakan ulasan dari berbagai macam produk kecantikan [26]. Female Daily awalnya hanya blog pribadi yang berisi mengenai berbagai macam konten *fashion* dan kecantikan yang dikelola oleh Hanifa Ambara pada tahun 2005 dan pada tahun 2007 Affi Assegaf bergabung dengan Hanifa Ambara untuk mengembangkan blog Female Daily, setelah mengamati potensi pengembangkan blog pada tahun 2009 Female Daily tidak hanya membahas kecantikan dan *fashion* tapi membahas *parenting*, tema keluaga, dan *shopping* [27].

2.2.5 Sociolla

Sociolla merupakan *online store* yang berdiri sejak Maret 2015 dan didirikan oleh Chrisanti Indiana di Indonesia yang memberikan kenyamanan pada tiap konsumen untuk belanja produk kecantikan, parfum, kosmetik, *skincare* dll karena terpercaya dan terlengkap [28]

2.2.6 *Web Scraping*

Web scraping merupakan Teknik pengambilan data dengan cara mengambil data *review* dari website target umumnya berupa halaman website seperti HTML. Fokus dari *web scraping* sendiri adalah untuk mengubah data yang tidak terstruktur menjadi lebih terstruktur [29].



Gambar 2.2 Proses *Web Scraping* [29].

2.2.7 Preprocessing

Preprocessing adalah tahapan untuk mempersiapkan dokumen dalam data mentah sehingga siap untuk dilakukan analisis sesuai dengan algoritma yang akan digunakan. *Preprocessing* bertujuan untuk meningkatkan hasil akurasi dari proses klasifikasi. Beberapa Teknik yang digunakan dalam melakukan *preprocessing* yaitu *normalisasi, case folding, tokenisasi, stopword removal* dan *stemming*. Untuk penjelasannya akan ditampilkan pada Tabel 2.2.

Tabel 2.2 Penjelasan *Preprocessing* [30].

Proses	Penjelasan
Normalization	Yaitu membersihkan data dari link URL, tanggal, dan lain-lain.
Case Folding	Yaitu mengubah huruf besar menjadi huruf kecil.
Tokenization	Yaitu kumpulan kalimat dipecah ke bagian terkecil.
Stopword Removal	Yaitu menghilangkan kata-kata yang sering muncul tanpa memberi arti yang signifikan.
Stemming	Yaitu merubah kata yang berimbuhan menjadi kata dasar.

2.2.8 Kamus Lexicon Inset

Kamus Lexicon Inset merupakan metode yang digunakan untuk melabeli data pada kelas tertentu dalam analisis sentimen. Metode ini memiliki cara kerja dengan menggunakan sebuah *corpus* atau kamus yang dilengkapi dengan bobot di setiap

katanya sebagai sumber bahasa atau leksikal. Hasil dari analisis menggunakan *lexicon* akan menjadi kategori positif, negatif, dan netral [31].

2.2.9 Term Frequency – Inverse Document Frequency (TF-IDF)

TF-IDF adalah metode pembobotan kata yang digunakan sebagai pembanding terhadap metode pembobotan baru. Pada metode ini, dilakukan perkalian antara nilai *term frequency* dengan nilai *inverse document frequency* yang menghasilkan bobot term(t) [32]. Pada *Term Frequency* (TF) yaitu merupakan pemberian bobot di setiap term yang muncul dalam suatu dokumen. Notasi dari *Term Frequency* yaitu $tf_{t,d}$. Sedangkan *Document Frequency* merupakan banyaknya dokumen yang mengandung istilah t yang dinotasikan dft. Untuk menghitung *Invers Document Frequency* (IDF) adalah sebagai berikut:

$$idf_t = log \frac{N}{dft} \quad (2.12)$$

Dimana :

idf_t : Nilai idf dari istilah t

N : Banyaknya dokumen yang ada

dft : Banyaknya jumlah kemunculan t dalam d

Inverse Document Frequency (idf) dihitung dengan menggunakan formula:

$$tfidf_{t,d} = tf_{t,d} \times idf_t \quad (2.13)$$

Dimana:

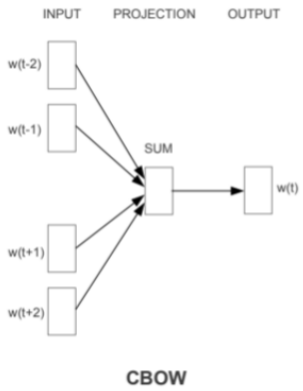
$tfidf_{t,d}$: bobot TFIDF dari t dalam dokumen d

$tf_{t,d}$: frekuensi munculnya t dalam dokumen d

idf_t : Nilai idf dari t

2.2.10 Word2Vec

Word2Vec adalah algoritma yang digunakan untuk mengubah kata menjadi vektor menggunakan *neural network* untuk mempelajari embedding kata. Teks yang diubah menjadi vektor supaya mesin dapat lebih memahami makna setiap kata lebih akurat. Sehingga kata-kata yang memiliki arti yang cukup mirip akan memiliki *output* yang berdekatan antara satu sama lain. *Input* dan *output* dalam proses *Word2Vec* ini ialah kumpulan vektor. *Pada penelitian ini Word2Vec* yang digunakan yaitu menggunakan model arsitektur *Continuos Bag Of Word* (CBOW) [33].

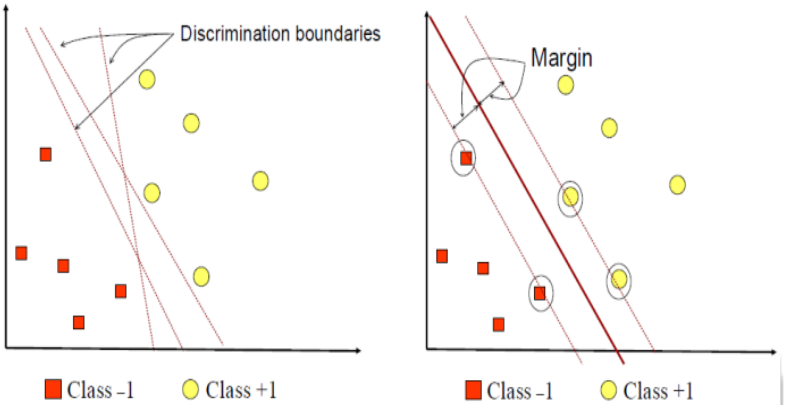


Gambar 2.3 Gambar Arsitektur *CBOW* [34].

2.2.11 Support Vector Machine (SVM)

Support Vector Machine (SVM) adalah metode *supervised learning* yang digunakan dalam mengklasifikasi data untuk menemukan *hyperplane* terbaik dengan membagi kelas pada *input space*. *Support Vector Machine* memiliki prinsip dasar yaitu pengklasifikasian linear [35]. Hal pertama yang mendasari untuk memahami klasifikasi dengan SVM adalah mencari garis (*hyperplane*) yang optimal. Hal ini untuk memisahkan dua kelas data yang berbeda, yaitu positif (+1) dan negatif (-1). Pada gambar 2.1 untuk data positif (+1) ditandai dengan warna

kuning dan data negatif (-1) disimbolkan dengan warna merah. Secara umum, penggambaran pada proses SVM dapat dilihat dalam gambar 2.4. Grafik sebelah kiri pada gambar 2.4 menggambarkan mengenai kemungkinan garis pemisah (*discrimination boundaries*) pada SVM untuk membuat *dataset.* Sedangkan pada grafik sebelah kanan menggambarkan *discrimination boundaries* dengan margin maksimum. Margin atau juga yang disebut juga batas pemisah adalah jarak antara dua kelas data terdekat pada bidang *hyperplane. Hyperplane* dengan margin terbaik yang akan menghasilkan generalisasi untuk mendapatkan hasil klasifikasi yang lebih baik.



Gambar 2.4 Proses SVM dalam Menemukan *Hyperplane* [36]

Data yang tersedia dilambangkan dengan x ∈ R d, sedangkan label terkait dilambangkan dengan yi ∈ {-1 +1} untuk I = 1, 2, …., 1, dimana 1 adalah jumlah dara. Diasumsikan bahwa kedua kelas -1 dan +1 dapat dipisahkan sepenuhnya oleh bidang *hyper* dengan dimensi d yang ditentukan. Diasumsikan bahwa kedua kelas -1 dan +1 dapat dipisahkan sepenuhnya oleh *hyperplane* berdimensi d, yang didefinisikan :

$$w.x + b = 0 \qquad (2.1)$$

Pola xi milik kelas -1 (sampel negative) dapat dirumuskan sebagai pola memenuhi pertidaksamaan:

$$w.x + b = -1 \quad (2.2)$$

Sedangkan pola yang termasuk dalam kelas +1 (sampel positif):

$$w.x + b = +1 \quad (2.3)$$

Margin terbesar dapat ditentuksn dengan memaksimalkan nilai jarak antara *hyperplane* dan titik terdekatnya, yaitu 1 / || di ||. Ini dapat dirumuskan sebagai masalah pemrograman kuadratik (QP), yaitu mencari titik minimum persamaan dengan Batasan persamaan:

$$\min \tau(w) = \frac{1}{2} ||w||^2 \quad (2.4)$$

$$yi(xi.w+b) -1 \geq 0 \quad (2.5)$$

Problem ini dapat dipecahkan dengan berbagai Teknik komputasi, diantaranya *Lagrange Multiplier* [37].

1. Mencari *Langrange Multipliers* (ai)

$$\tilde{L}(a) = \sum_{i=1}^{n} ai - \frac{1}{2} \sum_{i,j} aiajyiyjxi^{T}xj \quad (2.6)$$

Dikenakan (untuk setiap i=1, …., n)

Keterangan:

yi: kelas data latih (+1/-1)

yj: kelas data latih (+1/-1)

xi: vector bobot kalimat komentar

xj: vector bobot kalimat komentar

2. Mencari Nilai Bobot (w)

$$W = \sum_{i=1}^{n} \alpha\, i\, yixi) \quad (2.7)$$

Keterangan :

w : *vector* bobot

yi : kelas data latih (+1/-1).

xi : *vector* bobot kalimat komentar yang menjadi *vector* pendukung.

3. Mencari Nilai Bias (b)

$$b = \frac{1}{NSV} \sum_{i=1}^{NSV} (w\,.xi - yi) \quad (2.8)$$

Keterangan :

NSV : jumlah *vector* pendukung.

w : *vector* bobot

yi : kelas data latih (+1/-1)

xi : *vector* bobot kalimat komentar yang menjadi *vector* pendukung.

Proses pengklasifikasian (pengujian) dalam SVM menggunakan persamaan:

$$f(\overrightarrow{t}) = sgn\ (\sum_{i=1,xi\in SV}^{n} aiyi < \vec{t}.\vec{x}i > +b) \quad (2.9)$$

Keterangan :

t : *vector* bobot data uji

xi *: vector* pendukung

b : nilai bias

yi : kelas atau label dari *vector* pendukung (+1/-1)

αi adalah *Lagrange multipliers*, yang bernilai nol atau positif (αi 0). Nilai optimal dari persamaan ini dapat dihitung dengan meminimalkan L terhadap w dan b, dan memaksimalkan L terhadap αi. Dengan memperhatikan sifat bahwa pada titik optimal gradient L = 0, persamaan Langkah dapat juga dimodifikasi sebagai maksimalisasi problem yang hanya mengandung αi saja, sebagaimana pesamaan *Maximize*:

$$\sum_{i=1} ai - \frac{1}{2}\sum_{i,j=1} aiajyi.yj, \vec{x}i, \vec{x}j \quad (2.10)$$

2.2.12 Naïve Bayes

Naïve Bayes merupakan salah satu metode yang digunakan dalam melakukan klasifikasi berdasarkan perhitungan probabilitas. *Naïve Bayes* dapat menghasilkan akurasi yang tinggi dikarenakan memiliki cara kerja yang cepat dan sederhana. Hal ini membuat *Naïve Bayes* sangat populer dan

sering digunakan [38]. Rumus *Naïve Bayes* secara umum adalah sebagai berikut [39] :

$$P(H|X) = \frac{P(H|X)P(H)}{P(X)} \quad (2.11)$$

Keterangan :

X : Data dengan class yang belum diketahui

H : Hipotesis data X merupakan suatu class spesifik

P(H|X) : Probabilitas hipotesis H berdasarkan kondisi x (posteriori prob.)

P(H) : Probabilitas hipotesis H (prior prob)

P(X|H) : Probabilitas X berdasarkan kondisi tersebut

P(X) : Probabilitas dari X.

2.2.13 Confusion Matrix

Confusion matrix merupakan metode dalam perhitungan akurasi terhadap konsep yang ada pada *data mining*. Dalam mengklasifikasikan data menghasilkan jumlah data uji yang benar dan salah digambarkan dalam bentuk tabel [40]. Contoh *confusion matrix* untuk klasifikasi biner ditunjukan pada Tabel 2.3

Tabel 2.3 Confusion Matrix [41].

		Kelas Prediksi	
		1	0
Kelas sebenarnya	1	TP	FN
	0	FP	TN

Keterangan :

1. *True Positive* (TP), adalah jumlah dokumen dari kelas 1 yang benar dan diklasifikasikan sebagai kelas 1.

2. *True Negative* (TN), adalah jumlah dokumen dri kelas 0 yang benar diklasifikasikan sebagai kelas 0.
3. *False Positive* (FP), adalah jumlah dokumen dari kelas 0 yang salah diklasifikasikan sebagai kelas 1.
4. *False Negative* (FN), adalah jumlah dokumen dari kelas 1 yang salah diklasifikasikan sebagai kelas 0.
   a) *Accuracy* merupakan persentase dari total sentimen yang benar dikenali. Perhitungan akurasi dilakukan dengan cara membagi jumlah data sentimen yang benar dengan total data dan data uji. Untuk menghitung nilai akurasinya.

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (2.17)$$

   b) *Precision* merupakan perbandingan jumlah data relevan yang ditemukan terhadap jumlah data yang ditemukan. Perhitungan precision dilakukan dengan cara membagi jumlah data benar yang bernilai positif dibagi dengan jumlah data benar yang bernilai positif dan data salah yang bernilai positif. Nilai dari data salah bernilai positif diambil dari jumlah nilai selain true positif kolom yang sesuai tiap kelasnya.

$$Precision = \frac{TP}{TP+FP} \quad (2.18)$$

   c) *Recall* merupakan perbandingan jumlah materi relevan yang ditemukan terhadap jumlah materi yang relevan. Perhitugan recall dilakukan dengan cara membagi data benar bernilai positif dengan hasil penjumlahan dari data benar yang bernilai positif dan data salah yang bernilai negatif. Nilai dari data salah yang bernilai negatif diambil dari jumlah nilai selain true positif baris yang sesuai tiap kelasnya.

. $$Recall = \frac{TP}{TP+FN} \quad (2.19)$$

d) *F1 Score* merupakan parameter tunggal ukuran keberhasilan retrieval yang menggabungkan *recall* dan *precision.* Nilai *F1-Score* didapat dari perhitungan hasil perkalian *precision* dan *recall* dibagi dengan hasil penjumlahan precision dan recall kemudian dikalikan dua.

$$F1 - Score = 2\ x \frac{Precision\ x\ Recall}{Precision + Recall}$$