ANALISIS SENTIMEN ULASAN PRODUK *TONER* PADA *BEAUTY BRAND* "THE BODY SHOP" MENGGUNAKAN METODE *NAÏVE BAYES CLASSIFIER* DAN *SUPPORT VECTOR MACHINE*: STUDI KASUS DI FEMALE DAILY

LAPORAN TUGAS AKHIR

Diajukan Sebagai salah Satu Syarat Untuk Memperoleh Gelar Sarjana Strata-1
Pada Jurusan Teknik Industri Fakultas Teknologi Industri

Disusun oleh:
Asri Nabila
18522093

PROGRAM STUDI TEKNIK INDUSTRI
FAKULTAS TEKNOLOGI INDUSTRI
UNIVERSITAS ISLAM INDONESIA
YOGYAKARTA
2022

UNIVERSITAS ISLAM INDONESIA

FAKULTAS TEKNOLOGI INDUSTRI

Gedung KH. Mas Mansur
Kampus Terpadu Universitas Islam Indonesia
Jl. Kaliurang km 14,5 Yogyakarta 55584
T. (0274) 898444 ext. 4100, 4101
F. (0274) 895007
E. fti@uii.ac.id
W. fti.uii.ac.id

SURAT KETERANGAN PENELITIAN

Nomor : 240/Ka.Lab.Datmin/70/Lab.Datmin/X/2022

Assalamu'alaikum Warahmatullahi Wabarakatuh

Kami yang bertanda tangan dibawah ini, menerangkan bahwa mahasiswa dengan keterangan sebagai berikut :

Nama	:	Asri Nabila
No. Mhs	:	18522093
Dosen Pembimbing	:	Annisa Uswatun Khasanah, ST., M.B.A., M.Sc

Telah selesai melaksanakan penelitian yang berjudul "Komparasi Analisis Sentimen Ulasan Produk Toner Pada Beauty Brand "The Body Shop" Menggunakan Metode Naive Bayes Classifier dan Support Vector Machine: Studi Kasus di Female Daily" di Laboratorium Data Mining, Program Studi Teknik Industri, Fakultas Teknologi Industri, Universitas Islam Indonesia tercatat mulai tanggal 8 Januari 2022 sampai dengan tanggal 8 Oktober 2022.

Demikian surat keterangan kami keluarkan, agar dapat dipergunakan sebagaimana mestinya.

Wassalamu'alaikum Warahmatullahi Wabarakatuh

Yogyakarta, 25 Oktober 2022

Kepala Laboratorium
Data Mining

Annisa Uswatun Khasanah, ST., M.B.A., M.Sc.

PERNYATAAN KEASLIAN

Demi Allah saya akui bahwa karya ini merupakan karya saya sendiri kecuali kutipan dan ringkasan yang setiap salah satunya dicantumkan sumbernya. Jika ditemukan di kemudian hari ternyata terbukti pengakuan saya ini tidak benar dan melanggar peraturan yang sah dalam karya tulis dan hak kekayaan intelektual maka saya bersedia ijazah yang saya terima untuk ditarik oleh Universitas Islam Indonesia.

Yogyakarta, 24 Oktober 2022

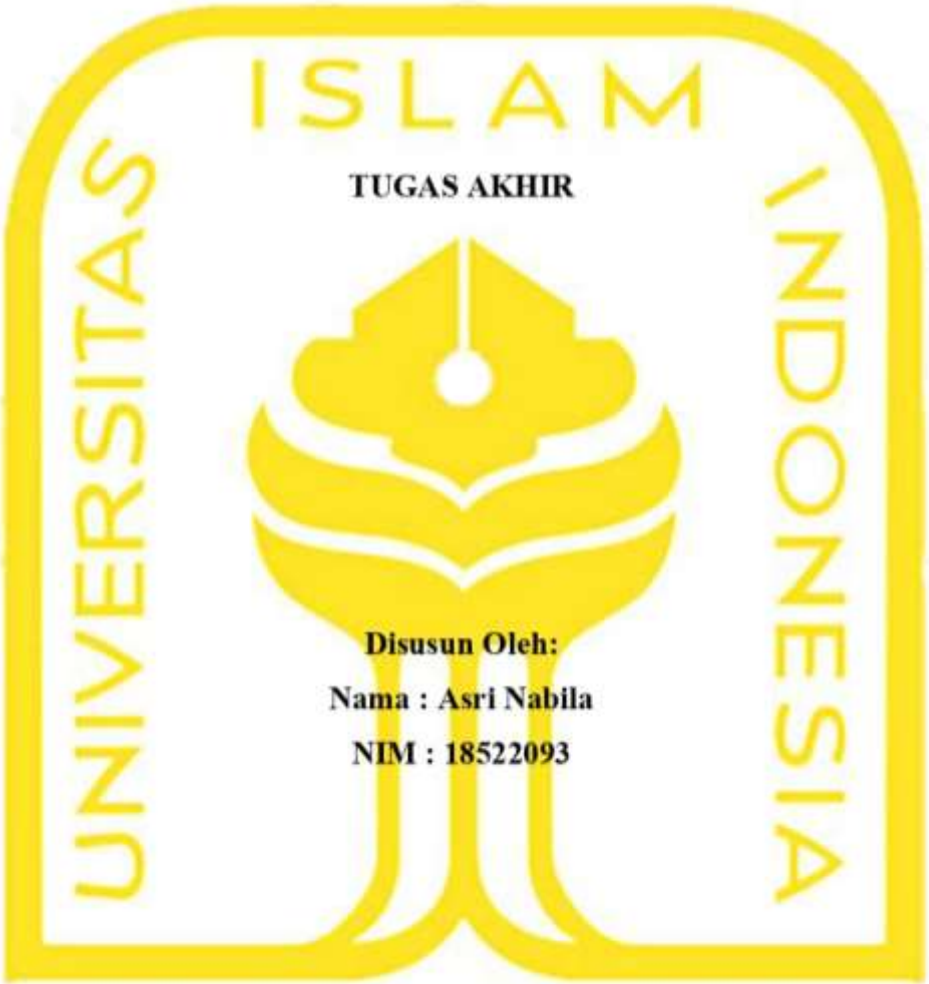
Peneliti

Asri Nabila

HALAMAN PENGESAHAN PEMBIMBING

KOMPARASI ANALISIS SENTIMEN ULASAN PRODUK *TONER* PADA *BEAUTY BRAND* "THE BODY SHOP" MENGGUNAKAN METODE *NAÏVE BAYES CLASSIFIER* DAN *SUPPORT VECTOR MACHINE*: STUDI KASUS DI FEMALE DAILY

TUGAS AKHIR



Disusun Oleh:

Nama : Asri Nabila

NIM : 18522093

Yogyakarta, 24 Oktober 2022

Dosen Pembimbing

Annisa Uswatun Khasanah, S.T., M.B.A., M.Sc.

NIP. 145220102

LEMBAR PENGESAHAN DOSEN PENGUJI

ANALISIS SENTIMEN ULASAN PRODUK *TONER* PADA *BEAUTY BRAND* "THE BODY SHOP" MENGGUNAKAN METODE *NAÏVE BAYES CLASSIFIER* DAN *SUPPORT VECTOR MACHINE*: STUDI KASUS DI FEMALE DAILY

TUGAS AKHIR

Disusun Oleh:

Nama : Asri Nabila

NIM : 18522093

Telah dipertahankan didepan sidang penguji sebagai salah satu syarat untuk memperoleh gelar sarjana Strata S-1 Teknik Industri

Tim Penguji

Annisa Uswatun Khasanah, S.T., M.B.A., M.Sc.	
Ketua	
Yuli Agusti Rochman, S.T., M.Eng.	
Anggota I	
Abdullah 'azzam, S.T., M.T.	
Anggota II	

Mengetahui,
Ketua Program Studi Teknik Industri
Fakultas Teknologi Industri
Universitas Islam Indonesia

Ir. Muhammad Kutwan Adni Purnomo, S.T. M.Sc., Ph.D., IPM.

HALAMAN PERSEMBAHAN

Dengan Menyebut Nama Allah yang Maha Pengasih lagi Maha Penyayang.
Puji Syukur atas Kehadiran Allah SWT atas nikmat yang telah beliau berikan,
Skripsi ini saya persembahkan kepada:

Bapak Abdus Shoheb dan Ibu Esny Shoheb selaku kedua orang tua saya yang tiada henti mendo'akan, mendidik, dan memberikan fasilitas serta semangat dari kecil hingga saat ini. Terima kasih yang selalu menjadi sosok utama dan segalanya didalam hidup saya yang tidak pernah tergantikan. Terima kasih atas pengorbanan dan segala hal yang tidak bisa diungkapkan kata.

Kakak saya Happy Pratiwi, adik saya Aida Balqis, dan Muhammad Faiz Zain, serta Nenek saya maupun keluarga besar lainnya, terima kasih telah memberikan dukungan, semangat, dan motivasi sampai hingga tahap ini dan akan seterusnya.

Serta kerabat, sahabat, dan teman-teman saya yang selalu mendukung, menemani dan membantu saya setiap hari dari awal hingga akhir masa perkulihan ini.

MOTTO

"*Indeed*, with every *difficulty*, comes *relief*."
(QS. Al-Insyirah: 5)

"Perfectionism is the enemy of an action. Perfectionism is a mask that we all wear when we afraid of failure."
(TED X Leiden University)

"You can be everything you need if you have patience, focus on yourself, create your own security, understand your emotions, and heal from your history. Don't rush into anything just so you have something."
(Steven Barlett)

"If you are striving of your dreams, you should believe in yourself and don't let anyone bring you down. Negativity doesn't exist, it's all about positivity anyways."
"The winner is the person who never gives up, we all go through those difficult phases, but the person who doesn't gives up like the one who's in the last, yet you'll make it anyways, so think about that way."
(Mark Lee)

"Writing down your goals won't help you achieve anything, Stop romanticizing your life. Go on and take an action. What matters is your presence, not your perfection."
(Asri Nabila)

"It can be said that the secret to success doesn't exist and even if there are, there are only two of them. First, endure to the end and do not give up. Second, if you want to give up, you have to go back to the first secret, which is to survive to the end."
(Zhong Chenle)

KATA PENGANTAR

Assalamu'alaikum Warahmatullahi Wabarakatuh

Alhamdulillah, puji serta syukur kehadirat Allah SWT yang telah melimpahkan nikmat, rahmat, dan hidayah-Nya, sehingga Tugas Akhir dengan judul "Analisis Sentimen Ulasan Produk *Toner* Pada *Beauty Brand* "The Body Shop" Menggunakan Metode *Naïve Bayes* dan *Support Vector Machine*: Studi Kasus di Female Daily" ini dapat diselesaikan. Selawat serta salam senantiasa tercurahkan kepada Nabi Muhammad SAW beserta keluarga, sahabat serta para pengikut beliau hingga akhir zaman.

Tugas Akhir merupakan salah satu syarat yang harus dipenuhi dalam menyelesaikan jenjang sarjana strata-1 (S-1) Teknik Industri, Fakultas Teknologi Industri, Universitas Islam Indonesia. Selama proses ini, penulis menyadari banyak pihak yang telah memberikan dukungan, bimbingan, dan bantuan materi maupun non materi, oleh karena itu penulis ingin mengucapkan terima kasih kepada:

1. Bapak Prof. Dr. Ir. Hari Purnomo, M.T., IPU., ASEAN Eng., selaku Dekan Fakultas Teknologi Industri, Universitas Islam Indonesia.
2. Dr. Drs. Imam Djati Widodo, M.Eng.Sc.selaku Ketua Jurusan Teknik Industri, Fakultas Teknologi Industri, Universitas Islam Indonesia.
3. Bapak Ir. Muhammad Ridwan Andi Purnomo, S.T. M.Sc., Ph.D., IPM. selaku Ketua Program Studi Teknik Industri Fakultas Teknologi Industri Universitas Islam Indonesia.
4. Ibu Annisa Uswatun Khasanah, S.T., M.B.A., M.Sc. selaku Dosen Pembimbing Tugas Akhir.
5. Keluarga besar saya, Ibu, Ayah, Kakak dan Adik saya yang senantiasa selalu memotivasi penulis dan memberikan doa, dukungan, serta semangat.
6. Keluarga Laboratorium Data Mining angkatan 2016, 2017, 2018, 2019 dan 2020 serta kepala laboratorium Ibu Annisa Uswatun Khasanah, S.T., M.B.A., M.Sc., laboran mas Bayu Hertanta, dan teman-teman Teknik Industri Angkatan 2018 lainnya yang telah memberikan doa dan dukungan.
7. Seluruh dosen pengajar dan staf Prodi Teknik Industri yang telah memberikan bekal ilmu dan atas bantuannya, semoga menjadi amal kebaikan Bapak/Ibu.
8. Semua pihak yang belum tersebutkan melainkan terlibat dari awal hingga akhir dalam pembuatan tugas akhir.

Penulis menyadari bahwa dengan keterbatasan wawasan serta pengalaman penulis, membuat Tugas Akhir ini masih jauh dari kesempurnaan. Oleh karena itu, penulis mengharapkan kritik dan saran yang bersifat membangun. Semoga Tugas Akhir ini dapat memberikan manfaat. Akhir kata, semoga kita semua mendapatkan rahmat dan selalu ada dalam lindungan Allah SWT.

Wassalamu'alaikum Warahmatullahi Wabarakatuh

Yogyakarta, 24 Oktober 2022

Penulis

Asri Nabila

ABSTRAK

Seiring pertumbuhan *e-commerce,* peranan media sosial saat ini telah bertransformasi menjadi *social commerce*. Pertumbuhan ini secara tidak langsung mengubah cara berinteraksi *word of mouth marketing* (WOMM) dari tradisional menjadi modern. Media sosial dapat memberikan ruang bagi konsumen untuk saling memberikan ulasan dan rekomendasi. Female Daily merupakan suatu komunitus *online* kecantikan terbesar di Indonesia yang menyediakan interaksi sosial untuk saling memberikan ulasan maupun rekomendasi. Mengevaluasi ulasan produk *The Tea Tree Skin Clearing Toner* dari *Beauty Brand* "THE BODY SHOP" dengan jumlah yang banyak membutuhkan analisis sentimen untuk mengelompokkan ulasan konsumen menjadi opini positif, negatif, atau netral. Data ulasan tersebut dilabeli dan dianalisis dengan menggunakan metode *Support Vector Machine* (SVM) dan *Naïve Bayes Classifier* (NBC) untuk mengklasifikasikan data ulasan. Sentimen tersebut dapat membantu baik perusahaan maupun individu untuk mengetahui kualitas produk secara detail. *Dataset* terdiri dari 1.050 ulasan terbagi menjadi 780 ulasan positif dan 270 ulasan negatif. Klasifikasi dengan metode SVM dengan *kernel linear* diperoleh tingkat akurasi sebesar 86% dengan nilai *Area Under Rate* (AUC) sebesar 0.91, yang lebih besar daripada tingkat akurasi menggunakan metode NBC yaitu sebesar 83% dengan nilai *Area Under Rate* (AUC) sebesar 0.82. Oleh karena itu, klasifikasi data ulasan produk *Tea Tree Skin Clearing Mattifying Toner* "The Body Shop" di Female Daily sebaiknya dilakukan menggunakan algoritma *Support Vector Machine* daripada *Naïve Bayes.*

Kata kunci: *Support Vector Machine* (SVM), *Naïve Bayes Classifier* (NBC), Klasifikasi, Analisis Sentimen, Female Daily, The Body Shop

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I

PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi saat ini sangat memudahkan masyarakat untuk menerima atau memperoleh informasi melalui media cetak, media elektronik, media *digital* dan media sosial. Menurut *We Are Social*, *total* pengguna media sosial di Indonesia mencapai 191 juta pengguna atau setara dengan 68.8% dari jumlah total penduduk Indonesia, dengan pengguna berbasis *mobile*-nya mencapai 370.1 juta pengguna. Pengguna media sosial di Indonesia paling banyak berada pada usia produktif yaitu usia 18-34 tahun baik laki-laki maupun perempuan.

Gambar 1. 1 Jumlah Pengguna Media Sosial di Indonesia
(*WeAreSocial*, 2022)

Perkembangan teknologi internet serta jumlah pengguna media sosial juga didampingi dengan perkembangan bisnis produk perawatan kulit baik secara *global* maupun domestik di Indonesia yang ditandai dengan bermunculannya banyak *brand* global yang merintis di Indonesia maupun pelaku bisnis domestik yang menciptakan *brand* perawatan kulit lokal Indonesia. Hal ini dapat dilihat dari membanjirnya kosmetik yang berasal dari merek domestik dan global, seperti The Body Shop, Wardah, Somethinc, Victoria Secret, Avoskin, Loreal, dll.

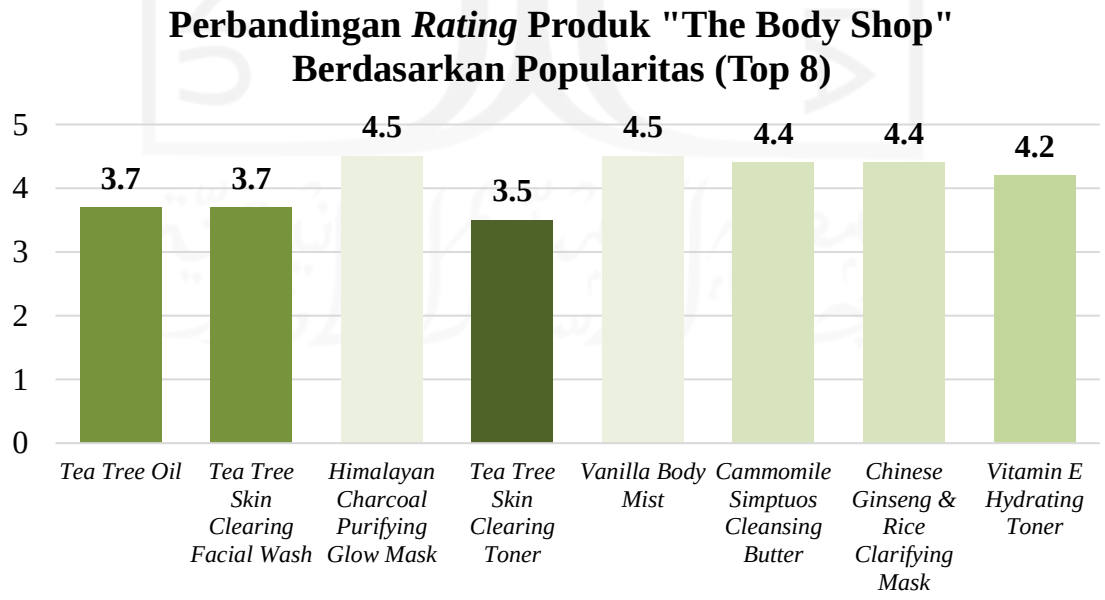
Di Indonesia sendiri The Body Shop cukup diminati oleh banyak orang, terbukti dari banyaknya toko yang dibuka di pusat perbelanjaan kota-kota besar (Wisnu, 2021). Bahan-bahan utama yang alami menjadi daya tarik sendiri bagi masyarakat Indonesia. The Body Shop Indonesia menjual berbagai macam produk kecantikan, diantaranya produk *skincare*, *make-up*, *bodycare*, produk pewangi, hingga produk *haircare*. Salah satu produk unggulan The Body Shop di Indonesia adalah produk pewangi berupa *body mist* dan produk *skincare* berupa *body cream*. Kedua produk tersebut selama 8 tahun berhasil meraih predikat *Top Brand Award* dalam kategorinya (Top Brand Award, 2022).

Tabel 1. 1 *Top Brand Index* "The Body Shop"

Jenis Produk	Tahun							
	2015	2016	2017	2018	2019	2020	2021	2022
Body Mist	32.0%	22.6%	14.0%	17.1%	35.0%	44.3%	49.6%	44.9%
Body Butter	29.0%	21.7%	14.4%	17.1%	30.9%	42.5%	44.4%	41.5%

(Top Brand Award, 2022)

Diantara banyaknya produk kosmetik yang dikeluarkan oleh The Body Shop, Tabel 1.1 menunjukkan bahwa hanya dua produk kosmetik jenis pewangi dan *skincare* saja yang dapat menguasai pasar (Surabagiarta & Purnaningrum, 2021). Sementara itu, produk-produk lainnya khususnya yang berada dalam rangkaian *skincare* pada kategori *toner*, salah satu contohnya yaitu *Tea Tree Skin Clearing Toner* yang dikeluarkan oleh The Body Shop belum dapat menguasai pasar dan masih kalah saing dengan *brand* lain.



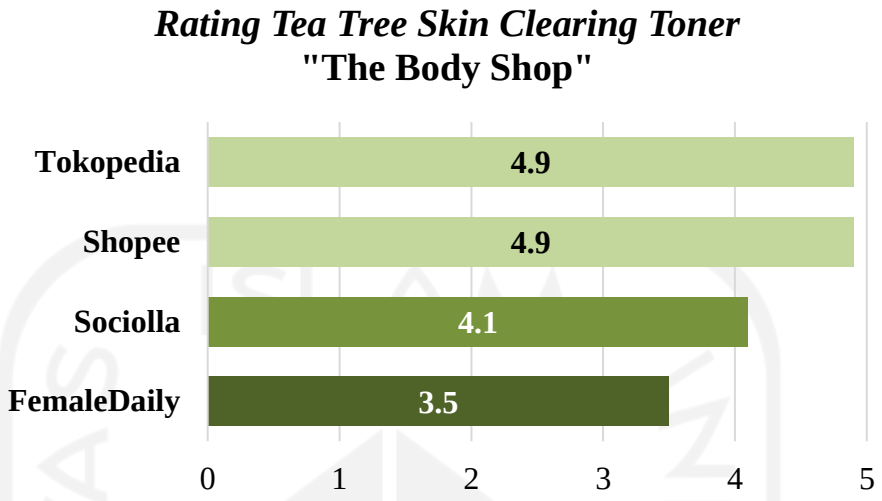
Gambar 1. 2 Perbandingan *Rating* Produk "The Body Shop" Berdasarkan Popularitas (Female Daily, 2022)

Hal ini dapat dibuktikan pada Gambar 1.2 (Female Daily, 2022), yang menyatakan bahwa diantara berbagai produk populer dari *brand* The Body Shop, terdapat produk *Tea Tree Clearing Skin Toner* di Female Daily yang mendapatkan *rating* terendah yaitu 3.5 dari skala 5. *Rating* yang rendah disertai berbagai ulasan positif dan negative pada produk *Tea Tree Clearing Skin Toner* di Female Daily menunjukkan bahwa performa produk yang ditawarkan oleh The Body Shop belum sepenuhnya memenuhi ekspektasi dan kebutuhan pengguna *skincare* tersebut. Oleh karena itu perlu adanya evaluasi bagi The Body Shop untuk dapat meningkatkan kualitas produk hingga memperbaiki reputasi produk menjadi produk yang berkualitas tinggi dan diterima baik di benak konsumen.

Dewasa ini, tingkah laku konsumen sebelum memutuskan untuk membeli produk, mereka akan melakukan penelitian tentang kosmetik, membandingkan harga, berdiskusi dengan konsumen lain tentang pendapat mereka, atau membaca ulasan produk. Jejaring sosial media diciptakan untuk membantu dan mendorong orang untuk berpartisipasi dalam mengirim komentar, dan memberikan *feedback* mereka (Amanatidou, 2022). Menurut hasil survei Desember 2019 yang dilakukan terhadap 1.005 pengguna internet AS, sebanyak 76% orang mempercayai *online review* sebagaimana rekomendasi dari keluarga dan teman (Murphy, 2019). Oleh karena itu, ulasan/*online review* dari pengguna dapat digunakan sebagai alat yang efektif dan efisien untuk mendapatkan informasi terhadap produk atau layanan yang diberikan.

Sebagai bagian dari *online review, Rating* adalah pendapat pelanggan pada skala tertentu. Sebuah sistem *rating* yang popular di *e-commerce* yang memberikan peringkat bintang, dimana semakin banyak bintang yang diberikan, semakin baik peringkat produk penjual (Lackermair et al, 2013). *Rating* disebut sebagai tipe lain dari opini yang diberikan oleh banyak orang dan menjadi evaluasi rata-rata dari para konsumen suatu produk berdasarkan pengalaman konsumen yang mengacu pada keadaan psikologis dan emosional konsumen saat menggunakan produk tersebut (Farki et al., 2016) serta menjadi representasi dari opini konsumen dengan skala yang spesifik (Lackermair et al, 2013). Fungsi *online rating* yang terdapat pada halaman produk toko *online* merupakan cara bagi konsumen untuk menilai kualitas produk. Banyaknya bintang yang diperoleh suatu produk tertentu dapat dikaitkan dengan kualitas produk tersebut (Auliya et al., 2017). Hal ini memudahkan calon konsumen untuk menilai produk tertentu, karena jumlah bintang dalam ulasan *online* dianggap mampu sebagai tolak ukur kualitas produk tertentu. *Rating* juga dapat membantu pembeli membuat keputusan pembelian dan menarik lebih banyak

calon pembeli yang berkualitas ke halaman produk atau situs *web* penjual (Latief & Ayustira, 2020).



Gambar 1. 3 *Rating Toner* "The Body Shop"

Berdasarkan Gambar 1.3, *platform* penyedia ulasan produk *toner* The Body Shop yang memiliki *rating* terendah adalah Female Daily. Pemilihan *platform* Female Daily juga didukung oleh kredibilitas Female Daily sebagai sosial media *community-based* pertama di Indonesia pada bidang kecantikan yang menyediakan *review* terpercaya mengenai *skincare, body care, hair care*, dan kebutuhan wanita lainnya (Ramadhanti, 2021). Female Daily dapat menciptakan peluang tidak hanya bagi pengguna yang dapat membagikan ulasan dan pendapat mereka tentang produk dan layanan, tetapi juga untuk *marketeers* maupun pihak berkepentingan lainnya. *Marketing experts* dapat memanfaatkan data pada media sosial untuk membantu mereka dalam membuat *digital campaign*, sedangkan bagi *business experts* ataupun seorang *data analyst* dapat mengambil *insight* dari pemanfaatan data di sosial media untuk memberikan strategi bisnis lain yang inovatif dan lebih efektif untuk segala jenis bisnis.

Analisis sentimen merupakan salah satu implementasi dalam pemanfaatan data di media sosial yang berfungsi untuk mengkategorikan ulasan menjadi opini positif dan negatif maupun netral. Kategorisasi ulasan memungkinkan konsumen untuk mengidentifikasi kualitas produk dan mencari produk yang sesuai dengan kebutuhan kulitnya. Bagi perusahaan The Body Shop, analisis sentimen mampu menganalisis perspektif *user* mengenai *brand skincare* yang dibicarakan di media sosial. Dikarenakan pada media sosial terdapat banyak data *User Generated Content* (UGC) yang diutarakan

user mengenai produk suatu *brand* sehingga menjadi data yang berharga bagi perusahaan. Ada banyak metode yang dapat digunakan untuk analisis sentimen, diantaranya *Decision Tree, k-Nearest Neighbor, Random Forest, Neural Network, Naïve Bayes, Support Vector Machine*, dsb. Dua diantara metode klasifikasi tersebut yang akan digunakan pada penelitian ini adalah metode *Naïve Bayes* dan *Support Vector Machine*.

Naive bayes adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas dalam suatu kelompok kelas dan terbukti memiliki tingkat akurasi dan kecepatan yang tinggi saat diaplikasikan kedalam *database* yang besar (Saleh, 2015). Algoritma *Support Vector Machine* adalah salah satu algoritma klasifikasi yang memiliki nilai akurasi yang tinggi di antara metode yang lainnya (Tuhuteru, 2020). Penelitian berkaitan dengan perbandingan metode *Naïve Bayes Classifier* maupun *Support Vector Machine* dengan metode klasifikasi yang lain pernah dilakukan sebelumnya. Penelitian tersebut mengangkat permasalahan perbandingan analisis klasifikasi menggunakan metode *Naïve Bayes Classifier* (NBC), *k-Nearest Neighbour* (k-NN), dan *Random Forest* pada data ulasan film di *web portal* penyedia ulasan film IMDb (*Internet Movie Database*) (Baid, et al., 2017). Penelitian tersebut mengklasifikasikan data ulasan film kedalam label positif dan negatif. Berdasarkan hasil dan analisis penelitian tersebut, diperoleh kesimpulan bahwa metode NBC lebih baik dibandingkan dengna k-NN dan *Random Forest*, dimana nilai akurasi tertinggi diperoleh pada metode NBC yaitu sebesar 81.45%.

Sementara itu, penelitian lain juga telah dilakukan yang membahas mengenai perbandingan estimasi evaluasi kinerja metode *machine learning* berupa *Support Vector Machine* (SVM) dengan *Stacked Denoising Autoencoder* (SDA), dan *Convolutional Neural Network* (CNN) pada ulasan produk *face-lotion* di @cosme, sebuah jejaring sosial informasi terintegrasi untuk kosmetik/kecantikan di Jepang (Ma, et al., 2020). Penelitian tersebut mengklasifikasikan data ulasan *face lotion* menjadi label *good, middle, poor,* dan *no mention*. Berdasarkan hasil dan analisis penelitian tersebut, diperoleh kesimpulan bahwa kinerja metode SVM adalah yang terbaik dibandingkan dengan kedua metode lainnya, dimana nilai presisi tertinggi diperoleh pada metode SVM (RBF) yaitu sebesar 70%.

Berdasarkan kedua penelitian terdahulu tersebut, metode NBC dan SVM pada masing-masing penelitian memberikan performa klasifikasi yang paling baik. Hal ini yang mendorong peneliti untuk melakukan perbandingan metode NBC dengan metode

klasifikasi lainnya, seperti SVM. Penelitian ini dilakukan untuk mengetahui manakah metode yang lebih tepat dalam mengklasifikasikan ulasan produk *Tea Tree Clearing Skin Toner* dari *brand* "The Body Shop", serta untuk mengetahui seberapa pentingnya analisis sentimen yang harus dilakukan dalam memberikan analisa sentiment atau isi perbincangan *user* di media sosial terkait produk dari *brand* The Body Shop. Berdasarkan uraian tersebut, maka peneliti mengajukan judul penelitian yaitu "Komparasi Analisis Sentimen Pada *Cosmetic Product* Menggunakan Metode *Naïve Bayes Classifier* dan *Support Vector Machine*: Studi Kasus Pada *Beauty Brand* "The Body Shop" di Female Daily*".*

1.2 Rumusan Masalah

Berikut merupakan rumusan masalah dalam penelitian ini:

1. Bagaimana hasil analisis penerapan metode dan kinerja model pengklasifikasian sentimen terhadap ulasan produk *Tea Tree Skin Clearing Mattifying Toner* dari *brand* kecantikan "The Body Shop" di Female Daily dengan menggunakan metode *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine (SVM)*?
2. Bagaimana pembentukan kata yang didapatkan berdasarkan hasil klasifikasi kata dalam bentuk *word cloud* dan asosiasi kata?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah tujuan dari penelitian ini sebagai berikut:

1. Mengetahui hasil analisis penerapan metode dan kinerja model pengklasifikasian *sentimen* terhadap ulasan produk *Tea Tree Skin Clearing Mattifying Toner* dari *brand* kecantikan "The Body Shop" di Female Daily dengan menggunakan metode *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine (SVM)*.
2. Mengetahui analisis pembentukan kata yang didapatkan berdasarkan hasil klasifikasi kata dalam bentuk *word cloud* dan asosiasi kata.

1.4 Batasan Masalah

Batasan masalah dari penelitian ini diperlukan supaya tugas khusus dalam pelaksanaan kerja praktik dapat berjalan dengan baik. Berikut merupakan batasan masalah yang diperlukan:

1. Metode yang digunakan dalam penelitian ini adalah *Naïve Bayes* dan *Support Vector Machine (SVM)*.
2. Penelitian ini akan fokus kepada data ulasan pengguna produk *Tea Tree Skin Clearing Mattifying Toner* dari *Beauty Brand* "The Body Shop" di Female Daily.
3. Pengambilan data dilakukan dengan melakukan proses *scrapping data* di Female Daily mulai dari tanggal 24 Juni 2018 – 11 September 2022.

1.5 Manfaat Penelitian

Dari hasil penelitian yang telah dilakukan diharapkan dapat memberikan kebermanfaatan bagi berbagai pihak diantaranya:

1. Bagi Perusahaan atau *Brand Owner*

   Hasil akhir penelitian berupa wawasan baru atau *insight* yang didapatkan dari data tidak terstruktur dengan jumlah yang besar yang harapannya dapat memberikan kemudahan pihak yang berkepentingan dalam mengetahui reputasi dan posisi *brand* di tengah masyarakat, serta me-*monitoring sentiment* dari isi perbincangan masyarakat terhadap produk keluaran *brand*, yang kemudian bisa dijadikan evaluasi bagi *brand owner* dalam merancang *campaign*, *launching* produk baru maupun memperbaiki kualitas produk yang sudah ada.
2. Bagi Peneliti

   Manfaat bagi peneliti dari hasil penelitian ini tentunya sebagai media pembelajaran dan implementasi keilmuan Teknik Industri di dalam sebuah perusahaan. Selain itu memberikan wawasan dan pengetahuan yang lebih mendalam mengenai *Text Classification* dengan algoritma *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM) menggunakan bahasa pemrograman Python.

1.6 Sistematika Penulisan

Sistematika penulisan dibuat dengan tujuan memudahkan pembaca dalam memahami isi penelitian ini. Sistematika penulisan dalam penelitian yaitu:

BAB I PENDAHULUAN

Pada bab ini memaparkan permasalahan yang mendorong penulis untuk melakukan penelitian ini, yang terdiri dari latar belakang, rumusan, batasan, tujuan, manfaat, dan cara penulisan penelitian secara sistematis.

BAB II KAJIAN LITERATUR

Bab ini terdiri dari kajian deduktif dan kajian induktif. Kajian deduktif merupakan uraian teori yang pernah dijelaskan oleh peneliti terdahulu berdasarkan penelitian terdahulu, sedangkan kajian induktif merupakan landasan teori yang dijadikan bahan pendukung dalam bentuk fakta dan kesimpulan yang masih memiliki hubungan dengan topik penelitian.

BAB III METODE PENELITIAN

Bab ini berisi mengenai uraian alur penelitian seperti *flowchart* penelitian, populasi dan sampel penelitian, jenis dan sumber data, variabel penelitian, metode pengumpulan data, dan metode analisis data.

BAB IV PENGUMPULAN DAN PENGOLAHAN DATA

Bab ini berisi mengenai proses pengumpulan hingga pengolahan data. Dimulai dari proses pengumpulan hingga bagaimana data yang telah didapatkan lalu diakumulasi, kemudian dilakukan proses *pre-processing.*

BAB V ANALISA DAN PEMBAHASAN

Bab ini menganalisis dan menjelaskan hasil penelitian ini. Analisis yang dilakukan disesuaikan dengan metode yang digunakan hingga sampai mendapatkan *output* berupa rekomendasi.

BAB VI KESIMPULAN DAN SARAN

Bab ini menguraikan ringkasan dari analisis berdasarkan penelitian yang telah dilakukan. Kesimpulan yang ditarik merupakan jawaban dari rumusan masalah masalah. Saran yang akan diberikan merupakan solusi yang dapat diteruskan kepada peneliti berikutnya maupun pihak terkait.

DAFTAR PUSTAKA

LAMPIRAN

BAB II

KAJIAN LITERATUR

2.1 Kajian Induktif

Beberapa penelitian telah dilakukan pada analisis sentimen sebelumnya. Penelitian tersebut sangatlah penting dalam penelitian ini sebagai kajian untuk mengetahui keterkaitan antara penelitian terdahulu dengan penelitian yang akan dilakukan, untuk menghindari terjadinya tindakan duplikasi yang dilakukan oleh penulis. Tujuan dari tinjauan pustaka ini adalah untuk mengkaji metode yang akan dipakai dalam penelitian ini. Dalam penelitian analisis sentimen terdapat 2 metode yang paling populer digunakan yaitu *Naive Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM). Masing-masing metode memiliki nilai akurasi yang berbeda-beda tergantung dengan konteks kasus yang digunakan. Pada penelitian terdahulu yang penulis kaji, penelitian dengan metode *Naive Bayes Classifier* dan *Support Vector Machine* memiliki nilai akurasi yang terbilang tinggi. Meskipun demikian, peneliti juga mengkaji metode lainnya sebagai pembanding metode *Naive Bayes Classifier* dan *Support Vector Machine.*

Contohnya pada penelitian tentang analisis sentimen pada ulasan produk di Amazon yang dilakukan oleh (Dey, et al., 2020) menggunakan metode SVM dan *Naïve Bayes* yang memiliki nilai akurasi 84% dan 82.87% pada masing-masing metode. Sementara itu, pada penelitian yang dilakukan oleh (Bayhaqy, et al., 2018) mengenai analisis sentimen tentang *e-commerce* dari *tweet* menggunakan metode *Decision Tree, K-Nearest Neighbor, and Naïve Bayes* memiliki nilai akurasi masing-masing sebesar 80%, 78%, dan 77%. Selain itu, pada penelitian yang dilakukan oleh (Siswanto, et al., 2022) mengenai analisa sentimen publik mengenai perekonomian indonesia pada pandemi Covid-19 di Twitter menggunakan *K-Nearest Neighbor* (KNN) dan SVM mendapatkan nilai akurasi masing-masing sebesar 76% dan 78%. Kemudian, pada penelitian yang dilakukan oleh (Tuhuteru, 2020) mengenai analisis sentimen masyarakat terhadap pembatasan sosial berksala besar menggunakan metode *Support Vector Machine* memiliki nilai akurasi sebanyak 87.33%.

Kemudian, pada penelitian yang dilakukan oleh (Ma, et al., 2020) yang membahas mengenai klasifikasi *spam email* menggunakan SVM dan *Naïve Bayes* menghasilkan nilai akurasi masing-masing sebesar 95.5% dan 94.5%. Pada penelitian tersebut, nilai akurasi dengan metode SVM lebih tinggi dari NBC dikarenakan *Naive Bayes* hanya menganggap setiap kata yang dimasukkan dalam email sebagai kata independen tanpa mempertimbangkan posisi atau urutan kata dalam klasifikasi. Berbeda dengan *Naive Bayes Classifier, Support Vector Machine* dapat memprediksi kelas dengan mencari *hyperplane optimal* di antara kemungkinan *hyperplane* yang memisahkan dua kelas yang berbeda. Penelitin lain yang dilakukan oleh (Wilis, et al., 2020) mengenai analisis sentimen sosial media pada rekomendasi *souvenir* menggunakan metode *Lexicon Based* pada *Support Vector Machine* menghasilkan akurasi sebesar 88%.

Penelitian lain juga telah dilakukan oleh (Afdhal, et al., 2022) yang membahas tentang penerapan algoritma *random forest* untuk analisis sentimen komentar di YouTube tentang islamofobia dengan menggunakan algoritma *Random Forest* menghasilkan akurasi sebesar 79%. Lalu, penelitian mengenai analisis sentimen publik terkait kebijakan pembelajaran daring yang dilakukan oleh (Isnain, et al., 2021) menggunakan metode *K-Nearest Neighbor* memiliki hasil akurasi sebesar 84.65%. Sementara itu, penelitian yang dilakukan oleh (Utami, 2018) mengenai *review* produk kosmetik memiliki nilai akurasi 80%. Penelitian lain yaitu yang dilakukan oleh (Fikria, 2018) membahas tentang analisis sentimen *review* aplikasi *e-ticketing* pada aplikasi KAI *access* dengan pembobotan TF-IDF yang memiliki nilai 89% dan pada aplikasi tiket.com memiliki nilai akurasi 84.68%.

Penelitian lain juga telah dilakukan yang mengangkat permasalahan perbandingan analisis klasifikasi menggunakan metode *Naïve Bayes Classifier* (NBC), *k-Nearest Neighbour* (k-NN), dan *Random Forest* pada data ulasan film di *web portal* penyedia ulasan film IMDb (*Internet Movie Database*) (Baid, et al., 2017). Penelitian tersebut mengklasifikasikan data ulasan film kedalam label positif dan negatif. Berdasarkan hasil dan analisis penelitian tersebut, diperoleh kesimpulan bahwa metode NBC lebih baik dibandingkan dengan k-NN dan *Random Forest*, dimana nilai akurasi tertinggi diperoleh pada metode NBC yaitu sebesar 81.45%. Sedangkan akurasi pada masing-masing metode *Random Forest* dan k-NN yaitu sebesar 78.65% dan 55.30%.

Pada penelitian kali ini, peneliti akan menggunakan metode *Support Vector Machine* dan *Naïve Bayes Classifier* dengan menggunakan data *review* dari suatu *platform* penyedia *review* produk perawatan kulit. Hasil kajian penelitian ini diharapkan

dapat menjadi rujukan metode pada analisis sentimen untuk penelitian terkait ulasan produk *skincare*. Berikut Tabel 2.1 yang berisi rangkuman perbandingan metode dan nilai akurasi dari penelitian terdahulu:

Tabel 2. 1 Kajian Induktif

No.	*Author,* Tahun	Objek	Metode	Hasil Akurasi	Industri / Topik
1.	Dey, et al., 2020	Analisis sentimen pada ulasan produk Amazon	Metode *Naive Bayes classifier* dan *Support Vector Machine*	SVM: 84% NBC: 82.87%	*Retail Product & E-commerce*
2.	Bayhaqy, et al., 2018	Analisis sentimen tentang *e-commerce* dari Twitter	Metode *Decision Tree, K-Nearest Neighbour, and Naïve Bayes*	*Dec Tree*: 80% KNN: 78% NBC: 77%	*E-commerce*
3.	Siswanto, et al., 2022	Analisa sentimen publik mengenai perekonomian indonesia pada masa pandemi covid-19 di Twitter	Metode *K-Nearest Neighbour* dan *Support Vector Machine*	KNN: 76% SVM: 78%	Ekonomi
4.	Tuhuteru, 2020	Analisis sentimen masyarakat terhadap pembatasan sosial berksala besar	*Support Vector Machine*	SVM: 87.33%	Kebijakan Sosial
5.	Ma, et al., 2020	Klasifikasi spam email	*Naïve Bayes Classifier* dan *Support Vector Machine*	SVM: 95.5% NBC: 94.5%	*Email Services*
6.	Wilis, et al., 2020	Analisis sentimen sosial media pada	*Lexicon Based dan Support Vector Machine*	*Lexicon*: 88%	Bisnis (Penjualan *Souvenir*)

No.	*Author,* Tahun	Objek	Metode	Hasil Akurasi	Industri / Topik
		rekomendasi *souvenir*			
7.	Afdhal, et al., 2022	Analisis sentimen komentar di YouTube tentang islamofobia	*Random Forest*	RF *Random Forest* 79%	Agama
8.	Isnain, et al., 2021	Sentimen publik terkait kebijakan pembelajaran daring	*K-Nearest Neighbour*	KNN: 84.65%	Pendidikan
9.	Fikria, 2018	Analisis sentimen *review* aplikasi *e-ticketing* KAI Access dan tiket.com menggunakan pembobotan TF-IDF	*Support Vector Machine*	KAI Access: 89.60% Tiket.com: 84.68%	Transportasi
10.	Baid, et al., 2017	Analisis klasifikasi ulasan film menggunakan *software* Weka	*Naïve Bayes, k-Nearest Neighbour,* dan *Random Forest*	NBC: 81.45% *Random Forest*: 78.65% KNN: 55.30%	*Entertainment/ Movie Review*
11.	Mariel, et al., 2018	Analisis sentimen: perbandingan antara metode *Deep Learning Neural Network,* SVM dan NBC pada data *tweets* mengenai institusi pemerintah Indonesia	*Naïve Bayes Classifier, Support Vector Machine* dan *Deep Learning Neural Network*	NBC: 85% SVM: 88% DLNN: 98%	Indonesia *Government Institution*

No.	*Author*, Tahun	Objek	Metode	Hasil Akurasi	Industri / Topik
12	Ababneh, 2019	Aplikasi *Naïve Bayes, Decision Tree,* dan KNN untuk klasifikasi text terautomasi pada data artikel dari Saudi Press Agency (SPA)	*Naïve Bayes, Decision Tree,* dan *K-Nearest Neighbour*	NBC: 88.1% KNN: 85.2% *Decision Tree*: 82.3%	*Press Agency*/ Artikel Saudi *Press Agency* (SPA)
13	Pranckevičius & Marcinkevičius, 2017	Komparasi klasifikasi *text review* prouk Amazon dengan *Naïve Bayes, Random Forest, Decision Tree, Support Vector Machines,* dan *Logistic Regression Classifiers*	*Naïve Bayes, Random Forest, Decision Tree, Support Vector Machines,* dan *Logistic Regression Classifiers*	NBC: 45.22% *Random Forest*: 43.93% SVM: 44.06% *Regression*: 58.5% *Decision Tree*: 34.58%	*Retail Product & E-commerce*
14	Hermansyah & Sarno, 2020	Analisis sentimen pada *tweets* produk dan jasa PT. Telekomunikasi Indonesia Tbk	*TextBlob, Naïve Bayes,* dan *K-Nearest Neighbour*	*TextBlob*: 54.67% NBC: 69.44% KNN: 75%	*Product & Service*
15	Hapsari, et al., 2021	Analisis sentimen terhadap *review* produk kosmetik (Bahasa Indonesia) di *e-commerce*	*Naïve Bayes Classifier* dan *Word2Vec*	NBC & *Word2Vec*: 68.17%	*Cosmetics/ Beauty Products*

No.	*Author*, Tahun	Objek	Metode	Hasil Akurasi	Industri / Topik
16	Utami, 2018	Analisis sentimen *review* produk kosmetik di *website beauty forum*	*Naïve Bayes Classifier* dan TF-IDF	NBC & TF-IDF: 80%	*Cosmetics/ Beauty Products*
17	Arthamevia & Purbolaksono, 2021	Analisis sentimen berbasis aspek pada review produk kecantikan di FemaleDaily	TF-IDF dan *Support Vector Machine*	SVM & TF-IDF: 88.35%	*Cosmetics/ Beauty Products*
18	Kirana & Faraby, 2021	Analisis sentimen pada *review* produk kecantikan di FemaleDaily	*K-Nearest Neighbour* dan Metode TF-IDF dengan *Chi-Square Feature Selection*	KNN: 71% TF-IDF dengan *Feature Selection*: 71% TF-IDF tanpa *Feature Selection*: 70.75%	*Cosmetics/ Beauty Products*
19	Indrayuni & Nurhadi, 2020	Optimisasi algoritma genetic untuk analisis sentiment pada *review* produk Apple	*Support Vector Machines* dan *Genetic Algorithms*	SVM: 70% SVM & GA: 85.76%	*Electronic Products*
20	Dhini & Kusumaningrum, 2018	Analisis sentimen terhadap *review customer* bandara	*Naïve Bayes* dan *Support Vector Machine*	NBC: 65.7% SVM: 67%	*Aviation Services*

Berdasarkan hasil penelitian terdahulu menunjukkan performansi metode *Support Vector Machine* dan *Naïve Bayes Classifier* yang terbilang cukup baik. Namun sampai saat ini belum ditemukan penelitian yang menggunakan metode *Support Vector Machine* dan *Naïve Bayes Classifier* untuk analisis sentimen pada industri kecantikan dengan produk perawatan kulit khususnya untuk produk *toner* dari *brand* “The Body Shop”

sebagai objeknya. Selain itu, berdasarkan Tabel 2.1, dapat diketahui bahwa metode *Support Vector Machine* dan *Naïve Bayes Classifier* memberikan rata-rata nilai akurasi yang cenderung cukup tinggi daripada metode lainnya khususnya untuk *use-case* mengenai *review* produk kecantikan atau penggunaan kasus pada industri kosmetik. Oleh karena itu, dengan didukungnya performa baik kedua metode tersebut pada *use-case* berupa data *review* produk kecantikan, maka perlu adanya penelitian ini untuk mendapatkan gambaran kinerja produk maupun *brand sentiment* yang kemudian dapat dijadikan sebagai bahan evaluasi pengembangan bisnis, khususnya pada produk *Tea Tree Skin Clearing Mattifying Toner* dari *brand* The Body Shop yang berada dibawah perusahaan Natura & Co Holding menggunakan metode *Support Vector Machine* dan *Naïve Bayes Classifier*.

2.2 Kajian Deduktif

2.2.1 *Brand* The Body Shop

Kisah The Body Shop dimulai di Brighton, Inggris pada tahun 1976. Dimulai dengan pendirinya, Dame Anita Roddick, dan keyakinannya pada sesuatu yang revolusioner: bahwa bisnis dapat menjadi kekuatan untuk kebaikan. Mengikuti visinya, The Body Shop telah melanggar aturan, tidak pernah berpura-pura dan membuat perubahan sejak saat itu. Sebagai B-Corp bersertifikat, The Body Shop mengoperasikan sekitar 2800 lokasi ritel di lebih dari 70 negara yang selalu menghadirkan perawatan kulit, perawatan tubuh, perawatan rambut, dan rias wajah berkualitas tinggi yang terinspirasi secara alami, diproduksi secara etis dan berkelanjutan (The Body Shop, 2021).

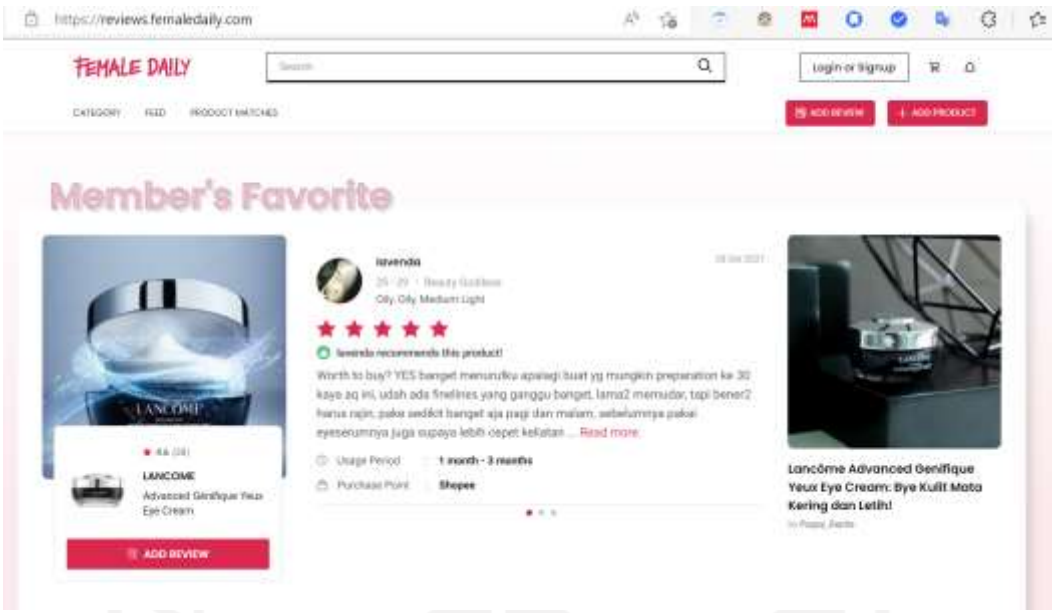
The Body Shop® adalah produsen produk kecantikan atau perawatan tubuh, wajah, rambut, perlengkapan mandi, hingga perharum tubuh dan ruangan yang berkelanjutan dan bertanggung jawab secara lingkungan. Seluruh produk The Body Shop diperkaya dengan bahan baku yang alami, tidak diujicobakan pada hewan, dan ramah lingkungan (The Body Shop, 2022).

Gambar 2. 1 *Logo Brand* "The Body Shop"

The Body Shop Indonesia mendorong para pekerjanya untuk memperlajari keterampilan baru melalui program *Learning is of Value to Everyone* (LOVE). Perusahaan mendanai berbagai program pelatihan, kegiatan kebersamaan, dan perawatan kesehatan untuk meningkatkan kesejahteraan staf kami. Perusahaan The Body Shop percaya untuk menjadi pribadi yang lebih baik hanya ada satu cara sederhana, yaitu berbuat kebaikan. Inilah yang menjadi alasan perusahaan untuk memiliki kebijakan global bagi staf perusahaan: meluangkan minimal 3 hari pertahun untuk melakukan kerja sukarela dengan prinsip, *small act can be a big impact for others*.

2.2.2 Female Daily

Female Daily merupakan media dan komunitas kecantikan terbesar di Indonesia yang sudah berdiri selama 15 tahun. Female Daily kini berada di bawah badan hukum PT Daily Dinamika Kreasi dengan memiliki 7,8 juta *page view* per bulan dan 18 ribu *forum thread* (Wardani, 2017). Layanan Female Daily dapat diakses melalui aplikasi android, ios, dan *web*. *Female Daily Network* berdiri untuk menemani para *beauty enthusiast* dalam merawat diri, mengenal produk-produk kecantikan, memperluas wawasan, bersosialisasi, bahkan membantu menciptakan sosok-sosok yang menginspirasi. Melalui Female Daily, para *beauty enthusiast* dapat mengenal berbagai produk kecantikan mulai dari makeup, *skincare, body care, hair care,* hingga *tools* melalui Female Daily. Female Daily seringkali membahas produk terkini, mengulas rekomendasi produk, *beauty news*, sampai *tips* dan trik melalui artikel.



Gambar 2. 2 *Website* Female Daily

Female Daily memiliki fitur ulasan produk yang berisikan *review* dan untuk mencari produk yang sesuai dengan kebutuhan pengguna. Fitur lain pada Female Daily yaitu *workshop* kecantikan, tutorial kecantikan, opini pakar, opini konsumen, dan forum yang membahas serta mengulas rekomendasi produk.

Ada lebih dari 1.500.000 *honest reviews*, lebih dari 50.000 produk dari sekitar 2.400 *brands* yang ada di FD *Beauty Review*. *Review* yang sangat banyak dan *reliable* ini bisa membantu pembaca dalam mencari sebuah produk. Saat ini, membaca *review* terpercaya adalah langkah yang sangatlah penting sebelum melakukan pembelian. Dengan begitu, calon konsumen dapat mengetahui produk yang paling sesuai dengan kebutuhan. Konsumen juga bisa lebih mudah memilah-milah mana yang harus dibeli duluan dan belakangan. Hal ini akan memudahkan konsumen menemukan produk yang tepat (Female Daily, 2021).

2.2.3 *Brand Management*

Definisi *brand* adalah nama, istilah, tanda lambang, desain atau kombinasi untuk mengidentifikasi barang atau jasa dari salah satu penjual atau kelompok penjual dan mendiferensiasikan mereka dari para pesaing (Zikmund, et al., 2009). Penentuan *brand* sudah ada selama berabad-abad yang menjadi sarana untuk membedakan suatu barang

atau produk yang dihasilkan oleh satu perusahaan dengan perusahaan lainnya. Saat ini *brand* memegang peranan cukup penting bagi suatu perusahaan yang dapat mempengaruhi nilai perusahaan dan dapat mempengaruhi konsumen untuk menentukan pilihan pada perusahaan tersebut.

2.2.4 *User Generated Content* (UGC)

Jejaring sosial media dinilai dapat mempengaruhi konsumen yang membutuhkan informasi hingga sampai pada tahap pembelian (Rachmania, 2021). Interaksi sosial antara konsumen melalui media sosial memberi konsumen kesempatan untuk berbagi ulasan dan untuk *sharing of user-generated review* di antara pengguna media sosial (Casaló, Flavián, & Guinalíu, 2011). Percakapan yang ditulis oleh pengguna di media sosial disebut *User Generated Content* (UGC). UGC adalah data yang diposting oleh pengguna dan dapat dilihat secara publik oleh orang lain (Moens, Li, & Chua, 2014). Pengguna ini termasuk pengguna individu dan pengguna dari perusahaan di berbagai industri. Menurut hasil survei Desember 2019 yang dilakukan terhadap 1.005 pengguna internet AS, sebanyak 76% orang mempercayai *online review* sebagaimana rekomendasi dari keluarga dan teman (Murphy, 2019).

2.2.5 Web Scrapping

Scraping web adalah teknik yang digunakan untuk mengumpulkan informasi secara otomatis dari situs *web* tanpa duplikasi manual. Tujuan dari teknologi ini adalah untuk menemukan informasi yang diperlukan dan kemudian mengumpulkannya ke dalam jaringan baru melalui pengambilan dan ekstraksi (Yani, et al., 2019). Proses *web scrapping* atau pengambilan sebuah dokumen semi-terstruktur dari internet atau pada halaman-halaman website menggunakan bahasa *markup* seperti HTML (*HyperText Markup Language*) atau XHTML (*Extensible HyperText Markup Language*) (Chapman, et al., 2010). Menurut Josi, et al. (2014) terdapat beberapa langkah dalam proses *web scraping*, yaitu:

1. *Create Scraping Template*

Pembuat program mempelajari dokumen HTML dari *website* yang akan diambil informasinya dari tag HTML.

2. *Explore Site Navigation*

   Pembuat program mempelajari teknis navigasi pada *website* yang akan diambil informasinya untuk ditirukan pada aplikasi *web scraper*.

3. *Automate Navigation and Extraction*

   Berdasarkan informasi yang didapatkan dari langkah 1 dan 2 diatas, aplikasi *web scraper* dibuat untuk mengotomatisasi pengambilan informasi dari *website* yang ditentukan.

4. *Extracted Data and Package History*

   Informasi yang didapat dari langkah 3 disimpan dalam tabel atau tabel-tabel *database*.

2.2.6 Machine Learning

Machine learning didefinisikan sebagai aplikasi computer dan penerapan algoritma matematika yang diadaptasi dengan cara pembelajaran yang diperoleh dari data dan memiliki output berupa prediksi di masa yang akan dating. Proses pembelajaran yang dimaksud adalah usaha yang dilakukan agar memperoleh kecerdasan melalui dua tahapan yaitu, pelatihan *(training)* dan pengujian *(testing)* (Roihan et al., 2020). Machine learning merupakan proses membangun program computer agar mampu meningkat secara otomatis melalui pembelajaran yang dilakukan yang didasarkan kepada pengalaman.

Machine learning terbagi menjadi 3 jenis yaitu *supervised learning*, *unsupervised learning* dan *reinforcement learning* (Somvanshi & Chavan, 2016). Dalam supervised learning Teknik yang digunakan adalah analisis klasifikasi yang mana data yagn digunakan telah diberikan label secara menyeluruh yang digunakan untuk mengklasifikasikan kelas yang tidak dikenal. Sedangkan *unsupervised learning* sering dikenal dan disebut sebagai *cluster*. Hal ini dikarenakan data yang digunakan tidak perlu untuk diberikan label dan hasil dari *cluster* sendiri tidak mengidentifikasi contoh dikelas yang telah ditentukan (Thupae, et al., 2018). Untuk jenis *machine learning* yang terakhir yaitu *reinforcement learning*, merupakan jenis *machine learning* yang mampu bekerja di lingkungan yang dinamis yang mana konsepnya harus menyelesaikan tujuan yang telah

ditentukan tanpa ada pemberitahuan dari komputer secara eksplisit apakah tujuan telah tercapai atau belum (Das & Nene, 2017).

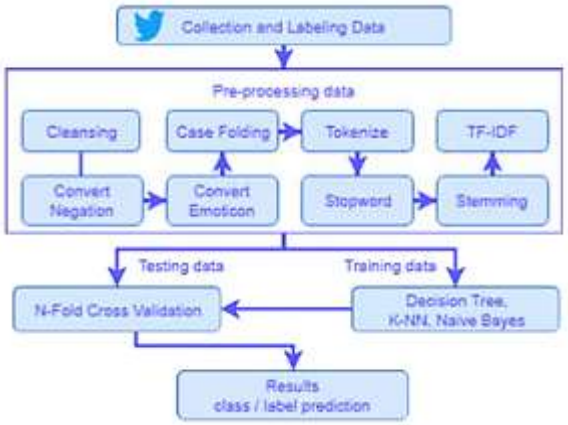
2.2.7 *Data Mining*

Data mining adalah proses yang memanfaatkan teknik pembelajaran komputer (*machine learning*) untuk menganalisis, mengekstraksi dan menemukan pola data penting dan menarik ataupun pengetahuan (*Knowledge*) secara otomatis dari basis data yang besar (*big data*) (Eska, 2018). Definisi lain dari data mining adalah pembelajaran berbasis induksi uang mana hal ini merupakan proses pembentukan berbagai definisi konsep umum yang dilakukan dengan cara mengamati beberapa contoh yang spesifik dari konsep-konsep yang dipelajari. Data mining sendiri masuk ke dalam *Knowledge Discovery in Database (KDD)*. *Knowledge Discovery in Database (KDD)* adalah penerapan metode yang digunakan untuk mencari pola dari suatu data yang memiliki sifat sah, baru, dapat bermanfaat, dan dapat dimengerti (Gilchrist & Mooers, 2012).
Berikut ini merupakan beberapa Teknik dan sifat dari *data mining*:

1) *Classification (Predictive)*
2) *Clustering (Descriptive)*
3) *Association Rules (Descriptive)*
4) *Regression (Descriptive)*
5) *Deviation Detection (Predictive)*

2.2.8 *Text Pre-processing*

Text-Preprocessing merupakan tahapan awal yang dilakukan untuk mempersiapkan teks agar dapat diolah lebih lanjut. *Pre-processing* secara umum bertujuan untuk mengubah informasi dari tiap-tiap sumber data ke dalam bentuk atau *format* yang baku sebelum menerapkan berbagai metode-metode pengambilan data terhadap dokumen yang akan diproses (Jaka, 2015).



Gambar 2. 3 Diagram Sistem Klasifikasi untuk Sentimen Analisis

Berdasarkan Gambar 2.1 dapat diketahui terdapat beberapa tahapan *pre-processing* untuk *sentiment* analisis, antara lain:

1. *Cleaning Data*, merupakan proses pembersihan kata dengan menghilangkan tanda baca yang bertujuan untuk mengurangi *noise*.
2. *Case Folding*, merupakan pengubahan semua karakter huruf diubah menjadi dokumen yang berisi kalimat menjadi huruf kecil atau sebaliknya. Pada tahap *case folding* juga menghilangkan yang dianggap tidak valid atau menghilangkan karakter selain bentuk huruf seperti angka dan lain-lain.
3. *Tokenizing*, adalah proses memotong suatu kalimat menjadi beberapa bagian berdasarkan kata perkata. Potongan kata perkata tersebut disebut dengan token.
4. *Filtering*, merupakan tahap pembersihan kalimat dengan membuang kata-kata yang tidak signifikan, seperti kata ganti, kata hubung, kata keterangan, dan lain-lain dengan menggunakan *stopword* yakni daftar kata yang akan dihapus pada dokumen, pada *filtering* juga dilakukan penghapusan terhadap spasi yang berlebih akibat dari penghapusan beberapa kata.

2.2.9 Pembobotan Kata *Term Frequency - Inverse Doc Frequency*

Tahapan penting setelah dilakukannya *preprocessing* data yaitu pembobotan kata, yang berguna untuk mengubah data yang belum terstruktur menjadi lebih terstruktur dan nantinya akan dikategorikan menggunakan metode *classifier*. Hasil pembobotan kata dapat berbeda berdasarkan metode yang digunakan dalam melakukan pembobotan kata

tersebut (Ramadhan, et al., 2021). Salah satu metode yang digunakan dalam pembobotan kata yaitu *Term Frequency - Inverse Doc Frequency* (TF-IDF).

Term Frequency - Inverse Doc Frequency merupakan metode penggabungan dua konsep untuk melakukan perhitungan bobot, yaitu frekuensi kemunculan sebuah kata dalam suatu dokumen dan *inverse* frekuensi dokumen yang mengandung kata tersebut (Nurjannah, et al., 2013). TF-IDF merupakan sebuah ukuran statistik yang digunakan untuk mengevaluasi seberapa penting sebuah kata didalam sebuah dokumen atau sekelompok kata. Metode ini dikenal sebagai metode pembobotan kata yang efisien, mudah, dan memiliki hasil yang akurat (Melita, et al., 2018). Terdapat beberapa metode pembobotan, seperti:

a. *Term Presence* (TP)

   Term Presence (TP) merupakan metode pembobotan pada suatu dokumen teks yang melihat keberadaan daftar kata-kata (*term*) atau fitur yang ada pada corpus terhadap suatu dokumen. Jika fitur yang ada pada daftar fitur acuan terdapat pada dokumen yang sedang diboboti maka nilai fitur tersebut pada feature vector akan diberi nilai 1 dan tidak menghiraukan jumlah kemunculan fitur tersebut. Jika fitur tersebut tidak ada pada dokumen maka nilai 0 pada *feature space* (O'Keefe & Koprinska, 2009).

b. *Term Frequency* (TF)

   Term Frequency (TF) memiliki kesamaan dengan TP yang telah dijelaskan sebelumnya, tapi yang membedakan adalah TF menghitung jumlah kemunculan fitur acuan pada suatu dokumen bukan hanya keberadaan fitur tersebut (O'Keefe & Koprinska, 2009).

2.2.10 Analisis Sentimen

Analisis sentimen merupakan salah satu metode yang berguna untuk mengekstrak data opini, memahami serta mengolah tekstual data secara otomatis untuk melihat sentimen yang terdapat dalam suatu opini (Sari & Wibowo, 2019). Analisis sentimen dapat menghasilkan satu daftar atribut produk (misal, kualitas, fitur, dan lain-lain) serta dapat menghitung agregasi dari opini masing-masing atribut (rendah, sedang, tinggi) (Ramdhani & Rahim, 2014).

Analisis sentimen dapat menyatakan perasaan emosional, sedih, gembira atau marah serta dapat mencari pendapat terkait produk, merek atau orang dan menentukan apakah mereka dilihat positif atau negatif di web (Saraswati, 2011). Pengkategorian analisis sentimen terbagi menjadi tiga *task,* yaitu *informative text detection, information extraction* dan *sentiment interestingness classification (emotional, polarity indentification)* (Habibi, et al., 2016).

2.2.11 Klasifikasi

Klasifikasi adalah bentuk analisis data yang menyajikan model prediksi untuk menggambarkan label atau kelas data (Han & Kamber, 2011). Model klasifikasi (*classifier*) digunakan untuk memprediksi dan mendeskripsikan label data kategori (*label* kelas), sedangkan model prediksi numerik (regresi) digunakan untuk memprediksi fungsi data kontinu. Klasifikasi memiliki fungsi untuk menemukan sekumpulan model yang menjelaskan dan membedakan kelas-kelas data, sehingga model tersebut dapat digunakan untuk memprediksi nilai suatu kelas yang belum diketahui pada sebuah objek (Gorunescu, 2011).

Klasifikasi termasuk dalam *supervised methods*, karena label kelas data telah tersedia sebelum dilakukan proses klasifikasi. *Supervised methods* merupakan suatu metode untuk mendapatkan hubungan antara atribut input (variabel independen) dan atribut target/kelas (variabel dependen). Hubungan yang ditemukan diwakili dalam struktur yang disebut sebagai model. Biasanya model menggambarkan dan menjelaskan fenomena yang tersembunyi dalam *dataset* dan dapat digunakan untuk memprediksi nilai atribut target mengetahui nilai-nilai atribut masukan (Maimon & Rokach, 2009).

Algoritma klasifikasi bermacam-macam dan digunakan secara luas, yaitu *Decision/Classification Trees, Bayesian Classifiers/ Naïve Bayes Classifiers, Neural Networks, Analisa Statistik, Algoritma Genetika, Rough sets, K-Nearest Neighbor, Metode Rule Based, Memory Based Reasoning*, dan *Support Vector Machines* (SVM) (Jananto, 2013).

2.2.12 Naïve Bayes Classifier

Naive Bayes merupakan salah satu algoritma yang digunakan dalam *text mining*. Metode NBC menempuh dua tahap dalam proses klasifikasi teks, yaitu tahap pelatihan dan tahap klasifikasi. Pada tahap pelatihan dilakukan proses analisis terhadap sampel dokumen berupa pemilihan *vocabulary*, yaitu kata yang mungkin muncul dalam koleksi dokumen sampel yang sedapat mungkin dapat menjadi representasi dokumen. Selanjutnya adalah penentuan probabilitas *prior* bagi tiap kategori berdasarkan sampel dokumen. Pada tahap klasifikasi ditentukan nilai kategori dari suatu dokumen berdasarkan term yang muncul dalam dokumen yang diklasifikasi (Rozi, et al., 2020). *Bayesian classification* diketahui mempunyai tingkat akurasi dan kecepatan yang tinggi ketika diaplikasikan pada *database* dengan data yang besar. Berikut adalah rumus Naïve Bayes (Utomo & Purba, 2021):

$$\boldsymbol{P(H|X) = \frac{P(X|H)P(H)}{P(X)}} \quad (1)$$

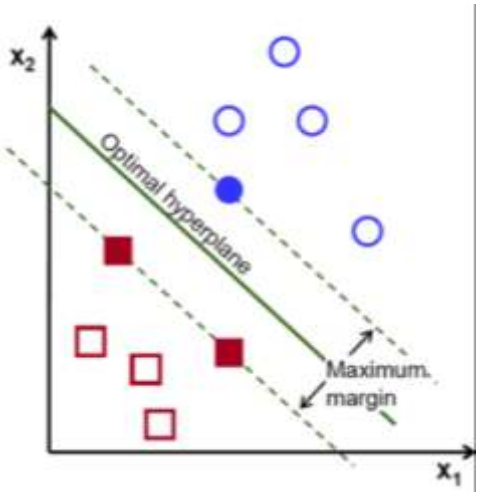
Keterangan:

X	= Data dengan *class* yang belum diketahui
H	= Hipotesis data x merupakan suatu class spesifik
P(H|X)	= Probabilitas hipotesis H berdasarkan kondisi x
P(H)	= Probabilitas hipotesis H
P(X|H)	= Probabilitas X berdasarkan kondisi tersebut
P(X)	= Probabilitas dari X

2.2.13 Support Vector Machine (SVM)

SVM merupakan salah satu metode *supervised learning* yang memiliki kemampuan untuk generalisasi dalam mengklasifikasi suatu pola dan memiliki kelebihan untuk mengidentifikasi *hyper lane* terpisah yang memaksimalkan margin antar kelas (Kristiyanti, 2015). SVM bekerja pada prinsip *Structural Risk Minimization* (SRM) dan memiliki konsep landasan teori yang lebih matang, sehingga dapat dianalisis dengan jelas secara matematis dibandingkan dengan klasifikasi lainnya. Namun metode ini memiliki kekurangan, yaitu sensitif terhadap pemilihan fitur. Selain itu, penggunaan parameter

akan mempengaruhi hasil akurasi klasifikasi. Inti dari proses pelatihan *Support Vector Machine* (SVM) adalah usaha untuk mencari lokasi *hyper lane* yang optimal.



Gambar 2. 4 Dua Kelas yang Dipisahkan oleh *Hyperlane1*

Hyperlane memiliki fungsi untuk memisahkan antara kedua kelas pada *input space*. Divisualisasikan garis solid seperti pada Gambar 2.3 merupakan hyperlane terbaik. Posisinya tepat berada di antara kedua kelas. *Hyperlane* terbaik adalah bidang yang memisahkan data dan memiliki *margin* yang besar. *Margin* merupakan jarak antara *hyperlane* dengan data dari masing-masing kelas. *Support vector* atau subset data training set merupakan data terluar yang berada paling dekat dengan bidang pembatas (garis putus-putus).

2.2.14 Confusion Matrix

Untuk menggambarkan seberapa baik sistem klasifikasi yang digunakan menggunakan pengukuran kinerja dari suatu sistem. Salah satu metode yang digunakan sebagai pengukuran kinerja klasifikasi yaitu *confusion matrix*. *Confusion matrix* dapat diartikan sebagai alat yang berfungsi untuk melakukan analisis apakah *classifier* tersebut dapat mengenali *tuple* dari kelas yang berbeda (Han & Kamber, 2011). *Confusion matrix* mengandung nilai *true positive, true negative, false positive,* dan *false negative*. Nilai dari *true positive* dan *true negative* memberikan informsai bahwa ketika *classifier* dalam melakukan klasifikasi data yang bernilai benar, dan sedangkan *false negative* dan *false*

positive memberikan informasi bahwa ketika *classifier* salah dalam melakukan pengklasifikasian data.

Pengukuran efektif dapat dilakukan dengan perhitungan perolehan atau recall, nilai ketepatan atau presisi, nilai akurasi, dan nilai *specificity*. *Recall* merupakan proporsi jumlah yang dapat ditemukan kembali dalam proses pencarian. Presisi merupakan proporsi jumlah dokumen yang ditemukan dan dianggap relevan untuk kebutuhan suatu informasi. Akurasi adalah nilai ketepatan suatu klasifikasi dalam bentuk persen dan *specificity* digunakan untuk mengukur proporsi negatif yang benar diidentifikasi (Wijaya & Santoso, 2016).

Tabel 2. 2 *Binary Confusion Matrix*

	Actual Positive	***Actual Negative***
Predicted Positive	*True Positive (TP)*	*False Positive (FP)*
Predicted Negative	*False Negative (FN)*	*True Negative (TN)*

1. *True Positive* (TP) merupakan kelas yang dihasilkan dari prediksi pada klasifikasi positif dan kelas sebenarnya positif.
2. *True Negative* (TN) merupakan kelas yang dihasilkan dari prediksi pada klasifikasi negatif padahal kelas sebenarnya positif.
3. *False Positive* (FP) merupakan kelas yang dihasilkan dari prediksi pada klasifikasi negatif padahal kelas sebenarnya positif.
4. *False Negative* (FN) meripakan kelas yang dihasilakan dari prediksi pada klasifikasi negatif dan kelas sebenarnya negatif.

Nilai *Area Under Curve* (AUC) digunakan untuk mengukur kinerja deskriminatif menggunakan perkiraan probabilitas hasil dari sampel yang telah dipilih secara acak dari suatu populasi negatif dan positif. Nilai AUC berkisar antara 0 sampai 1, klasifikasi dikatakan baik jika nilai AUC semakin tinggi.

Tabel 2. 3 Nilai *Area Under Curve* (AUC)

Nilai AUC	**Keterangan**
0,91 - 1,00	Klasifikasi Sangat Baik
0,81 - 0,90	Klasifikasi Baik
0,71 - 0,80	Klasifikasi Cukup
0,61 - 0,70	Kasifikasi Buruk

Nilai AUC	Keterangan
≤ 0,6	Klasifikasi Salah

(Gorunescu, 2011)

2.2.15 Word Cloud

Word cloud merupakan salah satu metode populer dalam *text mining* untuk memvisualisasikan data teks secara visual. *Word cloud* merupakan representasi grafis dari suatu dokumen dengan dilakukan *plotting* kata-kata yang sering muncul pada dokumen tersebut (Castella & Sutton, 2014). Dengan menggunakan *word cloud*, *output* dari suatu teks nantinya ditampilkan dalam bentuk menarik namun tetap informatif. Semakin banyak data yang digunakan maka akan semakin besar pula ukuran kata yang ditampilkan dalam *word cloud* (Pradana, 2020). Berikut merupakan contoh *word cloud:*

Gambar 2. 5 Visualisasi *Word Cloud*

2.2.16 Asosiasi Kata

Asosiasi kata dihitung dengan memperkirakan nilai korelasi. Umumnya nilai korelasi digunakan untuk menyatakan hubungan antara dua atau lebih variabel kuantitatif, tetapi dalam asosiasinya istilah nilai korelasi diartikan sebagai hubungan erat antara dua variabel kuantitatif atau lebih. (Ulwan, 2016). Jika nilainya mendekati 1 atau -1, hubungan antar kata akan semakin kuat, dan jika nilainya mendekati 0, hubungan antar kata akan semakin lemah. Ada beberapa kategori nilai korelasi yang digunakan sebagai berikut (Disa, 2019).

0	: Tidak ada korelasi antar dua variabel
0-0,25	: Korelasi lemah
0,25 – 0,5	: Korelasi cukup
0,5 0,75	: Korelasi kuat
1	: Korelasi sangat kuat

BAB III

METODE PENELITIAN

3.1 Objek Penelitian

Objek penelitian akan diperlukan dalam sebuah penelitian karena melalui hal tersebut dapat diketahui variabel-variabel penting yang akan mendukung penelitian. Pada penelitian yang akan dilakukan, objek penelitian yang akan ditujukan berasal dari ulasan pengguna produk *Tea Tree Skin Clearing Toner* dari *Beauty Brand* "The Body Shop" di Female Daily.

3.2 Populasi dan Sampel Penelitian

Populasi dalam penelitian ini adalah semua ulasan atau *review* pengguna produk *Tea Tree Skin Clearing Toner "*The Body Shop*"* dari *database website* Female Daily. Sedangkan untuk sampel yang digunakan adalah ulasan produk *Tea Tree Skin Clearing Toner* pada tanggal 24 Juni 2018 – 11 September 2022 sebanyak 1.075 data ulasan.

3.3 Metode Pengumpulan Data

Pada penelitian ini, metode pengumpulan data yang digunakan adalah metode *web scraping*. Alat yang digunakan dalam metode ini adalah "Data Miner" yang merupakan *extension* dari Google Chrome.

3.4 Jenis dan Sumber Data

Sumber data dalam penelitian ini yaitu data sekunder yang terdiri dari:

a. Sumber data sekunder yang digunakan dalam pengolahan data didapatkan melalui proses *scrapping data* secara *online* di Female Daily pada ulasan produk *Tea Tree Skin Clearing Toner* dari *Brand* "The Body Shop" menggunakan *extensions* Google Chrome yaitu Data Miner pada alamat *website* https://reviews.femaledaily.com/products/cleanser/toner/the-body-shop/tea-tree-toner-22?cat=&cat_id=0&age_range=&skin_type=&skin_tone=&skin_undertone=&hair_texture=&hair_type=&order=newest&page=1
b. Data sekunder lainnya yang digunakan untuk membantu proses penelitian didapatkan melalui referensi beberapa jurnal ilmiah mengenai teori-teori yang memiliki permasalahan yang sama dengan penelitian yang akan dilakukan. Penelitian yang akan dilakukan tidak hanya membutuhkan teori-teori, namun juga membutuhkan referensi dari penelitian terdahulu yang sudah pernah dilakukan. Selain melalui jurnal ilmiah, sumber referensi dapat diperoleh melalui beberapa artikel terpercaya.

3.5 Variabel Penelitian

Pada penelitian ini terdapat dua macam variabel yang digunakan, yaitu:

a. *Date*, merupakan tanggal dibuatnya ulasan produk oleh pengguna.
b. *Review*, merupakan isi ulasan pengguna.

3.6 Metode Analisis Data

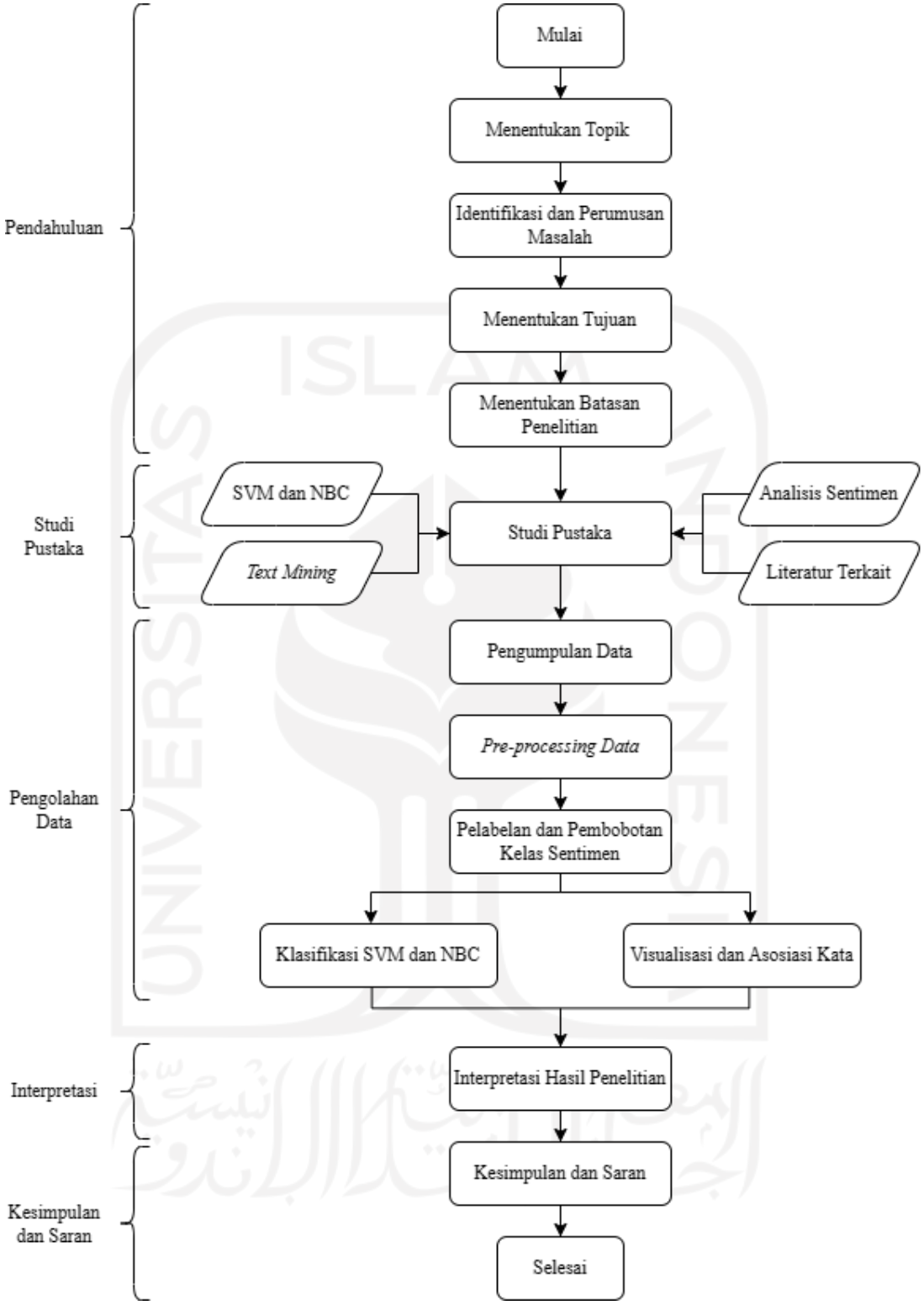
Dalam penelitian ini, digunakan bahasa pemrograman Python dengan bantuan *software Google Collaboratory* dan Microsoft Excel 2016 serta R Studio khususnya untuk proses pelabelan sentimen menggunakan *lexicon-based*. Terdapat beberapa metode analisis data yang digunakan dalam penelitian ini, antara lain:

a. Analisis sentimen, digunakan untuk melakukan pelabelan data ke dalam kelas sentimen positif dan negatif.

b. Metode *machine learning*, yaitu metode *Support Vector Machine* (SVM) dan *Naïve Bayes Classifier* (NBC) digunakan untuk mengklasifikasikan ulasan yang berbentuk positif dan negatif.
c. *Wordcloud*, digunakan untuk melakukan visualisasi kata yang paling sering muncul digunakan dalam ulasan.

3.7 Alur Penelitian

Pada sub-bab ini, akan dijelaskan tahapan penelitian tugas akhir yang telah dilakukan mulai dari penentuan topik, *scrapping data*, *pre-processing data* hingga mendapatkan kesimpulan dan saran. Berikut merupakan alur penelitian yang dilakukan dalam penelitian ini:



Gambar 3. 1 Alur Penelitian

Berdasarkan alur penelitian tahapan yang akan dilakukan sebagai berikut:

1. Mulai
2. Menentukan Topik

Tahapan penelitian diawali dengan menentukan topik penelitian. Topik penelitian yang diangkat yaitu perbandingan analisis sentimen pada produk perawatan kulit *Tea Tree Skin Clearing Toner* dari *brand* The Body Shop yang dibangun oleh perusahaan Natura & Co Holding.

3. Identifikasi dan Perumusan Masalah

   Pada tahap ini dilakukan identifikasi masalah dengan melihan kondisi dan situasi saat. Hasil identifikasi didapatkan informasi bahwa sebeum datangnya banyak *brand* baru yang bermunculan, The Body Shop memiliki reputasi *beauty brand* yang sangat baik ditandai dengan dikenalnya oleh banyak masyarakat. Dengan datangnya beberapa *brand* baru baik lokal maupun internasional, The Body Shop harus memiliki strategi yang *sustainable* agar bisa menjaga *loyalty* konsumen yang ada maupun menarik calon konsumen baru. Saah satunya adalah dengan pengembangan produk berdasarkan *sharing of user-generated review* di antara pengguna media sosial para pengguna produk The Body Shop. Berdasarkan *rating* pada Female Daily, terdapat produk dari The Body Shop yang menunjukkan performa produk yang rendah/buruk atau dapat dikatakan belum sesuai harapan *customer*. Selanjutnya dilakukan perumusan masalah yang akan menjadi pedoman dalam penelitian ini. Berdasarkan permasalahan yang didapatkan, peneliti merumuskan masalah menjadi bagaimana hasil analisis penerapan metode dan kinerja model pengklasifikasian *sentimen* terhadap ulasan produk, bagaimana pembentukan kata yang didapatkan berdasarkan hasil klasifikasi kata dalam bentuk *word cloud* dan asosiasi kata, dan bagaimana identifikasi faktor penyebab permasalahan yang dihadapi oleh pengguna produk *toner* "The Body Shop" serta solusi permasalahan atas keluhan tersebut?

4. Menentukan Tujuan

   Tujuan penelitian ini menjelaskan hal-hal yang ingin dicapai dalam penelitian ini. Tujuan dalam penelitian ini adalah mengetahui hasil analisis penerapan metode dan kinerja model pengklasifikasian *sentimen* terhadap ulasan produk toner The Body Shop, mengetahui analisis pembentukan kata yang didapatkan berdasarkan hasil klasifikasi kata dalam bentuk *word cloud* dan asosiasi kata, dan mengetahui identifikasi faktor penyebab permasalahan yang dihadapi oleh pengguna produk *Tea Tree Skin Clearing Mattifying Toner* "The Body Shop" serta solusi permasalahan atas keluhan tersebut.

5. Menentukan Batasan Masalah

Dalam penelelitian ini digunakan batasan penelitian agar penelitian tetap terarah dan pembahasan tidak melebar kepada hal yang tidak perlu. Adapun batasan dalam penelitian ini yaitu data yang digunakan merupakan data ulasan mengenai produk *Tea Trea Skin Clearing Toner* dari *brand* The Body Shop pada *website* Female Daily yang diambil pada tanggal 24 Juni 2018 – 11 September 2022 dan ulasan yang diambil merupakan ulasan yang menggunakan bahasa Indonesia

6. Studi Pustaka

   Pada tahap ini dilakukan studi pustaka yang berkaitan dengan topik analisis sentimen untuk dapat mengetahui metode yang sesuai untuk penelitian ini. Terdapat beberapa topik yang digunakan dalam studi pustaka, seperti *Text Mining*, Klasifikasi, dan *Wordcloud*.

7. Pengumpulan Data

   Pada tahap ini dilakukan pengumpulan secara *online* dengan teknik *scraping data* dengan bantuan ekstensi dari Google Chrome yaitu Data Scraper. Data yang digunakan yaitu data ulasan produk *Tea Trea Skin Clearing Toner* dari *brand* The Body Shop pada *website* Female Daily.

8. *Pre-processing* Data

   Pre-processing merupakan suatu proses yang memiliki tujuan untuk menyeleksi data-data yang tidak diperlukan agar data menjadi lebih terstruktur, selain itu informasi yang didapatkan akan menjadi jelas. Pada preprocessing ini terdapat beberapa proses, yaitu:

   a. *Data Cleaning*: tahap ini merupakan tahap penghapusan teks yang tidak diperlukan seperti simbol, emotikon, tanda baca, dan lain hal sebagainya.
   b. *Case Folding*: tahapan untuk mengubah teks menjadi lowercase, menghapus tanda baca pada teks, dan menghapus angka pada teks ulasan.
   c. *Tokenizing*: tahap ini merupakan tahap pembersihan dan ekstraksi kata.
   d. *Removing Stop words/Stemming*: tahap ini merupakan tahap *filtering* atau penyaringan kata yang terdapat pada suatu dokumen

9. Pelabelan dan Pembobotan Data

   a. Klasifikasi SVM dan NBC

      Data ulasan yang telah didapatkan dan diberikan pelabelan, selanjutnya akan dilakukan proses klasifikasi kedalam sentimen kelas positif dan negatif. Pada proses klasifikasi ini, metode machine learning yang digunakan dalam membantu

proses klasifikasi saat melakukan pengolahan data ulasan ini yaitu *Support Vector Machine* dan *Naïve Bayes Classifier*.

b. Visualisasi dan Asosiasi Kata

Visualisasi dan asosiasi kata merupakan proses ekstraksi seluruh informasi yang telah dilakukan pengolahan data, dimana informasi yang didapatkan berupa kumpulan kata yang sering muncul pada ulasan dari penggunaan aplikasi. Bentuk dari visualisasi dan asosiasi kata akan berbentuk *wordcloud* dan *barplot*.

10. Interpretasi Hasil Penelitian

Setelah melakukan visualisasi kata dan dapatkan informasi mengenai penggunaan maupun pemberian ulasan negatif dari pengguna dari pemakaian produk *toner* The Body Shop. Selanjutnya akan dilakukan analisis hasil penerapan metode SVM dan NBC serta hasil klasifikasi maupun asosiasi kata.

11. Kesimpulan dan Saran

Tahap terakhir dari penelitian ini yaitu dengan menarik kesimpulan berdasarkan pembahasan dan analisis yang telah dilakukan dan memberikan saran baik kepada penelitian berikutnya maupun kepada perusahan serta *stakeholder* yang membutuhkan / menggunakan pendekatan analisis yang sama dalam proses *problem solving*.

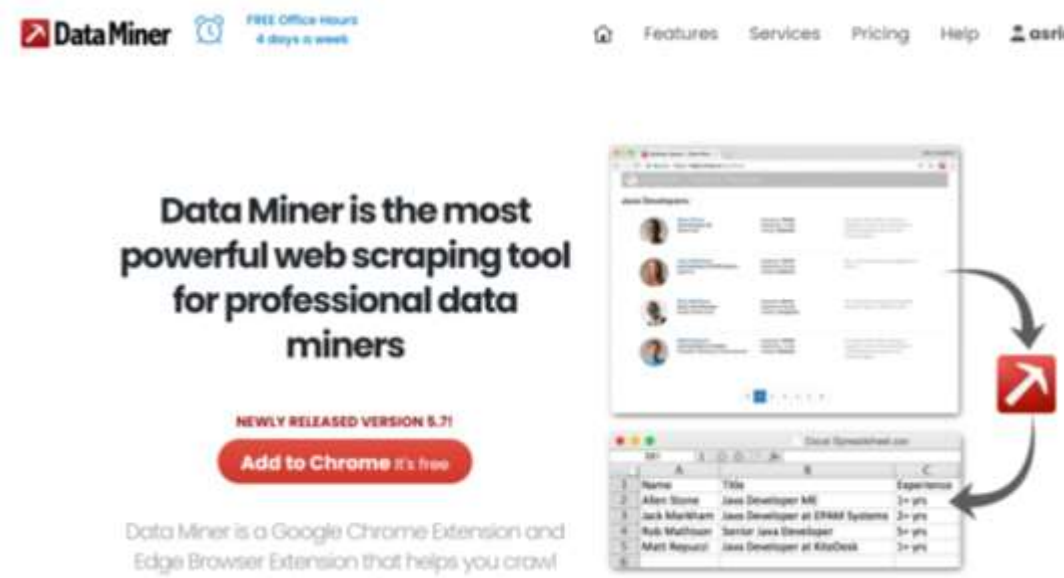
12. Selesai

BAB IV

PENGUMPULAN DAN PENGOLAHAN

4.1 Pengumpulan Data

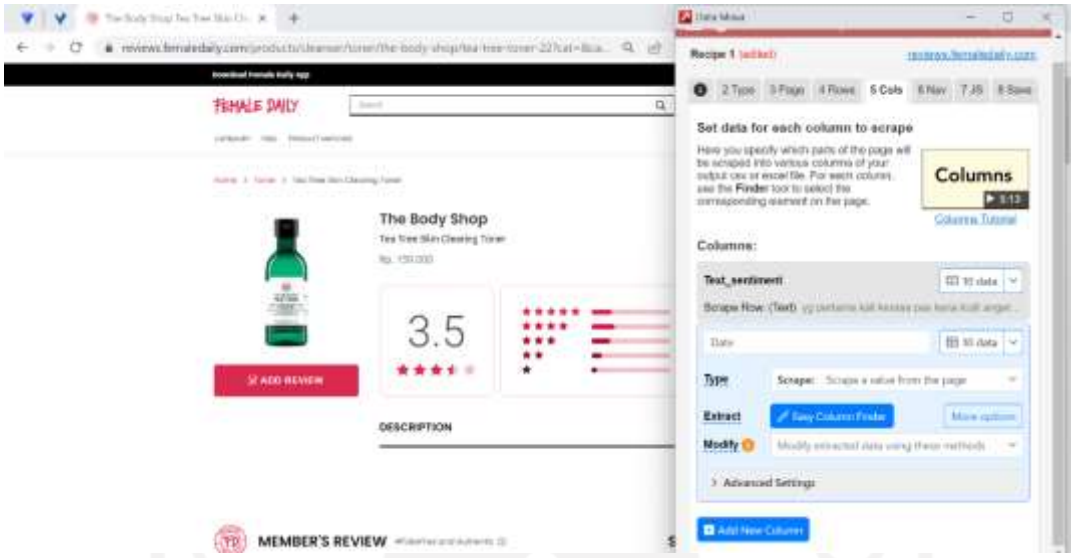
Pengambilan data dari FemaleDaily.com menggunakan teknik *scraping* dengan memanfaatkan ekstensi dari *Google Chrome* yaitu *Data Scraper*. Pada penelitian ini digunakan *Data Scraper* versi 5.2.74 untuk mengambil data ulasan yang selanjutnya diimpor menjadi spreadsheet Microsoft Excel atau CSV. Data ulasan produk yang diambil adalah data yang diunggah *user* dalam jangka waktu mulai dari 24 Juni 2018 – 11 September 2022.



Gambar 4. 1 Ektensi Data Scraper

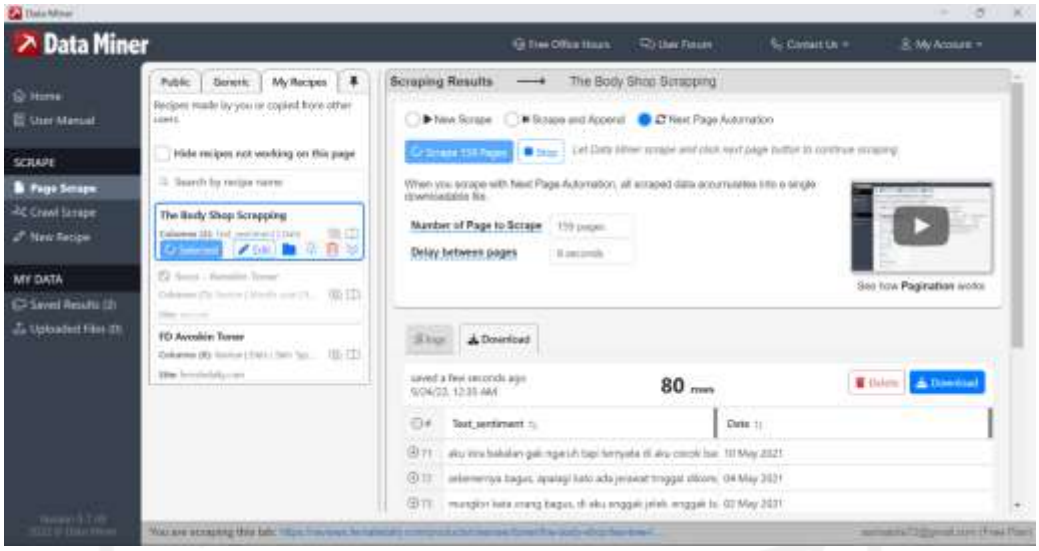
Langkah awal yang dilakukan adalah menambahkan atau mengaktifkan ekstensi *Data Scraper* pada *Google Chrome*. Kemudian dapat dilakukan pengambilan data sesuai alamat domain yang diinginkan. Pada penelitian ini, data yang digunakan diambil dari ulasan pengguna produk *Tea Tree Skin Clearing Mattifying Toner dari brand* "The Body Shop" dengan alamat domain:
https://reviews.femaledaily.com/products/cleanser/toner/the-body-shop/tea-tree-toner-

22?cat=&cat_id=0&age_range=&skin_type=&skin_tone=&skin_undertone=&hair_texture=&hair_type=&order=newest&page=1



Gambar 4. 2 Pengisian Kolom Data

Langkah selanjutnya adalah membuka alamat *domain* atau *website* seperti Gambar 4.2 kemudian memilih data ulasan sesuai dengan batasan yang telah ditentukan. Pertama pada *start* memilih sesuai dengan jenis data pada *website*. Selanjutnya pada menu *Cols* memilih sesuai dengan data yang akan diambil, dalam penelitian ini data yang diambil berupa data *review* dan tanggal (*day-month-year*). Meskipun diambil beberapa variable data sedemikian rupa, hanya data *review* dan tanggal saja yang akan diolah dalam menganalisis *sentiment* produk. Variabel data selain review dan tanggal akan dimanfaatkan untuk menarik *insight* pendukung yang bisa menjadi informasi pelengkap dalam memberikan rekomendasi. Pembuatan setiap *query* pada *Cols* akan terdapat tampilan seperti Gambar 4.2.



Gambar 4. 3 Proses *Scraping*

Gambar 4. 4 Hasil Proses *Scraping*

Hasil data yang telah diekstraksi dapat disimpan dalam bentuk xlsx atau csv. Pada penelitian ini data disimpan dalam bentuk xlsx. Gambar 4.4 menunjukkan hasil data yang didapat dari *scraping* sebesar 1075 *data sentiment* dari Female Daily yang selanjutnya akan digunakan untuk proses analisis berikutnya.

4.2 Pengolahan Data

Pengambilan data dilakukan dengan melakukan analisis deskriptif sebanyak 1.075 data ulasan produk *Tea Tree Skin Clearing Toner* oleh brand “The Body Shop” di Female Daily. Kemudian, data yang telah didapatkan akan dilakukan pre-processing data yang meliputi *cleaning*, *case folding*, *stemming*, dan *removing stop words.*

4.2.1 *Pre-processing* Data

Setelah data yang didapatkan sudah dalam bentuk “csv” kemudian dilakukan tahap *pre-processing*, tahap ini bertujuan untuk membersihkan data-data dari *noise* dan pembenahan bahasa seperti menghilangkan singkatan, bahasa gaul, serta menghapus kata yang tidak diperlukan. Dikarenakan data awal yang didapatkan berupa data yang tidak terstruktur maka dilakukan tahap *pre-processing* agar data tersebut dapat terstruktur sebelum dilakukan klasifikasi dan dianalisis. Tahap *pre-processing* dilakukan dengan bantuan *software* Google Collaboratory menggunakan bahasa pemrograman Python. Berikut merupakan tahapan-tahapan *pre-processing*:

4.2.2.1 Cleaning Data

Tahap ini adalah tahap di mana data disiapkan untuk menjadi data yang siap dianalisis. Hasil dari *scrapping* merupakan data mentah atau data yang diperoleh masih terdapat unsur simbol, URL, *date*, *user* dan sebagainya yang tidak mempunyai arti pada *sentiment* yang akan dianalisis pada tahap berikutnya. Dalam memudahkan para pembaca untuk menemukan topik atau pembahasan informasi terkait, maka diperlukan proses *cleaning* guna membersihkan data sehingga pembaca dapat mengetahui informasi dengan mudah. Proses *cleaning* data adalah proses untuk merapihkan dan membersihkan kalimat dari kata-kata yang tidak memiliki arti sehingga lebih mudah dan cepat dalam mendapatkan informasi dari data yang didapat.

Tabel 4. 1 Perbandingan Sebelum & Sesudah Proses *Cleaning*

No.	*Input*	*Output*
1	Panas. Kayaknya emang kulit aku ga cocok sm kandungan tea tree, kulit juga jadi jerawatan :((I don't think it works for me. Mahal padahal ya, masih banyak karena beli yg 250ml kali ada yg mau nyoba bisa DM ya	Panas. Kayaknya emang kulit aku ga cocok sm kandungan tea tree, kulit juga jadi jerawatan. I don't think it works for me. Mahal padahal ya, masih banyak karena beli yg 250ml kali ada yg mau nyoba bisa DM ya

No.	*Input*	*Output*
2	Punya ekspektasi tinggi bgt buat ngurangin minyak pas beli ini. wangi nya enak, bener2 tea tree. but sadly, ga cocok sama kulitku :(malah memperparah minyak di muka. awal nya masih mikir positif kalo itu cuma reaksi di awal pemakaian aja, tapi makin lama makin parah. sedih bgt ga cocok padahal review orang bagus2	Punya ekspektasi tinggi bgt buat ngurangin minyak pas beli ini. wangi nya enak, bener2 tea tree. but sadly, ga cocok sama kulitku malah memperparah minyak di muka. awal nya masih mikir positif kalo itu cuma reaksi di awal pemakaian aja, tapi makin lama makin parah. sedih bgt ga cocok padahal review orang bagus2

4.2.2.2 Case Folding

Pada tahap *case folding* merupakan tahap pengubahan huruf kapital menjadi huruf non kapital atau semuanya menjadi huruf kecil menggunakan bahasa pemrograman *python*.

Tabel 4. 2 Hasil *Case Folding*

No.	*Input*	*Output*
1	Panas. Kayaknya emang kulit aku ga cocok sm kandungan tea tree, kulit juga jadi jerawatan. I don't think it works for me. Mahal padahal ya, masih banyak karena beli yg 250ml kali ada yg mau nyoba bisa DM ya	panas kayaknya emang kulit aku ga cocok sm kandungan tea tree kulit juga jadi jerawatan i dont think it works for me mahal padahal ya masih banyak karena beli yg 250ml kali ada yg mau nyoba bisa dm ya
2	Punya ekspektasi tinggi bgt buat ngurangin minyak pas beli ini. wangi nya enak, bener2 tea tree. but sadly, ga cocok sama kulitku malah memperparah minyak di muka. awal nya masih mikir positif kalo itu cuma reaksi di awal pemakaian aja, tapi makin lama makin	punya ekspektasi tinggi bgt buat ngurangin minyak pas beli ini wangi nya enak bener2 tea tree but sadly ga cocok sama kulitku malah memperparah minyak di muka awal nya masih mikir positif kalo itu cuma reaksi di awal pemakaian aja tapi makin lama makin

No.	*Input*	*Output*
	parah. sedih bgt ga cocok padahal review orang bagus2	parah sedih bgt ga cocok padahal review orang bagus2

Pada Tabel 4.2 terdapat huruf kapital yang diubah menjadi huruf kecil atau *lowercase* semua pada proses *case folding*.

4.2.2.3 Stemming

Tahapan *stemming* atau *tokenizing* adalah proses untuk memisahkan kata di dalam dokumen menjadi potongan kata yang tidak saling berpengaruh yang disebut *token* untuk kemudian dapat diidentifikasi.

Tabel 4. 3 Hasil Proses *Tokenizing*

No.	*Input*	*Output*
1	panas kayaknya emang kulit aku ga cocok sm kandungan tea tree kulit juga jadi jerawatan i dont think it works for me mahal padahal ya masih banyak karena beli yg 250ml kali ada yg mau nyoba bisa dm ya	[‘panas’, 'kayaknya', 'emang', 'kulit', 'ga', 'cocok', 'sm', 'kandungan', 'tea', 'tree', 'kulit', ‘juga’, ‘jadi’, 'jerawatan', 'i', 'dont', 'think', 'it', 'works', 'for', 'me', 'mahal', 'ya', ‘banyak’, ‘karena’, 'beli', 'yg', '250ml', 'kali', ‘ada’, 'yg', ‘mau’, 'nyoba', 'dm', 'ya']
2	punya ekspektasi tinggi bgt buat ngurangin minyak pas beli ini wangi nya enak bener2 tea tree but sadly ga cocok sama kulitku malah memperparah minyak di muka awal nya masih mikir positif kalo itu cuma reaksi di awal pemakaian aja tapi makin lama makin parah sedih bgt ga cocok padahal review orang bagus2	[‘punya’, 'ekspektasi', ‘tinggi’, 'bgt', ‘buat’, 'ngurangin', 'minyak', 'pas', 'beli', ‘ini’, 'wangi', 'nya', 'enak', 'bener2', 'tea', 'tree', 'but', 'sadly', 'ga', 'cocok', ‘sama’, 'kulitku', ‘malah’, 'memperparah', 'minyak', ‘di’, 'muka', ‘awal’, 'nya', 'mikir', 'positif', 'kalo', ‘itu’, ‘cuma’, 'reaksi', ‘di’, ‘awal’, 'pemakaian', 'aja', ‘tapi’, ‘makin’, ‘lama’, ‘makin’, 'parah', 'sedih', 'bgt', 'ga', 'cocok', 'review', 'orang', 'bagus2']

Pada Tabel 4.3 tersebut merupakan contoh dari hasil akhir proses *tokenizing.*

4.2.2.4 Removing Stop Word

Tahap *filtering* yaitu tahapan untuk mengambil kata-kata yang penting. Proses *filtering* dapat menggunakan algoritma *stopword* (menghapus kata tidak penting). Contoh *stopword* yaitu “yang”, “dan”, “ke”, “dari”, ”oleh” dan lainnya. Kata-kata tersebut merupakan kata yang berfrekuensi tinggi dan dapat ditemukan di hampir setiap kalimat. *Stopword* atau menghapus kata dapat mengurangi ukuran indeks dan waktu pemrosesan serta dapat mengurangi *noise*.

Tabel 4. 4 Perbandingan Hasil Akhir Proses *Filtering*

No.	*Input*	*Output*
1	[‘panas’, 'kayaknya', 'emang', 'kulit', 'ga', 'cocok', 'sm', 'kandungan', 'tea', 'tree', 'kulit', ‘juga’, ‘jadi’, 'jerawatan', 'i', 'dont', 'think', 'it', 'works', 'for', 'me', 'mahal', 'ya', ‘banyak’, ‘karena’, 'beli', 'yg', '250ml', 'kali', ‘ada’, 'yg', ‘mau’, 'nyoba', 'dm', 'ya']	[‘panas’, 'kayaknya', 'emang', 'kulit', 'ga', 'cocok', 'sm', 'kandungan', 'tea', 'tree', 'kulit', 'jerawatan', 'i', 'dont', 'think', 'it', 'works', 'for', 'me', 'mahal', 'ya', 'beli', 'yg', '250ml', 'kali', 'yg', 'nyoba', 'dm', 'ya']
2	[‘punya’, 'ekspektasi', ‘tinggi’, 'bgt', ‘buat’, 'ngurangin', 'minyak', 'pas', 'beli', ‘ini’, 'wangi', 'nya', 'enak', 'bener2', 'tea', 'tree', 'but', 'sadly', 'ga', 'cocok', ‘sama’, 'kulitku', ‘malah’, 'memperparah', 'minyak', ‘di’, 'muka', ‘awal’, 'nya', 'mikir', 'positif', 'kalo', ‘itu’, ‘cuma’, 'reaksi', ‘di’, ‘awal’, 'pemakaian', 'aja', ‘tapi’, ‘makin’, ‘lama’, ‘makin’, 'parah', 'sedih', 'bgt', 'ga', 'cocok', 'review', 'orang', 'bagus2']	['ekspektasi', 'bgt', 'ngurangin', 'minyak', 'pas', 'beli', 'wangi', 'nya', 'enak', 'bener2', 'tea', 'tree', 'but', 'sadly', 'ga', 'cocok', 'kulitku', 'memperparah', 'minyak', 'muka', 'nya', 'mikir', 'positif', 'kalo', 'reaksi', 'pemakaian', 'aja', 'parah', 'sedih', 'bgt', 'ga', 'cocok', 'review', 'orang', 'bagus2']

4.2.2 Pemberian *Label* Sentimen

Peneliti menggunakan metode dengan menerapkan algoritma yang dapat mengekstrak kalimat opini secara otomatis. Salah satu algoritma yang umum digunakan adalah *lexicon based*. *Lexicon based* dapat mengekstrak kalimat opini dengan presisi yang sangat tinggi (Azhar, 2017) dan cepat karena dilakukan secara otomatis. Oleh karena itu dalam penelitian ini, proses perhitungan dilakukan menggunakan metode *Lexicon Based*. Metode lexicon based bekerja dengan cara membuat kamus kata opini (*lexicon*) terlebih dahulu. Kata-kata yang terdapat pada kamus tersebut akan digunakan untuk mengidentifikasi kata positif dan negatif pada suatu kalimat. Rumus umum yang digunkan untuk mendapatkan skor sentimen adalah sebagai berikut:

Skor = Jumlah kata positif – Jumlah kata negatif

Dalam penelitian ini, perhitungan skor semtimen dilakukan dengan bantuan aplikasi RStudio dengan bantuan *library* "tm" yang umum digunakan dalam proses *text mining*. Setelah mendapatkan skor sentimen, pengolahan data dilanjutkan dengan melakukan pelabelan kelas setimen. Pelabelan dilakukan dengan membagi data ulasan menjadi tiga kelas sentimen, yaitu sentimen positif, netral dan negatif dengan ketentuan sebagai berikut (Santoso & Nugroho, 2019):

- Sentimen negatif : skor < 0
- Sentimen netral : skor = 0
- Sentimen positif : skor > 0

Setelah melewati proses pembobotan dan pelabelan data ulasan menggunakan *software* R. Berikut ini adalah salah satu contoh *dataset* yang telah diberi *label*:

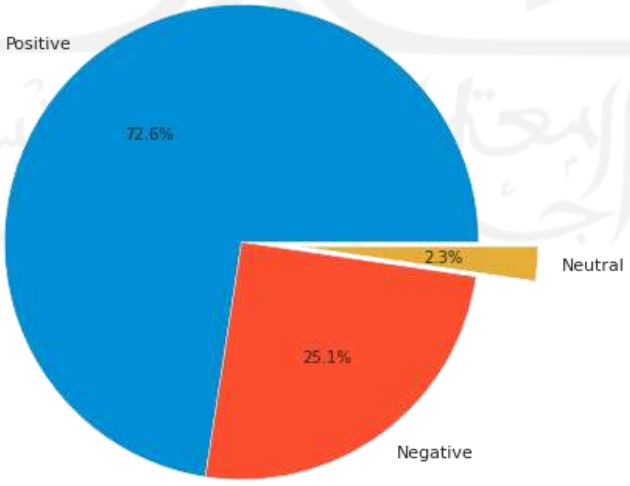
Tabel 4. 5 Hasil Pemberian *Label*

No.	*text_sentiment*	*Score*	*Analysis*
1	toner yang cocok banget untuk kulit kombinasi aku. Hampir tiap hari pake ini ada rasa seger2 nya bikin kulit fresh. Yang paling penting toner ini nahan minyak dan bikin muka less oily meskipun awal pemakaian ada rasa keset dan ketarik ganyaman. Selain itu toner ini bantu bgt untuk jerawat yg lg meradang	2	*Positive*

No.	*text_sentiment*	*Score*	*Analysis*
2	kurang nampol untuk ngilangin jerawat walaupun ada kandungan tea tree oil ngabisin satu botol tapi bener2 gak ada perubahan before after agak buang2 duit beli ini karna lumayan mahal tapi efeknya gak ketara lebih suka facial washnya	-1	*Negative*
3	the body shop di aku memang hit and miss. kalo bagus bagus bangeet, kalo biasa ya ok lah. klo produk ini in between lol. di pakai diwajah ada sensasi burning nya, beberapa hari timbul bruntusan di daerah dagu dan pipi bawah. akhirnya aku pakein ke badan. gaada sensasi burning sih, agak aneh jg. mungkin less sensitive kali ya. tapi jerawat badan mendadak kempes, but still, orang beli ini pertama kan untuk wajah yah, jadi yaa bingung jg mau kasih rating berapa	0	*Neutral*

4.2.3 Analisis Sentimen

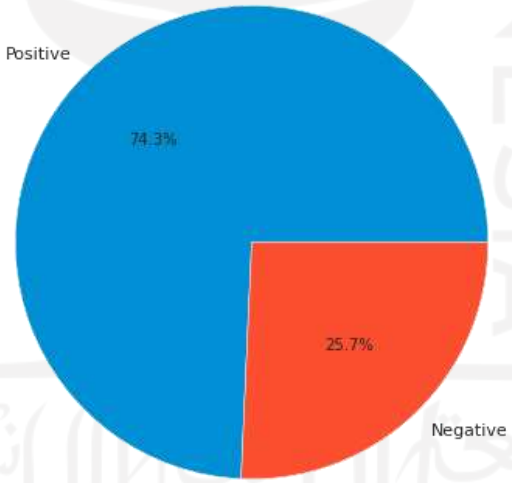
Berikut merupakan hasil perbandingan jumlah data dari pelabelan kelas sentimen pada Gambar 4.5:



Gambar 4. 5 Hasil Analisis Sentimen Sebelum Reduksi (*Pie Chart*)

Klasifikasi data pada penelitian ini dibagi menjadi sentimen positif, sentimen negatif, dan sentiment netral. Pada ulasan klasifikasi yang mengandung pernyataan positif seperti pernyataan kecocokan pemakaian produk, ungkapan kagum, pujian, dan lainnya. Untuk ulasan klasifikasi yang mengandung pernyataan negatif seperti ketidakpuasaan, ketidakcocokan, ketidaksesuaian, dan lainnya. Pada Gambar 4.5 didapatkan jumlah sentimen positif sebanyak 780 data atau sebesar 72.6%, dan jumlah sentimen negatif sebanyak 270 data atau sebesar 25.1%., serta jumlah sentimen netral sebanyak 25 atau sebesar 2.3%.

Setelah hasil pengkategorian kelas sentimen didapatkan, pada proses berikutnya peneliti menghilangkan *label* sentimen netral, dikarenakan jumlah data netral yang secara signifikan sangat kecil dan tidak memberikan pengaruh signifikan terhadap pemodelan klasifikasi kedepannya. Selain itu, data netral tidak memiliki *argument* yang mendukung (positif-negatif), sehingga kurang efektif dalam memberikan esensi sebuah informasi maupun manfaat kepada *Brand* "The Body Shop". Berikut merupakan grafik hasil analisis sentimen yang telah direduksi:



Gambar 4. 6 Analisis Sentimen Setelah Reduksi (*Pie Chart*)

Pada Gambar 4.6 didapatkan hasil reduksi dimana data awal yang digunakan yaitu 1075 menjadi 1050 data ulasan. Dari 1050 ulasan tersebut, didapatkan bahwa 780 ulasan (74.3%) merupakan kelas sentimen positif dan 270 ulasan (25.7%) merupakan kelas sentimen negatif. Dengan perolehan hasil pada Gambar 5.2 dapat dikatakan bahwa sentimen positif lebih banyak dibandingkan sentimen negatif.

4.2.4 Analisis Klasifikasi

Setelah melakukan tahap pelabelan kelas, proses pengolahan data dilakukan dengan analisis klasifikasi. Hasil data pelabelan dibagi menjadi dua, yaitu data *training* dan data *testing*. Proses klasifikasi menggunakan dua algoritma, yaitu *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM).

4.2.5.1 Pembuatan *Data Training* dan *Data Testing*

Dalam penelitian ini, pembuatan data latih sangat diperlukan. Data latih dapat mempengaruhi tingkat akurasi yang dihasilkan. Data uji merupakan data yang digunakan untuk mengujji tingkat akurasi dari model yang dibuat oleh data latih. Pembuatan data latih dilakukan dengan menentukan proporsi data yang telah melewati proses sebelumnya. Total ulasan review tentang produk *The Body Shop Tea Tree Skin Clearing Toner yang* telah didapatkan melalui proses *scraping data* secara keseluruhan sebanyak 1.075. data yang kemudian setelah melewati proses pre-processing data menjadi sebesar 1.050 data. Peneliti menggunakan tiga nilai perbandingan data *training* dan *testing*. Berikut merupakan perbandingan jumlah data *training* dan data *testing*:

1) Perbandingan data *training* sebesar 70% dan data *testing* 30%

Tabel 4. 6 Pembagian Data *Training* dan *Testing* (7:3)

Jenis Data	Presentase	Jumlah
Data *Training*	70%	735
Data *Testing*	30%	315

Berdasarkan Tabel 4.6, diketahui bahwa perbandingan data *training* dan data *testing* sebesar 70% : 30% dati total 1050 ulasan, terdapat sebanyak 735 ulasan sebagai data *training* dan 315 ulasan sebagai data *testing.*

2) Perbandingan data *training* sebesar 80% dan data *testing* 20%

Tabel 4. 7 Pembagian Data *Training* dan *Testing* (8:2)

Jenis Data	Presentase	Jumlah
Data *Training*	80%	840
Data *Testing*	20%	210

Berdasarkan Tabel 4.7, diketahui bahwa perbandingan data *training* dan data *testing* sebesar 80% : 20% dati total 884 ulasan, terdapat sebanyak 840 ulasan sebagai data *training* dan 210 ulasan sebagai data *testing.*

3) Perbandingan data *training* sebesar 90% dan data *testing* 10%

Tabel 4. 8 Pembagian Data *Training* dan *Testing* (9:1)

Jenis Data	Presentase	Jumlah
Data *Training*	90%	945
Data *Testing*	10%	105

Berdasarkan Tabel 4.8, diketahui bahwa perbandingan data *training* dan data *testing* sebesar 90% : 10% dati total 884 ulasan, terdapat sebanyak 945 ulasan sebagai data *training* dan 105 ulasan sebagai data *testing.*

4.2.5.2 Perbandingan Klasifikasi *Support Vector Machine* dan *Naïve Bayes*

Data latih yang telah dibentuk akan dipelajari melalui proses klasifikasi. Pada data latih terdapat ulasan negatif, ulasan netral, dan ulasan positif, kemudian pola data tersebut akan dipelajari menggunakan algoritma SVM berdasarkan ciri-ciri data pada masing-masing label kelas yang telah dibentuk. Selanjutnya akan dilakukan proses *machine learning* yaitu hasil pelatihan pada algoritma SVM akan diuji untuk mengetahui nilai akurasi dalam memprediksi data baru.

Model pendekatan atau sering disebut *kernel* dapat membantu mengatasi masalah ruang fitur (*feature space*), dan mempengaruhi akurasi yang akan dihasilkan (Diani, 2017). Guna memperoleh klasifikasi dengan akurasi terbaik, penelitian ini melakukan percobaan beberapa *kernel* berupa *Linear, Polynomial, Radial Basis Function (RBF),* dan *Sigmoid*. *Kernel linear* memiliki waktu pelatihan *kernel linier* lebih cepat dibandingkan dengan *kernel* lain dan cocok untuk data berdimensi kecil hingga besar (Praghakusma & Charibaldi, 2021). Pada *kernel polynomial* menggunakan derajat yang dapat disesuaikan untuk meningkatkan kemungkinan bahwa data dapat dipisahkan secara *linier* dalam ruang berdimensi tinggi tanpa memperlambat waktu pemodelan (Géron, 2019). Pada *kernel* RBF digunakan ketika data tidak dapat dipisahkan secara *linier*, dimana optimasi parameter *cost* dan *gamma* dilakukan ketika melakukan analisis dengan RBF (Widayani & Harliana, 2021). Pada *kernel sigmoid* penggunaan gamma dapat diatur untuk

meningkatkan nilai akurasi akan tetapi bergantung pada jumlah fitur yang digunakan (Al-Mejibli et al., 2020). Peneliti mengimplementasikan keempat *Kernel* SVM tersebut pada *Python's Scikit-learn library*. Berikut merupakan hasil perbandingan setiap *kernel* yang telah diuji coba dapat dilihat pada Tabel 4.9.

Tabel 4. 9 Akurasi Spesifikasi *Kernel*

Kernel	**Akurasi**
Linear	84%
Polynomial	71%
Gaussian RBF	77%
Sigmoid	68%

Pada Tabel 4.9 terlihat bahwa dari perbandingan keempat *kernel* yang telah diuji, *kernel linear* memiliki nilai akurasi yang paling tinggi. Dengan demikian kernel linear akan dipilih untuk proses klasifikasi.

Pada penelitian ini dilakukan tiga kali percobaan untuk setiap *dataset*. Perbedaan setiap percobaan terdapat pada berapa kali data set diacak dengan menggunakan rumus *random_state*=(n) pada *programming language* Python, dimana n merupakan banyak *random* n kali. Random_state yang dipilih akan dijadikan parameter dari fungsi *random*. Terdapat 3 nilai *random_state* yang digunakan dalam penelitian ini, yaitu 3 pada percobaan ke-1, 8 pada percobaan ke-2, dan 9 pada percobaan ke-3.

Tabel 4. 10 Perbandingan Akurasi Model

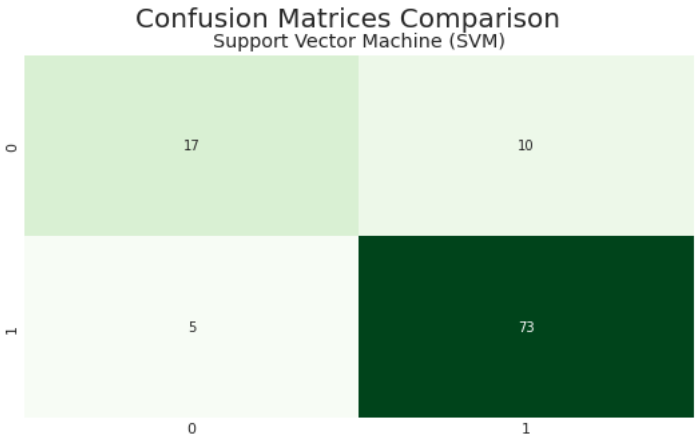
Percobaan	**Akurasi Model**	
	NBC	**SVM**
Perbandingan 70% : 30%		
Percobaan 1	81%	77%
Percobaan 2	80%	**83%**
Percobaan 3	**82%**	80%
Perbandingan 80% : 20%		
Percobaan 1	**82%**	82%
Percobaan 2	80%	**84%**
Percobaan 3	80%	**84%**
Perbandingan 90% : 10%		

Percobaan	Akurasi Model	
	NBC	SVM
Percobaan 1	81%	**86%**
Percobaan 2	**83%**	**86%**
Percobaan 3	74%	82%

Berdasarkan Tabel 4.10 diketahui bahwa nilai rata-rata total akurasi metode SVM lebih besar yaitu 83% dibandingkan metode NBC yang memiliki rata-rata total akurasi sebesar 80%. Oleh karena itu, dapat disimpulkan bahwa algoritma SVM memiliki kinerja lebih baik dalam melakukan klasifikasi data ulasan produk The Body Shop dibandingkan metode NBC.

4.2.5.3 Klasifikasi *Support Vector Machine*

Berdasarkan hasil tingkat akurasi pada proses sebelumnya, didapatkan bahwa pada metode *Support Vector Machine* dengan pembagian data latih sebanyak 90% dan data uji sebanyak 10% memiliki tingkat akurasi teringgi yaitu sebesar 86%, sehingga dengan menggunakan proporsi data tersebut akan digunakan untuk melakukan evaluasi kinerja model.



Gambar 4. 7 *Confusion Matrix* SVM

Dilihat dari Gambar 4.8 terdapat 73 data yang diprediksi oleh mesin atau model dengan tepat dan tidak terjadi *miss classification* atau disebut *True Positive* (TP). Selain itu, terdapat pula sebanyak 10 data sentimen positif yang terprediksi pada sentimen

negatif atau disebut *False Negative* (FN). Pada kategori sentimen negatif terdapat 17 data yang diprediksi sesuai dengan data actual atau disebut *True Negative* (TN). Tetapi terdapat 5 data sentimen negatif yang terprediksi kedalam sentimen positif atau disebut dengan data *False Positive* (FP).

	precision	recall	f1-score	support
Negative	0.77	0.63	0.69	27
Positive	0.88	0.94	0.91	78
accuracy			0.86	105
macro avg	0.83	0.78	0.80	105
weighted avg	0.85	0.86	0.85	105

Gambar 4. 8 Hasil *Confusion Matri*x SVM

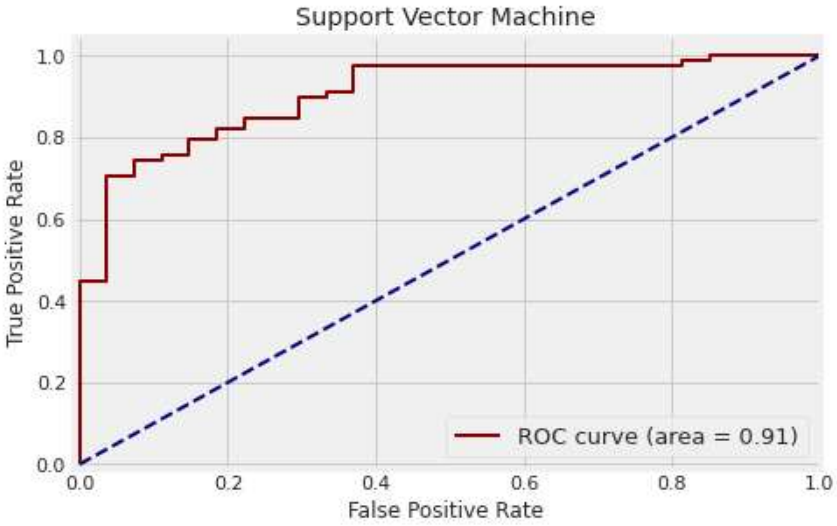
Precision (presisi) merupakan tingkat ketelitian atau ketepatan dalam klasifikasi, dimana nilai presisi didapatkan sebesar 77% untuk label *negative* dan 88% untuk label *positive*. Hal tersebut berarti tingkat ketelitian atau ketepatan dalam klasifikasi dapat dikatakan bagus dalam memprediksi label negatif dan sangat bagus untuk memprediksi label positif. *Recall* merupakan tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi. Dari perhitungan *recall* didapatkan hasil sebesar 63% untuk label negatif dan 94% untuk label positif, hal ini berarti tingkat keberhasilan sistem sangat bagus dalam menemukan kembali sebuah informasi.

Kemudian, pada nilai *f1-score*, yang merupakan nilai rata-rata tertimbang dari presisi dan *recall* untuk kelas itu. Secara umum memberikan gambaran yang lebih besar tentang bagaimana kinerja model untuk label itu dan semakin tinggi angka tersebut, maka semakin baik kinerja model yang dibangun, dimana nilai f1-score sebanyak 69% untuk label *negative* dan 91% untuk label *positive*.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} x\ 100\% \\ = \frac{73 + 17}{73 + 5 + 17 + 10} x\ 100\% \\ = 85{,}7\% \qquad (2)$$

Accuracy (akurasi) digunakan untuk mengetahui seberapa bagus model bias mengklasifikasikan data dengan benar. Pada perhitungan akurasi didapatkan nilai sebesar 86%, dari hasil tersebut dapat dikatakan bahwa model bias dapat mengklasifikasikan data

dengan benar. Dari hasil rangkuman evaluasi model, dapat dilihat bahwa nilai akurasi mendapatkan nilai 86%, *recall* dengan nilai 94%, *f1score* dengan nilai 91%, dan *precision* mendapatkan nilai 88%. Hasil evaluasi ini menunjukkan kinerja model yang dibuat sangat bagus.



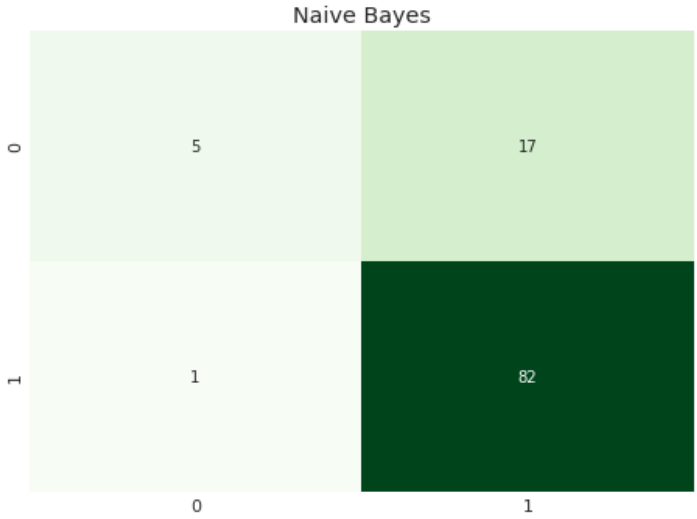
Gambar 4. 9 *Receiver Operating Characteristic* (ROC) SVM

Selain evaluasi menggunakan *confusion matrix*, evaluasi lain yang dapat dilakukan adalah membuat grafik *Receiver Operating Characteristic* **(**ROC). Dapat dilihat pada Gambar 4. 10 yang menunjukkan grafik ROC, dimana grafik terlihat bahwa garis merah mendekati sumbu Y yang artinya pemodelan yang dilakukan baik. Selain itu jika ingin menilai kinerja kurva garis merah dalam bentuk angka maka dapat dilakukan dengan membandingkan luas di bawah kurva atau *Area Under Curve* (AUC). Dari gambar tersebut, dapat diketahui hasil dari nilai *Area Under Curve* (AUC) yang menunjukkan luas area diatas garis pada grafik *Receiver Operating Characteristic* **(**ROC) menunjukkan angka 0.91. Klasifikasi dikatakan baik jika nilai AUC semakin tinggi. Dengan nilai AUC sebesar 91% tersebut, maka dapat diartikan bahwa klasifikasi tergolong sangat baik.

4.2.5.4 Klasifikasi *Naïve Bayes*

Berdasarkan hasil tingkat akurasi pada proses sebelumnya, didapatkan bahwa pada metode *Naïve Bayes* dengan pembagian data latih sebanyak 90% dan data uji sebanyak

10% memiliki tingkat akurasi teringgi yaitu sebesar 83%, sehingga dengan menggunakan proporsi data tersebut akan digunakan untuk melakukan evaluasi kinerja model.



Gambar 4. 10 *Confusion Matrix* NBC

Dilihat dari Gambar 4.11 terdapat 82 data yang diprediksi oleh mesin atau model dengan tepat dan tidak terjadi *miss classification* atau disebut *True Positive* (TP). Selain itu, terdapat pula sebanyak 17 data sentimen positif yang terprediksi pada sentimen negatif atau disebut *False Negative* (FN). Pada kategori sentimen negatif terdapat 5 data yang diprediksi sesuai dengan data actual atau disebut *True Negative* (TN). Lalu, terdapat 1 data sentimen negatif yang terprediksi kedalam sentimen positif atau disebut dengan data *False Positive* (FP).

Naive Bayes

	precision	recall	f1-score	support
Negative	0.83	0.23	0.36	22
Positive	0.83	0.99	0.90	83
accuracy			0.83	105
macro avg	0.83	0.61	0.63	105
weighted avg	0.83	0.83	0.79	105

Gambar 4. 11 Hasil *Confusion Matrix* NBC

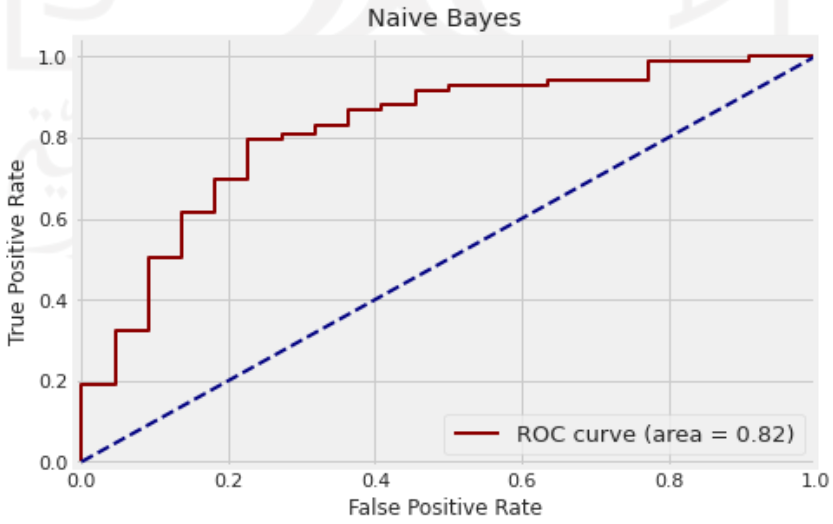
Precision (presisi) merupakan tingkat ketelitian atau ketepatan dalam klasifikasi, dimana nilai presisi didapatkan sebesar 83% untuk label *positive* maupun *negative*. Hal tersebut berarti tingkat ketelitian atau ketepatan dalam klasifikasi dapat dikatakan sangat bagus dalam memprediksi label positif maupun negatif. *Recall* merupakan tingkat

keberhasilan sistem dalam menemukan kembali sebuah informasi. Dari perhitungan *recall* didapatkan hasil sebesar 99% untuk label positif, hal ini berarti tingkat keberhasilan sistem sangat bagus dalam menemukan kembali sebuah informasi.

Kemudian, pada nilai *f1-score*, yang merupakan nilai rata-rata tertimbang dari presisi dan *recall* untuk kelas itu. Secara umum memberikan gambaran yang lebih besar tentang bagaimana kinerja model untuk label itu dan semakin tinggi angka tersebut, maka semakin baik kinerja model yang dibangun, dimana nilai *f1-score* sebanyak 90% untuk label *positive*.

$$\begin{aligned} Accuracy &= \frac{TP + TN}{TP + FP + TN + FN} x\ 100\% \\ &= \frac{82 + 5}{82 + 1 + 5 + 17} x\ 100\% \\ &= 82.8\% \end{aligned} \tag{3}$$

Accuracy (akurasi) digunakan untuk mengetahui seberapa bagus model bias mengklasifikasikan data dengan benar. Pada perhitungan akurasi didapatkan nilai sebesar 83%, dari hasil tersebut dapat dikatakan bahwa model bias dapat mengklasifikasikan data dengan benar. Dari hasil rangkuman evaluasi model, dapat dilihat bahwa nilai akurasi mendapatkan nilai 83%, *recall* dengan nilai 99%, *f1score* dengan nilai 90%, dan *precision* mendapatkan nilai 83%. Hasil evaluasi ini menunjukkan bahwa kinerja model yang dibuat cukup bagus.



Gambar 4. 12 *Receiver Operating Characteristic* (ROC) NBC

Selain evaluasi menggunakan *confusion matrix*, evaluasi lain yang dapat dilakukan adalah membuat grafik *Receiver Operating Characteristic* **(**ROC). Dapat dilihat pada Gambar 4.13 yang menunjukkan grafik ROC, dimana grafik terlihat bahwa garis merah mendekati sumbu Y yang artinya pemodelan yang dilakukan baik. Selain itu jika ingin menilai kinerja kurva garis merah dalam bentuk angka maka dapat dilakukan dengan membandingkan luas di bawah kurva atau *Area Under Curve* (AUC). Dari gambar tersebut, dapat diketahui hasil dari nilai *Area Under Curve* (AUC) yang menunjukkan luas area diatas garis pada grafik *Receiver Operating Characteristic* **(**ROC) menunjukkan angka 0.82. Klasifikasi dikatakan baik jika nilai AUC semakin tinggi. Dengan nilai AUC sebesar 82% tersebut, maka dapat diartikan bahwa klasifikasi tergolong baik.

4.2.5 Visualisasi *Word Cloud* dan Asosiasi Kata

Tahapan selanjutnya yang dilakukan adalah *word cloud dan* asosisi kata. *Word cloud* merupakan sebuah representasi untuk melihat visualisasi kata-kata pada data teks yang didapatkan setelah proses *pre-processing*. Fungsi dari *word cloud* untuk memunculkan kata-kata yang sering digunakan sehingga memudahkan pembaca dalam memahami informasi. Selain itu, perlu dilakukan mencari asosiasi kata yang paling sering muncul secara bersamaan. Proses ini digunakan untuk memperkuat informasi yang telah didapatkan dari proses visualisasi. Hasil *word cloud* dan asosiasi kata terbagi menjadi dua bagian, yaitu pada ulasan positif dan negatif.

4.2.6 Sentimen Positif

Berikut merupakan hasil *word cloud* pada sentiment positif dapat dilihat pada Gambar 4.14 berikut:

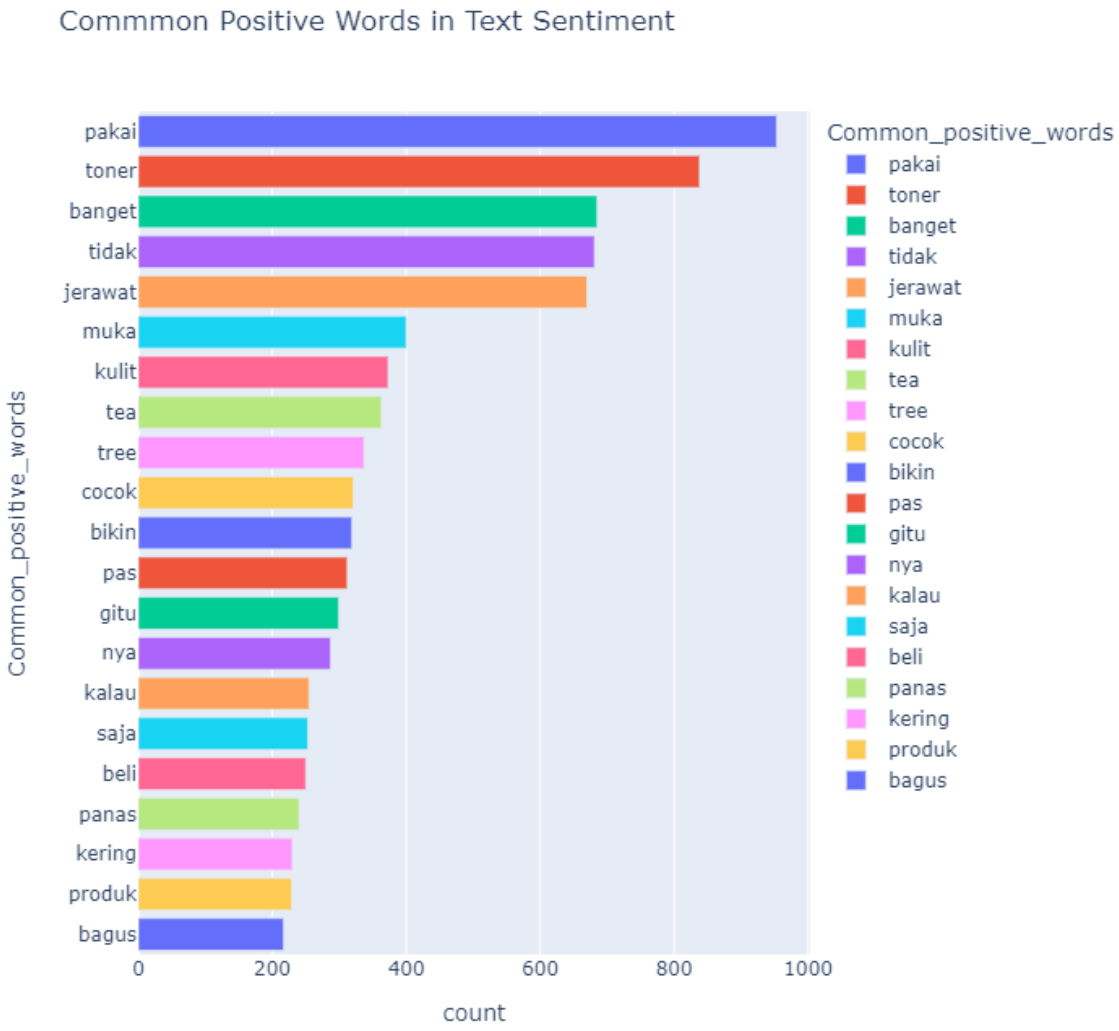
Gambar 4. 13 Tampilan *Word Cloud* Sentimen Positif

Berdasarkan Gambar 4.14 dapat dilihat hasil klasifikasi ulasan positif dari sejumlah 780 ulasan positif terhadap produk *Tea Tree Skin Clearing Toner* dari *brand* "The Body Shop" menghasilkan perolehan kata yang paling banyak muncul, diantaranya yaitu kata "pakai", "toner", "banget", "tidak", "jerawat", dan seterusnya. Semakin sering dan banyak kata digunakan maka semakin besar ukuran kata tersebut pada *word cloud*. Hasil yang menarik dari tampilan *word cloud* juga dapat memudahkan untuk membaca dan mengetahui informasi terkait. Selain itu, frekuensi kata-kata umum tersebut secara rinci dapat dilihat pada Tabel 4.11 dan *bar plot* yang ditampilkan di Gambar 4.15.

Tabel 4. 11 Frekuensi Kata pada *Word Cloud* Sentimen Positif

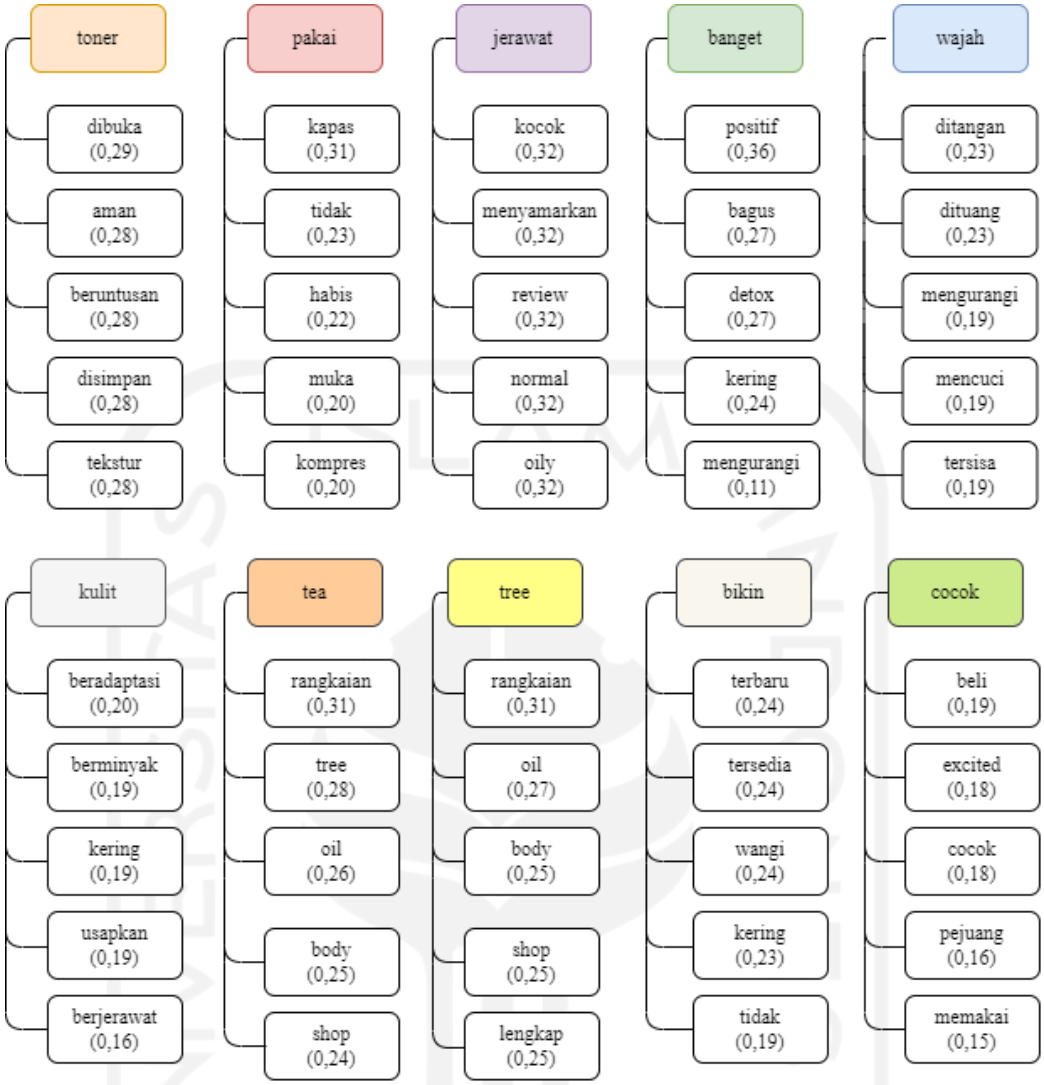
No.	***Common Positive Words***	***Count***
0	pakai	953
1	toner	838
2	banget	685
3	tidak	681
4	jerawat	670
5	muka	400
6	kulit	373
7	tea	363

No.	*Common Positive Words*	*Count*
8	tree	337
9	cocok	321
10	bikin	319
11	pas	312
12	gitu	299
13	nya	287
14	kalau	255
15	saja	253
16	beli	250
17	panas	240
18	kering	230
19	produk	229
20	bagus	217



Gambar 4. 14 Persebaran Kata yang Umum pada Sentimen Positif

Dapat dilihat juga pada Gambar 4.15 yang merupakan visualisasi *bar plot* dari 20 yang sering muncul dengan perolehan kata yang paling banyak muncul, diantaranya yaitu kata “pakai” dengan frekuensi 953 kali, “toner” sebanyak 838 kali, “banget” sebanyak 685 kali, “tidak” sebanyak 681 kali, “jerawat” sebanyak 670 kali, dan seterusnya. Kata-kata yang muncul pada Gambar 4.15 merupakan kata yang memiliki sentimen positif dan merupakan topik pembicaraan yang paling sering diulas oleh pengguna *skincare* tersebut. Kata-kata tersebut selanjutnya digunakan sebagai dasar untuk menemukan asosiasi dengan kata lainnya, sehingga dapat diperoleh informasi yang lebih baik. Selanjutnya, dilakukan pencarian asosiasi kata yang sering muncul secara bersamaan dan diperoleh hasil sebagai berikut:



Gambar 4. 15 Asosiasi Kata Pada Sentimen Positif

Berdasarkan Gambar 4.16 diperoleh asosiasi kata pada klasifikasi kelas positif. Proses ekstrasi informasi dengan asosiasi dilakukan secara berulang-ulang dengan cara menyaring kata-kata yang memiliki hubungan dengan kata lain dan didasarkan pada relevansi kata dengan topik yang diulas. Asosiasi kata yang berkatian dengan kata "*toner*", dapat diperoleh informasi tentang pengguna *Tea Tree Skin Clearing* "The Body Shop" ini bahwa produk ini mampu mengurangi beruntusan, dimana dengan tektur produk yang cair dan daya tahan toner yang selama 12 bulan setelah produk dibuka, sehingga aman untuk disimpan dalam waktu cukup lama. Kata-kata yang berasosiasi dengan kata "pakai" memberikan informasi bahwa pengguna produk ini sering memakai produk setelah/sehabis cuci wajah, kemudian melakukan *double cleansing* dengan mengusapkan *toner* menggunakan kapas ke seluruh wajah, dan sebagian pengguna juga menggunakannya dengan cara mengompres bagian jerawat saja. Kata-kata yang

berasosiasi dengan kata “jerawat” memberikan informasi bahwa hasil review dari pemakaian produk oleh pengguna yaitu untuk memastikan toner telah dikocok sebelum dipakai, sehingga residu pada toner tercampur sepenuhnya. Selain itu, pengguna merasakan toner ini membantu menyamarkan bekas jerawat, serta pengguna merasakan kulit yang semula berminyak/*oily* menjadi lebih normal/*less oily.* Kata-kata yang berasosiasi dengan kata “banget” memberikan informasi bahwa produk *Tea Tree Skin Clearing Toner* “The Body Shop” ini memberikan efek positif dan bagus dipakai pada pemilik kulit yang *acne prone*. Pengguna merasakan efek *detox* dan hasil berupa berkurangnya kemerahan, meskipun memberikan efek kering pula pada wajah.

Kata-kata yang berasosiasi dengan kata “wajah” memberikan informasi bahwa biasanya pengguna memakai produk dengan cara dituangkan *toner* tersebut di tangan lalu diusapkan ke wajah. Pengguna juga merasakan efek dari produk *toner* yang dapat mengurangi minyak di wajah dan mengangkat residu yang tersisa setelah mencuci muka. Kata-kata yang berasosiasi dengan kata “kulit” memberikan informasi bahwa kulit pengguna beradapatsi seiring bertambahnya penggunaan toner ini, yang semula wajah terasa sedikit panas hingga menjadi ebih normal dan berkurangnya minyak di wajah. Pengguna juga mendapatkan kondisi kulit yang kering apabila toner digunakan setiap hari, dan mensiasatinya dengan mengusapkan toner khusus bagian yang berjerawat saja. Kata-kata yang berasosiasi dengan kata “tea” memberikan informasi bahwa sebagian pengguna seringkali membeli sekaligus satu rangkaian “The Body Shop” yang *series tea tree,* salah satunya adalah *toner* dan *acne spot tea tree oil*.

Kata-kata yang berasosiasi dengan kata “tree” memberikan informasi bahwa sebagian pengguna seringkali membeli satu rangkaian lengkap dari “The Body Shop” khususnya series *tea tree,* dimana series ini terdiri dari *facial wash, acne spot tea tree oil, primer, moisturizer,* dsb. Kata-kata yang berasosiasi dengan kata “bikin” memberikan informasi bahwa produk *Tea Tree Skin Clearing Toner* “The Body Shop” ini tidak menyebabkan beruntusan, kemerahan dan sensasi teralu kering di waajah. Pengguna juga menyampaikan aroma *tea tree* dari produk cukup kuat, serta memberikan informasi bahwa produk tersedia dengan ukuran 250 ml dan 400 ml dengan harga terbarunya. Kata-kata yang berasosiasi dengan kata “cocok” memberikan informasi bahwa bagi pejuang atau pengguna yang melawan jerawat, produk ini sangat direkomendasikan untuk dibeli. Pengguna juga mengungkapkan rasa cocok dan *excited* dengan hasil yang didapatkan setelah memakai *toner* ini.

4.2.7 Sentimen Negatif

Hasil *word cloud* pada sentiment negatif dapat dilihat pada Gambar 5.10 berikut:

Gambar 4. 16 Tampilan *Word Cloud* Sentimen Negatif

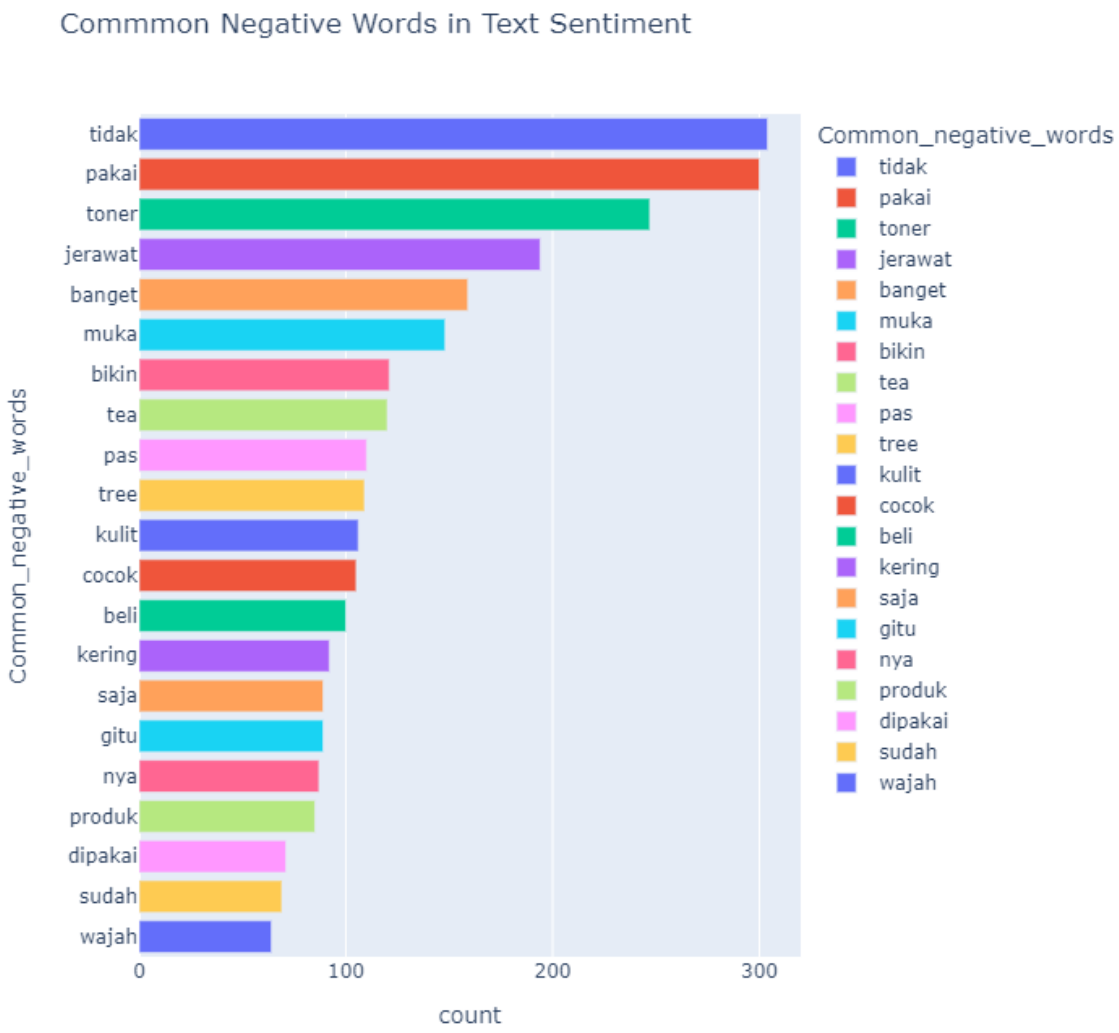
Hasil *word cloud* pada Gambar 4.17 dapat memberikan informasi kata-kata yang sering atau banyak digunakan pada *tweet* atau sentimen negatif Beberapa kata yang sering dibahas pengguna diantaranya adalah mengenai "tidak", "pakai", "toner", "jerawat", "banget", "muka", "bikin", dan seterusnya. Semakin sering dan banyak kata digunakan maka semakin besar ukuran kata tersebut pada *word cloud*. Hasil yang menarik dari *word cloud* juga dapat memudahkan untuk membaca dan mengetahui informasi terkait. Frekuensi kata-kata umum tersebut secara rinci dapat dilihat pada Tabel 4.12 dan *bar plot* yang ditampilkan di Gambar 4.17.

Tabel 4. 12 Jumlah Kata pada *Word Cloud* Sentimen Negatif

Index	***Common Negative Words***	***Count***
0	tidak	304
1	pakai	300
2	*toner*	247
3	jerawat	194
4	banget	159
5	muka	148

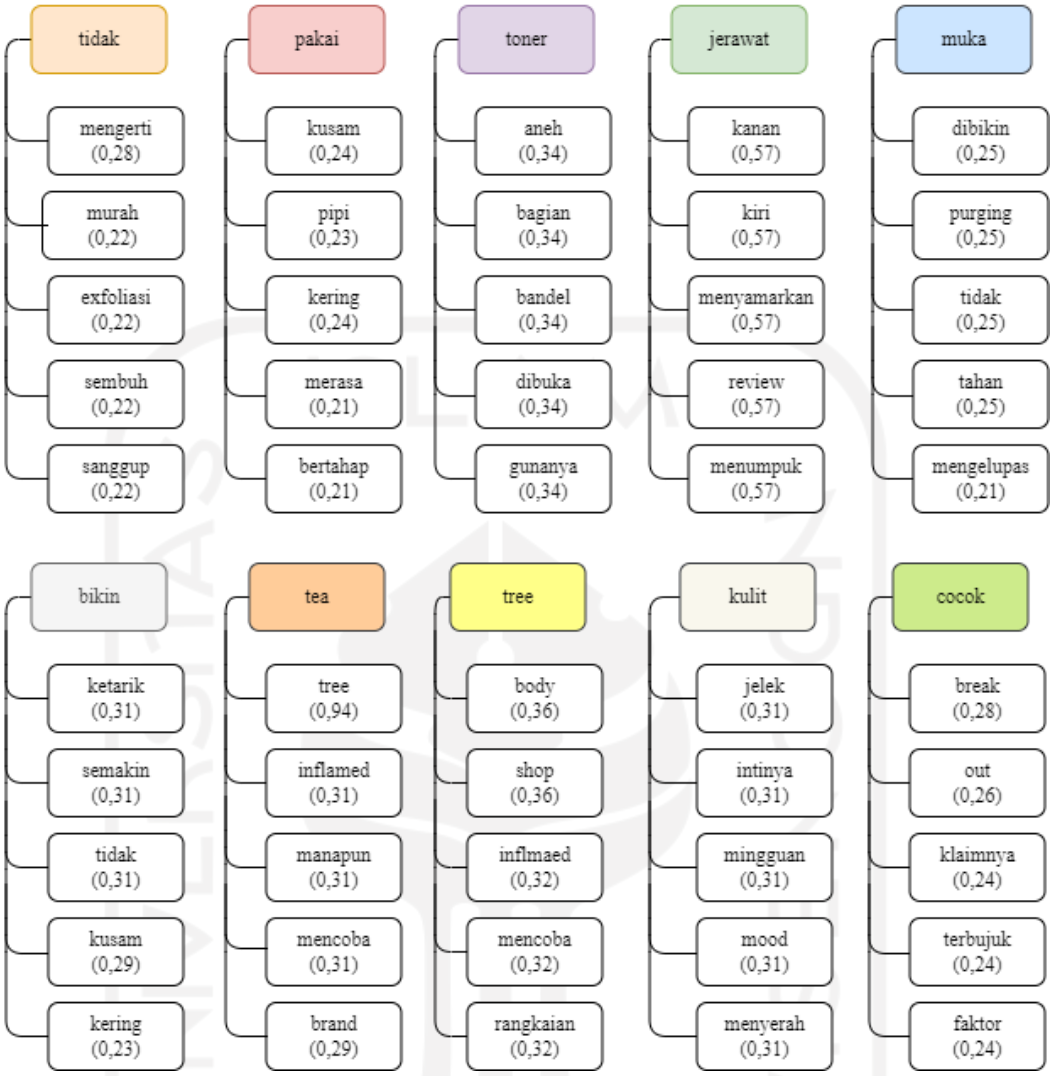
Index	*Common Negative Words*	*Count*
6	bikin	121
7	*tea*	120
8	pas	110
9	*tree*	109
10	kulit	106
11	cocok	105
12	beli	100
13	kering	92
14	saja	89
15	gitu	89
16	nya	87
17	produk	85
18	dipakai	71
19	sudah	69
20	wajah	64

Ekstrasi informasi pada ulasan negatif dilakukan secara berulang-ulang hingga mendapatkan informasi tentang ulasan negatif pengguna *skincare Tea Tree Skin Clearing Toner "*The Body Shop" yang paling sering diulas/dibicarakan. Dari total ulasan sebanyak 1050 ulasan, teridentifikasi sebanyak 270 ulasan negatif. Hasil ekstrasi informasi berupa ulasan negatif diidentifikasi berdasarkan frekuensi kata dalam ulasan, selain itu juga didasarkan pada relevansi kata dengan topik yang mengacu pada sentimen negatif. Berikut merupakan visualisasi hasil ekstrasi informasi yang didapatkan dari ulasan pengguna dengan klasifikasi negatif:



Gambar 4. 17 Persebaran Kata Umum pada Sentimen Negatif

Selain dari tampilan *word cloud* sentiment negataif Gambar 4.17 dapat dilihat juga pada Gambar 4.18 yang merupakan bar plot 20 kata teratas yang sering atau banyak digunakan pada sentimen negatif terkait dengan produk *"*The Body Shop" *Tea Tree Skin Clearing Toner*. Pada hasil ulasan negatif tersebut diperoleh beberapa kata yang paling banyak muncul dengan topik yang dianggap relevan sebagai sentimen negatif, seperti kata "tidak" dengan frekuensi sebanyak 304 kali, "pakai" sebanyak 300 kali, "toner" sebanyak 247 kali, dan seterusnya. Kata-kata yang muncul seperti pada Gambar 4.18 merupakan kata yang memiliki sentimen negatif dan merupakan topik pembicaraan yang paling banyak diulas oleh pengguna *skincare brand "*The Body Shop" tersebut. Kata-kata tersebut selanjutnya digunakan sebagai dasar untuk menemukan asosiasi kata yang lain. Berikut hasil pencarian asosiasi antar kata yang sering muncul secara bersamaan:



Gambar 4. 18 Asosiasi Kata Pada Sentimen Negatif

Berdasarkan Gambar 4.19 diperoleh asosiasi kata pada klasifikasi kelas negatif. Proses ekstrasi informasi dengan asosiasi dilakukan secara berulang-ulang dengan cara menyaring kata-kata yang memiliki hubungan dengan kata lain dan didasarkan pada relevansi kata dengan topik yang diulas. Asosiasi kata yang berkatian dengan kata "tidak", dapat diperoleh informasi tentang pengguna *Tea Tree Skin Clearing Toner* "The Body Shop" ini tidak mengerti dengan efek pada kulitnya yang cukup berbeda dengan efek yang dirasakan pengguna lain. Pengguna mengalami efek *purging* dan tidak kunjung sembuh dari efek tersebut, sehingga pengguna tidak sanggup melanjutkan pemakaian produk karena menimbulkan beruntusan serta kemerahan pada wajah pengguna. Selain itu, pengguna juga kurang merekomendasikan *toner* ini dikarenakan sensasi kering/dehidrasi pada kulit yang diberikan oleh *toner* sama seperti cara kerjanya *toner*

exfoliasi, sementara itu *toner* ini merupakan bukan sebuah *toner* exfoliasi. Selain itu, pengguna merasa bahwa masih banyak *toner* dari *brand* lain yang dinilai harganya lebih murah daripada *toner* ini.

Kata-kata yang berasosiasi dengan kata "pakai" memberikan informasi bahwa pengguna produk ini merasakan efeknya secara bertahap yaitu berkurangnya kusam di wajah, tetapi di area pipi pengguna sering merasa kering dan wangi yang sedikit menyengat serta sensasi perih di bagian mata. Pengguna menyimpukan bahwa pengguna kurang cocok dengan produk *toner* ini. Kata-kata yang berasosiasi dengan kata "*toner*" memberikan informasi bahwa pengguna produk merasakan manfaaat berupa berkurangnya minyak di wajah, tetapi setelah pemakaian hampir 2 bulan secara rutin, pengguna tidak menemukan efek yang signifikan berpengaruh baik ke kulit. Pengguna juga menemukan munculnya jerawat baru di bagian pipi, padahal pengguna tidak mengonsumsi makanan yang aneh-aneh. Serta, pengguna menilai toner dengan harga Rp 179.000 ini seolah kurang berguna, dikarenakan membuat dehidrasi kulit wajah, dan terasa sia-sia mengeluarkan uang sedemikian.

Kata-kata yang berasosiasi dengan kata "jerawat" memberikan informasi bahwa pengguna merasakan produk ini membantu menyembuhkan jerawat yang menumpuk pada pipi kanan dan kiri, tetapi disaat bersamaan juga datanglah beruntusan kembali, sehingga pengguna menilai produk ini tidak menjamin mampu menghambat jerawat/beruntusan untuk datang kembali dan kurang efektif dalam menyamarkan bekas jerawatnya. Kata-kata yang berasosiasi dengan kata "muka" memberikan informasi bahwa beberapa pengguna mendapatkan munculnya jerawat- jerawat kecil selama pemakaian. Pengguna berasumsi bahwa hal tersebut merupakan efek *purging,* hingga kemudian munculnya jerawat baru lagi membuat pengguna berhenti memakai *toner* ini. Pengguna juga merasakan kulit sedikit mengelupas dan tidak tahan dengan sensasi dehidrasi/kering pada kulit.

Kata-kata yang berasosiasi dengan kata "bikin" memberikan informasi bahwa pengguna ketika pertama kali memakai *toner*, kulit wajah rasanya dibuat seperti ketarik dan terasa panas. Setelah selama sebulan memakai *toner*, kulit pengguna semakin kering, kusam, dan munculnya jerawat maupun beruntusan. Kata-kata yang berasosiasi dengan kata "*tea*" memberikan informasi bahwa pengguna kecewa karena merasa tidak cocok dengan hasil pemakaian serangkaian produk *tea tree* dari "The Body Shop", hal ini dikarenakan kulit pengguna yang menjadi *super inflamed*, dan jerawatnya semakin parah.

Pengguna juga mengakui tidak tertarik lagi untuk mencoba produk yang mengandung *tea tree* dari *brand* manapun. Kata-kata yang berasosiasi dengan kata "*tree*" memberikan informasi bahwa pengguna kecewa karena merasa tidak cocok dengan hasil pemakaian serangkaian produk *tea tree* dari "The Body Shop", hal ini dikarenakan kulit pengguna yang menjadi *super inflamed*, dan jerawatnya semakin parah.

Pengguna juga mengakui tidak tertarik lagi untuk mencoba produk yang mengandung *tea tree* dari *brand* manapun. Kata-kata yang berasosiasi dengan kata "kulit" memberikan informasi bahwa pengguna menemukan jerawat yang muncul pada pemakaian seminggu pertama. Pengguna mengasumsikan bahwa hal tersebut wajar sebagai reaksi awal, kemudian setelah pemakaian 3 minggu-an, kulit tidak semakin membaik, sehingga pengguna menyerah untuk melanjutkan pemakaian produk. Pengguna juga mengungkapkan bahwa *mood*/suasana hati pengguna untuk meneruskan pemakaian produk menjadi jelek/buruk akibat hasil performa produk yang tidak sesuai dengan ekspektasi. Kata-kata yang berasosiasi dengan kata "cocok" memberikan informasi bahwa sebagian pengguna terbujuk membeli *toner* ini dikarenakan pernah cocok dengan produk *tea tree facial wash* sebelumnya. Setelah pemakaian *toner*, produk merasa kecewa karena produk tidak bisa menyembuhkan *break out*/jerawat-jerawat pengguna, serta pengguna beropini bahwa penggunaan *toner* ini menjadikan *breakout* pengguna semakin parah.

BAB V

ANALISA DAN PEMBAHASAN

5.1 Gambaran Umum Persepsi Pengguna Produk

Produk *skincare* berupa *Tea Tree Skin Clearing Toner* dari *brand* “The Body Shop” di Female Daily memiliki sebanyak 1.075 pengguna yang telah mengulas produk tersebut (11/09/2022). Rating aplikasi ini hanya mencapai 3,5 dari 5 bintang, dengan jumlah presentase pengguna yang merekomendasikan produk sebanyak 22%. Nilai tersebut menunjukkan bahwa produk yang berasal dari *Brand* “The Body Shop” tersebut belum sepenuhnya memiliki performa yang memuaskan maupun menjawab kebutuhan kulit pengguna.

Proses analisis sentimen yang telah dilakukan menunjukkan bahwa jumlah ulasan yang termasuk dalam kelas sentimen negatif sebanyak 270 ulasan dan kelas positif sebanyak 780 ulasan dari total 1.050 ulasan. Hasil tersebut menunjukkan bahwa eror performa dalam produk “The Body Shop” *Tea Tree Skin Clearing Toner* masih cukup tinggi, yaitu mencapai 25% yang ditunjukkan dengan jumlah ulasan negatif. Berdasarkan penjelasan diatas, dapat didapatkan informasi bahwa hasil analisis sentimen relevan dan telah menginterpretasikan rating yang didapatkan produk dari *brand* “The Body Shop”. Hasil analisis sentimen dan rating produk “The Body Shop” *Tea Tree Skin Clearing Toner* menunjukkan bahwa eror kinerja pada produk masih sangat tinggi. Oleh karena itu perlu adanya evaluasi dan perbaikan performa produk “The Body Shop” *Tea Tree Skin Clearing Toner*.

5.2 Hasil Penerapan Metode *Support Vector Machine* dan *Naïve Bayes*

Proses klasifikasi diawali dengan membagi data uji dan data latih. Pada penelitian ini terdapat tiga perbandingan presentase data *training* dan data *testing* yang berbeda. Setiap bentuk perbandingan akan dilakukan percobaan sebanyak tiga kali percobaan.

Berdasarakan ketiga pembagian data tersebut, metode SVM maupun NBC memiliki akurasi terbaik pada pembagian data jenis ketiga yang dilakukan dengan *data training* 90% (945 ulasan) dan *data testing* 10% (105 ulasan). Dari tiga kali percobaan, metode SVM memiliki akurasi terbaik pada percobaan kedua yaitu sebesar 86% dengan nilai presisi kelas positif 88% dan negatif sebesar 77%. Sementara itu, pengukuran lain berupa nilai *Area Under Rate* (AUC) pada metode SVM yaitu sebesar 0.91 yang artinya bahwa klasifikasi yang digunakan tergolong sangat baik. Berdasarkan dari pengolahan data menggunakan metode *Support Vector Machine* tersebut, dapat dianalisa bahwa semakin banyak porsi data untuk data *training* maka tingkat akurasinya juga akan meningkat. Penelitian yang telah dilakukan oleh (Hendriyanto, et al., 2022) mengenai "Analisis Sentimen Ulasan Aplikasi Mola Pada Google *Play Store* Menggunakan Algoritma *Support Vector Machine*" juga menyatakan hal serupa. Penelitian tersebut memaparkan perbandingan proporsi pembagian data sebesar 70:10, 80:20, dan 90:10, dengan perolehan hasil berupa proporsi pembagian data sebesar 90:10 yang mendapatkan nilai akurasi, presisi, recall, dan *f1-score* tertinggi.

Sedangkan pada metode NBC, akurasi tertinggi juga didapatkan pada percobaan kedua yaitu dilakukan dengan *data training* 90% (945 ulasan) dan *data testing* 10% (105 ulasan, dimana memperoleh nilai akurasi sebesar 83%. Selain itu, nilai *Area Under Rate* (AUC) pada metode NBC yaitu sebesar 0.82 yang artinya bahwa klasifikasi yang digunakan tergolong baik. Berdasarkan dari pengolahan data menggunakan metode *Naïve Bayes* tersebut, dapat dianalisa bahwa semakin tinggi proporsi pembagian data untuk data *training* atau data latih, maka semakin tinggi nilai akurasi yang didapatkan. Hal ini didukung dengan hasil yang sama pada hasil penelitian pada jurnal (Gunawan, et al., 2018) mengenai "Sistem Analisis Pada Ulasan Produk Menggunakan Metode *Naïve Bayes*" dengan menggunakan proporsi pembagian data sebesar 80:20 dan 90:10 yang menghasilkan kesimpulan bahwa proporsi pembagian data sebesar 90:10 yang mendapatkan nilai akurasi tertinggi.

Sementara itu, berdasarkan rata-rata yang didapatkan, metode SVM memiliki performa lebih baik dengan rata-rata akurasi sebesar 83% dibandingkan metode NBC yang memiliki rata-rata akurasi sebesar 80%. Oleh karena itu dapat disimpulkan bahwa metode SVM memiliki kemampuan yang lebih baik dalam mengklasifikasikan ulasan pengguna *toner* The Body Shop dibandingkan metode NBC. Hal ini bisa disebabkan oleh SVM yang mempunyai fungsi yang bisa mentransformasikan data ke ruang dimensi

yang lebih tinggi yaitu ruang *kernel* yang disebut dengan fungsi *kernel tricks* sehingga data dapat dipisahkan dengan lebih baik dibandingkan metode Naïve Bayes (Mukarramah, et al., 2021).

Pada nilai rata-rata total akurasi dari metode SVM dan NBC menunjukkan *error* lebih dari 17%, yang mana nilai ini harusnya dapat lebih diminimalkan. Nilai *error* ini dipengaruhi oleh jenis pelabelan kelas sentimen yang digunakan. Pendeteksian sentimen antar kata pada penelitian ini menggunakan kamus *lexicon* atau *bag of words*. Dalam sistem ini, kata-kata yang terkandung dalam sumber dipelajari tanpa ada klasifikasi kata berdasarkan struktur kalimat, seperti kata benda, kata kerja, kata sifat, dll. Inilah sebabnya mengapa kesalahan dipicu dalam sistem pelabelan karena kerancuan padanan kata. Sebagai contoh kata "tidak cocok", padanan kata ini memiliki kata "tidak" yang bersentimen negatif sedangkan kata "cocok" bersentimen positif, dimana keduanya bertolak belakang.

5.3 Hasil Klasifikasi dan Asosiasi Kata

Berdasarkan hasil klasifikasi ulasan produk "The Body Shop" *Tea Tree Skin Clearing Toner*, terdapat 780 ulasan positif. Dari ulasan tersebut, didapatkan beberapa kata yang paling sering muncul diantaranya adalah kata "pakai" dengan frekuensi 953 kali, "toner" sebanyak 838 kali, "banget" sebanyak 685 kali, "tidak" sebanyak 681 kali, "jerawat" sebanyak 670 kali, dan seterusnya seperti yang tertera pada Gambar 4.9. Kata-kata tersebut merupakan kata-kata yang memiliki sentimen positif dan menjadi topik pembicaraan yang paling banyak diulas oleh pengguna produk.

Hasil klasifikasi ulasan menunjukkan bahwa terdapat 270 ulasan negatif. Dari ulasan tersebut, didapatkan beberapa kata yang paling sering muncul seperti kata "tidak" dengan frekuensi sebanyak 304 kali, "pakai" sebanyak 300 kali, "*toner*" sebanyak 247 kali, dan seterusnya seperti yang terlihat pada Gambar 4.12. Kata-kata yang paling sering muncul dalam setiap kelas sentimen ini perlu dilakukan analisis asosiasi kata. Hal tersebut bertujuan agar peneliti mendapatkan informasi dan kesimpulan yang jelas atau tidak rancu.

Hasil asosiasi kata pada klasifikasi kelas sentimen positif seperti pada Gambar 4.16. menunjukkan proses ekstraksi informasi dengan asosiasi dilakukan secara berulang-

ulang dengan cara menyaring kata-kata positif yang memiliki hubungan dengan kata lain dan didasarkan pada relevansi kata dengan topik yang diulas. Asosiasi kata yang diulas berkaitan dengan kata "toner", "pakai", "jerawat", "banget", "wajah", "kulit", "tea", "tree", "buat", dan "cocok". Pengguna produk *Tea Tree Skin Clearing* "The Body Shop" merasakan manfaat produk berupa mampu mengurangi beruntusan, dimana dengan tektur produk yang cair dan daya tahan toner yang selama 12 bulan setelah produk dibuka, sehingga aman untuk disimpan dalam waktu cukup lama. Dalam hal cara pemakaian, pengguna *toner* ini sering memakai *toner* setelah mencuci wajah, kemudian melakukan *double cleansing* dengan mengusapkan *toner* menggunakan kapas ke seluruh wajah, dan sebagian pengguna juga menggunakannya dengan cara mengompres bagian jerawat saja. Selain itu, pengguna juga menyampaikan informasi bahwa sebeum memakai *toner*, pastikan *toner* telah dikocok sebelum dipakai, sehingga residu pada toner tercampur sepenuhnya. Berdasarkan *feedback* pengguna, pengguna merasakan *toner* ini membantu menyamarkan bekas jerawat, serta pengguna merasakan kulit yang semula berminyak/*oily* menjadi lebih normal/*less oily.* Selain bagus digunakan oleh pemilik kulit yang *acne prone,* efek positif lainnya yang dirasakan oleh pengguna adalah efek *detox* dan berkurangnya kemerahan, meskipun memberikan efek kering pula pada wajah. Informasi lainnya yang didapatkan dari asosisai kata sentiment positif ini adlaah bahwa sebagian pengguna tidak hanya membeli produk *Tea Tree Skin Cearing Toner* "The Body Shop" ini saja, melainkan seringkali membeli satu rangkaian lengkap dari "The Body Shop" khususnya series *tea tree,* dimana series ini terdiri dari *facial wash, acne spot tea tree oil, primer, moisturizer,* dsb.

Hasil asosiasi kata pada klasifikasi kelas sentimen negatif seperti pada Gambar 4.19. menunjukkan asosiasi antar kata pada ulasan negatif, yaitu "tidak", "pakai", "toner", "jerawat", "muka", "bikin", "tea", "tree", "kulit", dan "cocok". Sebagian pengguna produk *Tea Tree Skin Clearing Toner* "The Body Shop" mengalami efek *purging,* dan munculnya beruntusan serta kemerahan pada wajah pengguna. Selain itu, pengguna juga kurang merekomendasikan *toner* ini dikarenakan sensasi kering/dehidrasi pada kulit. Informasi lain perihal persepsi pengguna diantaranya adalah pengguna menilai wangi *tea tree* pada produk cukup kuat atau sedikit menyengat serta. Selain diketahui adanya ketidakcocokan pengguna dengan produk *toner* ini, diketahui pula bahwa terdapat ketidakpuasan pengguna terhadap kinerja produk bekerja. Pengguna yang rutin memakai produk toner tidak menemukan efek yang signifikan berpengaruh baik ke kulit. Dari segi

harga, sebagian pengguna menilai produk tergolong cukup mahal, apabila dibandingkan dengan *brand* lain masih banyak ditemukan *toner brand* lain yang dinilai harganya lebih murah daripada *Tea Tree Skin Clearing Toner* "The Body Shop" ini. Dengan dibandrol harga Rp 179.000, beberapa pengguna menyayangkan uang yang mereka keluarkan untuk produk yang tidak bisa memenuhi ekpektasi dan kebutuhan kulit pengguna. Secara segi kinerja, produk toner ini mampu mengurang minyak pada wajah, tetapi dilain sisi membuat wajah dehidrasi dan mendatangkan beruntusan maupun jerawat muncul kembali, sehingga pengguna menilai produk ini tidak menjamin mampu menghambat jerawat/beruntusan untuk datang kembali dan kurang efektif dalam menyamarkan bekas jerawatnya. Efek *purging* yang diberikan hingga munculnya jerawat baru serta kulit yang mengelupas akibat dehidrasi/terlalu kering/kurangnya hidrasi pada kulit menyebabkan kulit tampak kusam bagi pengguna. Sebagian pengguna juga memakai serangkaian produk lengkap dari *series tea tree* "The Body Shop", dimana kulit sebagian pengguna menjadi *super inflamed,* dan jerawatnya semakin parah. Kekecewaan lain juga pengguna ungkapkan, seperti produk tidak bisa menyembuhkan *break out*/jerawat-jerawat pengguna, hingga pengguna beropini bahwa penggunaan *toner* ini menjadikan *breakout* pengguna semakin parah.

BAB VI

PENUTUP

6.1 Kesimpulan

Berdasarkan hasil analisis dan rumusan masalah, diperoleh hasil kesimpulan untuk menjawab rumusan masalah tersebut, yaitu:

1. Analisis sentimen dengan algoritma *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM) berhasil mengklasifikasikan 1.050 data bersih sentimen kedalam sentimen positif (74.3% atau 780 data) dan sentimen negatif (25.7% atau 270 data). Berikut hasil performa kinerja model pada masing-masing algoritma:
    a. Berdasarkan hasil klasifikasi dari model *sentiment* menggunakan SVM pada *Kernel Linier* dengan pembagian data latih 90% dan data uji 10% diperoleh hasil akurasi sebesar 86%, maka dapat dikatakan bahwa kinerja model klasifikasi yang dibuat sudah sangat bagus. Nilai AUC sebesar 0.91 yang artinya nilai AUC sudah baik atau klasifikasi tergolong sangat baik.
    b. Berdasarkan hasil klasifikasi dari model *sentiment* menggunakan NBC diperoleh hasil akurasi sebesar 83%, sehingga dapat dikatakan bahwa kinerja model klasifikasi yang dibuat sudah bagus. Nilai AUC sebesar 0.82 yang artinya nilai tersebut sudah baik atau klasifikasi tergolong baik.
    c. Dapat disimpulkan bahwa algoritma SVM *memiliki* tingkat akurasi dan nilai AUC yang lebih tinggi daripada algoritma NBC, sehingga klasifikasi data ulasan produk *Tea Tree Skin Clearing Toner* "The Body Shop" di FemaleDaily sebaiknya dilakukan dengan algoritma SVM daripada NBC.
2. Berdasarkan proses asosiasi kata yang telah dilakukan diketahui topik ulasan yang sering dibicarakan pengguna produk *Tea Tree Skin Clearing Toner Brand* "The Body Shop". Kata yang paling sering muncul dalam topik ulasan pada kelas sentimen positif yaitu "toner", "pakai", "jerawat", "banget", "wajah", "kulit", "tea", "tree", bikin", dan "cocok". Sedangkan kata yang sering muncul kelas sentimen negatif yaitu "tidak", "pakai", "toner", "jerawat", "wajah", "bikin", "tea", "tree", "kulit", dan "cocok".

6.2 Saran

Berikut saran yang dapat diberikan melalui hasil penelitian yang telah dilakukan:

1. Bagi perusahaan Natura & Co Holding yang menaungi *brand* The Body Shop
   Saran peneliti untuk *Brand* The Body Shop, khususnya bagian *Product Development* pada produk *Miraculous Refining Toner* yaitu tetap menjaga dan mempertahankan kualitas bagus pada setiap produknya, serta mempertimbangkan dan memperbaiki hal-hal yang dikeluhkan oleh pengguna pada komentar ulasan produk, salah satunya yaitu adanya sensasi sedikit panas ketika pemakaian, efek kulit kering/dehidrasi setelah pemakaian, efek yang tidak siginifikan berpengaruh dalam menyembuhkan jerawat, hingga aroma *tea tree* yang terlalu menyengat. Oleh karena itu, perlu adanya pengembangan produk terutama pada komposisi formula untuk lebih efektif dalam menangani permasalahan kulit yang dialami pengguna, seperti jerawat, kulit menjadi kering, dan keluhan lainnya. Selain itu, berdasarkan keluhan dari aspek harga, diperlukanya strategi penentuan harga produk yang lebih optimal maupun pemberian *voucher* atau promo *good deals* untuk menjaga *loyalty* dari konsumen. Hal-hal tersebut dapat dijadikan acuan untuk perusahaan agar terus berkembang dan dapat memuaskan pengguna.
2. Bagi peneliti berikutnya
   a. Diharapkan pada penelitian selanjutnya, pada *dataset* yang digunakan baik jumlah data positif, negatif maupun netral memiliki perbandingan kelas yang seimbang.
   b. Metode yang digunakan juga masih perlu dikembangkan dengan tujuan untuk meningkatkan akurasi hasil klasifikasi, misalnya dengan menggunakan fitur atau operator lain seperti *k-fold cross validation*, *n-Gram* dan sebagainya.
   c. Bagi peneliti selanjutnya, dapat menggunakan pendekatan metode *machine learning* lain seperti *Neural Network*, *Random Forest*, dsb sebagai pembanding performa algoritma *Naïve Bayes Classifier* dan *Suport Vector Machine*.

DAFTAR PUSTAKA

Ababneh, J. (2019). Application of Naïve Bayes, decision tree, and K-nearest neighbors for automated text classification. *Modern Applied Science*, *13*(11), 31.

Afdhal, I., Kurniawan, R., Iskandar, I., Salambue, R., Budianita, E., & Syafria, F. (2022). Penerapan Algoritma Random Forest Untuk Analisis Sentimen Komentar Di YouTube Tentang Islamofobia. *Penerapan Algoritma Random Forest Untuk Analisis Sentimen Komentar Di YouTube Tentang Islamofobia*, *5*(1), 122-130.

Al-Mejibli, I. S., Alwan, J. K., & Abd Dhafar, H. (2020). The effect of gamma value on support vector machine performance with different kernels. *International Journal of Electrical and Computer Engineering*, *10*(5), 5497.

Amanatidou, A. (2022). Sentiment analysis of cosmetic industry YouTube video campaigns.

Arthamevia, N. P., & Purbolaksono, M. D. (2021, August). Aspect-Based Sentiment Analysis in Beauty Product Reviews Using TF-IDF and SVM Algorithm. In *2021 9th International Conference on Information and Communication Technology (ICoICT)* (pp. 197-201). IEEE.

Auliya, Z. F., Umam, M. R. K., & Prastiwi, S. K. (2017). Online costumer reviews (OTRs) dan rating: Kekuatan baru pada pemasaran online di Indonesia. *Ebbank*, *8*(1), 89-98.

Azhar, Y. (2017). Metode Lexicon-Learning Based Untuk Identifikasi Tweet Opini Berbahasa Indonesia. *Jurnal Nasional Pendidikan Teknik Informatika: JANAPATI*, *6*(3), 237-242.

Baid, P., Gupta, A., & Chaplot, N. (2017). Sentiment analysis of movie reviews using machine learning techniques. *International Journal of Computer Applications*, *179*(7), 45-49.

Bayhaqy, A., Sfenrianto, S., Nainggolan, K., & Kaburuan, E. R. (2018, October). Sentiment analysis about E-commerce from tweets using decision tree, K-nearest neighbor, and naïve bayes. In *2018 international conference on orange technologies (ICOT)* (pp. 1-6). IEEE.

Casaló, L., Flavián, C., & Guinalíu, M. (2011). Understanding the intention to follow the advice obtained in an online travel community. *Computers in Human Behavior*, *27(2)*(https://doi.org/10.1016/j.chb.2010.04.013), 622-633.

Castella, Q., & Sutton, C. (2014). Word Storm: Multiples of Word Clouds for Visual Comparison of Documents.

Chapman, A. D., Turland, N. J., & Watson, M. F. (2010). Report of the special committee on electronic publication. *Taxon*, *59*(6), 1853-1862.

Das, S., & Nene, M. J. (2017). A survey on types of machine learning techniques in intrusion prevention systems. 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2296–2299. https://doi.org/10.1109/WiSPNET.2017.8 300169

Dey, S., Wasif, S., Tonmoy, D. S., Sultana, S., Sarkar, J., & Dey, M. (2020, February). A comparative study of support vector machine and Naive Bayes classifier for sentiment analysis on Amazon product reviews. In *2020 International Conference on Contemporary Computing and Applications (IC3A)* (pp. 217-220). IEEE.

Dhini, A., & Kusumaningrum, D. A. (2018, December). Sentiment analysis of airport customer reviews. In *2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 502-506). IEEE.

Diani, R. (2017). Analisis Pengaruh Kernel Support Vector Machine (SVM) pada Klasifikasi Data Microarray untuk Deteksi Kanker. *Indonesia Journal on Computing (Indo-JC)*, *2*(1), 109-118.

Diba, S. F. (2019). *Implementasi Metode Naive Bayes Classifier Dalam Analisis Sentimen Pada Opini Masyarakat Terhadap RKUHP (Studi Pada: Data Komentar Twitter Mengenai RKUHP Tahun 2019)* (Doctoral dissertation, Universitas Islam Indonesia).

Eska, J. (2018). *Penerapan Data Mining Untuk Prediksi Penjualan Wallpaper Menggunakan Algoritma C4.5*. *2*. https://doi.org/10.31227/osf.io/x6svc

Farki, A. (2016). *Pengaruh online customer review dan rating terhadap kepercayaan dan minat pembelian pada online marketplace di Indonesia* (Doctoral dissertation, Institut Teknology Sepuluh Nopember).

Female Daily. (2021). *4 Alasan Kenapa Female Daily Network adalah #YourBeautySupportSystem!* Retrieved October 07, 2022, from https://editorial.femaledaily.com/blog/2021/08/20/4-alasan-kenapa-female-daily-network-adalah-yourbeautysupportsystem

Fikria, N. (2018). Analisis Klasifikasi Sentimen Review Aplikasi E-Ticketing Menggunakan Metode Support Vector Machine dan Asosiasi.

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Unsupervised learning techniques*. O'Reilly Media, Incorporated.

Gilchrist, M., Lehmann Mooers, D., Skrubbeltrang, G., & Vachon, F. (2012). Knowledge discovery in databases for competitive advantage. Journal of Management and Strategy, 3(2).

Good News from Indonesia. (2021). *Berkenalan dengan Produk Skincare Lokal Terpopuler*. Retrieved July 28, 2022, from https://www.goodnewsfromindonesia.id/2021/04/27/berkenalan-dengan-produk-skincare-lokal-terpopuler

Gorunescu, F. (2011). Data Mining: Concepts, models and techniques. Springer-Verlag Berlin Heidelberg, Intelligent Systems References Library, Vol. 12 ISBN: 978-3-642-19720-8.

GULTOM, F. (2016). *Analisa Bad Hike Pada Kran Lavatory Tipe S11234r Menggunakan Metode Nominal Group Technique Dan Metode Fishbone di PT. Surya Toto Indonesia, TBK* (Doctoral dissertation, Universitas Gadjah Mada).

Gunawan, B., Sastypratiwi, H., & Pratama, E. E. (2018). Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes. *JEPIN (Jurnal Edukasi dan Penelitian Informatika)*, *4*(2), 113-118.

Habibi, R., Setyohadi, D. B., & Ernawati. (2016). Analisis Sentimen pada Twitter Mahasiswa Menggunakan Metode Backpropagation. *Jurnal Informatika, Vol. 12, No. 1*, 103-109.

Han, J., & Kamber, M. (2011). *Data Mining Concepts and Techniques Third Edition.* Waltham: Elsevier Inc.

Hendriyanto, M. D., Ridha, A. A., & Enri, U. (2022). Analisis Sentimen Ulasan Aplikasi Mola Pada Google Play Store Menggunakan Algoritma Support Vector Machine. *Journal of Information Technology and Computer Science*, *5*(1), 1-7.

Hapsari, C. C. P., Astuti, W., & Purbolaksono, M. D. (2021, October). Naive Bayes Classifier and Word2Vec for Sentiment Analysis on Bahasa Indonesia Cosmetic

Product Reviews. In *2021 International Conference on Data Science and Its Applications (ICoDSA)* (pp. 22-27). IEEE.

Hermansyah, R., & Sarno, R. (2020, September). Sentiment Analysis about Product and Service Evaluation of PT Telekomunikasi Indonesia Tbk from Tweets Using TextBlob, Naive Bayes & K-NN Method. In *2020 International Seminar on Application for Technology of Information and Communication (iSemantic)* (pp. 511-516). IEEE.

Indrayuni, E., & Nurhadi, A. (2020). Optimizing Genetic Algorithms for Sentiment Analysis of Apple Product Reviews Using SVM. *Sinkron: jurnal dan penelitian teknik informatika*, *4*(2), 172-178.

Isnain, A. R., Supriyanto, J., & Kharisma, M. P. (2021). Implementation of K-Nearest Neighbor (K-NN) Algorithm For Public Sentiment Analysis of Online Learning. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *15*(2), 121-130.

Jaka, A. T. (2015). Preprocessing Text untuk Meminimalisir Kata yang Tidak Berarti dalam Proses Text Mining. *Informatika UPGRIS Vol. 1 Edisi Juni 2015*, 1-2.

Jananto, A. (2013). Algoritma Naive Bayes untuk Mencari Perkiraan Waktu Studi Mahasiswa. *Dinamik*, *18*(1).

Josi, A., & Abdillah, L. A. (2014). Penerapan teknik web scraping pada mesin pencari artikel ilmiah. *arXiv preprint arXiv:1410.5777*.

Kirana, Y. D., & Al Faraby, S. (2021). Sentiment Analysis of Beauty Product Reviews Using the K-Nearest Neighbor (KNN) and TF-IDF Methods with Chi-Square Feature Selection. *Journal of Data Science and Its Applications*, *4*(1), 31-42.

Kristiyanti, D. (2015). Analisis Sentimen Review Produk Kosmetik menggunakan Algoritma Support Vector Machine dan Particle Swarm Optimization sebagai Metode Seleksi Fitur. *SNIT, 1(1)*, 134-141.

Lackermair, G., Kailer, D., & Kanmaz, K. (2013). Importance of online product reviews from a consumer's perspective. *Advances in economics and business*, *1*(1), 1-5.

Latief, F., & Ayustira, N. (2020). Pengaruh Online Costumer Review Dan Customer Rating Terhadap Keputusan Pembelian Produk Kosmetik Di Sociolla. *Jurnal Mirai Management*, *5*(3), 139-154.

Ma, T. M., Yamamori, K., & Thida, A. (2020, October). A comparative approach to Naïve Bayes classifier and support vector machine for email spam classification. In *2020 IEEE 9th Global Conference on Consumer Electronics (GCCE)* (pp. 324-326). IEEE.

Ma, Q., Tsukagoshi, M., & Murata, M. (2020, December). Estimating Evaluation of Cosmetics Reviews with Machine Learning Methods. In *2020 International Conference on Asian Language Processing (IALP)* (pp. 259-263). IEEE.

Maimon, O., & Rokach, L. (2009). Introduction to knowledge discovery and data mining. In Data mining and knowledge discovery handbook (pp. 1-15). Springer, Boston, MA.

Mariel, W. C. F., Mariyah, S., & Pramana, S. (2018, March). Sentiment analysis: a comparison of deep learning neural network algorithm with SVM and naïve Bayes for Indonesian text. In *Journal of Physics: Conference Series* (Vol. 971, No. 1, p. 012049). IOP Publishing.

Melita, R., Amrizal, V., Suseno, H. B., & Dirjam, T. (2018). Penerapan Metode Term Frequency Inverse Document Frequency (TF-IDF) dan Cosine Similarity pada Sistem Temu Kembali Informasi untuk mengetahui Syarah Hadits berbasis Web

(Studi Kasus: Syarah Umdatil Ahkam). *Jurnal Teknik Informatikam Vol. 11, No. 2,* 149-164.

Moens, M., Li, J., & Chua, T. (2014). *Mining User Generated Content.* Boca Raton: CRC Press.

Mukarramah, R., Atmajaya, D., & Ilmawan, L. B. (2021). Performance comparison of support vector machine (SVM) with linear kernel and polynomial kernel for multiclass sentiment analysis on twitter. *ILKOM Jurnal Ilmiah, 13*(2), 168-174.

Murphy, R. (2019). *Local Consumer Review Survey*. Retrieved July 28, 2022, from https://www.brightlocal.com/research/local-consumer-review-survey/

Nurjannah, M., Hamdani, & Astuti, I. F. (2013). Penerapan Algoritma Term Frequency-Inverse Document Frequency)TF-IDF) untuk Text Mining. *Jurnal Informatika Mulawarman, Vol. 8, No. 3,* 110-113.

O'Keefe, T., & Koprinska, I. (2009, December). Feature selection and weighting methods in sentiment analysis. In *Proceedings of the 14th Australasian document computing symposium, Sydney* (pp. 67-74).

Pradana, M. G. (2020). Penggunaan Fitur Wordcloud dan Dokumen Term Matrix dalam Text Mining. *Jurnal Ilmiah Informatika (JIF), Vol. 8, No. 1*, 38-43.

Praghakusma, A. Z., & Charibaldi, N. (2021). Komparasi Fungsi Kernel Metode Support Vector Machine untuk Analisis Sentimen Instagram dan Twitter (Studi Kasus: Komisi Pemberantasan Korupsi). *Jurnal Sarjana Teknik Informatika ISSN*, *2338*(5197), 33.

Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing*, *5*(2), 221.

Rachmania, R. (2021). Pengaruh Perceived Social Media Marketing Instagram Shopee Indonesia Terhadap Niat Beli pada E-Commerce. *Syntax Literate; Jurnal Ilmiah Indonesia*, *6*(6), 2998-3006.

Ramadhan, R., Sari, Y. A., & Adikara, P. P. (2021). Perbandingan Pembobotan Term Frequency-Inverse Document Frequency dan Tern Frequency-Relevance Frequency terhadap Fitur N-Gram pada Analisis Sentimen. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, Vol. 5, No. 11*, 5075-5079.

Ramadhanti, D., 2021. *PERAN E-WOM DALAM FEMALE DAILY TERHADAP MINAT BELI PRODUK SKINCARE (Studi Kasus Produk Sun Protection L'Oreal Paris)* (Doctoral dissertation, Universitas Pendidikan Indonesia).

Ramdhani, M. A., & Rahim, O. N. (2014). Analisis Sentimen untuk Mengukur Popularitas Tokoh Publik Berdasar Data pada Media Sosial Twitter Menggunakan Algoritma Data Mining Teknik Klasifikasi. *Jurnal Informasi, Vol. VI, No. 2*, 56-87.

Roihan, A., Sunarya, P. A., & Rafika, A. S. (2020). Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper. *IJCIT (Indonesian Journal on Computer and Information Technology), 5*(1), 75–82. https://doi.org/10.31294/ijcit.v5i1.7951

Rozi, I., Firdausi, A., & Islamiyah, K. (2020). Analisis Sentimen Pada Twitter Mengenai Pasca Bencana Menggunakan Metode Naïve Bayes Dengan Fitur N-Gram,. *J. Inform. Polinema, vol. 6, no. 2,* 33–39.

Saleh, A. (2015). Implementasi Metode Klasifikasi Naive Byes Dalam Memprediksi Besrnya Penggunaan Listrik Rumah Tangga. *Citec Journal*, 209-210.

Santoso, E. B., & Nugroho, A. (2019). Analisis sentimen calon presiden indonesia 2019 berdasarkan komentar publik di facebook. *Jurnal Eksplora Informatika*, *9*(1), 60-69.

Saraswati, M. K., & Waluyo, E. A. (2022). Pemindahan Ibu Kota Negara Ke Provinsi Kalimantan Timur Berdasarkan Analisis Swot. *Jurnal Ilmu Sosial dan Pendidikan (JISIP)*.

Sari, F. V., & Wibowo, A. (2019). Analisis Sentimen Pelanggan Toko Online JD.ID Menggunakan Metode Naive Bayes Classifier Berbasis Konversi Ikon Emosi. *Jurnal Simetris, Vol. 10 No. ,* 681-686.

Siswanto, D., Nijal, L., & Rajab, S. (2022). Analisa Sentimen Publik Mengenai Perekonomian Indonesia Pada Masa Pandemi Covid-19 Di Twitter Menggunakan Metode Klasifikasi K-NN Dan Svm. *Jurnal Pustaka AI (Pusat Akses Kajian Teknologi Artificial Intelligence)*, *2*(1), 1-9.

Somvanshi, M., & Chavan, P. (2016). A review of machine learning techniques using decision tree and support vector machine. 2016 International Conference on Computing Communication Control and Automation (ICCUBEA), 1–7. https://doi.org/10.1109/ICCUBEA.2016.7 860040

Surabagiarta, I. K., & Purnaningrum, E. (2021). Pengaruh Brand Image, Brand Awareness, dan Persepsi Kualitas Terhadap Keputusan Pembelian The Body Shop. *Journal of Sustainability Bussiness Research (JSBR), 2*(2), 294-301.

The Body Shop. (2021). *The Body Shop Sustainability Report 2021*. The Body Shop.

The Body Shop. (2022). *About Us*. Retrieved September 30, 2022, from https://www.thebodyshop.co.id/about-us

Thupae, R., Isong, B., Gasela, N., & Abu-Mahfouz, A. M. (2018). Machine Learning Techniques for Traffic Identification and Classifiacation in SDWSN: A Survey. IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society, 4645–4650. https://doi.org/10.1109/IECON.2018.859 1178

Top Brand Award. (2022). *Top Brand Award "The Body Shop"*. Retrieved October 08, 2022, from https://www.topbrand-award.com/top-brand%20index/?tbi_find=the%20body%20shop

Tuhuteru, H. (2020). Analisis Sentimen Masyarakat Terhadap Pembatasan Sosial Berksala Besar Menggunakan Algoritma Support Vector Machine. *INFORMATION SYSTEM DEVELOPMENT (ISD), 5*.

Ulwan, M. N. (2016). Pattern Recognition pada Unstructured Data Teks Menggunakan Support Vector Machine dan Association. *Skripsi: Program Studi Statistika Universitas Islam Indonesia*.

Utami, P. D. (2018). Analisis Sentimen Review Kosmetik Bahasa Indonesia Menggunakan Algoritma Naïve Bayes Classifier.

Utomo, D., & Purba, B. (2021). Penerapan Datamining pada Data Gempa Bumi Terhadap Potensi Tsunami di Indonesia. *e-Proceeding of Engineering: Vol.8, No.5 Oktober 2021* (p. 846). Bandung: Prosiding Seminar Nasional Riset Information Science.

Wardani, A. K. (2017). Pengaruh Electronic Word of Mouth Pada Forum Online. 4(2), 1–15.

WeAreSocial. (2022, February 11). Hootsuite (We are Social): Indonesian Digital Report 2022. *Data Reportal Reports*, p. 17.

Widayani, W., & Harliana, H. (2021). Analisis Support Vector Machine Untuk Pemberian Rekomendasi Penundaan Biaya Kuliah Mahasiswa. *Jurnal Sains dan Informatika*, *7*(1), 20-27.

Wijaya, A. P., & Santoso, H. A. (2016). Naive Bayes Classification pada Klasifikasi Dokumen Untuk Identifikasi Konten E-Government. *Journal of Applied Intelligent System, Vol.1, No. 1, Februari 2016*, 48-55.

Wilis, K., Hidayatulah, H., & Parasian, S. (2020). Peer Review The Accuracy Comparison of Social Media Sentiment Analysis Using Lexicon Based and Support Vector Machine on Souvenir Recommendations.

Wisnu, H. (2021). Analisis Keputusan Pembelian Konsumen "The Body Shop". *Analisis Keputusan Pembelian Konsumen "The Body Shop"*.

Yadav, M., & Rahman, Z. (2018). The influence of social media marketing activities on

Yani, D. D. A., Pratiwi, H. S., & Muhardi, H. (2019). Implementasi web scraping untuk pengambilan data pada situs marketplace. *JUSTIN (Jurnal Sistem dan Teknologi Informasi)*, *7*(4), 257-262.

Zikmund, W., Babin, B., Carr, J., & Grif, M. (2009). *Business Research Methods.* South Western: Cengage Learning.

LAMPIRAN

```
# Mengidentifikasi apakah terdapat duplicate value pada tabel
df[df["Text"].duplicated(keep=False)].sort_values("Text")
# Menghilangkan duplicate value pada tabel
df = df.drop_duplicates(subset=['Text']).reset_index()
# Mendeteksi pattern dengan Regex
# Format html
html_tag = re.compile(r'<.*?>')

# Format URL
http_link = re.compile(r'https://\S+')
www_link = re.compile(r'www\.\S+')
useless = re.compile(r'\n')
#text = ' you @warui  and @madawar '
tags = re.compile(r'@\w*')
#tags == ['@warui', '@madawar']

# Tanda baca yang tidak diperlukan
punctuation = re.compile(r'[^\w\s]') #penggunaan ^ merupakan notasi
 untuk menunjukkan negasi dimana dalam perintah ini berarti karakte
r yang bukan huruf dan bukan spasi

# Function untuk memproses cleaning teks data
def data_cleaning(text):
  # menghilangkan html tag
  text = re.sub(html_tag, r'', text)

  # menghilangkan url
  text = re.sub(http_link, r'', text)
  text = re.sub(www_link, r'', text)
  text = re.sub(useless, r'', text)
  text = re.sub(tags, r'', text)

  # menghilangkan tanda baca
  text = re.sub(punctuation, r'', text)

  return text
# Memanggil function data cleaning untuk kemudian diterapkan dalam
dataframe
df["cleaned_text"] = df["Text"].apply(lambda x: data_cleaning(x))
df.head()
df["remove_punc"] = df["Text"].apply(lambda x: x.lower())

#perintah lower() digunakan untuk merubah upper case menjadi lower
case

df["remove_punc"] = df["cleaned_text"].apply(lambda x: data_cleanin
g(x))

df.head()
```

```
start_time = datetime.now()

df["text_clean"] = df["remove_punc"].apply(lambda x: word_tokenize(
x))

end_time = datetime.now()

print('Waktu yang diperlukan: {}'.format(end_time - start_time))

df.head()
```

```
# stopword using nltk
start_time = datetime.now()
stop_words = set(stopwords.words('indonesian'))
df["text_clean"] = df["text_clean"].apply(lambda x: [word for word
in x if word not in stop_words])
end_time = datetime.now()
print('Waktu yang diperlukan: {}'.format(end_time - start_time))
```

Pseudo code yang digunakan sebagai berikut:

```
library(readxl)
kalimat <- read.csv(file = "C:/Users/asrin/Downloads/Career Prep n
TA/Sentiment R/need_labelling.csv", header = TRUE)
View(kalimat)
library(tm)
positif <- scan("C:/Users/asrin/Downloads/Career Prep n
TA/Sentiment R/positive.txt", what = "character", comment.char =
";")
negatif <- scan("C:/Users/asrin/Downloads/Career Prep n
TA/Sentiment R/negative.txt", what = "character", comment.char =
";")
kata.positif <- c(positif)
kata.negatif <- c(negatif)
score.sentiment = function(kalimat, kata.positif,
kata.negatif,.progress='none')
{
  require(plyr)
  require(stringr)
  scores = laply(kalimat, function(kalimat, kata.positif,
kata.negatif)
  {
    kalimat = gsub('[[:punct:]]','', kalimat)
    kalimat = gsub('[[:cntrl:]]','', kalimat)
    kalimat = gsub('\\d+','', kalimat)
    kalimat = tolower(kalimat)
    list.kata = str_split(kalimat, '\\s+')
    kata2 = unlist(list.kata)
```

```
    positif.matches = match(kata2, kata.positif)
    negatif.matches = match(kata2, kata.negatif)
    positif.matches = !is.na(positif.matches)
    negatif.matches = !is.na(negatif.matches)
    score = sum(positif.matches) - (sum(negatif.matches))
    return(score)
  }, kata.positif, kata.negatif, .progress= .progress)
  scores.df = data.frame(score=scores, Kalimat = kalimat)
  return(scores.df)
}
hasil = score.sentiment(kalimat$text_sentiment, kata.positif,
kata.negatif)
head(hasil)
View(hasil)
```

```
# function to analyze the reviews
def analysis(score):
    if score < 0:
        return 'Negative'
    elif score == 0:
        return 'Neutral'
    else:
        return 'Positive'
df = pd.DataFrame(df[['Text', 'remove_punc', 'text_clean', 'text_se
ntiment','score']])
df['Analysis'] = df['score'].apply(analysis)
df['Analysis'] = df['score'].apply(analysis)
df.head(10)

posneg_counts = df.Analysis.value_counts()
posneg_counts
temp = df.groupby('Analysis').count()['Text'].reset_index().sort_va
lues(by='Text',ascending=False)
temp.style.background_gradient(cmap='Purples')
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
plt.figure(figsize=(12,6))
sns.countplot(x='Analysis',data=df)
label_sentiment = pd.DataFrame(df[['remove_punc', 'score', 'Analysi
s']])
# Menampilkan 5 data teratas dari label positive
label_sentiment[label_sentiment["Analysis"]=='Positive'][:15]
# Menampilkan 5 data teratas dari label negative
label_sentiment[label_sentiment["Analysis"]=='Negative'][:20]
# Menampilkan 5 data teratas dari label neutral
label_sentiment[label_sentiment["Analysis"]=='Neutral'][20:]
import matplotlib.pyplot as plt
%matplotlib inline

posneg_counts= df.Analysis.value_counts()
```

```
plt.figure(figsize=(10, 7))
plt.pie(posneg_counts.values, labels = posneg_counts.index, explode
 = (0, 0, 0.25), autopct='%1.1f%%', shadow=False)
# plt.legend()
# Plot sentiment dari user
sentiment_label = df["Analysis"].value_counts()
sentiment_label.plot(kind="bar")
plt.show()
from collections import Counter
import plotly.express as px
import re
import string
import numpy as np
import random
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
from plotly import graph_objs as go
import plotly.express as px
import plotly.figure_factory as ff
from collections import Counter

from PIL import Image
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator


import nltk
from nltk.corpus import stopwords

from tqdm import tqdm
import os
import nltk
import spacy
import random
from spacy.util import compounding
from spacy.util import minibatch

import warnings
warnings.filterwarnings("ignore")

import os
fig1 = go.Figure(go.Funnelarea(
    text =temp.Analysis,
    values = temp.Text,
    title = {"position": "top center", "text": "Funnel-
Chart of Sentiment Distribution"}
    ))
fig1.show()
def random_colours(number_of_colors):
    '''
    Simple function for random colours generation.
    Input:
```

```
        number_of_colors - integer value indicating the number of c
olours which are going to be generated.
    Output:
        Color in the following format: ['#E86DA4'] .
    '''
    colors = []
    for i in range(number_of_colors):
        colors.append("#"+''.join([random.choice('0123456789ABCDEF'
) for j in range(6)]))
    return colors

def display_word(data, color):plt.subplots(figsize=(15,15))
word_cloud = WordCloud(width = 500, height = 300, background_color=
"white", contour_color=color,max_words=2000, random_state=42, collo
cations=False)
  word_cloud.generate(' '.join(data))
  plt.imshow(word_cloud)
  plt.axis('off')
  plt.show()
#hanya menampilkan sentiment positive
positive = df[(df.Analysis == 'Positive')]
positive.head()
positive['temp_list'] = positive['text_sentiment'].apply(lambda x:s
tr(x).split())
top = Counter([item for sublist in positive['temp_list'] for item i
n sublist])
temp = pd.DataFrame(top.most_common(21))
temp.columns = ['Common_positive_words','count']
temp.style.background_gradient(cmap='Blues')
fig2 = px.bar(temp, x="count", y="Common_positive_words", title='Co
mmmon Positive Words in Text Sentiment', orientation='h',
             width=700, height=700,color='Common_positive_words')
fig2.show()
# wordcloud review positif
display_word(positive["text_sentiment"],'blue')

Evaluating Models
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns

posneg_counts = df.Analysis.value_counts()
posneg_counts
# split data
train_data,test_data = train_test_split(df,train_size=0.80,random_s
tate=8)
# select the columns and
# prepare data for the models
X_train = vectorizer.fit_transform(train_data['text_sentiment'])
y_train = train_data['Analysis']
X_test = vectorizer.transform(test_data['text_sentiment'])
```

```
y_test = test_data['Analysis']
start=dt.datetime.now()
nb = MultinomialNB()
nb.fit(X_train,y_train)
print('Elapsed time: ',str(dt.datetime.now()-start))

start=dt.datetime.now()
svm = SVC()
svm.fit(X_train,y_train)
print('Elapsed time: ',str(dt.datetime.now()-start))
from sklearn.svm import SVC
svclassifier = SVC(kernel='linear')
svclassifier.fit(X_train, y_train)
y_pred = svclassifier.predict(X_test)
from sklearn.metrics import classification_report, confusion_matrix
print(confusion_matrix(y_test,y_pred))
print(classification_report(y_test,y_pred))
# preparation for the confusion matrix
svm_cm=confusion_matrix(y_test.values, svclassifier.predict(X_test)
)
from sklearn.svm import SVC
svclassifier = SVC(kernel='poly', degree=8)
svclassifier.fit(X_train, y_train)
y_pred = svclassifier.predict(X_test)
from sklearn.metrics import classification_report, confusion_matrix
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
from sklearn.svm import SVC
svclassifier = SVC(kernel='rbf')
svclassifier.fit(X_train, y_train)
y_pred = svclassifier.predict(X_test)
from sklearn.metrics import classification_report, confusion_matrix
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
from sklearn.svm import SVC
svclassifier = SVC(kernel='sigmoid')
svclassifier.fit(X_train, y_train)
y_pred = svclassifier.predict(X_test)
from sklearn.metrics import classification_report, confusion_matrix
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))

plt.figure(figsize=(8,5))
plt.suptitle("Confusion Matrices Comparison",fontsize=20)

plt.subplot(1,1,1)
plt.title("Support Vector Machine (SVM)")
sns.heatmap(svm_cm, annot = True, cmap="Greens",cbar=False);
pred_svm = svclassifier.decision_function(X_test)
fpr_svm,tpr_svm,_ = roc_curve(y_test, pred_svm[:,1], pos_label='Neu
tral')
roc_auc_svm = auc(fpr_svm,tpr_svm)
f, axes = plt.subplots(2, 2,figsize=(15,10))
```

```
axes[0,1].plot(fpr_svm, tpr_svm, color='darkred', lw=2, label='ROC
curve (area = {:0.2f})'.format(roc_auc_svm))
axes[0,1].plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
axes[0,1].set(xlim=[-0.01, 1.0], ylim=[-0.01, 1.05])
axes[0,1].set(xlabel ='False Positive Rate', ylabel = 'True Positiv
e Rate', title = 'Support Vector Machine')
axes[0,1].legend(loc='lower right', fontsize=13)

# split data
train_data,test_data = train_test_split(df,train_size=0.90,random_s
tate=8)
# select the columns and
# prepare data for the models
X_train = vectorizer.fit_transform(train_data[text_sentiment'])
y_train = train_data['Analysis']
X_test = vectorizer.transform(test_data[text_sentiment'])
y_test = test_data['Analysis']

start=dt.datetime.now()
nb = MultinomialNB()
nb.fit(X_train,y_train)
print('Elapsed time: ',str(dt.datetime.now()-start))

pred_nb = nb.predict_proba(X_test)
fpr_nb,tpr_nb,_ = roc_curve(y_test, pred_nb[:,1], pos_label='Neutra
l')
roc_auc_nb = auc(fpr_nb,tpr_nb)

f, axes = plt.subplots(2, 2,figsize=(15,10))

axes[0,0].plot(fpr_nb, tpr_nb, color='darkred', lw=2, label='ROC cu
rve (area = {:0.2f})'.format(roc_auc_nb))
axes[0,0].plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
axes[0,0].set(xlim=[-0.01, 1.0], ylim=[-0.01, 1.05])
axes[0,0].set(xlabel ='False Positive Rate', ylabel = 'True Positiv
e Rate', title = 'Naive Bayes')
axes[0,0].legend(loc='lower right', fontsize=13)

print("\n Naive Bayes")
print(mt.classification_report(y_test, nb.predict(X_test)))
# preparation for the confusion matrix
nb_cm=confusion_matrix(y_test.values, nb.predict(X_test))

plt.figure(figsize=(15,12))
plt.suptitle("Confusion Matrices Comparison",fontsize=24)

plt.subplot(2,2,1)
plt.title("Naive Bayes")
sns.heatmap(nb_cm, annot = True, cmap="Greens",cbar=False);
```

```
#Assosiasi Kata Dengan R
```

```
library(NLP)
library(tm)
library(SnowballC)
library(RColorBrewer)
library(wordcloud)
library(stringr)
library(wordcloud2)
docs2 <- read.csv(file = "C:/Users/asrin/Downloads/Career Prep n
TA/Sentiment R/CSV/negatif2.csv", header = TRUE)
View(docs2)
docs2 <- Corpus(VectorSource(docs2$text_clean))
docs2 <- tm_map(docs2, stripWhitespace)
inspect(docs2)
dtm<- TermDocumentMatrix(docs2)
m <- as.matrix(dtm)
v2 <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v2),freq=v2)
head(d, 10)
set.seed(1234)
wordcloud(words = d$word, freq = d$freq, min.freq = 1,
          max.words=100, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
findFreqTerms(dtm, lowfreq = 4)
v2<-as.list(findAssocs(dtm, terms
                       =c("pakai", "cekit", "banget", "toner",
"kulit", "exfoliasi", "jerawat", "produk", "coba", "cocok",
"tidak"),
                       corlimit =
c(0.2,0.2,0.2,0.2,0.2,0.2,0.2,0.2,0.2,0.2,0.2)))
View(v2)
k<-barplot(d[1:20,]$freq, las = 2, names.arg =
             d[1:20,]$word,cex.axis=1.2,cex.names=1.2,
           main ="Most Frequent Negative Words",
           ylab = "Word frequencies",col =topo.colors(20))
```

```
k<-barplot(d[1:15,]$freq, las = 2, names.arg =
             d[1:15,]$word,cex.axis=1.2,cex.names=1.2,
           main ="Most Frequent Negative Words",
           ylab = "Word frequencies",col = terrain.colors(20))
termFrequency<- rowSums(as.matrix(dtm))
termFrequency<- subset(termFrequency, termFrequency>=18)
text(k,sort(termFrequency, decreasing = T)-
       1,labels=sort(termFrequency, decreasing = T),pch = 6, cex
= 1)
```