IMPLEMENTASI ALGORITMA SUPPORT VECTOR MACHINE PADA ANALISIS SENTIMEN DATA TWITTER

(Studi Kasus : Ulasan Tentang Indihome Pada *Platform* Twitter)

TUGAS AKHIR

Diajukan Sebagai Salah Satu Syarat Untuk Memperoleh Gelar Sarjana Program Studi Statistika

Disusun Oleh:

Desy Rizki Ramadhanty

17611036

PROGRAM STUDI STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS ISLAM INDONESIA

YOGYAKARTA

2021

HALAMAN PERSETUJUAN PEMBIMBING

TUGAS AKHIR

Judul	:	Implementasi Algoritma Support Vector Machine Pada Analisis Sentimen Data Twitter (Studi Kasus : Ulasan Tentang Indihome Pada *Platform* Twitter)
Nama Mahasiswa	:	Desy Rizki Ramadhanty
NIM	:	17611036

TUGAS AKHIR INI TELAH DIPERIKSA DAN DISETUJUI UNTUK DIUJIKAN

Yogyakarta, 8 Agustus 2021

Pembimbing

(Dr.techn. Rohmatul Fajriyah, S.Si., M.Si.)

HALAMAN PENGESAHAN

TUGAS AKHIR

IMPLEMENTASI ALGORITMA SUPPORT VECTOR MACHINE PADA ANALISIS SENTIMEN DATA TWITTER

(Studi Kasus : Ulasan Tentang Indihome Pada *Platform* Twitter)

Nama Mahasiswa : Desy Rizki Ramadhanty

NIM : 17611036

TUGAS AKHIR INI TELAH DIUJIKAN

PADA TANGGAL : 23 AGUSTUS 2021

Nama Penguji	Tanda Tangan
1. Dina Tri Utari, S.Si., M.Sc.	.....
2. Sekti Kartika Dini,S.Si., M.Si.	.....
3. Dr.techn. Rohmatul Fajriyah, S.Si., M.Si.	.....

Mengetahui,

Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

(Prof. Riyanto, S.Pd., M.Si., Ph.D.)

KATA PENGANTAR

بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ

Assalamu alaikum wa rahmatullahi wa barakaatuh

Puji syukur kepada Allah SWT. yang telah melimpahkan rahmat, taufik, serta hidayah-Nya sehingga penulis dapat menyelesaikan Tugas Akhir yang berjudul "**Implementasi Algoritma *Support Vector Machine* Pada Analisis Sentimen Data Twitter (Studi Kasus: Ulasan Tentang Indihome pada *Platform* Twitter)**" ini. Tugas Akhir ini digunakan sebagai salah satu syarat untuk mendapatkan gelar Sarjana di Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia.

Penulis menyadari bahwa penulisan laporan ini banyak memperoleh, bimbingan, arahan, serta dorongan dari berbagai pihak. Maka dari itu, pada kesempatan ini penulis mengucapkan terima kasih kepada:

1. Bapak Prof. Riyanto, S.Pd., M.Si., Ph.D. selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia, Yogyakarta beserta seluruh jajarannya
2. Bapak Dr. Edy Widodo, S.Si., M.Si. selaku Ketua Jurusan Statistika beserta seluruh jajarannya.
3. Ibu Dr.techn. Rohmatul Fajriyah, S.Si., M.Si., selaku dosen pembimbing yang telah memberikan bimbingan serta arahan selama pengerjaan laporan Tugas Akhir ini hingga dapat terselesaikan.
4. Seluruh dosen pengajar prodi Statistika yang telah menginspirasi, mendidik, serta membekali ilmu kepada penulis.
5. Kedua orang tua tercinta, Bapak Muslih Asha dan Ibu Ida Kurniyati yang telah merawat, mendukung, serta menjaga dengan penuh rasa kasih sayang kepada penulis selama ini.
6. Kakak Berly Kurniawan serta Keluarga Besar, yang selalu memotivasi, menyemangati, dan mendoakan yang terbaik.
7. Teman-teman seperjuangan Galuh, Lala, Tamara, Eggy, Kirana, Shilma, Dini yang selalu ada dalam suka maupun duka selama perkuliahan di Universitas Islam Indonesia.

8. Teman-teman satu bimbingan Tugas Akhir, yang telah berbagi cerita, ilmu, dan pengalamannya.
9. Seluruh teman Statistika angkatan 2017 yang telah memberikan semangat serta dukungannya selama pengerjaan Tugas Akhir ini.
10. Semua pihak yang mungkin belum disebutkan, yang telah membantu dalam penyusunan Tugas Akhir ini.

Penulis menyadari bahwa Tugas Akhir ini tidak luput dari kesalahan dan masih banyak kekurangan. Oleh karena itu, penulis sangat mengharapkan berbagai kritik dan saran yang bersifat membangun sehingga dapat membawa penulis ke arah yang lebih baik. Semoga laporan ini dapat bermanfaat bagi penulis dan bagi semua yang membutuhkan.

Wassalamu alaikum wa rahmatullahi wa barakaatuh.

Yogyakarta, 5 Agustus 2021

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam Tugas Akhir ini tidak terdapat karya-karya yang sebelumnya pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan orang lain, kecuali yang diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Yogyakarta, 5 Agustus 2021

Penulis

INTISARI

IMPLEMENTASI ALGORITMA SUPPORT VECTOR MACHINE PADA ANALISIS SENTIMEN DATA TWITTER

(Studi Kasus : Ulasan Tentang Indihome Pada *Platform* Twitter)

Desy Rizki Ramadhanty
Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Islam Indonesia

Teknik pengambilan data yang bersumber dari internet terutama sebuah *website* dikenal sebagai teknik *web scraping*. Penelitian ini menggunakan teknik tersebut untuk mengumpulkan data ulasan mengenai Indihome yang bersumber dari *platform* Twitter. Indihome merupakan salah satu *provider* internet di Indonesia yang cukup banyak penggunanya. Proses pengklasifikasian pada analisis sentimen data Twitter ini menggunakan algoritma *Support Vector Machine* (SVM). Tujuan dilakukannya penelitian ini agar mengetahui gambaran umum sentimen yang diberikan oleh pelanggan terhadap Indihome, seberapa akurat hasil sentimen menggunakan algoritma SVM, serta informasi yang didapatkan dari hasil analisis sentimen. Pengkategorian kelas sentimen akan dibagi menjadi tiga kelas, yaitu sentimen positif, negatif, dan netral. Dataset diambil pada rentang bulan Maret dan April 2021. Klasifikasi dengan algoritma SVM mendapatkan tingkat akurasi tertinggi menggunakan metode *kernel Radial Basis Function* (RBF), yakni sebesar 88,47% pada bulan Maret, dan 98,06% pada bulan April. Pelanggan Indihome banyak memberikan ulasan negatif dibandingkan dengan ulasan positif dan netral. Penilaian negatif yang diberikan, diantaranya mengenai sinyal internet yang lambat, hilang, atau mati. Sedangkan, pada penilaian positif maupun netral berisi tentang respon dari pihak Indihome dalam menangani keluhan dan malasah yang dialami pelanggannya, serta berisi beberapa tips atau pertanyaan tentang Indihome.

Kata Kunci : Analisis Sentimen, Indihome, *Support Vector Machine* (SVM), Twitter, *Web Scraping*.

ABSTRACT

IMPLEMENTATION OF SUPPORT VECTOR MACHINE ALGORITHM IN TWITTER DATA SENTIMENT ANALYSIS
(Case Study : Review About Indihome on Twitter Platform)

Desy Rizki Ramadhanty
Department of Statistics, Faculty of Matematics and Natural Sciences
Universitas Islam Indonesia

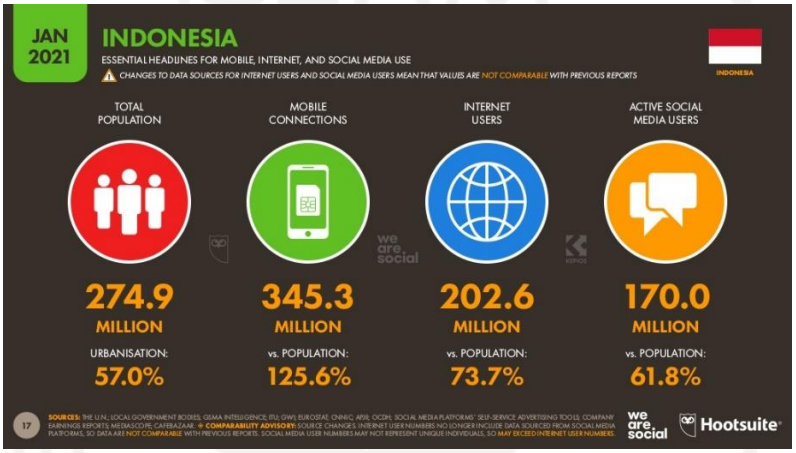
The process of retrieving data sourced from the internet, especially from a website, is known as web scraping. This research used web scraping to collect data of customer reviews about Indihome that sourced from Twitter platform. Indihome itself is one of the internet providers in Indonesia that has quite a lot of users. The classification process in this Twitter sentiment analysis used the Support Vector Machine (SVM) algorithm. The purpose of this research is to find out the general description of the sentiment that is given by customers to Indihome, how accurate the sentiment results are using the SVM algorithm, and information obtained from the results of sentiment analysis. The categorization of sentiment class will be divided into three classes, that are positive, negative, and neutral sentiment. The dataset was taken in the range of March and April 2021. Classification with the SVM algorithm gets the highest accuracy using the Radial Basis Function (RBF) kernel, which gets 88.47% in March, and 98.06% in April. In general, Indihome customers give a lot of negative reviews compared to positive and neutral reviews. The negative sentiments are about the internet signal being slow, lost connection, and not working. Meanwhile, the positive and neutral sentiment are related to the response from Indihome admin in dealing with complaints and problems experienced by its customers. There are some tips and questions about Indihome as well.

Keywords: *Sentiment Analysis, Indihome, Support Vector Machine (SVM), Twitter, Web Scraping.*

BAB I
PENDAHULUAN

1.1. Latar Belakang Masalah

Perkembangan teknologi saat ini sangatlah pesat, terutama berkaitan dengan teknologi informasi, yaitu internet. Internet menjadi salah satu teknologi yang sering digunakan untuk menerima ataupun memberi suatu informasi secara mudah dan dalam waktu yang cepat dari seluruh penjuru dunia.



Gambar 1.1 Jumlah Pengguna Internet Indonesia (Kemp, 2021)

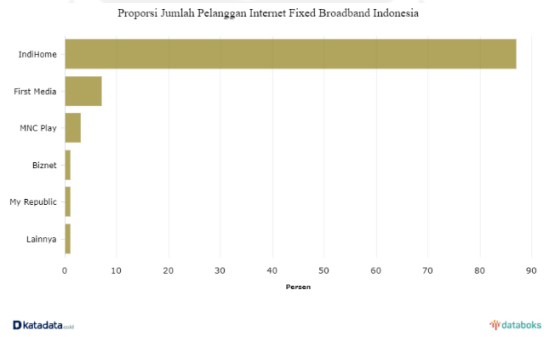
Berdasarkan **Gambar 1.1**, jumlah pengguna internet di Indonesia hingga Januari 2021 mencapai angka 202,6 juta pengguna dari total penduduk Indonesia sebanyak 274,9 juta penduduk. Tentunya, jumlah ini akan terus bertambah seiring berjalannya waktu, begitu pula dengan perusahaan-perusahaan yang akan berlomba-lomba untuk bergerak pada bidang usaha teknologi informasi yang menyediakan jasa layanan internet. Penggunaan internet dapat diakses menggunakan beberapa *provider* internet yang telah tersedia. Salah satu *provider* internet di Indonesia adalah Indihome (Indonesia *Digital Home*) yang mana merupakan salah satu produk layanan yang dimiliki oleh PT. Telekomunikasi Indonesia (Jihaderajad, 2017).

Indihome merupakan layanan digital yang menyediakan internet, telepon rumah dan TV Interaktif (Indihome TV) dengan beragam pilihan paket serta layanan tambahan yang bisa dipilih sesuai kebutuhan. Saat ini, jaringan Indihome sudah tersebar di seluruh wilayah Indonesia, dan terus berinovasi untuk memenuhi kebutuhan internet yang lebih baik bagi masyarakat. Mengawali tahun 2020,

Indihome semakin menguatkan posisinya sebagai *market leader fixed broadband* dengan target 8,3 juta pelanggan serta pertumbuhan *revenue* hingga 20% (IndiHome, 2020).

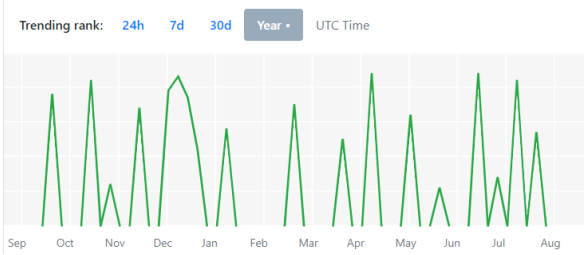
Setiap tahunnya, Indihome terus menerus berinovasi dengan menawarkan berbagai macam promo dan penawaran paket internet bagi masyarakat Indonesia. Namun demikian, hal tersebut terkadang masih belum membuat pelanggan merasa terpuaskan, baik dari segi pelayanan, sistem, maupun jaringan yang ditawarkan oleh Indihome. Maka dari itu, akan muncul tanggapan berupa ulasan ataupun keluhan dari berbagai pelanggan terhadap *provider* ini.

Munculnya berbagai tanggapan, baik itu positif, negatif, ataupun netral bisa terjadi karena beberapa faktor yang perlu diperbaiki atau ditingkatkan oleh Indihome. Hal ini diharapkan dapat memunculkan umpan balik dari pengguna itu sendiri.



Gambar 1.2 Proporsi Pelanggan Provider di Indonesia (Annur, 2021)

Pada **Gambar 1.2** menampilkan perbandingan persentase dari beberapa *provider* internet di Indonesia. Didapatkan data hingga pertengahan tahun 2021, bahwa Indihome menduduki puncak tertinggi sebagai *provider* internet *fixed broadband* dengan pelanggan terbanyak di Indonesia, dibandingkan dengan *provider* internet lainnya.



Gambar 1.3 Grafik Indihome *Trending* Pada Twitter Indonesia (GetDayTrends, 2021)

Begitu pula yang ditampilkan pada **Gambar 1.3** mengenai grafik seberapa sering kata Indihome masuk pada *trending* Twitter Indonesia dalam setahun terakhir. Dapat dikatakan bahwa Indihome cukup sering masuk *trending* yang di mana sering menjadi bahan cuitan oleh pengguna *platform* Twitter atau pelanggan Indihome itu sendiri, terutama pada rentang bulan Maret dan April 2021 yang akan menjadi fokus pada penelitian ini. Berdasarkan beberapa alasan inilah, penulis memandang penting untuk meneliti lebih lanjut mengenai tanggapan tersebut sehingga menghasilkan sentimen yang nantinya akan berguna untuk mengetahui tingkat kepuasan pelanggan terhadap Indihome.

Saat ini peran komunikasi dari mulut ke mulut (*word of mouth*) sangatlah penting bagi konsumen karena dapat memberikan referensi kepada calon konsumen lainnya untuk memutuskan proses pembelian mereka. Jika melihat ke belakang, komunikasi dari mulut ke mulut hanya bisa dengan bertemu secara langsung (tatap muka). Tetapi, seiring dengan perkembangan teknologi informasi yang begitu pesat, peran *word of mouth* bisa dilakukan melalui internet. Beberapa studi mengindikasikan bahwa pesan dari elektronik *word of mouth* merupakan sarana yang penting bagi konsumen, karena akan mendapatkan informasi mengenai kualitas produk dan jasa yang ditawarkan oleh suatu perusahaan (Wibowo, 2015).

Hal ini secara tidak langsung dapat menaikkan atau sebaliknya, yaitu menjatuhkan nama *brand* dari produk dan jasa tersebut. Opini-opini pelanggan yang tersalurkan di media sosial memiliki jumlah yang terlalu banyak dan tidak terorganisir untuk diproses secara manual. Oleh karena itu, diperlukan suatu pengklasifikasian untuk membagi opini tersebut, apakah masuk ke dalam kategori positif, negatif, atau netral dengan bantuan metode tertentu.

Twitter adalah salah satu media sosial yang banyak digunakan orang untuk berbagi informasi dalam periode *real-time*, melalui komentar singkat tentang pengalaman dan pemikiran yang akan mereka tuangkan (Maclean, et al., 2013). Jumlah pengguna aktif harian Twitter di Indonesia meningkat 34 persen menjadi 186 juta pengguna pada kuartal kedua 2020 dibandingkan dengan tahun sebelumnya (Josina, 2020). Twitter sering kali digunakan untuk menyampaikan opini atau saran terhadap suatu produk, *public figure*, layanan publik, dan lain sebagainya, tidak terkecuali dengan Indihome.

Banyaknya ulasan yang diberikan pelanggan terhadap Indihome melalui *platform* Twitter belum mampu untuk memberikan keterangan-keterangan terkait kepuasan atau problematika yang dialami oleh pelanggan itu sendiri. Oleh karena itu, diperlukan bantuan pengklasifikasian data ulasan tersebut dengan melakukan analisis sentimen pada *text mining* melalui metode yang mampu mengklasifikasikan secara akurat. Analisis sentimen adalah studi yang bertujuan untuk menganalisis opini, sentimen dan emosi yang terdapat pada dokumen atau data (Rizkia, et al., 2019). Adanya analisis sentimen ini dapat menjadi jembatan komunikasi terhadap pelanggan dan pelaku bisnis itu sendiri.

Metode pengklasifikasian untuk menganalisis sentimen mengenai Indihome yang akan digunakan pada penelitian ini adalah metode *Support Vector Machine* (SVM). Algoritma SVM sendiri memiliki prinsip bekerja secara linier, dan dikembangkan untuk dapat diterapkan pada masalah nonlinier. Hal ini dilakukan dengan menggunakan metode *kernel trick* untuk mencari *hyperplane* dengan cara mentransformasi *dataset* ke ruang vektor yang berdimensi lebih tinggi (*feature space*), kemudian proses klasifikasi dilakukan pada *feature space* tersebut (Muis & Affandes, 2015). Berdasarkan (Muis & Affandes, 2015), penelitian-penelitian terdahulu menyatakan bahwa, metode SVM secara umum memberikan solusi yang lebih baik dibandingkan metode lainnya.

Selain itu, menurut penelitian yang dilakukan (Nugroho, et al., 2003), SVM memiliki beberapa kelebihan diantaranya, yaitu mampu mengklasifikasikan suatu pola, di mana datanya tidak termasuk dalam tahapan pembelajaran data *training*, serta SVM dapat diimplementasikan secara lebih mudah. Oleh karena itu, peneliti memilih menggunakan metode SVM untuk mengklasifikasikan ulasan-ulasan tentang Indihome.

Proses pengklasifikasian ini akan membagi ulasan-ulasan pelanggan Indihome menjadi tiga kategori, yaitu apakah memiliki kecenderungan ulasan yang positif, negatif, atau netral, yang diambil dalam bahasa Indonesia melalui cuitan pelanggan pada *platform* Twitter serta tingkat keakuratannya. Harapan dibuatnya penelitian ini, agar penulis mampu mengklasifikasikan ulasan teks dengan baik sehingga informasi yang didapatkan dari kumpulan ulasan mengenai Indihome dapat diolah dengan baik serta dapat memberikan informasi yang bermanfaat bagi

pihak yang membutuhkan, yaitu Indihome itu sendiri, pelanggan dan calon pelanggan.

1.2. Rumusan Masalah

Berdasarkan latar belakang masalah yang telah diuraikan, maka rumusan masalah dalam penelitian ini adalah sebagai berikut :

1. Bagaimana gambaran umum pendapat pelanggan terhadap pelayanan Indihome pada *platform* Twitter?
2. Bagaimana hasil klasifikasi dan tingkat akurasi yang dihasilkan menggunakan metode *Support Vector Machine* (SVM)?
3. Apa saja informasi yang diperoleh dari ulasan pelanggan Indihome berdasarkan hasil klasifikasi yang didapatkan?

1.3. Batasan Masalah

Adapun batasan masalah yang digunakan penulis agar pembahasan dalam penelitian ini tidak menyimpang adalah :

1. Data yang digunakan dalam penelitian ini adalah ulasan pengguna *provider* Indihome di *platform* Twitter pada bulan Maret 2021 dan April 2021.
2. Ulasan yang akan diklasifikasi adalah ulasan yang berbahasa Indonesia.
3. Metode yang digunakan penulis untuk analisis sentimen adalah algoritma *Support Vector Machine* (SVM).
4. *Software* yang digunakan dalam penelitian ini adalah *Python 3.8* dan *RStudio 3.6*.

1.4. Jenis Penelitian dan Metode Analisis

Jenis penelitian yang dilakukan berupa pengklasifikasian sentimen pelanggan *provider* Indihome—yang memberikan ulasannya pada salah satu *platform* media sosial, yaitu Twitter—menjadi tiga kategori sentimen, yaitu netral, positif, dan negatif menggunakan metode *kernel* yang terdapat pada algoritma *Support Vector Machine* (SVM). Metode kernel tersebut diantaranya, yaitu *kernel linear*, *kernel polynomial*, *kernel radial basis function* (RBF), dan *kernel sigmoid*.

1.5. Tujuan Penelitian

Berdasarkan rumusan masalah, dalam penelitian ini ada beberapa tujuan yang ingin dicapai yaitu :

1. Untuk mengetahui gambaran umum pendapat pelanggan terhadap pelayanan Indihome pada *platform* Twitter.
2. Untuk mengetahui pembagian klasifikasi serta tingkat akurasi pendapat pelanggan terhadap pelayanan IndiHome menjadi kelas positif, negatif, dan netral menggunakan metode *Support Vector Machine* (SVM).
3. Untuk mendapatkan informasi penting dari ulasan pengguna Indihome berdasarkan hasil klasifikasi.

1.6. Manfaat Penelitian

Adapun manfaat dari penelitian ini adalah sebagai berikut :

1. Mengetahui gambaran umum dari tanggapan masyarakat tentang *provider* Indihome.
2. Hasil klasifikasi ulasan dari pelanggan dapat membantu serta mempermudah pihak perusahaan dalam mengetahui persepsi pengguna yang dibagi menjadi opini positif, opini negatif, dan opini netral. Sehingga, hal ini dapat dilakukan evaluasi serta dijadikan acuan dalam meningkatkan kualitas serta memperbaiki kekurangan dari produk tersebut.

BAB II
TINJAUAN PUSTAKA

Penelitian yang telah dilakukan sebelumnya sangatlah penting bagi penulis, karena dapat mengetahui hubungan antara penelitian sebelumnya dengan penelitian yang akan dilakukan agar tidak terjadi duplikasi. Berikut ini adalah beberapa penelitian sebelumnya yang dilakukan terhadap data dan metode yang digunakan sebelumnya. Beberapa jurnal dan studi yang dikutip oleh penulis adalah sebagai berikut.

Penelitian tentang *provider* internet Indihome sebelumnya pernah dilakukan oleh Rizkia, dkk (2019) yang membahas opini pelanggan terhadap perusahaan jasa tersebut guna menjadi tolak ukur untuk memperbaiki layanan yang diberikan. Pengklasifikasian dilakukan menggunakan metode *Decision Tree* dengan 3 skenario tokenisasi, yaitu unigram, bigram, dan trigram, serta pembobotan dengan algoritma *Term Frequency-Inverse Document Frequency* (TF-IDF). Pengukuran dilakukan dengan membandingkan hasil pelabelan yang dibagi menjadi 3 kelas menggunakan metode *confussion matrix*.

Pada penelitian yang dilakukan Haranto dan Sari (2019) menggunakan teknik *text mining* dengan menerapkan algoritma SVM, yang digunakan untuk menganalisis sentimen pengguna Twitter pada layanan Telkom dan Biznet. Penelitian ini menggunakan metode 10-*Fold Cross Validation* untuk membagi *dataset* menjadi data latih dan data uji, serta *confusion matrix* yang digunakan untuk menemukan nilai akurasi, presisi, *recall*, dan F1-*score*.

Penelitian yang dilakukan oleh Mailo dan Lazuardi (2019), menerapkan metode *data mining*, yaitu a *Naive Bayes classifier* untuk mengidentifikasi analisis sentimen terkait obesitas di Indonesia melalui *tweet* berbahasa Indonesia yang mengandung kata-kata kunci, seperti obesitas, gemuk, kegemukan, gendut, atau kegendutan. Hasil akurasi yang didapatkan pada suatu metode analisis sentimen sangat dipengaruhi dalam beberapa hal, yakni jumlah data latih dan data uji, jumlah *dataset* yang digunakan, serta komposisi jumlah data positif dan negatif. Hasil akurasi yang didapatkan pada penelitian ini berhasil memprediksi kategori sentimen dengan baik.

Penelitian menggunakan metode SVM lainnya, juga pernah dilakukan oleh Prayoginingsih dan Kusumawardani (2018) untuk mengklasifikasikan data Twitter pelanggan berdasarkan kategori myTelkomsel (layanan *web* dari Telkomsel). Dataset didapatkan dengan teknik *crawling* menggunakan *Streaming* API. Dilakukan percobaan sebanyak 6 kali, di mana masing-masing percobaan memiliki performa SVM yang baik dalam mengklasifikasikan *tweet* dengan atau tanpa *stemming*. *Kernel linear* menghasilkan akurasi terbaik untuk klasifikasi SVM menggunakan *stemming* dan *stopword special* pada percobaan ke-3 dengan nilai *cost* 1, sehingga menghasilkan akurasi sebesar 98,79%.

Deviyanto dan Wahyudi (2018) pernah melakukan penelitian pada *dataset* yang didapatkan dari *platform* Twitter, yang pada tahun 2017 menjadi *trending topic* di Twitter, yaitu isu politik tentang pemilihan gubernur Jakarta (Pilkada DKI). *Tweet* yang diambil berbahasa Indonesia dan akan dilakukan klasifikasi menggunakan algoritma KNN (K-*Nearest Neighbor*) untuk menentukan apakah *tweet* tersebut bersentimen positif atau negatif, dengan cara mengelompokkan data ke dalam suatu kelas sebanyak "k", serta melihat nilai jarak terdekat nya menggunakan data latih. Setelah melakukan percobaan membagi kelas hingga jumlah k=15, didapatkan hasil akurasi tertinggi sebesar 67,2% dan presisi tertinggi sebesar 56,94% pada k=5. Sedangkan, nilai *recall* tertinggi sebesar 78,24% ketika jumlah k=15.

Buntoro (2016) melakukan penelitian sentimen berbahasa Indonesia terhadap tagar *Hatespeech* (*#HateSpeech*) di *platform* Twitter yang diklasifikasi menjadi dua kategori, yaitu *HateSpeech* dan *GoodSpeech* menggunakan metode *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM) dengan *preprocessing* data menggunakan tokenisasi, *cleansing*, dan *filtering*. Terdapat kekurangan saat pengambilan *dataset* di Twitter, dikarenakan terlihat masih banyak *tweet* opini yang tidak termasuk dalam kategori *Hatespeech*, namun diberi tagar *Hatespeech*. Hasil yang diberikan menggunakan metode SVM dengan tokenisasi unigram, *stopword list* Bahasa Indonesia dan *emoticons*, memiliki nilai rata-rata akurasi mencapai 66,6%, nilai presisi 67,1%, nilai *recall* 66,7%, nilai TP *rate* 66,7%, serta nilai TN *rate* 75,8%. Meskipun model yang dihasilkan memiliki akurasi yang cukup baik,

akan tetapi masih terdapat kesalahan pada proses klasifikasi untuk kelas positif yang ditunjukkan dengan nilai TN *Rate* lebih besar dari nilai TP *Rate*.

Penelitian yang dilakukan Muis dan Affandes (2015) menerapkan metode *Support Vector Machine* (SVM) menggunakan *kernel Radial Basis Function* (RBF) dengan parameter *c* dan *γ* untuk mengklasifikasikan *tweet* yang dibagi ke dalam dua kelas, yaitu kelas iklan yang memuat kata-kata promosi, jual, harga, pembelian, dan memungkinkan orang untuk fokus melihat dan mendengarkan iklan, sehingga ingin membeli produk yang diiklankan. Serta, kelas tidak iklan yang tidak mengandung semua kata yang terdapat pada unsur iklan. Pada penelitian ini dilakukan pengujian pada data yang sudah dilakukan pemilihan fitur, serta data yang belum dilakukan pemilihan fitur dengan cara merangking nilai kemunculan kata. Pada data yang sudah dilakukan pemilihan fitur, nilai akurasi terbaik terdapat pada titik c=20, γ=6 dengan nilai akurasi 99,12%. Sedangkan, data yang belum dilakukan pemilihan fitur memiliki nilai akurasi sebesar 97,54% pada titik c=2.5, γ=0.3 dan c=3, γ=0.3. Pada **Tabel 2.1** merupakan tabel perbandingan dengan penelitian sebelumnya yang berkaitan tentang *text mining* khususnya metode *support vector machine*.

Tabel 2.1 Tabel Penelitian Sebelumnya

Tahun	Nama	Judul	Hasil Penelitian
2019	(Rizkia, et al.)	Analisis Sentimen Kepuasan Pelanggan Terhadap Internet Provider Indihome di Twitter Menggunakan Metode Decision Tree dan Pembobotan TF-IDF	Mengklasifikasikan opini pelanggan di Twitter terhadap layanan Indihome ke dalam kelas positif, negatif dan netral. Hasil eksperimen menunjukkan nilai akurasi maksimal yang didapatkan sebesar 80,1 % dengan menggunakan bigram dan pembobotan TF-IDF.
2019	(Haranto & Sari)	Implementasi *Support Vector Machine* untuk Analisis Sentimen Pengguna Twitter Terhadap Pelayanan Telkom dan Biznet	Pada penelitian ini digunakan teknik *text mining* dengan algoritma *Support Vector Machine*, untuk menganalisis sentimen pengguna Twitter terhadap pelayanan Telkom dan Biznet. Hasil pengujian

Tahun	Nama	Judul	Hasil Penelitian
			model dengan algoritma SVM pada masing-masing Telkom dan Biznet mendapatkan nilai akurasi sebesar 79,6% dan 83,2%.
2019	(Mailo & Lazuardi)	Analisis Sentimen Data Twitter Menggunakan Metode *Text Mining* Tentang Masalah Obesitas di Indonesia	Penelitian ini menggunakan metode *Naïve Bayes Classifier* untuk mengklasifikasikan sentimen pengguna Twitter menjadi tiga kelas, yaitu positif, netral, dan negatif. Didapatkan hasil akurasi yang masuk ke dalam kategori "*Excellent Classification*" sebesar 94%.
2018	(Prayoginingsih & Kusumawardani)	Klasifikasi Data Twitter Pelanggan Berdasarkan Kategori myTelkomsel Menggunakan Metode *Support Vector Machine* (SVM) Studi Kasus: Telekomunikasi Selular	Mengklasifikasikan data *tweet* kasus aduan pelanggan untuk layanan telekomunikasi pada aplikasi myTelkomsel. Hasil penelitian menggunakan kernel RBF yang dioptimasi dengan metode *grid search*, mendapatkan nilai akurasi dan f-*measure* sebesar 84,84% dan 84,88%.
2018	(Deviyanto & Wahyudi)	Penerapan Analisis Sentimen pada Pengguna Twitter Menggunakan Metode K-Nearest Neighbor	Menganalisis sentimen positif dan negatif pengguna Twitter tentang topik Pilkada DKI 2017 menggunakan algoritma KNN dengan pembobotan kata TF-IDF dan fungsi *Cosine Similarity*. Didapatkan hasil pengujian nilai akurasi yaitu 67,2% ketika k=5.
2016	(Buntoro)	Analisis Sentimen Hatespeech pada Twitter dengan Metode *Naïve Bayes Classifier* dan *Support Vector Machine*	Menganalisis tagar *#HateSpeech* menggunakan metode klasifikasi *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine*

Tahun	Nama	Judul	Hasil Penelitian
			(SVM). Pada hasil eksperimen, didapatkan metode klasifikasi SVM menghasilkan nilai akurasi lebih tinggi dibandingkan dengan menggunakan algoritma NBC.
2015	(Muis & Affandes)	Penerapan Metode *Support Vector Machine* (SVM) Menggunakan *Kernel Radial Basis Function* (RBF) pada Klasifikasi Tweet	Pada penelitian ini dilakukan pengklasifikasian data *tweet* yang dibagi menjadi dua kelas, berupa kelas iklan, dan kelas tidak iklan menggunakan metode *support vector machine* (SVM) yang bersifat non linear, yaitu menggunakan *kernel* RBF (*Radial Basis Function*). Didapatkan nilai akurasi tertinggi 97,54% untuk data yang belum dilakukan pemilihan *feature*, sedangkan untuk data yang sudah dilakukan pemilihan terhadap *feature* mencapai nilai akurasi tertinggi 99,12%.

Berdasarkan pemaparan berbagai ide penelitian pada **Tabel 2.1**, penulis akan melakukan penelitian untuk mengimplementasikan proses *scraping* dalam mengumpulkan data ulasan pengguna *provider* Indihome melalui *platform* Twitter. Selanjutnya akan dilakukan analisis sentimen yang diklasifikasikan menjadi tiga kelompok, yaitu positif, negatif, dan netral menggunakan metode-metode *kernel* pada algoritma *Support Vector Machine* (SVM). Selanjutnya, dari hasil pengelompokan tersebut akan dilakukan proses visualisasi menggunakan *Word Cloud* yang dilanjutkan dengan asosiasi kata.

Jika dibandingkan dengan penelitian-penelitian sebelumnya, pada penelitian yang penulis lakukan, kelas sentimen akan dibagi menjadi tiga kelompok, yaitu sentimen positif, negatif, dan netral. Penelitian ini lebih berfokus pada percobaan

setiap metode *kernel* pada algoritma SVM. Hal ini dilakukan agar dapat mengetahui metode *kernel* terbaik pada algoritma SVM yang dapat mengklasifikasikan sentimen dengan lebih akurat. Pada penelitian ini juga melakukan uji lanjutan untuk merepresentasikan hasil klasifikasi dalam bentuk *Word Cloud* dan juga asosiasi kata agar lebih menarik.

BAB III
LANDASAN TEORI

3.1. *Online Review*

Review merupakan penilaian terhadap suatu karya, seperti film, buku dan karya lainnya yang memiliki tujuan untuk mengetahui kualitas, kelebihan dan kekurangan yang ada pada karya tersebut, serta untuk melakukan kritik terhadap suatu peristiwa atau karya seni bagi khalayak (Pristiyanti, et al., 2018). *Review* biasanya tertuang dalam bentuk teks yang membagi informasi dalam bentuk evaluasi atau opini pelanggan terhadap produk yang digunakan. Maka dari itu, *online review* dapat diartikan sebagai ulasan yang dapat berupa kritik atau saran terhadap suatu hal yang tertuang melalui media yang menggunakan jaringan internet, seperti media sosial.

3.2. Twitter

Twitter didirikan oleh 3 orang, yaitu: Jack Dorsey, Biz Stone, dan Evan Williams pada bulan Maret tahun 2006, dan diluncurkan pada bulan Juli di tahun yang sama. Twitter adalah jejaring sosial dan *micro-blogging*, yang memfasilitasi seseorang, sebagai pengguna, dapat memberikan *update* (perbaruan) informasi tentang diri, bisnis, dan lain sebagainya. Teknologi yang melatar belakangi Twitter berupa teknologi *web* berbasis *Internet Relay Chat* (IRC). Pemrograman *web interface* pada Twitter memakai pemrograman *Ruby on Rails Framework*. Sejak 2009, Twitter memiliki sebuah layanan aplikasi pemrograman antarmuka (*Application Programming Interface*/API), yang memungkinkan layanan *web* dan aplikasi lainnya untuk saling berintegrasi dengan Twitter (Waloeyo, 2010).

Twitter merupakan salah satu media sosial di mana dapat mengetahui apa yang sedang terjadi di dunia serta yang sedang hangat dibicarakan saat ini, melalui cuitan-cuitan yang disalurkan di *platform* tersebut. Hal ini dapat membantu dalam melakukan analisis sentimen terhadap ulasan-ulasan mengenai suatu hiburan, berita, politik, dan lain sebagainya yang berisi *review* dari para pengguna.

3.3. Statistik Deskriptif

Statistik deskriptif merupakan cara pengumpulan dan penyajian data yang menguraikan atau memberikan keterangan-keterangan mengenai suatu data. Berdasarkan ruang lingkup bahasannya, statistik deskriptif terdiri atas (Nasution, 2017):

1. Distribusi frekuensi, seperti:
    a. Grafik distribusi (*histogram*, ogif, dan poligon frekuensi);
    b. Ukuran nilai pusat (*mean*, *median*, modus, kuartil, dan sebagainya);
    c. Ukuran dispersi (jangkauan, variasi, simpangan rata-rata, simpangan baku, dan sebagainya);
    d. Kemencengan dan keruncingan kurva.
2. Angka indeks.
3. *Time series*/deret waktu.
4. Korelasi dan regresi sederhana.

Data yang telah dikumpulkan, nantinya akan dilakukan pengolahan data untuk mendapatkan ringkasan (seperti: jumlah, rata-rata, persentase, dan lainnya) dari data tersebut. Kemudian, untuk mempermudah dalam pemahaman hasil olah data tersebut, perlu dilakukan penyajian data dalam bentuk tertentu, yaitu tabel data (seperti tabel frekuensi, tabel klasifikasi, tabel kontigensi, dan tabel korelasi), dan grafik atau diagram data (seperti piktogram, grafik batang, grafik garis, grafik lingkaran, dan kartogram) (Nasution, 2017).

3.4. *Web Scraping*

Web scraping merupakan proses pengambilan dokumen semi-terstruktur yang bersumber dari internet, biasanya berupa halaman-halaman *web* dalam bahasa *markup* seperti HTML atau XHTML, serta menganalisis dokumen tersebut guna mengambil data tertentu dari halaman *web* untuk digunakan dalam konteks yang lain (Turland, 2010). Singkatnya, *web scraping* digunakan untuk mengambil atau mendapatkan data dari suatu *website* ke dalam format yang lebih mudah dipahami.

Pada penelitian yang dilakukan oleh (Josi, et al., 2014) dijelaskan bahwa terdapat beberapa langkah dalam melakukan *web scraping*, yaitu sebagai berikut:

1. *Create Scraping Template*: Mempelajari dokumen HTML dari *website* yang memuat informasi berupa *tag* HTML yang akan diambil atau dikumpulkan menjadi suatu data.
2. *Explore Site Navigation*: Mempelajari teknik navigasi pada *website* yang akan diambil informasinya untuk ditirukan pada aplikasi *web scraper*.
3. *Automate Navigation and Extraction*: Berdasarkan informasi yang didapat pada langkah 1 dan 2, 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*.

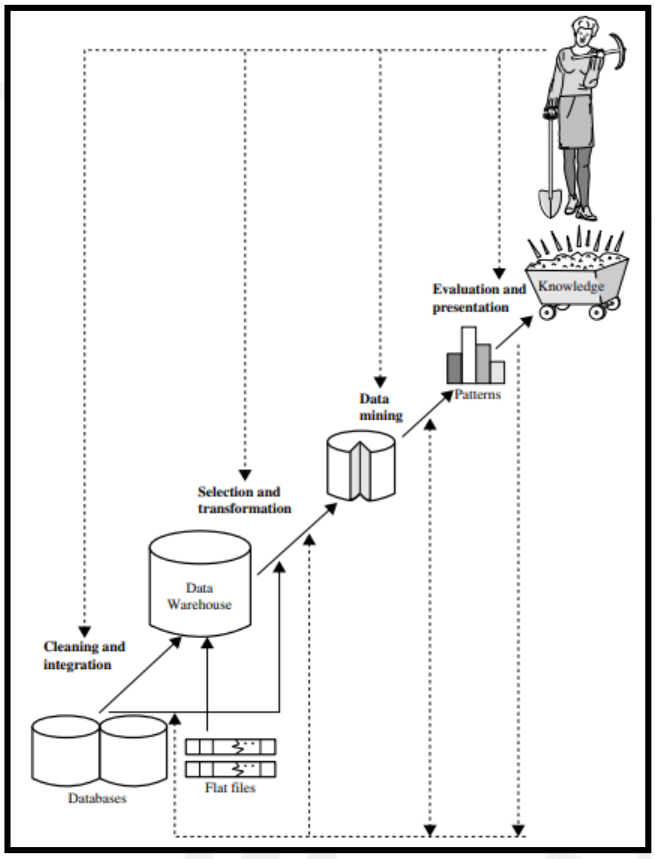
3.5. *Data Mining*

Data mining adalah proses menggunakan teknik pengenalan pola (seperti statistik dan matematika) untuk menemukan hubungan, pola, dan tren baru yang bermakna dengan menyaring sejumlah besar data yang disimpan dalam media penyimpanan (Larose, 2005).

Menurut (Larose, 2005) *data mining* memiliki beberapa tugas yang dapat dilakukan, yaitu:

1. *Description* (Deskripsi): Model *data mining* harus setransparan mungkin yang artinya, hasil dari model tersebut harus mendeskripsikan pola yang jelas, sesuai dengan interpretasi dan penjelasan yang intuitif.
2. *Estimation* (Estimasi): Estimasi hampir sama dengan klasifikasi, hanya saja variabel targetnya lebih ke arah numerik daripada kategorik. Model dibangun menggunakan *record* lengkap yang menyajikan nilai variabel target maupun prediksi.
3. *Prediction* (Prediksi): Prediksi hampir sama dengan klasifikasi dan estimasi, hanya saja hasil untuk prediksi berada di masa yang akan datang.
4. *Classification* (Klasifikasi): Dalam klasifikasi, terdapat variabel target yang bersifat kategori. Contohnya, pembagian kelompok untuk kategori pendapatan, yang dapat dibagi menjadi tiga kelas atau kategori, yaitu berpenghasilan tinggi, berpenghasilan menengah, dan berpenghasilan rendah.

5. *Clustering* (Pengelompokan): Kluster mengacu pada pengelompokan *record*, observasi, atau kasus ke dalam kelas objek yang memiliki kemiripan.
6. *Association* (Asosiasi): Tugas asosiasi dalam *data mining* adalah menemukan atribut mana yang akan muncul dalam suatu waktu.



Gambar 3.1 Tahapan Proses *Knowledge Discovery from Data* (KDD) (Han, et al., 2012)

Pada **Gambar 3.1** menampilkan tahapan proses dari *data mining* atau banyak orang ketahui sebagai *Knowledge Discovery from Data* (KDD), namun ada juga yang berpendapat bahwa *data mining* merupakan salah satu langkah penting dalam proses KDD. Berikut ini adalah langkah-langkah tahapan dari proses KDD secara rinci (Han, et al., 2012):

1. *Data cleaning* merupakan proses untuk menghilangkan data *noise* serta data yang tidak konsisten.
2. *Data integration* merupakan proses penggabungan beberapa data dari berbagai sumber ke dalam suatu penyimpanan yang saling terhubung.

3. *Data selection* merupakan proses untuk menyeleksi atau memilih data yang relevan terhadap analisis dari sebuah *database*.
4. *Data transformation* merupakan proses mentransformasikan atau mengubah data menjadi suatu bentuk yang tepat untuk proses *mining*.
5. *Data mining* merupakan proses penting di mana menggunakan metode-metode serta algoritma tertentu yang diterapkan untuk menghasilkan suatu pola data.
6. *Pattern evaluation* merupakan proses untuk mengidentifikasi pola data yang mewakili pengetahuan berdasarkan tindakan yang menarik.

3.6. *Machine Learning*

Machine learning (ML) adalah bidang ilmu yang berkembang mempelajari pola dan teori komputasi dalam *Artificial Intelligence* (AI) atau kecerdasan buatan. ML merupakan pembelajaran dan pembangunan algoritma yang dapat dipelajari serta dapat membuat prediksi pada suatu kumpulan data. Prosedur ini dioperasikan oleh model dari suatu *input* untuk membuat prediksi atau pilihan berdasarkan data daripada mengikuti instruksi program statis. ML memiliki dua jenis tugas, yaitu (Simon, et al., 2015):

1. *Supervised machine learning*: Pembelajaran yang terbentuk dari kumpulan algoritma yang telah dilatih atau ditentukan sebelumnya, kemudian dengan kemampuannya dalam mempelajari pola data latih tersebut dapat menggambarkan kesimpulan yang akurat saat diberi data baru.
2. *Unsupervised machine learning*: Pembelajaran untuk mendeteksi pola berdasarkan kemiripan yang tidak memerlukan data latih dalam mencari model yang tepat.

3.7. *Text Mining*

Text mining merupakan proses penambangan atau pengambilan data berupa teks yang bersumber dari suatu dokumen yang formatnya masih berantakan. Biasanya, *text mining* membutuhkan tahapan yang terstruktur (misalnya, penguraian kalimat, penambahan beberapa karakter dan penghapusan hal lainnya). Tugas dari *text mining* sendiri meliputi kategorisasi teks, pengelompokan teks, analisis sentimen, peringkasan dokumen, dan sejenisnya (Han, et al., 2012).

Proses yang dilakukan oleh *text mining* untuk mengubah data yang awalnya tidak hanya berupa teks yang tidak terstruktur menjadi lebih terstruktur, dilakukan dalam beberapa tahapan, yaitu (Putri & Setiadi, 2014):

1. *Tokenization* (Pengolahan teks): Proses ini digunakan untuk memecah suatu kalimat menjadi kata per kata.
2. *Case folding* dan menghilangkan tanda baca: Proses untuk merubahan huruf kapital menjadi huruf kecil, dan juga untuk menghilangkan tanda baca.
3. *Stemming* (Perubahan teks): Proses merubah kata yang memiliki imbuhan menjadi kata dasar.
4. *Filtering* (Pemilahan teks): Proses untuk melakukan perhitungan dan pengelompokkan kata per kata serta proses untuk membuang kata atau tanda yang tidak bermakna.

3.8. Analisis Sentimen

Analisis sentimen atau disebut juga *opinion mining* merupakan salah satu bidang ilmu dari *text mining*. Terdapat perbedaan definisi antara sentimen dan opini, di mana sentimen diartikan sebagai sikap, pemikiran, atau penilaian yang melibatkan perasaan. Sedangkan, opini berupa pandangan, anggapan, atau penilaian yang terbentuk oleh pikiran tentang suatu hal tertentu. Namun kedua hal tersebut masih dalam satu kesatuan.

Tujuan dari analisis sentimen sendiri adalah untuk menentukan *tools* yang dapat mengutip informasi subjektif dari teks dalam *natural language* (seperti opini, dan sentimen), sehingga dapat menciptakan pengetahuan yang terstruktur dan dapat ditindaklanjuti untuk pengambilan keputusan. Analisis sentimen dapat diklasifikasikan ke dalam dua bagian, yakni kalimat objektif dan subjektif. Jika kalimat yang diberikan diklasifikasikan sebagai kalimat objektif, maka tidak ada hal yang perlu dilakukan. Sedangkan, jika kalimat tersebut diklasifikasikan sebagai kalimat subjektif, maka polaritasnya perlu diestimasi. Klasifikasi polaritas (*polarity classification*) digunakan untuk membedakan antara kalimat yang mengekspresikan polaritas positif, negatif, atau netral (Pozzi, et al., 2017).

Teknik yang digunakan dalam analisis sentimen dapat dibagi menjadi pendekatan *lexicon based*, *machine learning*, dan *hybrid* (Medhat, et al., 2014).

Berikut adalah sedikit penjelasan terkait masing-masing teknik yang ada pada analisis sentimen:

1. *Lexicon-based approaches*

   Pendekatan berbasis *lexicon* bergantung pada *sentiment lexicon*, yakni kumpulan kata-kata sentimen yang telah diketahui dan dikumpulkan sebelumnya. Pendekatan ini dibagi menjadi *dictionary-based approach* (pendekatan berbasis kamus) dan *corpus-based approach* (pendekatan berbasis korpus) yang menggunakan metode statistik atau semantik untuk menemukan polaritas sentimen (Medhat, et al., 2014). Penggunaan pendekatan *lexicon* memiliki dua keuntungan utama, yaitu tidak perlu memberi *tag* (positif, netral, negatif) pada teks untuk pelatihan, dan dapat mencegah *overfitting* serta memungkinkan penggunaannya pada beberapa *dataset*. Terdapat beberapa sentimen *lexicon* yang cukup terkenal, seperti SWN (*SentiWordNet*), *Multi-Perspective Question Answering* (MPQA), *General Inquirer* (GI), dan *Opinion lexicon* (OL) (Han, et al., 2018). Beberapa penelitian menyatakan kata-kata positif dinotasikan (+1) atau skor positif, dan kata-kata negatif dinotasikan (-1) atau skor negatif. Jika hasil perhitungan skor sentimen mendapatkan skor 0, maka dapat diklasifikasikan sebagai kelas netral. **Persamaan 3.1** merupakan perhitungan skor sentimen *lexicon* yang biasa digunakan:

   $$Skor = (\sum kata\ positif) - (\sum kata\ negatif) \quad (3.1)$$

2. *Machine learning approaches*

   Pendekatan ini menerapkan algoritma *machine learning* untuk menganalisis sentimen dari suatu teks. Kekuatan utama dari pendekatan ini terletak pada kemampuan untuk menganalisis teks dari berbagai *domain* dan menghasilkan model klasifikasi yang disesuaikan dengan masalah yang dihadapi. Selain itu, pendekatan ini mampu menggabungkan sumber informasi tambahan dalam proses keputusannya. Beberapa penelitian mengusulkan bahwa pendekatan *machine learning* dapat dibagi menjadi dua kelompok, kelompok pertama yaitu algoritma *features-focused* yang mengusulkan fitur baru menggunakan metode pengklasifikasian, seperti *Support Vector Machine* (SVM), Naïve Bayes,

dan lainnya. Kemudian, kelompok kedua yaitu algoritma *model-focused* yang mengusulkan model klasifikasi baru, seperti model probabilitas yang dikombinasikan dengan regresi, menggunakan asosiasi *lexical* dan analisis distribusi teks untuk menyimpulkan sentimen, atau lainnya (Katz, et al., 2015).

3. *Hybrid approaches*

   Pendekatan *hybrid* merupakan pendekatan yang menggabungan antara pendekatan *lexicon* dan *machine learning* (Medhat, et al., 2014). Menurut penelitian yang dilakukan oleh (D'Andrea, et al., 2015), kombinasi antara kedua pendekatan tersebut dapat berpotensi meningkatkan kinerja klasifikasi sentimen. Keuntungan utama dari pendekatan *hybrid*, diantaranya simbiosis antara *lexicon*/*learning*, mampu mendeteksi dan mengukur sentimen pada tingkat konsep, dan sensitivitas yang lebih rendah terhadap perubahan dalam *topic domain*. Namun, terdapat kelemahan dari pendekatan ini, yaitu ulasan-ulasan yang memiliki banyak *noise* (kata-kata yang tidak relevan untuk subjek ulasan) sering diberi skor netral, karena metode ini gagal mendeteksi sentimen apa pun.

3.9. Klasifikasi

Menurut (Han, et al., 2012) klasifikasi adalah proses menemukan sebuah model (fungsi) yang mendeskripsikan dan membedakan kelas atau konsep data. Model tersebut didapatkan berdasarkan analisis dari sekumpulan data pelatihan (objek data yang label kelasnya diketahui). Model ini digunakan untuk memprediksi label kelas dari objek yang tidak diketahui label kelasnya. Klasifikasi memiliki dua tahapan, yaitu:

1. *Learning step* (Langkah pembelajaran): Langkah ini dibangun dari algoritma untuk pengklasifikasian dengan cara menganalisis atau "belajar dari" data pelatihan yang dibuat dari *database tuples* dan label kelas yang terkait.
2. *Classification step* (Langkah pengklasifikasian): Pada langkah ini, model yang telah didapatkan dari langkah sebelumnya, digunakan untuk memprediksi atau mengestimasi label kelas pada data yang diberikan (data pengujian).

3.9.1 Ukuran Evaluasi Model Klasifikasi

Hasil dari klasifikasi untuk data uji (*test data*) yang dilakukan menggunakan model klasifikasi dari data latih (*training data*), tentunya perlu diketahui seberapa baik atau akurat hasil pengklasifikasian label kelas dari suatu *tuple*.

Tabel 3.1 *Confusion Matrix* (Alrajak, et al., 2020)

		Prediksi		
	Kelas	**Positif**	**Netral**	**Negatif**
Aktual	Positif	*True Positive (TP)*	*False Neutral*	*False Negative (FN)*
	Netral	*False Positive (FP)*	*True Neutral*	*False Negative (FN)*
	Negatif	*False Positive (FP)*	*False Neutral*	*True Negative (TN)*

Pada **Tabel 3.1** merupakan *confusion matrix* (tabel kontingensi) 3 x 3 untuk klasifikasi *multi-class*, berguna untuk menganalisis seberapa baik hasil pengklasifikasian dapat mengenali *tuple* dari berbagai kelas. Berikut adalah penjelasan istilah-istilah dalam menghitung ukuran evaluasi (Han, et al., 2012):

a. *True positive* (TP): Hal ini berarti bahwa *tuple* positif dikategorikan benar oleh hasil pengklasifikasian.
b. *True negative* (TN): Hal ini berarti bahwa *tuple* negatif dikategorikan benar oleh hasil pengklasifikasian.
c. *True neutral*: Hal ini berarti bahwa *tuple* netral dikategorikan benar oleh hasil pengklasifikasian.
d. *False positive* (FP): Hal ini berarti *tuple* negatif dan *tuple* netral salah dikategorikan sebagai *tuple* positif.
e. *False negative* (FN): Hal ini berarti *tuple* positif dan *tuple* netral salah dikategorikan sebagai *tuple* negatif.
f. *False neutral*: Hal ini berarti *tuple* positif dan *tuple* negatif salah dikategorikan sebagai *tuple* netral.

Terdapat beberapa rumus yang dapat digunakan untuk mengukur performa hasil klasifikasi, yaitu sebagai berikut:

a. *Accuracy*

Akurasi dari hasil klasifikasi pada data uji yang diberikan merupakan persentase *tuple* data uji yang diprediksi dengan benar. Rumus perhitungan akurasi dapat dilihat pada **persamaan 3.2**.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (3.2)$$

b. *Precision*

Presisi dapat dianggap sebagai ukuran ketepatan (yaitu, berapa persentase *tuple* yang sebenarnya dikategorikan positif). Lihat rumus pada **persamaan 3.3**.

$$Precision = \frac{TP}{TP + FP} \quad (3.3)$$

c. *Recall*

Recall merupakan ukuran kelengkapan (yaitu, berapa persentase *tuple* yang diprediksi sebagai label positif). Lihat rumus pada **persamaan 3.4**.

$$Recall = \frac{TP}{TP + FN} \quad (3.4)$$

3.9.2 *Imbalanced Classification*

Menurut (Brownlee, 2019) *imbalanced classification* merupakan permasalahan pemodelan prediksi klasifikasi di mana distribusi untuk setiap label kelasnya tidak seimbang, atau mendekati sama, dan bahkan bias atau condong sebelah. Ketidakseimbangan ini terjadi ketika satu atau lebih kelas memiliki proporsi yang sangat rendah pada data *training* dibandingkan dengan kelas lainnya. Terdapat beberapa nama umum lainnya yang dapat digunakan untuk menggambarkan jenis masalah klasifikasi ini, seperti *Rare event prediction* (Prediksi kejadian yang langka), *Extreme event prediction* (Prediksi kejadian yang ekstrim), dan *Severe class imbalance* (Ketidakseimbangan kelas yang parah).

Permasalahan dari ketidakseimbangan ditentukan oleh distribusi kelas pada kumpulan data *training* tertentu. Biasanya, untuk menggambarkan ketidakseimbangan dari suatu kelas dalam *dataset*, ditentukan berdasarkan rasio atau persentase dari kumpulan data *training*. Misalnya, permasalahan klasifikasi kelas biner memiliki ketidakseimbangan 1:100, yang berarti bahwa untuk setiap satu contoh dalam satu kelas, terdapat 100 contoh di kelas lainnya. Contoh lainnya, pada permasalahan klasifikasi *multiclass* (banyak kelas) yang mungkin memiliki

ketidakseimbangan 80% pada kelas pertama, 18% pada kelas kedua, dan 2% pada kelas ketiga.

Ketidakseimbangan dari distribusi kelas akan bervariasi pada beberapa masalah. Permasalahan klasifikasi mungkin akan sedikit condong, seperti jika adanya sedikit ketidakseimbangan. Pada permasalahan lainnya mungkin memiliki ketidakseimbangan yang parah. Berikut sedikit penjelasan mengenai permasalahan rasio ketidakseimbangan distribusi kelas:

1. *Slight Imbalance*: Permasalahan klasifikasi yang tidak seimbang dimana distribusi kelasnya tidak merata dengan jumlah yang kecil pada data *training* (misalnya 4:6).
2. *Severe Imbalance*: Permasalahan klasifikasi yang tidak seimbang dimana distribusi kelasnya tidak merata dalam jumlah yang besar pada data *training* (misalnya 1:100 atau lebih).

Slight imbalance seringkali tidak terlalu menjadi perhatian, dan permasalahannya dapat sering diperlakukan seperti masalah pemodelan prediksi klasifikasi yang normal. Sedangkan, *severe imbalance* dapat menjadi tantangan untuk dimodelkan dan mungkin memerlukan penggunaan teknik khusus, seperti *undersampling method*, *oversampling method*, dan SMOTE (*Synthetic Minority Oversampling Technique*).

3.9.3 *K-Fold Cross Validation*

Cross-validation atau validasi silang merupakan suatu teknik validasi model yang digunakan untuk mengevaluasi bagaimana hasil statistik akan menggeneralisasi kumpulan data independen. Teknik ini biasanya digunakan untuk membuat prediksi model dan memperkirakan keakuratan model prediksi tersebut (Tempola, et al., 2018).

K-*fold cross validation* merupakan salah satu teknik dari validasi silang yang dilakukan untuk mengevaluasi kinerja suatu *classifier*. Metode ini dapat digunakan jika memiliki jumlah data yang terbatas (tidak banyak jumlah *instance*). K-*fold cross validation* juga digunakan untuk menentukan rata-rata tingkat keberhasilan sistem dengan mengacak atribut *input* untuk melakukan redudansi, sehingga sistem dapat menguji beberapa atribut *input* yang acak (Sitefanus, 2020).

Pada K-*Fold Cross Validation*, *dataset* akan dibagi menjadi sejumlah k-*folds*, di mana partisi tersebut terdiri atas data D_1, data D_2, …, D_k, masing-masing memiliki ukuran yang sama. Pelatihan dan pengujian data akan dilakukan sebanyak *k*-kali. Pada iterasi ke-*i*, partisi *Di* akan dijadikan sebagai data untuk pengujian, dan partisi yang tersisa lainnya secara kolektif digunakan untuk melatih model (Han, et al., 2012). Penggambaran iterasi model tersebut dapat dilihat pada **Gambar 3.2**.

Gambar 3.2 Contoh Model 5-*Folds Cross Validation* (Tempola, et al., 2018)

3.10. Pembobotan Kata (*Term Weighting*)

Pembobotan kata adalah proses untuk memberikan bobot pada suatu kata. Suatu *term* dalam dokumen akan dihitung frekuensi kemunculannya sebagai pembobotan dasar. Frekuensi kemunculan (*term frequency*) merupakan banyaknya *term* yang muncul dalam suatu dokumen. Semakin besar kemunculan suatu *term* dalam dokumen, maka nilai kesesuaian pun semakin besar. Faktor lain yang diperhatikan dalam pemberian bobot adalah kejarangmunculan kata (*term scarcity*) dalam dokumen. Pada pembobotan kata, kata yang jarang muncul pada beberapa dokumen harus dipandang sebagai kata yang lebih penting (*uncommontems*) dibandingkan kata yang sering muncul pada banyak dokumen. Pembobotan akan memperhitungkan faktor kebalikan frekuensi dokumen yang mengandung suatu kata tersebut (*inverse document frequency*) (Karmayasa, 2012).

Metode TF-IDF (*Term Frequency-Inverse Document Frequency*) merupakan metode pembobotan *term* yang banyak digunakan sebagai metode pembanding terhadap metode pembobotan baru. Pada metode ini, bobot *term t* dalam dokumen dapat dihitung dengan mengalikan nilai *Term Frequency* dengan *Inverse Document Frequency*. Terdapat beberapa jenis formula yang dapat digunakan pada *term frequency* (TF), yaitu sebagai berikut (Zafikri, 2008):

1. TF biner (*binary* TF): Hal yang diperhatikan, yaitu apakah terdapat suatu kata dalam dokumen, jika ada diberi nilai satu, jika tidak diberi nilai nol.
2. TF murni (*raw* TF): Nilai TF diberikan berdasarkan jumlah kemunculan suatu kata dalam dokumen. Misalnya, jika kata muncul sebanyak lima kali, maka kata tersebut akan bernilai lima.
3. TF logaritmik: Menghindari dominansi dokumen yang memiliki sedikit kata dalam *query*, tetapi mempunyai frekuensi yang tinggi.

$$tf = 1 + \log(f_{t,d}) \quad (3.5)$$

4. TF normalisasi: Menggunakan perbandingan antara frekuensi sebuah kata dengan jumlah keseluruhan kata pada dokumen.

$$tf = 0{,}5 + 0{,}5 \times \left(\frac{f_{t,d}}{\max\limits_{\{t' \in d\}} f_{t',d}} \right) \quad (3.6)$$

Inverse Document Frequency (IDF) dihitung dengan menggunakan rumus seperti pada **persamaan 3.7**.

$$idf_j = log\left(\frac{D}{df_j}\right) \quad (3.7)$$

dimana:

D : jumlah semua dokumen dalam koleksi

df_j : jumlah dokumen yang mengandung *term* t_j

Rumus umum yang digunakan dalam TF-IDF adalah penggabungan rumus antara *raw* TF dan IDF dengan cara mengalikan nilai *term frequency* (TF) dengan nilai *inverse document frequency* (IDF):

$$w_{ij} = tf_{ij} \times idf_i$$
$$w_{ij} = tf_{ij} \times \log\left(\frac{D}{df_j}\right) \quad (3.8)$$

dimana:

w_{ij} : bobot *term* t_j terhadap dokumen d_i

tf_{ij} : jumlah kemunculan *term* t_j dalam dokumen d_i

D : jumlah semua dokumen yang ada dalam *database*

df_j : jumlah dokumen yang mengandung *term* t_j (minimal ada satu kata yaitu *term* t_j)

Berikut adalah contoh perhitungan TF-IDF pada suatu kalimat yang telah dilakukan *pre-processing* pada penelitian (Deviyanto & Wahyudi, 2018), yaitu "partai politik percaya partai utama penting partai butuh rakyat".

Tabel 3.2 Contoh Perhitungan TF

***Term*/Kata**	**TF**					
	D_1	**D_2**	**D_3**	**D_4**	**D_5**	**D_6**
partai	1	1	3	2	0	0
politik	0	1	1	0	1	0
percaya	0	0	1	0	0	0
utama	0	0	1	0	0	0
penting	0	0	1	0	0	0
butuh	0	0	1	0	0	0
rakyat	0	0	1	0	1	0

Pada **Tabel 3.2** menampilkan nilai TF dari masing-masing *term*/kata, di mana TF merupakan banyaknya *term*/kata yang muncul dalam suatu dokumen. Dapat diketahui bahwa kata "partai" muncul sebanyak 1 kali pada dokumen 1, 1 kali pada dokumen 2, 3 kali pada dokumen 3, dan 2 kali pada dokumen 4. Begitu pula perhitungan TF untuk kata lainnya dalam dokumen. Tahapan selanjutnya adalah menghitung nilai IDF.

Tabel 3.3 Contoh Perhitungan IDF

***Term*/Kata**	**TF**						**DF**	**IDF =** $log\left(\frac{D}{df}\right)$
	D_1	**D_2**	**D_3**	**D_4**	**D_5**	**D_6**		
partai	1	1	3	2	0	0	4	$\log\left(\frac{6}{4}\right) = 0{,}176$
politik	0	1	1	0	1	0	3	$\log\left(\frac{6}{3}\right) = 0{,}301$
percaya	0	0	1	0	0	0	1	$\log\left(\frac{6}{1}\right) = 0{,}778$
utama	0	0	1	0	0	0	1	$\log\left(\frac{6}{1}\right) = 0{,}778$
penting	0	0	1	0	0	0	1	$\log\left(\frac{6}{1}\right) = 0{,}778$
butuh	0	0	1	0	0	0	1	$\log\left(\frac{6}{1}\right) = 0{,}778$
rakyat	0	0	1	0	1	0	2	$\log\left(\frac{6}{1}\right) = 0{,}778$

Pada perhitungan IDF, sebelumnya perlu diketahui terlebih dahulu nilai *D* dan DF nya, di mana *D* merupakan jumlah semua dokumen yang ada pada *dataset*, sedangkan DF merupakan jumlah dokumen yang mengandung *term* (*t*). Berdasarkan pada **Tabel 3.3**, misalnya pada kata "politik", diketahui bahwa nilai DF-nya adalah 3, yang berarti bahwa kata tersebut muncul pada 3 dokumen. Sehingga didapatkan nilai IDF untuk kata "politik" adalah 0,301.

Setelah didapatkan nilai TF dan IDF, selanjutnya akan dihitung nilai TF-IDF atau nilai bobot dengan mengalikan nilai TF dengan nilai IDF.

Tabel 3.4 Contoh Perhitungan TF-IDF

***Term*/Kata**	**TF**					**IDF**	**Bobot (W)=TF*IDF**				
	D_1	**…**	**D_3**	**…**	**D_6**		**D_1**	**…**	**D_3**	**…**	**D_6**
partai	1	…	3	…	0	0,176	0,176	…	0,528	…	0
politik	0	…	1	…	0	0,301	0	…	0,301	…	0
percaya	0	…	1	…	0	0,778	0	…	0,778	…	0
utama	0	…	1	…	0	0,778	0	…	0,778	…	0
penting	0	…	1	…	0	0,778	0	…	0,778	…	0
butuh	0	…	1	…	0	0,778	0	…	0,778	…	0
rakyat	0	…	1	…	0	0,778	0	…	0,778	…	0

Berdasarkan **Tabel 3.4**, didapatkan bahwa pada kata "percaya" pada dokumen 3 memiliki nilai TF yaitu 1, dan nilai IDF yaitu 0,778. Sehingga hasil perhitungan nilai bobot atau TF-IDF kata tersebut pada dokumen 3 adalah 0,778.

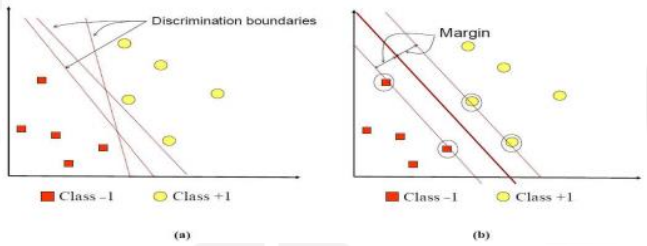
3.11. *Support Vector Machine* (SVM)

Support Vector Machine (SVM) dipresentasikan pertama kali pada tahun 1992 yang dikembangkan oleh Boser, Guyon, Vapnik. SVM bekerja dengan menemukan *hyperplane* yang terbaik pada *input space*. *Hyperplane* merupakan *affine subspace* dengan dimensi d-1 yang membagi ruang vektor menjadi dua bagian, di mana masing-masing bagian sesuai dengan kategori yang berbeda. *Linear classifier* merupakan prinsip dasar dari SVM, yang selanjutnya dikembangkan agar dapat bekerja pada masalah non-*linear* dengan cara memasukkan konsep *kernel trick* pada ruang kerja yang berdimensi tinggi. Saat ini, SVM telah berhasil diterapkan pada masalah di dunia nyata, dan biasanya memberikan solusi yang lebih baik

dibandingkan dengan metode konvensional lainnya, seperti *artificial neural network* (Nugroho, et al., 2003).

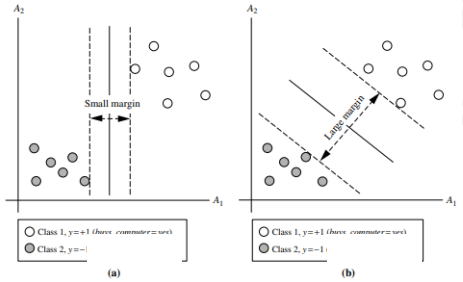
3.11.1 SVM *with Linearly Separable Data*

SVM dapat menangani suatu kasus di mana kelompok dipisahkan secara linier, dengan mengukur *margin hyperplane* dan menemukan titik maksimumnya. *Margin* merupakan jarak antara *hyperplane* dengan pola terdekat di setiap kelompok.



Gambar 3.3 Menemukan *Hyperplane* Terbaik (Drajana, 2017)

Pada **Gambar 3.3 bagian (a)** menampilkan beberapa pola anggota yang termasuk dalam dua kategori: +1 dan -1. Pola yang termasuk dalam kategori +1 diwakili dengan warna kuning, sedangkan pola dalam kategori - 1 diwakili dengan warna merah. Masalah klasifikasi dapat dilakukan dengan mencari garis (*hyperplane*) yang memisahkan kedua kelompok tersebut. **Gambar 3.3 bagian (a)** menunjukkan berbagai batas perbedaan alternatif (*discrimination boundaries*). Garis *solid* pada **Gambar 3.3 bagian (b)** menampilkan *hyperplane* optimal yang terletak di tengah antara dua kelompok, serta titik merah dan kuning di dalam lingkaran hitam merupakan *support vector* (Drajana, 2017). Pada data yang dapat dipisahkan secara linier, *support vector* adalah bagian dari data pelatihan yang sebenarnya.



Gambar 3.4 Kemungkinan *Hyperplane* dan Jarak *Margin* (Han, et al., 2012)

Gambar 3.4 menunjukkan dua kemungkinan dari *hyperplane* dan juga jarak *margin*-nya. Kedua *hyperplane* dapat mengklasifikasikan kelompok data secara

benar. *Hyperplane* dengan jarak *margin* yang lebih besar akan lebih akurat dalam mengklasifikasikan kelompok data dibandingkan *hyperplane* dengan jarak *margin* yang lebih kecil. Pada tahapan pelatihan, biasanya SVM mencari *hyperplane* dengan jarak *margin* terbesar, yaitu *maximum marginal hyperplane* (MMH). Formula *hyperplane* dapat ditulis seperti **persamaan 3.9**.

$$w \cdot x_i + b = 0 \quad (3.9)$$

Keterangan:

w : vektor bobot

x_i : data ke-*i*

b : nilai bias

Jika nilai *b* dianggap sebagai bobot tambahan, w_0, **persamaan 3.9** dapat ditulis kembali seperti **persamaan 3.10**.

$$w_0 + w.x_i = 0 \quad (3.10)$$

Jika setiap titik berada di atas *hyperplane*, maka digunakan **persamaan 3.11**. Demikian pula, jika setiap titik berada di bawah *hyperplane*, dapat digunakan **persamaan 3.12**.

$$w_0 + w.x_i > 0 \quad (3.11)$$

$$w_0 + w.x_i < 0 \quad (3.12)$$

Metode SVM membagi *dataset* menjadi 2 kelompok, sehingga dapat dirumuskan seperti **persamaan 3.13**. Di mana, y_i merupakan kelas data ke-*i*.

$$H_1 : w_0 + w.x_i \geq 1 \; untuk \; y_i = +1$$
$$H_2 : w_0 + w.x_i \leq -1 \; untuk \; y_i = -1 \quad (3.13)$$

(Han, et al., 2012)

Margin maksimum dapat diukur dengan memaksimalkan nilai jarak antara *hyperplane* dan titik terdekatnya, yaitu $\frac{1}{\|w\|}$. Hal ini dapat dirumuskan sebagai masalah *Quadratic Programming* (QP), yang akan mempertimbangkan *constraint* pada **persamaan 3.15**, untuk menemukan titik minimum pada **persamaan 3.14**.

$$\min_{w} \tau(w) = \frac{1}{2} \| w \|^2 \quad (3.14)$$

$$y_i(x_i.w + b) - 1 \geq 0, \qquad \forall_i \quad (3.15)$$

Masalah ini dapat diatasi dengan berbagai teknik perhitungan, di antaranya *Lagrange Multiplier* seperti pada **persamaan 3.16**.

$$L(w, b, \alpha) = \frac{1}{2} \| w \|^2 - \sum_{i=1}^{l} \alpha_i \left(y_i \left((x_i . w + b) - 1 \right) \right) \quad (3.16)$$

dengan $i = 1,2,\ldots,l$

Notasi α_i merupakan *Lagrange multipliers* yang memiliki nilai nol atau positif $(\alpha_i \geq 0)$. Nilai optimal pada **persamaan 3.16** dapat diukur dengan meminimalkan L terhadap w dan b, serta dapat memaksimalkan L terhadap α_i. Dengan memperhatikan sifat bahwa pada titik optimal *gradient* $L = 0$, **persamaan 3.16** dapat dimodifikasi sebagai maksimalisasi *problem* yang hanya mengandung α_i, sebagaimana pada **persamaan 3.17**.

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i x_j \quad (3.17)$$

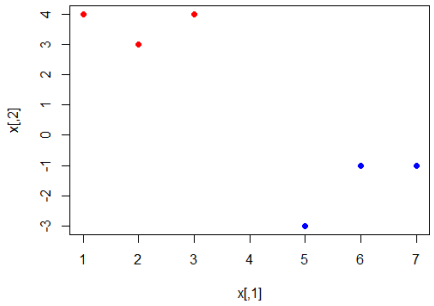
dimana $\alpha_i \geq 0 \ (i = 1,2, \ldots, l) \ \sum_{i=1}^{l} \alpha_i y_i = 0$

Hasil perhitungan α_i yang diperoleh, kebanyakan bernilai positif, dimana data yang berkorelasi dengan α_i ini disebut sebagai *support vector* (Nugroho, et al., 2003). Berikut ini akan dijelaskan simulasi perhitungan SVM linier.

Tabel 3.5 Contoh Data SVM *linear* (Sicotte, 2015)

x_1	x_2	y
3	4	-1
1	4	-1
2	3	-1
6	-1	1
7	-1	1
5	-3	1

Diberikan contoh data SVM *linear* seperti pada **Tabel 3.5**, dimana variabel x_1 dan x_2 merupakan kumpulan titik data, sedangkan variabel y merupakan kelas data yang menunjukkan bahwa angka (1) menyatakan kelas positif, dan angka (-1) menyatakan kelas negatif. Jika data tersebut dimasukkan ke dalam sebuah grafik, maka tampilannya akan seperti pada **Gambar 3.5**.



Gambar 3.5 Grafik Contoh Data

Pada **Gambar 3.5** dapat dilihat bahwa terdapat dua kelas yang terdiri dari kelas negatif dimana ditandai dengan warna merah, dan kelas positif yang ditandai dengan warna biru. Fomulasi untuk menemukan titik minimum persamaan $\frac{1}{2} \| w \|^2$ dengan memperhatikan pembatas, didapatkan persamaan *constraint* seperti berikut.

$$-3w_1 - 4w_2 - b \geq 1 \quad (3.18)$$

$$-w_1 - 4w_2 - b \geq 1 \quad (3.19)$$

$$-2w_1 - 3w_2 - b \geq 1 \quad (3.20)$$

$$6w_1 - w_2 + b \geq 1 \quad (3.21)$$

$$7w_1 - w_2 + b \geq 1 \quad (3.22)$$

$$5w_1 - 3w_2 + b \geq 1 \quad (3.23)$$

Diperoleh eliminasi pada **persamaan 3.18** dan **persamaan 3.21** seperti berikut.

$$\begin{array}{r} -3w_1 - 4w_2 - b = 1 \\ 6w_1 - w_2 + b = 1 \\ \hline 3w_1 - 5w_2 = 2 \end{array} +$$

$$w_1 = \frac{2 + 5w_2}{3} \quad (3.24)$$

Eliminasi kembali **persamaan 3.19** dan **persamaan 3.22**, kemudian substitusi **persamaan 3.24**, sehingga diperoleh seperti berikut.

$$\begin{array}{r} -w_1 - 4w_2 - b = 1 \\ 7w_1 - w_2 + b = 1 \\ \hline 6w_1 - 5w_2 = 2 \end{array} +$$

$$6\left(\frac{2 + 5w_2}{3}\right) - 5w_2 = 2$$

$$4 + 10w_2 - 5w_2 = 2$$

$$w_2 = 0{,}4$$

Diperoleh $w_2 = 0{,}4$, maka $w_1 = \frac{2+5w_2}{3} = 1{,}3$

Substitusi nilai w_1 dan w_2 yang telah diperoleh ke dalam **persamaan 3.20** seperti berikut.

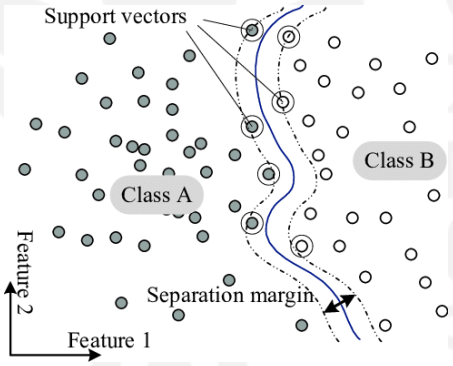
$$-2\mathrm{w}_1 - 3w_2 - b = 1$$
$$-2(1{,}3) - 3(0{,}4) - b = 1$$
$$-2{,}6 - 1{,}2 - b = 1$$
$$b = -4{,}8$$

Berdasarkan hasil nilai w_1, w_2, dan b yang telah diperoleh, maka didapatkan persamaan *hyperplane* seperti berikut.

$$\mathrm{f}(x) = 1{,}3x_1 + 0{,}4x_2 - 4{,}8$$

Pengujian selanjutnya untuk klasifikasi pada data baru, dapat disubstitusikan ke dalam fungsi *hyperplane* yang telah didapatkan sebelumnya. Jika semua nilai yang dihasilkan $f(x) > 0$, maka termasuk kelas positif, dan nilai $f(x) < 0$ masuk kelas negatif.

3.11.2 SVM *with Non-Linearly Separable Data*



Gambar 3.6 *Hyperplane* Nonlinier (Alsheikh, et al., 2014)

Pada kasus seperti pada **Gambar 3.6**, tidak ada garis lurus yang dapat ditemukan untuk memisahkan setiap kelompok. Penelitian yang dilakukan (Han, et al., 2012) menghasilkan perluasan metode SVM agar dapat mengolah data *nonlinear* dengan pendekatan untuk SVM pada data *linear*. Terdapat dua langkah dalam melakukan hal tersebut, yaitu menggunakan pemetaan *nonlinear* untuk mengubah data masukan asli menjadi ruang berdimensi lebih tinggi, dan mencari *hyperplane* pemisah linier di ruang baru.

Memilih pemetaan *nonlinear* ke ruang dimensi yang lebih tinggi dapat dilakukan dengan memecahkan masalah optimasi kuadrat dari SVM linier (yaitu, ketika mencari SVM linier di ruang dimensi baru yang lebih tinggi), data pelatihan

hanya muncul dalam bentuk perkalian, $\phi(X_i).\phi(X_j)$, di mana φ (X) merupakan fungsi pemetaan *nonlinear* yang diterapkan untuk mentransformasi data pelatihan. Secara matematis, hal ini setara dengan menerapkan *kernel function* $K(X_i, X_j)$ ke data asli, dibandingkan dengan menghitung perkalian pada data yang ditransformasikan. Lihat **persamaan 3.25**.

$$K(X_i, X_j) = \phi(X_i).\phi(X_j) \quad (3.25)$$

Setelah mengimplementasikan fungsi *kernel* seperti pada **persamaan 3.25**, kemudian dapat dilanjutkan untuk menemukan *hyperplane* pemisah maksimal. Berikut adalah beberapa rumus *kernel* yang umum digunakan, yaitu (Yan, et al., 2014):

1. *Linear kernel*

$$K(X_i, X_j) = (X_i.X_j) \quad (3.26)$$

2. *Polynomial kernel*

$$K(X_i, X_j) = (X_i.X_j + 1)^d \quad (3.27)$$

3. *Gaussian radial basis function kernel* (RBF)

$$K(X_i, X_j) = \exp(X_i, X_j) = \exp\left(-\gamma\|X_i.X_j\|^2\right) \quad (3.28)$$

4. *Sigmoid kernel*

$$K(X_i, X_j) = \tanh(\gamma(X_i.X_j) + c) \quad (3.29)$$

Keterangan:

$K(X_i,X_j)$: fungsi kernel
$X_i.X_j$: *dot product* vektor X_i dan X_j
d : derajat *polynomial*
γ : parameter *gamma*
c : koefisien

Tidak ada aturan khusus yang menentukan fungsi *kernel* mana yang akan menghasilkan SVM paling akurat. Faktanya, *kernel* yang dipilih biasanya tidak membuat perbedaan yang besar dalam keakuratan hasilnya. Tujuan utama penelitian SVM adalah meningkatkan kecepatan pelatihan dan pengujian sehingga SVM dapat menjadi pilihan yang lebih layak untuk kumpulan data yang sangat besar (misalnya, jutaan *support vector*). Masalah lainnya termasuk menentukan

kernel terbaik untuk kumpulan data tertentu, dan menemukan metode yang lebih efisien untuk berbagai jenis situasi (Han, et al., 2012).

3.12. *Word Cloud*

Word cloud merupakan representasi visual dari frekuensi kemunculan kata. Semakin sering suatu istilah/kata muncul dalam teks yang dianalisis, semakin besar kata tersebut ditampilkan dalam *word cloud* (Atenstaedt, 2012). *Word cloud* atau dikenal juga sebagai *tag cloud*, biasanya dengan memeriksa frekuensi kata, instruktur dapat mencari pola kata dan frasa tertentu, atau kekurangannya, dalam data teks (DePaolo & Wilkinson, 2014). Pada **gambar 3.7** merupakan contoh tampilan dari *word cloud*.

Gambar 3.7 *Word cloud* data tanggapan siswa tentang "Apa itu statistik?" (DePaolo & Wilkinson, 2014)

3.13. *Association Rules*

Association rules atau aturan asosiasi merupakan salah satu tugas data *mining* deskriptif yang digunakan untuk menemukan aturan asosiasi antara elemen data. Hal yang dilakukan dalam aturan asosiasi, yaitu untuk mengetahui seberapa sering suatu kombinasi elemen muncul secara bersamaan dalam *dataset*, atau biasa disebut sebagai *frequent patterns* (Aprianti, et al., 2017). Pada penelitian ini digunakan pendekatan asosiasi agar mengetahui hubungan antar kata yang cukup sering dibicarakan secara bersamaan pada masing-masing klasifikasi sentimen netral, positif, dan negatif. Hal ini bertujuan agar mendapatkan informasi yang dapat ditinjau oleh pihak Indihome maupun pelanggannya. Formula yang digunakan untuk menghitung nilai korelasi tersebut tertera pada **persamaan 3.30**.

$$r = \frac{n\sum X_i Y_i - (\sum X_i)(\sum Y_i)}{\sqrt{\left(n\sum X_i^2 - (\sum X_i)^2\right)\left(n\sum Y_i^2 - (\sum Y_i)^2\right)}} \quad (3.30)$$

Keterangan:

n : Banyaknya pasangan data X dan Y

ΣX_i : Total jumlah variabel X, $i = 1,2,3,\ldots,n$

ΣY_i : Total jumlah variabel Y, $i = 1,2,3,\ldots,n$

ΣX_i^2 : Kuadrat dari total jumlah variabel X, $i = 1,2,3,\ldots,n$

ΣY_i^2 : Kuadrat dari total jumlah variabel Y, $i = 1,2,3,\ldots,n$

$\Sigma X_i Y_i$: Jumlah dari hasil perkalian antara variabel X dan variabel Y

BAB IV
METODOLOGI PENELITIAN

4.1. Populasi Penelitian

Populasi dalam penelitian ini, yaitu semua *tweet* di *platform* Twitter pada setiap tahunnya. Sedangkan, sampel yang digunakan dalam penelitian ini menggunakan *tweet* yang mengandung kata "Indihome" dalam rentang bulan Maret dan April 2021.

4.2. Tempat dan Waktu Penelitian

Penelitian dilakukan dengan memanfaatkan fasilitas internet yang tersedia di setiap tempat. Rentang waktu yang dilakukan untuk penelitian ini dimulai dari bulan Desember 2020 hingga bulan Juli 2021.

4.3. Variabel penelitian

Pada **Tabel 4.1** menunjukkan definisi dari variabel yang digunakan dalam penelitian ini:

Tabel 4.1 Variabel Penelitian

Variabel	Definisi Variabel
Text	*Tweet*/ulasan pengguna *provider* Indihome pada *platform* Twitter.
Label	Klasifikasi atau kategori kelas dari ulasan yang diperoleh. Penulis mengkategorikan ulasan menjadi tiga kelas, yaitu kelas positif, kelas negatif, dan kelas netral.

4.4. Metode Analisis Data

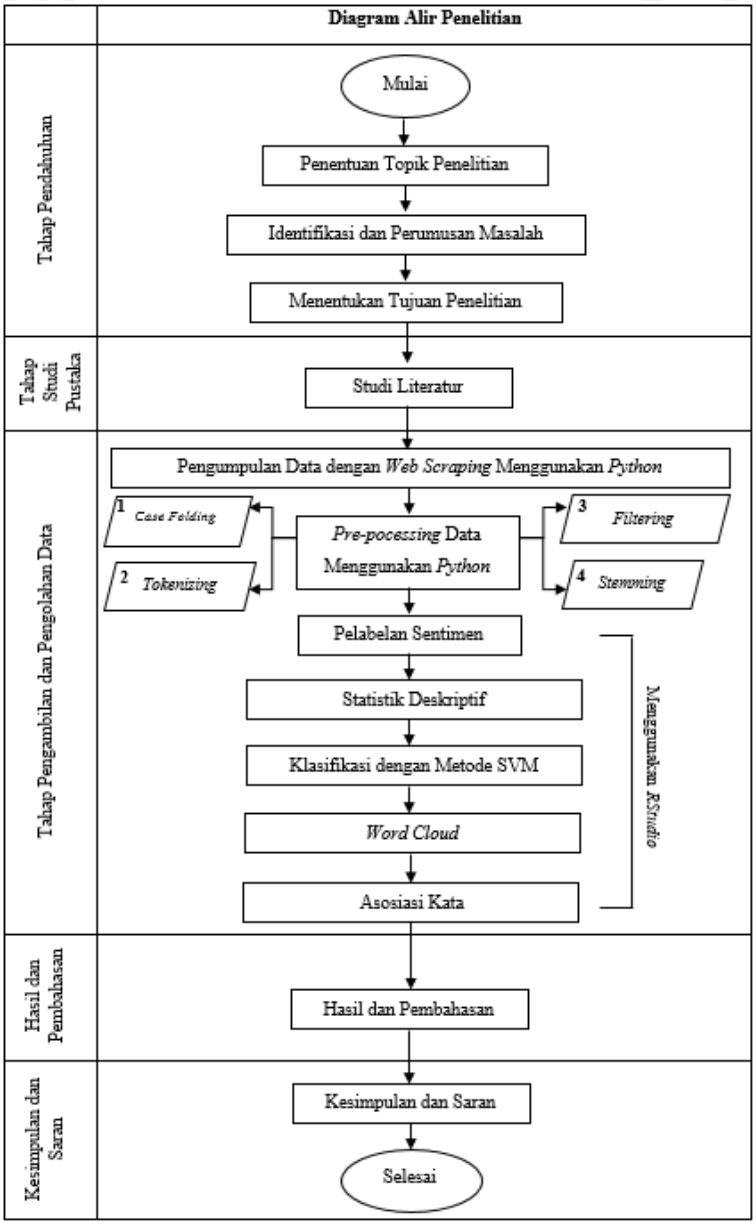
Pada penelitian ini akan dilakukan klasifikasi sentimen data *tweet* tentang *provider* Indihome di *platform* Twitter dengan menggunakan:

1. Statistik deskriptif data Twitter untuk mengetahui gambaran secara umum.

2. Metode *kernel* pada algoritma *Support Vector Machine* (SVM) untuk mengetahui pembagian klasifikasi netral, positif dan negatif.
3. *WordCloud*, untuk memvisualisasikan kata yang paling banyak/sering muncul dalam suatu ulasan.
4. *Association*, untuk mengetahui pola kata yang sering muncul secara bersamaan yang dilihat dari nilai korelasinya.

4.5. Tahapan Penelitian

Adapun tahapan penelitian yang dilakukan pada penelitian ini dapat dilihat pada **Gambar 4.1**.

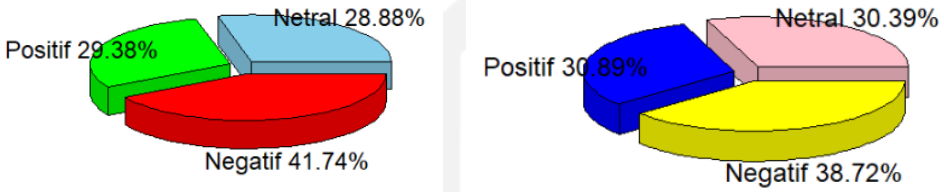


Gambar 4.1 Diagram Alir Penelitian

BAB V
HASIL DAN PEMBAHASAN

5.1. Statistik Deskriptif

Statistik deskriptif pada penelitian ini dilakukan untuk mengetahui gambaran umum terkait data ulasan pelanggan provider Indihome yang telah didapatkan. Data ulasan yang didapatkan, terdiri dari 58311 ulasan pada rentang bulan Maret 2021, dan 51988 ulasan pada rentang bulan April 2021. Namun, penulis memutuskan untuk melakukan penghapusan teks ulasan yang bersifat di luar konteks serta ulasan yang kosong setelah dilakukan *preprocessing data*, sehingga jumlah data menjadi 56944 ulasan pada bulan Maret, dan 51063 ulasan pada bulan April. Data ulasan tersebut dikategorikan menjadi tiga kelas, yaitu sentimen netral, sentimen positif, dan sentimen negatif. **Gambar 5.1** menunjukkan gambaran keseluruhan mengenai masing-masing kategori yang diperoleh dari hasil analisis pada bulan Maret dan April 2021.



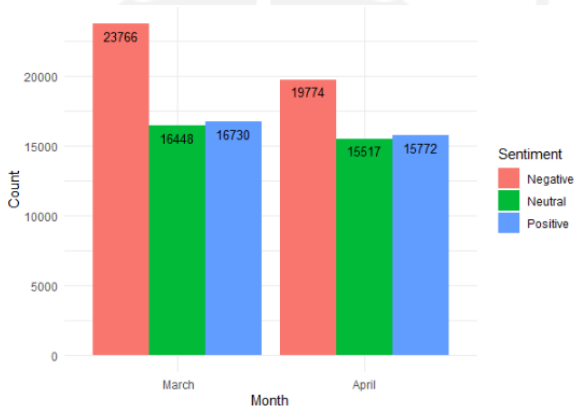
(a) **(b)**

Gambar 5.1 Persentase Sentimen Tentang Indihome

Berdasarkan **Gambar 5.1** bagian **(a)**, dapat dilihat bahwa bagian sentimen positif ditandai dengan warna hijau, sentimen netral ditandai dengan warna biru langit, dan sentimen negatif ditandai dengan warna merah. Sentimen negatif lebih dominan dibandingkan sentimen positif dan netral pada rentang bulan Maret, begitu pula pada rentang bulan April yang ditunjukkan pada **Gambar 5.1** bagian **(b)**. Sentimen negatif pada bulan April yang ditandai dengan warna kuning lebih dominan dibandingkan dengan sentimen positif yang ditandai dengan warna biru dan sentimen netral yang ditandai dengan warna merah muda.

Pada rentang bulan Maret, terdapat sebesar 29,38% ulasan merupakan sentimen positif yang mengacu pada ulasan yang mendukung terhadap pelayanan

atau kinerja *provider* Indihome, lalu sebesar 28,88% ulasan lainnya merupakan sentimen netral yang berarti mengandung ulasan yang tidak mengacu kepada kepuasan maupun ketidaksukaan terhadap Indihome, misalnya berupa suatu pertanyaan. Serta, sisanya sebesar 41,74% ulasan merupakan sentimen negatif yang mengarah pada ulasan yang bersifat menjatuhkan atau tidak menyukai suatu hal yang berhubungan dengan *provider* Indihome. Sedangkan, pada rentang bulan April, terdapat sebesar 30,89% ulasan merupakan sentimen positif, 30,39% ulasan merupakan sentimen netral, dan sisanya 38,72% ulasan merupakan sentimen negatif. Perbedaan persentase yang dihasilkan pada bulan April, memiliki pembagian yang hampir sama rata. Pada **Gambar 5.2**, merupakan perbandingan jumlah dari sentimen positif dan sentimen negatif yang telah dihasilkan pada bulan Maret dan April 2021.



Gambar 5.2 Perbandingan Jumlah Sentimen Tentang Indihome

Berdasarkan **Gambar 5.2**, dari jumlah sentimen negatif yang dihasilkan pada rentang bulan Maret dan April, yaitu sebanyak 23766 ulasan dan 19774 ulasan, dapat dikatakan bahwa sebagian besar pelanggan Indihome menggunakan *platform* Twitter untuk memberikan ulasannya terkait pelayanan atau kinerja dari Indihome itu sendiri, cenderung ke arah tidak menyukai atau tidak puas. Sedangkan, sisanya menyebar menjadi sentimen positif dan netral yang perbedaan nya tidak terlalu signifikan. Hal ini berarti masih banyak pelanggan yang merasa cukup puas terhadap pelayanan maupun kinerja dari Indihome, yang diutarakan melalui *platform* Twitter.

Selain penurunan jumlah ulasan yang dihasilkan pada bulan April, jumlah dari setiap sentimennya pun mengalami penurunan. Baik pada bulan Maret maupun April, jumlah sentimen positif dan netral tidak terlalu memiliki perbedaan yang

signifikan. Meskipun lebih banyak pelanggan yang memberikan sentimen negatif, tidak dipungkiri juga masih banyak yang memberikan sentimen positif ataupun netral mengenai Indihome.

5.2. *Preprocessing* Data

Data ulasan pada penelitian ini merupakan *tweet* atau cuitan yang diambil dari *platform* Twitter berupa teks yang tidak terstruktur, dikarenakan masih terdapat banyak *noise*, sehingga sebelum melakukan tahapan klasifikasi, data perlu diubah menjadi lebih terstruktur. Pada tahapan *preprocessing*, akan digunakan metode *text mining* untuk melakukan pembersihan data teks. Tahapan-tahapan tersebut diantaranya, yaitu *case folding*, *tokenizing, filtering* dan *normalization*, serta *stemming*. Pada **Tabel 5.1** merupakan contoh tampilan beberapa data asli yang diambil secara acak, sebelum dilakukan tahapan *preprocessing* data.

Tabel 5.1 Contoh Beberapa Ulasan Pada Bulan Maret dan April 2021

Bulan	No	*Tweet* Ulasan
Maret	1	@jarvis8_ keren kak , indihome memang yang terbaik ðŸ¥°ðŸ¥°
	2	hadeh indihome kenapa lagi ini.
	3	@xxasth Hallo, Sobat. Guna kami bantu pengecekan terlebih dahulu, mohon untuk menginformasikan nomor internet yang terdiri dari 12 digit angka, dan nomor HP aktif yang dapat dihubungi melalui DM kami di https://t.co/yx4WQQ5l5l ya Sobat. Terima kasih. -Ari
	4	@IndiHome hello, tengah malam gangguan kah?
	5	"@IndiHome Lanjutkan min. Tips yang bagus. Bisa lebih gampang deh. #InternetnyaIndonesia #MajuTerusBersamaIndiHome"
April	6	@IndiHome lemot banget jaringan!!!!
	7	Min saya mau berhenti berlangganan indihome @IndiHomeCare @IndiHome mohon ditindaklanjuti,
	8	Indihome knp bermasalah terus sih:(
	9	indihome menyial https://t.co/XyuVoPGZnG
	10	"Anti penanganan Lambat Terimakasih pak Indihome yang selalu benahi jaringan Indihome kita sehingga jaringanya stabil ..... https://t.co/Fc6hXKCSlD"

5.2.1 *Case Folding* dan *Cleaning Text*

Proses *case folding* merupakan tahapan mengubah huruf kapital menjadi huruf kecil. Contoh hasil *tweet* yang telah dilakukan proses *case folding*

ditunjukkan pada **Tabel 5.2**. Perubahan yang dihasilkan, ditunjukkan dengan karakter yang diberi warna merah.

Tabel 5.2 Ulasan Mengenai Indihome Setelah Melalui Proses *Case Folding*

Bulan	No	*Tweet* Ulasan	*Tweet* Setelah Melalui *Case Folding*
Maret	1	@jarvis8_ keren kak , indihome memang yang terbaik ðŸ¥°ðŸ¥°	@jarvis8_ keren kak , indihome memang yang terbaik ðÿ¥°ðÿ¥°
	2	hadeh indihome kenapa lagi ini.	hadeh indihome kenapa lagi ini.
	3	@xxasth Hallo, Sobat. Guna kami bantu pengecekan terlebih dahulu, mohon untuk menginformasikan nomor internet yang terdiri dari 12 digit angka, dan nomor HP aktif yang dapat dihubungi melalui DM kami di https://t.co/yx4WQQ5l5l ya Sobat. Terima kasih. -Ari	@xxasth hallo, sobat. guna kami bantu pengecekan terlebih dahulu, mohon untuk menginformasikan nomor internet yang terdiri dari 12 digit angka, dan nomor hp aktif yang dapat dihubungi melalui dm kami di https://t.co/yx4wqq5l5l ya sobat. terima kasih. -ari
	4	@IndiHome hello, tengah malam gangguan kah?	@indihome hello, tengah malam gangguan kah?
	5	"@IndiHome Lanjutkan min. Tips yang bagus. Bisa lebih gampang deh. #InternetnyaIndonesia #MajuTerusBersamaIndiHome"	"@indihome lanjutkan min. tips yang bagus. bisa lebih gampang deh. #internetnyaindonesia #majuterusbersamaindihome"
April	6	@IndiHome lemot banget jaringan!!!!	@indihome lemot banget jaringan!!!!
	7	Min saya mau berhenti berlangganan indihome @IndiHomeCare @IndiHome mohon ditindaklanjuti,	min saya mau berhenti berlangganan indihome @indihomecare @indihome mohon ditindaklanjuti,
	8	Indihome knp bermasalah terus sih:(	indihome knp bermasalah terus sih:(
	9	indihome menyial https://t.co/XyuVoPGZnG	indihome menyial https://t.co/xyuvopgzng
	10	"Anti penanganan Lambat Terimakasih pak Indihome yang selalu benahi jaringan Indihome kita sehingga jaringanya stabil ..... https://t.co/Fc6hXKCSlD"	"anti penanganan lambat terimakasih pak indihome yang selalu benahi jaringan indihome kita sehingga jaringanya stabil ..... https://t.co/fc6hxkcsld"

Pada tahapan *cleaning text*, dilakukan pembersihan dokumen untuk mengurangi *noise*, seperti tanda baca, *username*/*mention*, url, angka, *emoticon*,

hashtag, *special character*, dan sebagainya. Hasil *tweet* yang dilakukan *cleaning text* ditunjukkan pada **Tabel 5.3**.

Tabel 5.3 Ulasan Mengenai Indihome Setelah Melalui Proses *Cleaning Text*

Bulan	No	*Tweet* Setelah Melalui *Case Folding*	*Tweet* Setelah Melalui *Cleaning Text*
Maret	1	@jarvis8_ keren kak , indihome memang yang terbaik ðÿ¥°ðÿ¥°	keren kak indihome memang yang terbaik
	2	hadeh indihome kenapa lagi ini.	hadeh indihome kenapa lagi ini
	3	@xxasth hallo, sobat. guna kami bantu pengecekan terlebih dahulu, mohon untuk menginformasikan nomor internet yang terdiri dari 12 digit angka, dan nomor hp aktif yang dapat dihubungi melalui dm kami di https://t.co/yx4wqq5l5l ya sobat. terima kasih. -ari	hallo sobat guna kami bantu pengecekan terlebih dahulu mohon untuk menginformasikan nomor internet yang terdiri dari digit angka dan nomor hp aktif yang dapat dihubungi melalui dm kami di ya sobat terima kasih ari
	4	@indihome hello, tengah malam gangguan kah?	hello tengah malam gangguan kah
	5	"@indihome lanjutkan min. tips yang bagus. bisa lebih gampang deh. #internetnyaindonesia #majuterusbersamaindihome"	lanjutkan min tips yang bagus bisa lebih gampang deh
April	6	@indihome lemot banget jaringan!!!!	lemot banget jaringan
	7	min saya mau berhenti berlangganan indihome @indihomecare @indihome mohon ditindaklanjuti,	min saya mau berhenti berlangganan indihome mohon ditindaklanjuti
	8	indihome knp bermasalah terus sih:(	indihome knp bermasalah terus sih
	9	indihome menyial https://t.co/xyuvopgzng	indihome menyial
	10	"anti penanganan lambat terimakasih pak indihome yang selalu benahi jaringan indihome kita sehingga jaringanya stabil ..... https://t.co/fc6hxkcsld"	anti penanganan lambat terimakasih pak indihome yang selalu benahi jaringan indihome kita sehingga jaringanya stabil

5.2.2 *Tokenizing*

Langkah selanjutnya, yaitu *tokenizing* yang merupakan proses pemecahan suatu kalimat atau teks menjadi kata per kata yang disebut token. *Tokenizing* dapat

mempermudah dalam perhitungan frekuensi kemunculan kata tersebut. Hasil *tweet* yang telah dilakukan *tokenizing*, ditampilkan pada **Tabel 5.4**.

Tabel 5.4 Ulasan Mengenai Indihome Setelah Melalui Proses *Tokenizing*

Bulan	No	*Tweet* Setelah Melalui *Cleaning Text*	*Tweet* Setelah Melalui *Tokenizing*
Maret	1	keren kak indihome memang yang terbaik	‘keren’, ‘kak’, ‘indihome’, ‘memang’, ‘yang’, ‘terbaik’
	2	hadeh indihome kenapa lagi ini	‘hadeh’, ‘indihome’, ‘kenapa’, ‘lagi’, ‘ini’
	3	hallo sobat guna kami bantu pengecekan terlebih dahulu mohon untuk menginformasikan nomor internet yang terdiri dari digit angka dan nomor hp aktif yang dapat dihubungi melalui dm kami di ya sobat terima kasih ari	‘hallo’, ‘sobat’, ‘guna’, ‘kami’, ‘bantu’, ‘pengecekan’, ‘terlebih’, ‘dahulu’, ‘mohon’, ‘untuk’, menginformasikan’, ‘nomor’, ‘internet’, ‘yang’, ‘terdiri’, ‘dari’, ‘digit’, ‘angka’, ‘dan’, ‘nomor’, ‘hp’, ‘aktif’, ‘yang’, ‘dapat’, ‘dihubungi’, ‘melalui’, ‘dm’, ‘kami’, ‘di’, ‘ya’, ‘sobat’, ‘terima’, ‘kasih’, ‘ari’
	4	hello tengah malam gangguan kah	‘hello’, ‘tengah’, ‘malam’, ‘gangguan’, ‘kah’
	5	lanjutkan min tips yang bagus bisa lebih gampang deh	‘lanjutkan’, ‘min’, ‘tips’, ‘yang’, ‘bagus’, ‘bisa’, ‘lebih’, ‘gampang’, ‘deh’
April	6	lemot banget jaringan	‘lemot’, ‘banget’, ‘jaringan’
	7	min saya mau berhenti berlangganan indihome mohon ditindaklanjuti	‘min’, ‘saya’, ‘mau’, ‘berhenti’, ‘berlangganan’, ‘indihome’, ‘mohon’, ‘ditindaklanjuti’
	8	indihome knp bermasalah terus sih	‘indihome’, ‘knp’, ‘bermasalah’, ‘terus’, ‘sih’
	9	indihome menyial	‘indihome’, ‘menyial’
	10	anti penanganan lambat terimakasih pak indihome yang selalu benahi jaringan indihome kita sehingga jaringanya stabil	‘anti’, ‘penanganan’, ‘lambat’, ‘terimakasih’, ‘pak’, ‘indihome’, ‘yang’, ‘selalu’ ‘benahi’, ‘jaringan’, ‘indihome’, ‘kita’, ‘sehingga’, ‘jaringanya’, ‘stabil’

5.2.3 *Filtering* dan *Normalization*

Pada tahap *filtering*, dilakukan penyaringan dan penghapusan kata yang tidak bermakna menggunakan kamus *stopwords*, serta tambahan kata yang tidak terdapat pada *stopwords*. Hasil *filtering* dapat dilihat pada **Tabel 5.5**.

Tabel 5.5 Ulasan Mengenai Indihome Setelah Melalui Proses *Filtering*

Bulan	No	*Tweet* Setelah Melalui *Tokenizing*	*Tweet* Setelah Melalui *Filtering*
Maret	1	‘keren’, ‘kak’, ‘indihome’, ‘memang’, ‘yang’, ‘terbaik’	‘keren’, ‘indihome’, ‘terbaik’
	2	‘hadeh’, ‘indihome’, ‘kenapa’, ‘lagi’, ‘ini’	‘hadeh’, ‘indihome’
	3	‘hallo’, ‘sobat’, ‘guna’, ‘kami’, ‘bantu’, ‘pengecekan’, ‘terlebih’, ‘dahulu’, ‘mohon’, ‘untuk’, ‘menginformasikan’, ‘nomor’, ‘internet’, ‘yang’, ‘terdiri’, ‘dari’, ‘digit’, ‘angka’, ‘dan’, ‘nomor’, ‘hp’, ‘aktif’, ‘yang’, ‘dapat’, ‘dihubungi’, ‘melalui’, ‘dm’, ‘kami’, ‘di’, ‘ya’, ‘sobat’, ‘terima’, ‘kasih’, ‘ari’	sobat’, ‘bantu’, ‘pengecekan’, ‘mohon’, ‘menginformasikan’, ‘nomor’, ‘internet’, ‘digit’, ‘angka’, ‘nomor’, ‘hp’, ‘aktif’, ‘dihubungi’, ‘ya’, ‘sobat’, ‘ari’
	4	‘hello’, ‘tengah’, ‘malam’, ‘gangguan’, ‘kah’	‘malam’, ‘gangguan’, ‘kah’
	5	‘lanjutkan’, ‘min’, ‘tips’, ‘yang’, ‘bagus’, ‘bisa’, ‘lebih’, ‘gampang’, ‘deh’	‘lanjutkan’, ‘tips’, ‘bagus’, ‘gampang’
April	6	‘lemot’, ‘banget’, ‘jaringan’	‘lemot’, ‘banget’, ‘jaringan’
	7	‘min’, ‘saya’, ‘mau’, ‘berhenti’, ‘berlangganan’, ‘indihome’, ‘mohon’, ‘ditindaklanjuti’	‘berhenti’, ‘berlangganan’, ‘indihome’, ‘mohon’, ‘ditindaklanjuti’
	8	‘indihome’, ‘knp’, ‘bermasalah’, ‘terus’, ‘sih’	‘indihome’, ‘knp’, ‘bermasalah’
	9	‘indihome’, ‘menyial’	‘indihome’, ‘menyial’
	10	‘anti’, ‘penanganan’, ‘lambat’, ‘terimakasih’, ‘pak’, ‘indihome’, ‘yang’, ‘selalu’, ‘benahi’, ‘jaringan’, ‘indihome’, ‘kita’, ‘sehingga’, ‘jaringannya’, ‘stabil’	‘anti’, ‘penanganan’, ‘lambat’, ‘indihome’, ‘benahi’, ‘jaringan’, ‘indihome’, ‘sehingga’, ‘jaringannya’, ‘stabil’

Tahap *normalization* merupakan tahapan untuk memperbaiki kata-kata yang disingkat, salah pengejaan, atau dalam bahasa gaul. Misalnya kata ‘lambat’ memiliki banyak bentuk penulisan seperti lmbt, lelet, lemot, dan sebagainya. Hasil *normalization* ditunjukkan pada **Tabel 5.6**.

Tabel 5.6 Ulasan Mengenai Indihome Setelah Melalui Proses *Normalization*

Bulan	No	*Tweet* Setelah Melalui *Filtering*	*Tweet* Setelah Melalui *Normalization*
Maret	1	‘keren’, ‘indihome’, ‘terbaik’	‘keren’, ‘indihome’, ‘terbaik’
	2	‘hadeh’, ‘indihome’	‘aduh’, ‘indihome’
	3	‘sobat’, ‘bantu’, ‘pengecekan’, ‘mohon’, ‘menginformasikan’, ‘nomor’, ‘internet’, ‘digit’, ‘angka’, ‘nomor’, ‘hp’, ‘aktif’, ‘dihubungi’, ‘ya’, ‘sobat’, ‘ari’	‘sobat’, ‘bantu’, ‘pengecekan’, ‘mohon’, ‘menginformasikan’, ‘nomor’, ‘internet’, ‘digit’, ‘angka’, ‘nomor’, ‘handphone’, ‘aktif’, ‘dihubungi’, ‘ya’, ‘sobat’, ‘ari’
	4	‘malam’, ‘gangguan’, ‘kah’	‘malam’, ‘gangguan’, ‘kah’
	5	‘lanjutkan’, ‘tips’, ‘bagus’, ‘gampang’	‘lanjutkan’, ‘tips’, ‘bagus’, ‘gampang’
April	6	‘lemot’, ‘banget’, ‘jaringan’	‘lambat’, ‘banget’, ‘jaringan’
	7	‘berhenti’, ‘berlangganan’, ‘indihome’, ‘mohon’, ‘ditindaklanjuti’	‘berhenti’, ‘berlangganan’, ‘indihome’, ‘mohon’, ‘ditindaklanjuti’
	8	‘indihome’, ‘knp’, ‘bermasalah’	‘indihome’, ‘kenapa’, ‘bermasalah’
	9	‘indihome’, ‘menyial’	‘indihome’, ‘menyial’
	10	‘anti’, ‘penanganan’, ‘lambat’, ‘indihome’, ‘benahi’, ‘jaringan’, ‘indihome’, ‘sehingga’, ‘jaringannya’, ‘stabil’	‘anti’, ‘penanganan’, ‘lambat’, ‘indihome’, ‘benahi’, ‘jaringan’, ‘indihome’, ‘sehingga’, ‘jaringannya’, ‘stabil’

5.2.4 *Stemming*

Tahapan terakhir dalam *preprocessing data,* yaitu proses *stemming*. *Stemming* merupakan proses mengubah kata yang memiliki imbuhan menjadi kata dasar. Hal ini dilakukan supaya kata tersebut dapat direpresentasikan sama rata dengan kata lainnya yang tidak memiliki imbuhan.

Tabel 5.7 Ulasan Mengenai Indihome Setelah Melalui Proses *Stemming*

Bulan	No	*Tweet* Setelah Melalui *Normalization*	*Tweet* Setelah Melalui *Stemming*
Maret	1	‘keren’, ‘indihome’, ‘terbaik’	‘keren’, ‘indihome’, ‘baik’
	2	‘aduh’, ‘indihome’	‘aduh’, ‘indihome’
	3	‘sobat’, ‘bantu’, ‘pengecekan’, ‘mohon’, ‘menginformasikan’, ‘nomor’, ‘internet’, ‘digit’,	‘sobat’, ‘bantu’, ‘kece’, ‘mohon’, ‘informasi’, ‘nomor’, ‘internet’, ‘digit’, ‘angka’, ‘nomor’,

Bulan	No	*Tweet* Setelah Melalui *Normalization*	*Tweet* Setelah Melalui *Stemming*
Maret		‘angka’, ‘nomor’, ‘handphone’, ‘aktif’, ‘dihubungi’, ‘ya’, ‘sobat’, ‘ari’	‘handphone’, ‘aktif’, ‘hubung’, ‘ya’, ‘sobat’, ‘ari’
	4	‘malam’, ‘gangguan’, ‘kah’	‘malam’, ‘ganggu’, ‘kah’
	5	‘lanjutkan’, ‘tips’, ‘bagus’, ‘gampang’	‘lanjut’, ‘tips’, ‘bagus’, ‘gampang’
April	6	‘lambat’, ‘banget’, ‘jaringan’	‘lambat’, ‘banget’, ‘jaring’
	7	‘berhenti’, ‘berlangganan’, ‘indihome’, ‘mohon’, ‘ditindaklanjuti’	‘henti’, ‘berlanganan’, ‘indihome’, ‘mohon’, ‘ditindaklanjuti’
	8	‘indihome’, ‘kenapa’, ‘bermasalah’	‘indihome’, ‘kenapa’, ‘masalah’
	9	‘indihome’, ‘menyial’	‘indihome’, ‘sial’
	10	‘anti’, ‘penanganan’, ‘lambat’, ‘indihome’, ‘benahi’, ‘jaringan’, ‘indihome’, ‘sehingga’, ‘jaringannya’, ‘stabil’	‘anti’, ‘tangan’, ‘lambat’, ‘indihome’, ‘benah’, ‘jaring’, ‘indihome’, ‘sehingga’, ‘jaringannya’, ‘stabil’

Tahapan *preprocessing data* telah selesai dilakukan hingga menjadi lebih terstruktur seperti pada hasil terakhir yang ditunjukkan pada **Tabel 5.7**. Hal ini dapat mempermudah pada tahapan selanjutnya untuk melakukan proses klasifikasi.

5.3. Pelabelan Kelas Sentimen

Pada penelitian ini, penulis menggunakan teknik *hybrid approach*, yakni penggabungan antara pendekatan *lexicon* dan *machine learning*, di mana pendekatan *lexicon* digunakan untuk membantu dalam melabeli kelas sentimen, yang selanjutnya akan digunakan pendekatan *machine learning* untuk menguji kinerja dari analisis model yang dihasilkan. Penggunaan *hybrid approach* ini dilakukan, karena data yang dikumpulkan sebelumnya belum terlabeli dengan kelas sentimen apapun sehingga menggunakan bantuan *lexicon approach*, serta penulis ingin mengetahui lebih lanjut kinerja model yang dihasilkan menggunakan metode SVM pada *machine learning approach*. Sehingga, ketika kedua pendekatan tersebut digabungkan dapat berpotensi meningkatkan kinerja klasifikasi sentimen.

Setelah membuat data *tweet* menjadi lebih terstruktur, tahapan selanjutnya yang penting dilakukan adalah pelabelan kelas sentimen yang dibagi menjadi 3 kelas, yaitu sentimen positif, sentimen netral, dan sentimen negatif. Pelabelan atau

pemberian skor terhadap data *tweet* ini bertujuan sebagai proses pembelajaran dari data latih yang kemudian akan dipelajari oleh *machine learning* dengan algoritma SVM untuk melakukan klasifikasi. Proses pelabelan tersebut dilakukan dengan pembobotan *lexicon based* yang mempertimbangkan skor sentimen secara otomatis. Cara melakukan perhitungan skor sentimen tersebut, yaitu mencari jumlah kata dengan sentimen positif dan jumlah kata dengan sentimen negatif. Skor sentimen diperoleh dengan mengurangkan jumlah polaritas sentimen negatif dari jumlah polaritas sentimen positif (Luqyana, et al., 2018). Hasil perbandingan jumlah pelabelan sentimen ditunjukkan pada **Tabel 5.8**.

Tabel 5.8 Perbandingan Hasil Sentimen Bulan Maret dan April 2021

Sentimen	**Bulan**	
	Maret	**April**
Positif	16730	15772
Negatif	23766	19774
Netral	16448	15517
Jumlah	56944	51063

Berdasarkan **Tabel 5.8**, didapatkan hasil pelabelan kelas pada bulan Maret 2021, yaitu sentimen positif sebanyak 16730 ulasan, sentimen negatif sebanyak 23766 ulasan, dan sisanya 16448 ulasan merupakan sentimen netral. Sedangkan pada bulan April 2021, sentimen positif sebanyak 15772 ulasan, sentimen negatif sebanyak 19774 ulasan, dan sentimen netral sebanyak 15517 ulasan. Hal ini menunjukkan sentimen negatif lebih dominan dihasilkan, dibandingkan sentimen lainnya. Contoh beberapa ulasan yang telah diberi skor serta label ditunjukkan oleh **Tabel 5.9**.

Tabel 5.9 Hasil Pelabelan *Tweet* Ulasan

Bulan	**No**	**Sentimen**	**Skor**	**Tweet Ulasan**
Maret	1	Positif	2	@jarvis8_ keren kak , indihome memang yang terbaik ðŸ¥°ðŸ¥°
	2	Negatif	-1	hadeh indihome kenapa lagi ini.
	3	Positif	2	"@IndiHome Lanjutkan min. Tips yang bagus. Bisa lebih gampang deh. #InternetnyaIndonesia #MajuTerusBersamaIndiHome"
April	4	Negatif	-1	@IndiHome lemot banget jaringan!!!!
	5	Negatif	-1	Indihome knp bermasalah terus sih:(
	6	Positif	1	"Anti penanganan Lambat

Bulan	No	Sentimen	Skor	Tweet Ulasan
April				Terimakasih pak Indihome yang selalu benahi jaringan Indihome kita sehingga jaringan nya stabil ..... https://t.co/Fc6hXKCSlD"

Hasil perhitungan skor sentimen yang didapatkan pada **Tabel 5.9** menggunakan perhitungan sentimen *lexicon* seperti pada **persamaan 3.1** yang telah dipaparkan pada bab sebelumnya. Pada **Tabel 5.10** ditampilkan simulasi perhitungan skor sentimen.

Tabel 5.10 Contoh Perhitungan Skor Sentimen

Ulasan	Kata Positif	Kata Negatif
keren kak indihome memang yang terbaik	keren	-
	baik	-
Jumlah Kata	2	0

Berdasarkan **Tabel 5.10**, maka diperoleh hasil dari perhitungan skor sentimen sebagai berikut:

Skor = (jumlah kata positif) – (jumlah kata negatif)

Skor = 2 – 0 = 2

Dikarenakan hasil akhir mendapatkan skor 2 atau positif 2 di mana > 0, maka ulasan tersebut dapat diklasifikasikan sebagai sentimen positif.

5.4. Klasifikasi

Klasifikasi yang dilakukan pada penelitian ini, dilakukan dengan menggunakan algoritma *Support Vector Machine* (SVM). Klasifikasi menggunakan algoritma SVM ini, akan dicoba menggunakan beberapa *kernel* (*Linear*, *Polynomial*, RBF, *Sigmoid*), untuk mengetahui hasil akurasi terbaik pada SVM.

5.4.1 Pembagian *Data Training* dan *Data Testing*

Setelah data *tweet* diberi label, selanjutnya akan dilakukan *split data*, yakni membagi data menjadi 2 bagian, yaitu data *training* dan data *testing*. Data *training* atau data latih digunakan untuk melatih algoritma hingga mendapatkan suatu model, sedangkan data *testing* atau data uji digunakan untuk menguji model yang dibentuk oleh data *training* ketika menemukan data baru. Data *tweet* yang akan dibagi menjadi data *training* dan data *testing*, merupakan data yang telah diberi label sentimen, dengan rasio perbandingan jumlah data 80% : 20%. Berdasarkan

prinsip Pareto (*Pareto Principle*), perbandingan rasio yang umumnya digunakan adalah 80% : 20%. Prinsip ini dapat membantu dalam pembagian data *training* dan data *testing* untuk banyak kejadian. Perbandingan jumlah data *training* dan data *testing* ditunjukkan pada **Tabel 5.11**.

Tabel 5.11 Pembagian Data *Training* dan Data *Testing* Bulan Maret 2021

Bulan	Sentimen	Data *Training* (80%)	Data *Test* (20%)	Jumlah
Maret	Netral	13158	3290	16448
	Positif	13384	3346	16730
	Negatif	19013	4753	23766
Jumlah		**45555**	**11389**	**56944**

Berdasarkan **Tabel 5.11** diketahui bahwa untuk data ulasan terhadap Indihome pada bulan Maret, terdapat 45555 ulasan digunakan sebagai data *training*, yang terdiri dari 13158 sentimen netral, 13384 sentimen positif, dan 19013 sentimen negatif. Sedangkan, sisanya sebanyak 11389 digunakan sebagai data *testing*, yang terdiri dari 3290 sentimen netral, 3346 sentimen positif, serta 4753 sentimen negatif.

Tabel 5.12 Pembagian Data *Training* dan Data *Testing* Bulan April 2021

Bulan	Sentimen	Data *Training* (80%)	Data *Test* (20%)	Jumlah
April	Netral	12414	3103	15517
	Positif	12618	3154	15772
	Negatif	15819	3955	19774
Jumlah		**40851**	**10212**	**51063**

Berdasarkan **Tabel 5.12** diketahui bahwa untuk data ulasan terhadap Indihome pada bulan April, terdapat 40851 ulasan digunakan sebagai data *training*, yang terdiri dari 12414 sentimen netral, 12618 sentimen positif, dan 15819 sentimen negatif. Sedangkan, sisanya sebanyak 10212 digunakan sebagai data *testing*, yang terdiri dari 3103 sentimen positif, 3154 sentimen positif, serta 3955 sentimen negatif.

Rasio perbandingan kelas sentimen pada data *training* baik bulan Maret maupun April, tidak terlalu memiliki perbedaan yang cukup jauh atau ekstrim, yakni dengan perbandingan 13158:13384:19013 untuk bulan Maret, sedangkan untuk bulan April rasio perbandingannya adalah 4138:4206:5273. Hal ini dapat dikatakan sebagai *slight imbalance* (ketidakseimbangan yang sedikit). Sehingga,

penulis memutuskan hal tersebut tidak terlalu menjadi perhatian, seperti yang dikatakan oleh (Brownlee, 2019).

5.4.2 Klasifikasi Algoritma *Support Vector Machine*

Tahapan klasifikasi yang dilakukan pada penelitian ini, menggunakan algoritma *Support Vector Machine* (SVM). Cara kerja algoritma SVM, yaitu dengan mempelajari data *training* yang telah dibentuk atau ditentukan sebelumnya. Algoritma SVM akan mempelajari pola yang dibentuk dari data *training* itu sendiri, yang kemudian akan dilakukan pelatihan untuk mengklasifikasikan karakteristik dari masing-masing kelas/kategori pada data tersebut. Tahapan ini merupakan salah satu tugas dari *machine learning* (ML) yang dikenal sebagai *supervised machine learning*.

Hasil pelatihan dari algoritma SVM akan mendapatkan suatu model, di mana model tersebut akan dilakukan pengujian dengan menggunakan data *testing*. Hal ini dilakukan agar mengetahui nilai akurasi yang dihasilkan oleh model tersebut, serta dapat memprediksi jika diberi data baru. Pada penelitian ini, dilakukan beberapa percobaan menggunakan metode *kernel* pada SVM, yaitu *kernel Linear*, *Polynomial*, *Radial Basis Function* (RBF), dan *Sigmoid*. Diantara keempat metode tersebut, akan dipilih metode terbaik yang memiliki nilai akurasi tertinggi. Berikut adalah hasil dan pembahasan dari percobaan masing-masing metode *kernel* pada algoritma SVM:

1. *Kernel Linear*

Percobaan pertama pada penelitian ini menggunakan algoritma SVM dengan metode *kernel linear*. Berikut adalah hasil perhitungan manual nilai akurasi, *precision*, dan *recall* untuk metode *kernel linear*:

$$Akurasi_{linear} = \frac{jumlah\ data\ diprediksi\ benar}{jumlah\ seluruh\ data} \times 100\%$$

$$= \frac{3965 + 2770 + 2936}{11389} \times 100\% = 0{,}8492 = 84{,}92\%$$

Kelas Negatif

$$Precision_{Negatif} = \frac{TP}{TP + FP} = \frac{3965}{3965 + 282} = 0{,}9336$$

$$Recall_{Negatif} = \frac{TP}{TP + FN} = \frac{3965}{3965 + 788} = 0{,}8342$$

Kelas Netral

$$Precision_{Netral} = \frac{TP}{TP + FP} = \frac{2770}{2770 + 1152} = 0{,}7063$$

$$Recall_{Netral} = \frac{TP}{TP + FN} = \frac{2770}{2770 + 520} = 0{,}8419$$

Kelas Positif

$$Precision_{Positif} = \frac{TP}{TP + FP} = \frac{2936}{2936 + 284} = 0{,}9118$$

$$Recall_{Positif} = \frac{TP}{TP + FN} = \frac{2936}{2936 + 410} = 0{,}8775$$

$$Precision_{linear} = \frac{Precision\ Negatif + Netral + Positif}{Jumlah\ Kelas} \times 100\%$$

$$= \frac{0{,}9336 + 0{,}7063 + 0{,}9118}{3} \times 100\% = 0{,}8506 = 85{,}06\%$$

$$Recall_{linear} = \frac{Recall\ Negatif + Netral + Positif}{Jumlah\ Kelas} \times 100\%$$

$$= \frac{0{,}8342 + 0{,}8419 + 0{,}8775}{3} \times 100\% = 0{,}8512 = 85{,}12\%$$

Berdasarkan perhitungan di atas, tingkat akurasi yang dihasilkan dengan menggunakan metode *kernel linear* untuk bulan Maret, yaitu sebesar 84,92%, yang berarti bahwa ada sebanyak 84,92% data uji, berhasil diprediksi secara benar sebagai sentimen positif, negatif, dan netral pada keseluruhan data. Nilai rata-rata *precision* sebesar 85,06%, yang berarti bahwa rata-rata ada sebanyak 85,06% dari data klasifikasi yang memang sebenarnya dikategorikan sebagai sentimen yang relevan. Sedangkan nilai rata-rata *recall* sebesar 85,12%, yang berarti bahwa rata-rata ada sebanyak 85,12% dari data klasifikasi, mampu memprediksi sentimen yang relevan. Hasil *confusion matrix* untuk bulan Maret dan April 2021 dengan metode *linear* ditampilkan pada **Tabel 5.13** dan **Tabel 5.14**.

Tabel 5.13 *Confusion Matrix* Metode *Kernel Linear* Ulasan Bulan Maret 2021

Prediksi	**Aktual**			***Precision***
	Negatif	Netral	Positif	
Negatif	3965	267	15	93,36%
Netral	757	2770	395	70,63%
Positif	31	253	2936	91,18%
Recall	83,42%	84,19%	87,75%	
Akurasi	84,92%			

Berdasarkan **Tabel 5.13**, pada *confusion matrix* metode *kernel linear* untuk bulan Maret menghasilkan 3965 data sentimen negatif yang berhasil diklasifikasikan secara benar, dari total aktual sentimen negatif sebanyak 4753 data, dan sisanya sebanyak 757 data diklasifikasikan sebagai sentimen netral serta 31 data sebagai sentimen positif. Pada total 3346 data yang bersentimen positif, 2936 data diantaranya berhasil diklasifikasikan secara tepat sebagai sentimen positif, dan sisanya sebanyak 15 data diklasifikasikan sebagai sentimen negatif, lalu 395 data lainnya sebagai sentimen netral.

Sedangkan, dari total aktual sentimen netral sebanyak 3290 data, 2770 data diantaranya berhasil diklasifikasikan secara tepat, serta sisanya sebanyak 267 data diklasifikasikan sebagai sentimen negatif dan 253 data lainnya diklasifikasikan sebagai sentimen positif .

Tabel 5.14 *Confusion Matrix* Metode *Kernel Linear* Ulasan Bulan April 2021

Prediksi	**Aktual**			***Precision***
	Negatif	Netral	Positif	
Negatif	3449	204	6	94,26%
Netral	497	2779	214	79,63%
Positif	9	120	2934	95,79%
Recall	87,21%	89,56%	93,02%	
Akurasi	89,72%			

Berdasarkan **Tabel 5.14**, pada *confusion matrix* metode *kernel linear* untuk bulan April menghasilkan 3449 data sentimen negatif yang berhasil diklasifikasikan secara benar, dari total aktual sentimen negatif sebanyak 3955 data, serta sisanya 9 data diklasifikasikan sebagai sentimen positif dan 497 data sebagai sentimen netral. Pada total 3154 data yang bersentimen positif, 2934 data diantaranya berhasil diklasifikasikan secara tepat sebagai sentimen positif, lalu sisanya sebanyak 6 data diklasifikasikan sebagai sentimen negatif dan 214 data sebagai sentimen netral.

Sedangkan, dari total aktual sentimen netral sebanyak 3103 data, 2779 data diantaranya berhasil diklasifikasikan secara tepat, serta sisanya sebanyak 204 data diklasifikasikan sebagai sentimen negatif dan 120 data lainnya diklasifikasikan sebagai sentimen positif. Tingkat akurasi yang dihasilkan dengan menggunakan metode *kernel linear* untuk bulan April, yaitu sebesar 89,72%, yang berarti bahwa ada sebanyak 89,72% data uji, berhasil diprediksi secara benar sebagai sentimen positif, negatif, dan netral pada keseluruhan data. Nilai rata-rata *precision* sebesar

89,89%, yang berarti bahwa rata-rata ada sebanyak 89,89% dari data klasifikasi yang memang sebenarnya dikategorikan sebagai sentimen yang relevan. Sedangkan nilai rata-rata *recall* sebesar 89,93%, yang berarti bahwa rata-rata ada sebanyak 89,93% dari data klasifikasi, mampu memprediksi sentimen yang relevan.

2. *Kernel Polynomial*

Hasil klasifikasi untuk bulan Maret dan April 2021 menggunakan metode *kernel polynomial*, ditampilkan pada **Tabel 5.15** dan **Tabel 5.16**.

Tabel 5.15 *Confusion Matrix* Metode *Kernel Polynomial* Ulasan Bulan Maret 2021

Prediksi	**Aktual**			***Precision***
	Negatif	Netral	Positif	
Negatif	4509	2952	1594	49,80%
Netral	47	104	44	53,33%
Positif	197	234	1708	79,85%
Recall	94,87%	3,16%	51,05%	
Akurasi	55,50%			

Berdasarkan **Tabel 5.15**, *confusion matrix* metode *kernel polynomial* untuk bulan Maret menunjukkan dari total sentimen negatif yang sebenarnya sebanyak 4753 data, didapatkan hasil klasifikasi sebanyak 4509 data yang diklasifikasikan secara tepat dan benar, sedangkan sisanya 197 data diklasifikasikan sebagai sentimen positif dan 47 data lainnya sebagai sentimen netral. Lalu, untuk total sentimen positif secara aktual sebanyak 3346 data, 1708 data diantaranya berhasil diklasifikasikan secara tepat sebagai sentimen positif, serta sisanya sebanyak 1594 data diklasifikasikan sebagai sentimen negatif dan 44 data lainnya sebagai sentimen netral.

Sedangkan, dari total aktual sentimen netral sebanyak 3290 data, 104 data diantaranya berhasil diklasifikasikan secara tepat, serta sisanya sebanyak 2952 data diklasifikasikan sebagai sentimen negatif dan 234 data lainnya diklasifikasikan sebagai sentimen positif. Tingkat akurasi yang dihasilkan dengan menggunakan metode *kernel polynomial* untuk bulan Maret, yaitu sebesar 55,50%, yang berarti bahwa ada sebanyak 55,50% data uji, berhasil diprediksi secara benar sebagai sentimen positif, negatif, dan netral pada keseluruhan data. Nilai rata-rata *precision* sebesar 60,99%, yang berarti bahwa rata-rata ada sebanyak 60,99% dari data klasifikasi yang memang sebenarnya dikategorikan sebagai sentimen yang relevan.

Sedangkan nilai rata-rata *recall* sebesar 49,70%, yang berarti bahwa rata-rata ada sebanyak 49,70% dari data klasifikasi, mampu memprediksi sentimen yang relevan.

Tabel 5.16 *Confusion Matrix* Metode *Kernel Polynomial* Ulasan Bulan April 2021

Prediksi	**Aktual**			***Precision***
	Negatif	Netral	Positif	
Negatif	3680	2696	1341	47,69%
Netral	225	346	817	24,93%
Positif	50	61	996	89,97%
Recall	93,05%	11,15%	31,58%	
Akurasi	49,18%			

Berdasarkan **Tabel 5.16**, *confusion matrix* metode *kernel polynomial* untuk bulan April menunjukkan dari total sentimen negatif yang sebenarnya sebanyak 3955 data, didapatkan hasil klasifikasi sebanyak 3680 data yang diklasifikasikan secara tepat dan benar, sedangkan sisanya 50 data diklasifikasikan sebagai sentimen positif dan 225 data lainnya sebagai sentimen netral. Kemudian, untuk total sentimen positif secara aktual sebanyak 3154 data, 996 data diantaranya berhasil diklasifikasikan secara tepat sebagai sentimen positif, serta sisanya sebanyak 1341 data diklasifikasikan sebagai sentimen negatif dan 817 data lainnya diklasifikasikan sebagai sentimen netral.

Sedangkan, dari total aktual sentimen netral sebanyak 3103 data, 346 data diantaranya berhasil diklasifikasikan secara tepat, serta sisanya sebanyak 2696 data diklasifikasikan sebagai sentimen negatif dan 61 data lainnya diklasifikasikan sebagai sentimen positif. Tingkat akurasi yang dihasilkan dengan menggunakan metode *kernel polynomial* untuk bulan April, yaitu sebesar 49,18%, yang berarti bahwa ada sebanyak 49,18% data uji, berhasil diprediksi secara benar sebagai sentimen positif, negatif, dan netral pada keseluruhan data. Nilai rata-rata *precision* sebesar 54,20%, yang berarti bahwa rata-rata ada sebanyak 54,20% dari data klasifikasi yang memang sebenarnya dikategorikan sebagai sentimen yang relevan. Sedangkan nilai rata-rata *recall* sebesar 45,26%, yang berarti bahwa rata-rata ada sebanyak 45,26% dari data klasifikasi, mampu memprediksi sentimen yang relevan.

3. *Kernel Radial Basis Function* (RBF)

Hasil klasifikasi untuk bulan Maret dan April 2021 menggunakan metode *kernel* RBF, ditampilkan pada **Tabel 5.17** dan **Tabel 5.18**.

Tabel 5.17 *Confusion Matrix* Metode *Kernel* RBF Ulasan Bulan Maret 2021

Prediksi	**Aktual**			***Precision***
	Negatif	Netral	Positif	
Negatif	4162	251	10	94,10%
Netral	571	2778	200	78,28%
Positif	20	261	3136	91,78%
Recall	87,57%	84,44%	93,72%	
Akurasi	88,47%			

Berdasarkan **Tabel 5.17**, pada *confusion matrix* metode *kernel* RBF untuk bulan Maret menghasilkan 3136 data sentimen positif yang berhasil diklasifikasikan secara benar, dari total aktual sentimen positif sebanyak 3346 data, serta sisanya 10 data diklasifikasikan sebagai sentiment negatif dan 200 data lainnya sebagai sentimen netral. Lalu, untuk total 4753 data yang bersentimen negatif, 4162 data diantaranya berhasil diklasifikasikan secara tepat sebagai sentimen negatif, serta sisanya sebanyak 20 data diklasifikasikan sebagai sentimen positif dan 571 data lainnya sebagai sentimen netral.

Sedangkan, dari total aktual sentimen netral sebanyak 3290 data, 2778 data diantaranya berhasil diklasifikasikan secara tepat, serta sisanya sebanyak 251 data diklasifikasikan sebagai sentimen negatif dan 261 data lainnya diklasifikasikan sebagai sentimen positif. Tingkat akurasi yang dihasilkan dengan menggunakan metode *kernel* RBF untuk bulan Maret, yaitu sebesar 88,47%, yang berarti bahwa ada sebanyak 88,47% data uji, berhasil diprediksi secara benar sebagai sentimen positif, negatif, dan netral pada keseluruhan data. Nilai rata-rata *precision* sebesar 88,05%, yang berarti bahwa rata-rata ada sebanyak 88,05% dari data klasifikasi yang memang sebenarnya dikategorikan sebagai sentimen yang relevan. Sedangkan nilai rata-rata *recall* sebesar 88,58%, yang berarti bahwa rata-rata ada sebanyak 88,58% dari data klasifikasi, mampu memprediksi sentimen yang relevan.

Tabel 5.18 *Confusion Matrix* Metode *Kernel* RBF Ulasan Bulan April 2021

Prediksi	**Aktual**			***Precision***
	Negatif	Netral	Positif	
Negatif	3873	25	0	99,36%

Prediksi	Aktual			*Precision*
	Negatif	Netral	Positif	
Netral	79	3033	46	96,04%
Positif	3	45	3108	98,48%
Recall	97,93%	97,74%	98,54%	
Akurasi	98,06%			

Berdasarkan **Tabel 5.18**, pada *confusion matrix* metode *kernel* RBF untuk bulan April menghasilkan 3108 data sentimen positif yang berhasil diklasifikasikan secara benar, dari total aktual sentimen positif sebanyak 3154 data, serta sisanya 46 data diklasifikasikan sebagai sentimen netral. Kemudian, untuk total 3955 data yang bersentimen negatif, 3873 data diantaranya berhasil diklasifikasikan secara tepat sebagai sentimen negatif, serta sisanya 79 data diklasifikasikan sebagai sentimen netral dan 3 data diklasifikasikan sebagai sentimen positif.

Sedangkan, dari total aktual sentimen netral sebanyak 3103 data, 3033 data diantaranya berhasil diklasifikasikan secara tepat, serta sisanya sebanyak 25 data diklasifikasikan sebagai sentimen negatif dan 45 data lainnya diklasifikasikan sebagai sentimen positif. Tingkat akurasi yang dihasilkan dengan menggunakan metode *kernel* RBF untuk bulan April, yaitu sebesar 98,06%, yang berarti bahwa ada sebanyak 98,06% data uji, berhasil diprediksi secara benar sebagai sentimen positif, negatif, dan netral pada keseluruhan data. Nilai rata-rata *precision* sebesar 97,96%, yang berarti bahwa rata-rata ada sebanyak 97,96% dari data klasifikasi yang memang sebenarnya dikategorikan sebagai sentimen yang relevan. Sedangkan nilai rata-rata *recall* sebesar 98,07%, yang berarti bahwa rata-rata ada sebanyak 98,07% dari data klasifikasi, mampu memprediksi sentimen yang relevan.

4. *Kernel Sigmoid*

Hasil klasifikasi untuk bulan Maret dan April 2021 menggunakan metode *kernel sigmoid*, ditampilkan pada **Tabel 5.19** dan **Tabel 5.20**.

Tabel 5.19 *Confusion Matrix* Metode *Kernel Sigmoid* Ulasan Bulan Maret 2021

Prediksi	Aktual			*Precision*
	Negatif	Netral	Positif	
Negatif	4161	250	8	94,16%
Netral	573	2752	222	77,59%
Positif	19	288	3116	91,03%
Recall	87,54%	83,65%	93,13%	

Prediksi	Aktual			*Precision*
	Negatif	Netral	Positif	
Akurasi	88,06%			

Berdasarkan **Tabel 5.19**, *confusion matrix* metode *kernel sigmoid* untuk bulan Maret menunjukkan dari total sentimen positif yang sebenarnya sebanyak 3346 data, didapatkan hasil klasifikasi sebanyak 3116 data yang diklasifikasikan secara tepat dan benar, serta sisanya 8 data diklasifikasikan sebagai sentimen negatif dan 222 data lainnya sebagai sentimen netral. Lalu, untuk total sentimen negatif secara aktual sebanyak 4753 data, 4161 data diantaranya berhasil diklasifikasikan secara tepat sebagai sentimen negatif, serta sisanya sebanyak 19 data diklasifikasikan sebagai sentimen positif dan 573 data lainnya diklasifikasikan sebagai sentimen netral.

Sedangkan, dari total aktual sentimen netral sebanyak 3290 data, 2752 data diantaranya berhasil diklasifikasikan secara tepat, serta sisanya sebanyak 250 data diklasifikasikan sebagai sentimen negatif dan 288 data lainnya diklasifikasikan sebagai sentimen positif. Tingkat akurasi yang dihasilkan dengan menggunakan metode *kernel sigmoid* untuk bulan Maret, yaitu sebesar 88,06%, yang berarti bahwa ada sebanyak 88,06% data uji, berhasil diprediksi secara benar sebagai sentimen positif, negatif, dan netral pada keseluruhan data. Nilai rata-rata *precision* sebesar 87,59%, yang berarti bahwa rata-rata ada sebanyak 87,59% dari data klasifikasi yang memang sebenarnya dikategorikan sebagai sentimen yang relevan. Sedangkan nilai rata-rata *recall* sebesar 88,11%, yang berarti bahwa rata-rata ada sebanyak 88,11% dari data klasifikasi, mampu memprediksi sentimen yang relevan.

Tabel 5.20 *Confusion Matrix* Metode *Kernel Sigmoid* Ulasan Bulan April 2021

Prediksi	Aktual			*Precision*
	Negatif	Netral	Positif	
Negatif	3854	35	0	99,10%
Netral	97	3022	54	95,24%
Positif	4	46	3100	98,41%
Recall	97,45%	97,39%	98,29%	
Akurasi	97,69%			

Berdasarkan **Tabel 5.20**, *confusion matrix* metode *kernel sigmoid* untuk bulan April menunjukkan dari total sentimen positif yang sebenarnya sebanyak 3154 data,

didapatkan hasil klasifikasi sebanyak 3100 data yang diklasifikasikan secara tepat dan benar, sedangkan sisanya 54 data diklasifikasikan sebagai sentimen netral. Sedangkan, untuk total sentimen negatif secara aktual sebanyak 3955 data, 3854 data diantaranya berhasil diklasifikasikan secara tepat sebagai sentimen negatif, serta sisanya sebanyak 97 data diklasifikasikan sebagai sentimen netral dan 4 data lainnya diklasifikasikan sebagai sentimen positif.

Sedangkan, dari total aktual sentimen netral sebanyak 3103 data, 3022 data diantaranya berhasil diklasifikasikan secara tepat, serta sisanya sebanyak 35 data diklasifikasikan sebagai sentimen negatif dan 46 data lainnya diklasifikasikan sebagai sentimen positif. Tingkat akurasi yang dihasilkan dengan menggunakan metode *kernel sigmoid* untuk bulan April, yaitu sebesar 97,69%, yang berarti bahwa ada sebanyak 97,69% data uji, berhasil diprediksi secara benar sebagai sentimen positif, negatif, dan netral pada keseluruhan data. Nilai rata-rata *precision* sebesar 97,58%, yang berarti bahwa rata-rata ada sebanyak 97,58% dari data klasifikasi yang memang sebenarnya dikategorikan sebagai sentimen yang relevan. Sedangkan nilai rata-rata *recall* sebesar 97,71%, yang berarti bahwa rata-rata ada sebanyak 97,71% dari data klasifikasi, mampu memprediksi sentimen yang relevan.

Perbandingan kinerja dari keempat metode *kernel* pada SVM yang telah dijalankan pada masing-masing bulan Maret dan April, akan dipilih kinerja terbaik dengan melihat tingkat akurasi yang tertinggi. Pada **Tabel 5.21** ditampilkan hasil perbandingan klasifikasi percobaan empat *kernel* menggunakan data *testing*.

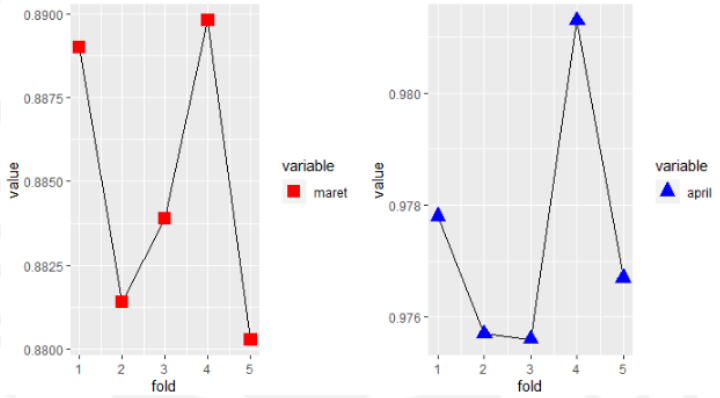
Tabel 5.21 Perbandingan Nilai Akurasi Metode *Kernel* pada SVM

Kernel	**Akurasi**	
	Maret	**April**
Linear	84,92%	89,72%
Polynomial	55,50%	49,18%
RBF	88,47%	98,06%
Sigmoid	88,06%	97,69%

Berdasarkan **Tabel 5.21**, dapat dilihat bahwa metode *kernel Radial Basis Function* (RBF) dengan algoritma *Support Vector Machine* baik pada bulan Maret maupun April, mendapatkan hasil tingkat akurasi tertinggi dibandingkan metode *kernel* lainnya, yaitu sebesar 88,47% dan 98,06%. Setelah mendapatkan model dari metode dengan tingkat akurasi tertinggi, akan dilakukan validasi dari model

tersebut. Hal ini bertujuan agar dapat mengetahui keefektifan serta konsistensi performa yang dihasilkan model tersebut. Maka dari itu, pada penelitian ini akan menggunakan metode *kernel* RBF untuk menuju pada tahapan *cross validation*.

Pada percobaan untuk menilai kinerja model dari metode *kernel* RBF, dibentuk 5-*fold cross validation*. Hal ini berarti akan dilakukan sebanyak 5 kali iterasi yang menghasilkan nilai akurasi. Hasil perbandingan nilai akurasi dari masing-masing percobaan tersebut pada bulan Maret dan April ditampilkan pada **Gambar 5.3**.



Gambar 5.3 Grafik Perbandingan Hasil *Cross Validation*

Berdasarkan **Gambar 5.3**, hasil akurasi yang dihasilkan setelah dilakukan 5-*fold cross validation* cukup konsisten. Terlihat pada grafik untuk bulan Maret yang ditandai dengan bentuk persegi warna merah, hasil akurasi yang dihasilkan masih dalam rentang 0,88 atau 88%. Begitu pula pada grafik untuk bulan April yang ditandai dengan bentuk segitiga warna biru, hasil akurasi masih dalam rentang 0,97 hingga 0,98 atau 97% hingga 98%. Sehingga, dapat dikatakan bahwa, baik dalam bulan Maret maupun April, performa dari model yang dihasilkan oleh metode *kernel* RBF cukup konsisten. Untuk melihat rincian hasil akurasi dari 5-*fold cross validation*, dapat dilihat pada **Tabel 5.22**.

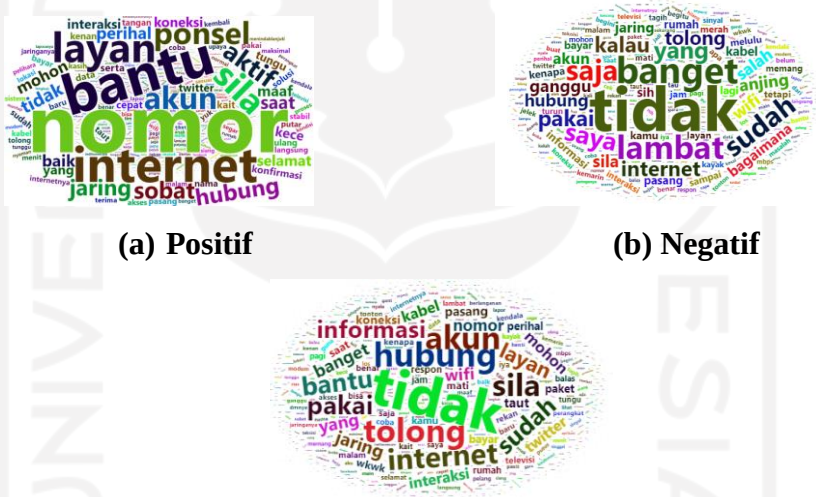
Tabel 5.22 Perbandingan Nilai Akurasi *Cross Validation*

Fold	**Akurasi**	
	Maret	**April**
1	88,90%	97,78%
2	88,14%	97,57%
3	88,39%	97,56%
4	88,98%	98,13%
5	88,03%	97,67%
Rata-Rata	88,49%	97,74%

Pada **Tabel 5.22** merupakan hasil akurasi 5-*fold cross validation* untuk bulan Maret dan April 2021. Rata-rata hasil akurasi yang dihasilkan dengan menggunakan metode kernel RBF pada bulan Maret yaitu sebesar 88,49%. Sedangkan hasil rata-rata akurasi pada bulan April yaitu sebesar 97,74%

5.5. *Word Cloud*

Word cloud merupakan representasi data berupa teks dengan memeriksa frekuensi kemunculan kata pada dokumen. Pada penelitian ini, akan diketahui kata-kata yang banyak diperbincangkan mengenai Indihome. Tampilan *word cloud* yang berisis ulasan mengenai Indihome pada bulan Maret berdasarkan kelas sentimennya, yaitu kelas positif, kelas negatif, dan kelas netral, tertera pada **Gambar 5.4**.



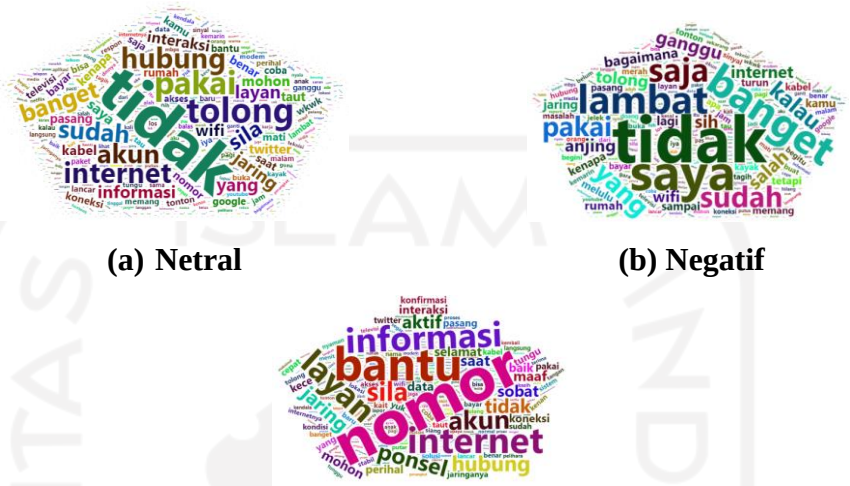
(a) Positif **(b) Negatif**

(c) Netral

Gambar 5.4 *Word Cloud* Data Bulan Maret

Pada **Gambar 5.4** bagian **(a)** merupakan tampilan *word cloud* bulan Maret untuk kelas positif, di mana kata yang banyak dibicarakan tentang IndiHome pada *platform* Twitter berupa kata "nomor", "internet", "bantu", "informasi", dan lain sebagainya. Pada **Gambar 5.4** bagian **(b)** yang merupakan tampilan *word cloud* untuk kelas negatif. Ulasan negatif yang sering dibicarakan mengenai Indihome banyak mengandung kata "tidak", "lambat", "banget", "sudah", dan lain sebagainya. Sedangkan pada **Gambar 5.4** bagian **(c)**, menampilkan *word cloud* untuk kelas netral. Ulasan netral banyak mengandung kata, seperti "tidak", "tolong", "hubung", "akun", dan lain sebagainya.

Word cloud untuk bulan April memiliki kemiripan dengan hasil *word cloud* pada bulan Maret dan tidak terlalu berbeda jauh. Tampilan *word cloud* untuk sentimen positif, negatif, dan netral pada bulan April, dapat dilihan pada **Gambar 5.5**.

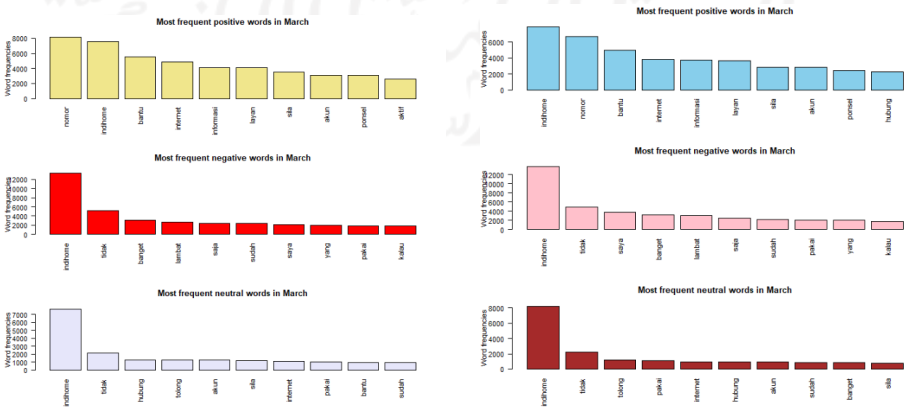


(a) Netral **(b) Negatif**

(c) Positif

Gambar 5.5 *Word Cloud* Data Bulan April

Pada **Gambar 5.5** bagian **(a)** merupakan tampilan *word cloud* bulan April untuk kelas netral, bagian **(b)** untuk kelas negatif, dan bagian **(c)** untuk kelas positif. Hasil *word cloud* yang ditampilkan untuk sentimen positif, negatif, dan netral pada bulan April tidak terlalu memiliki perbedaan yang signifikan dengan bulan sebelumnya. Untuk melihat visualisasi banyaknya frekuensi kata yang paling banyak disebutkan, dapat dilihat pada **Gambar 5.6**. Sedangkan untuk mengetahui jumlah kata-kata tersebut muncul pada keseluruhan dokumen, dapat dilihat pada **Tabel 5.23** dan **Tabel 5.24**.



(a) Maret **(b) April**

Gambar 5.6 *Barplot Most Frequency Word* Bulan Maret dan April

Tabel 5.23 *Most Frequency Word* Bulan Maret

Maret					
Positif		**Negatif**		**Netral**	
Word	***Freq***	***Word***	***Freq***	***Word***	***Freq***
nomor	8125	indihome	13350	indihome	7661
indihome	7541	tidak	5204	tidak	2106
bantu	5509	banget	3078	hubung	1279
internet	4828	lambat	2735	tolong	1251
informasi	4121	saja	2408	akun	1241

Tabel 5.24 *Most Frequency Word* Bulan April

April					
Positif		**Negatif**		**Netral**	
Word	***Freq***	***Word***	***Freq***	***Word***	***Freq***
indihome	7882	indihome	13735	indihome	8229
nomor	6681	tidak	4828	tidak	2197
bantu	4972	saya	3821	tolong	1170
internet	3851	banget	3203	pakai	1089
informasi	3723	lambat	2967	internet	969

Berdasarkan pada **Gambar 5.6** baik pada bulan Maret dan April, kata yang paling banyak disebutkan adalah "indihome". Hal ini sudah seharusnya dikarenakan penelitian ini berfokus untuk menganalisis tentang indihome. Namun, dengan mengesampingkan kata "indihome", penulis akan menggunakan beberapa kata lainnya untuk dilanjutkan ke tahap selanjutnya.

5.6. Asosiasi Kata

Asosiasi kata merupakan analisis lanjutan yang akan penulis lakukan setelah membuat *word cloud*. Asosiasi kata itu sendiri ialah analisis hubungan antara istilah atau kata yang sering muncul dengan melihat nilai korelasinya. Semakin besar nilai korelasi yang dihasilkan antara dua buah kata, maka kemungkinan kedua kata tersebut sering disebutkan secara bersamaan dalam suatu ulasan pun semakin besar.

Asosiasi kata diambil dari beberapa kata yang sering disebutkan baik pada sentimen netral, positif, maupun negatif di mana telah diketahui melalui hasil dari *word cloud* yang telah ditampilkan sebelumnya. Asosiasi yang dibentuk berupa sentimen positif, sentimen negatif, dan sentimen netral pada masing-masing bulan Maret dan April 2021, ditampilkan pada **Tabel 5.25** dan **Tabel 5.26**.

Tabel 5.25 Asosiasi Kata Bulan Maret 2021

Nomor (positif)		**Banget (negatif)**		**Bantu (netral)**	
ponsel	0,81	lambat	0,29	segar	0,29
aktif	0,67	jelek	0,16	respon	0,27
informasi	0,46	kesal	0,14	rekan	0,26
serta	0,34	indihome	0,11	kait	0,25
nama	0,31	parah	0,09	interaksi	0,25
Internet (positif)		**Hubung (negatif)**		**Informasi (netral)**	
ponsel	0,49	interaksi	0,46	layan	0,27
aktif	0,47	akun	0,42	akun	0,27
informasi	0,32	twitter	0,27	rinci	0,23
akses	0,27	perangkat	0,26	nomor	0,22
maksimal	0,26	kurang	0,18	nama	0,22
Bantu (positif)		**Sinyal (negatif)**		**Internet (netral)**	
informasi	0,24	hilang	0,23	nomor	0,28
perihal	0,22	jelek	0,15	mati	0,18
sistem	0,22	memuat	0,07	akses	0,18
aktif	0,21	telkomsel	0,07	lampu	0,18
akun	0,21	hujan	0,07	koneksi	0,15

Berdasarkan **Tabel 5.25**, penulis mengambil beberapa kata yang cukup sering disebutkan untuk sentimen positif yaitu “nomor”, “internet”, dan “bantu” yang menyangkut tentang Indihome pada bulan Maret 2021. Berdasarkan nilai asosiasi atau korelasi yang diperoleh untuk ulasan positif tersebut didapatkan suatu informasi antara lain, pelanggan Indihome melalui *platform* Twitter menyampaikan bahwa pelayanan yang diberikan oleh akun Indihome di Twitter dapat membantu memberikan solusi terhadap akses internet yang bermasalah seperti misalnya internet yang tidak aktif. Tidak hanya itu, sesekali informasi yang diberikan oleh akun Indihome cukup informatif.

Pada sentimen negatif, banyak ulasan yang berkaitan dengan sinyal yang hilang, jelek dan juga lambat saat diakses atau bahkan tidak bisa diakses sama sekali. Jika pelanggan mengalami hal tersebut, biasanya pelanggan akan menggunakan *platform* Twitter untuk meluapkan curahan rasa kekecewaannya terhadap Indihome. Hal ini selain dikarenakan *admin* akun Indihome yang cukup aktif menanggapi keluhan pelanggannya, bisa juga karena Indihome sering kali disalahkan jika sinyal internet sedang bermasalah.

Lalu pada sentimen netral, bisa dibilang memiliki kemiripan dengan sentimen positif, hanya saja ulasan nya tidak mengarah ke sentimen positif maupun negatif.

Admin dari akun Twitter Indihome seringkali berinteraksi dengan pengguna Twitter lainnya dan merespon cuitannya guna menanggapi serta melayani kendala yang dialami oleh pelanggannya tersebut. Biasanya *admin* dari akun Twitter Indihome akan mengarahkan kepada pengguna Twitter lainnya untuk memberikan rincian seperti nomor internet dan juga nama pelanggan melalui salah satu fitur dari Twitter, yaitu *direct message*.

Tabel 5.26 Asosiasi Kata Bulan April 2021

Nomor (positif)		**Bayar (negatif)**		**Twitter (netral)**	
ponsel	0,76	telat	0,28	akun	0,33
aktif	0,61	tagih	0,28	taut	0,27
data	0,34	mahal	0,15	facebook	0,26
nama	0,29	denda	0,15	interaksi	0,22
bantu	0,27	uang	0,13	indihomecare	0,20
Informasi (positif)		**Pakai (negatif)**		**Akun (netral)**	
ponsel	0,48	lancar	0,23	interaksi	0,47
aktif	0,45	tonton	0,16	taut	0,36
akun	0,33	netflix	0,16	alih	0,28
data	0,31	tidak	0,15	informasi	0,22
bantu	0,25	indihome	0,13	indihomecare	0,21
Layan (positif)		**Sinyal (negatif)**		**Hubung (netral)**	
aktif	0,29	hilang	0,22	akun	0,44
ponsel	0,24	jelek	0,15	interaksi	0,42
bantu	0,22	alir	0,15	twitter	0,25
pertanyan	0,19	telkomsel	0,09	perangkat	0,19
langsung	0,19	tidak	0,08	layan	0,16

Berdasarkan **Tabel 5.26**, ulasan positif tentang Indihome pada bulan April 2021 sering dikaitkan dengan layanan yang dilakukan *admin* akun Twitter Indihome dalam membantu menangani akses internet secara langsung dengan tenang. Beberapa kata yang sering disebutkan dari ulasan positif pada bulan April, tidak terlalu berbeda jauh dengan ulasan positif pada bulan Maret. Sama halnya seperti, akun Twitter Indihome yang masih memberikan informasi-informasi yang bermanfaat bagi para pengikutnya.

Pada sentimen negatif bulan April, ulasan yang diberikan masih tetap sama seperti bulan sebelumnya, di mana banyak keluhan yang membicarakan bahwa sinyal internet dari Indihome sangat jelek, dan juga sinyalnya sering hilang. Begitu pula ketika internet sedang dipakai pelanggan untuk menonton pada salah satu *platform*, seperti Netflix, dimana sinyalnya sering tidak lancar. Tatkala hal tersebut

terjadi, pelanggan Indihome juga mengeluhkan tentang biaya atau pembayaran yang harus dilakukan begitu mahal. Padahal yang dinikmati dari kinerja dan layanan Indihome itu sendiri masih sering telat, namun sudah ditagih.

Sentimen netral pada bulan April berisi pertanyaan atau pernyataan yang dilontarkan melalui cuitan dari akun Twitter Indihome maupun pelanggan yang memiliki akun Twitter. Pertanyaan yang sering dikaitkan, biasanya meliputi arahan jika terjadi masalah pada jaringan internet, maka dapat menghubungi akun Twitter maupun Facebook yang disediakan oleh Indihome. Jika terjadi permasalahan pada kondisi jaringan internet, maka akan dilihat dari lokasi yang bersangkutan apakah sedang dilakukan pemeliharaan jaringan atau tidak.

Secara keseluruhan perbandingan hasil asosiasi kata pada bulan Maret dan April cukup memiliki kemiripan yang signifikan. Seperti contoh pada sentimen positif bulan Maret yang memberikan informasi dan tips yang informatif, serta pelayanan dari *admin* Indihome yang cukup solutif, sama halnya pada bulan April, dimana juga memberikan tips yang informatif serta membahas kecakapan *admin* IndiHome dalam memberikan pelayanan. Pada sentimen netral juga memiliki kemiripan, baik pada bulan Maret maupun April berisi pertanyaan dan pernyataan tentang arahan jika terjadi masalah pada koneksi jaringan internet. Begitu pula pada sentimen negatif yang selalu membahas permasalahan yang sama mengenai sinyal internet yang sering hilang koneksi, lambat, dan sejenisnya. Namun pada bulan April terdapat tambahan keluhan, seperti tagihan yang diberikan cukup mahal, dimana tidak sesuai dengan kinerja internet yang dinikmati oleh pelanggan Indihome itu sendiri.

Berdasarkan hasil dan pembahasan dari penelitian ini, penulis dapat memberikan beberapa *consumer insight* kepada pihak Indihome terkait sentimen-sentimen yang dilontarkan oleh pelanggan Indihome itu sendiri. Hasil dari sentimen positif, dapat dijadikan acuan bagi Indihome untuk tetap memberikan informasi maupun tips yang bermanfaat mengenai internet, serta kecakapan dalam menanggapi keluhan oleh pelanggan dapat dipertahankan. Begitu pula jika dilihat dari hasil sentimen netral, dimana pihak Indihome dapat memiliki gambaran terhadap pertanyaan maupun pernyataan yang mungkin akan diajukan kembali oleh pelanggan di masa yang akan datang. Sedangkan hasil untuk sentimen negatif,

pihak Indihome dapat menjadikannya sebagai bahan evaluasi untuk lebih ditingkatkan, baik itu kinerja maupun kecepatan pelayanan dalam menghadapi keluhan pelanggan melalui *platform* Twitter, serta performa internet yang diberikan dari *provider* Indihome sendiri.

BAB VI
PENUTUP

6.1. Kesimpulan

Berdasarkan hasil analisis dan pembahasan mengenai rumusan masalah yang telah dipaparkan, didapatkan kesimpulan sebagai berikut:

1. Secara umum mengenai tanggapan pelanggan terhadap pelayanan dan juga kinerja Indihome pada *platform* Twitter, didapatkan bahwa dari total 56944 ulasan pada rentang bulan Maret 2021, sebesar 29,38% diantaranya memberikan ulasan positif mengenai Indihome. Lalu, sisanya sebesar 41,74% memberikan ulasan negatif, dan 28,88% diantaranya merupakan sentimen netral. Sedangkan pada bulan April 2021, 30,89% dari total 51063 ulasan yang diperoleh memberikan sentimen positif, dan sisanya memberikan sentimen negatif serta sentimen netral, yaitu sebesar 38,72% dan 30,39%.
2. Hasil klasifikasi menggunakan algoritma *Support Vector Machine* (SVM) dengan percobaan beberapa metode *kernel*, didapatkan metode *kernel* terbaik yaitu metode *Radial Basis Function* (RBF) baik pada bulan Maret maupun April 2021. Berikut adalah rincian hasil analisis sentimen pada bulan Maret dan April 2021:
    a. Pada bulan Maret, pembagian data menjadi 45555 data *training* dan 11389 data *testing*. Diperoleh dari data *testing*, sebanyak 3346 data merupakan kelas positif, 3290 data merupakan kelas netral, dan sisanya 4753 data merupakan kelas negatif. Tingkat akurasi yang didapatkan oleh metode RBF adalah sebesar 88,47%, di mana 10076 data diantaranya terklasifikasikan secara tepat pada kelas sentimennya.
    b. Pada bulan April, pembagian data menjadi 40851 data *training* dan 10212 data *testing*. Diperoleh dari data *testing*, sebanyak 3154 data merupakan kelas positif, 3103 data merupakan kelas netral, dan sianya 3955 data merupakan kelas negatif. Tingkat akurasi yang didapatkan oleh metode RBF adalah sebesar 98,06%, di mana 10014 data diantaranya terklasifikasikan secara tepat pada kelas sentimennya.

3. Dilihat dari hasil asosiasi kata yang ditinjau pada setiap sentimen pada bulan Maret dan April 2021, tidak terlalu berbeda jauh. Seperti halnya untuk sentimen positif dan sentimen netral yang masih membahas terkait informasi yang diberikan pihak Indihome seperti sebuah tips yang informatif, dan juga pelayanan yang disalurkan melalui kecakapan dalam membalas keluhan serta pertanyaan yang dilontarkan oleh pelanggan melalui cuitan di Twitter. Begitu pula untuk sentimen negatif, permasalahan yang dihadapi selalu sama, yaitu internet yang sering hilang sinyal atau sinyalnya mati, serta kecepatan internet yang lambat, sehingga pelanggan terus menerus memberikan keluhannya kepada Indihome.

6.2. Saran

Berdasarkan hasil analisis dan kesimpulan, penulis dapat memberikan saran sebagai berikut:

1. Pelabelan kelas sentimen pada penelitian ini belum mampu mendeteksi kata-kata negasi, sehingga diharapkan pada penelitian selanjutnya dapat dilakukan penanganan terhadap kata-kata negasi agar mendapat hasil yang lebih akurat.
2. Diharapkan pada penelitian selanjutnya, mampu menambah rentang waktu dalam pengumpulan data agar mendapat hasil yang lebih representatif.
3. Bagi penelitian selanjutnya, sebaiknya menggunakan algoritma-algoritma klasifikasi lain dengan melakukan perbandingan antar klasifikasi tersebut, sehingga dapat mengetahui algoritma dengan kinerja terbaik. Dapat juga menggunakan performa algoritma *Support Vector Machine* yang dibandingkan dengan algoritma lainnya.
4. Bagi pihak Indihome, diharapkan dari hasil analisis sentimen pada penelitian ini, baik berupa ulasan-ulasan sentimen positif, sentimen netral, dan terutama dari sentimen negatif, dapat dijadikan bahan evaluasi untuk dipertahankan ataupun lebih ditingkatkan di masa yang akan datang.

DAFTAR PUSTAKA

Alrajak, M. S., Ernawati, I. & Nurlaili, I., 2020. *Analisis Sentimen Terhadap Pelayanan PT PLN di Jakarta Pada Twitter dengan Algoritma K-Nearest Neighbor (K-NN).* Jakarta, s.n.

Alsheikh, M. A., Lin, S., Niyato, D. & Tan, H.-P., 2014. *Machine Learning in Wireless Sensor Networks: Algorithms, Strategies, and Applications.* s.l., Institute of Electrical and Electronics Engineers.

Annur, C. M., 2021. *databoks.* [Online] Available at: https://databoks.katadata.co.id/datapublish/2021/07/30/indihome-mendominasi-pasar-internet-fixed-broadband-di-indonesia#:~:text=Proporsi%20Jumlah%20Pelanggan%20Internet%20Fixed%20Broadband%20Indonesia&text=Berdasarkan%20laporan%20Bank%20Dunia%2C%20sebanya [Diakses 21 Agustus 2021].

Aprianti, W., Hafizd, K. A. & Rizani, M. R., 2017. Implementasi Association Rules dengan Algoritma Apriori pada Dataset Kemiskinan. *Journal Mathematics and its Aplications.*

Atenstaedt, R., 2012. Word cloud analysis of the BJGP. *British Journal of General Practice,* pp. 148-148.

Brownlee, J., 2019. *A Gentle Introduction to Imbalanced Classification.* [Online]
Available at: https://machinelearningmastery.com/what-is-imbalanced-classification/ [Diakses 31 Juli 2021].

Buntoro, G. A., 2016. Analisis Sentimen Hatespeech pada Twitter dengan Metode Naïve Bayes Classifier dan Support Vector Machine. *Jurnal Dinamika Informatika.*

D'Andrea, A., Ferri, F., Grifoni, P. & Guzzo, T., 2015. Approaches, Tools and Applications for Sentiment Analysis Implementation. *International Journal of Computer Applications.*

DePaolo, C. A. & Wilkinson, K., 2014. Get Your Head into the Clouds: Using Word Clouds for Analyzing Qualitative Assessment Data. *TechTrends.*

Deviyanto, A. & Wahyudi, M. D. R., 2018. Penerapan Analisis Sentimen pada Pengguna Twitter Menggunakan Metode K-Nearest Neighbor. *JISKa (Jurnal Informatika Sunan Kalijaga).*

Drajana, I. C. R., 2017. Metode Support Vector Machine dan Forward Selection Prediksi Pembayaran Pembelian Bahan Baku Kopra. *ILKOM Jurnal Ilmiah.*

GetDayTrends, 2021. *getdaytrends.* [Online] Available at: https://getdaytrends.com/indonesia/trend/Indihome/ [Diakses 21 Agustus 2021].

Han, H. et al., 2018. Improving the performance of lexicon-based review sentiment analysis method by reducing additional introduced sentiment bias. *PLOS ONE.*

Han, J., Kamber, M. & Pei, J., 2012. *Data Mining Concepts and Techniques Third Edition.* Waltham USA: Morgan Kaufmann.

Haranto, F. F. & Sari, B. W., 2019. Implementasi Support Vector Machine untuk Analisis Sentimen Pengguna Twitter Terhadap Pelayanan Telkom dan Biznet. *Jurnal PILAR Nusa Mandiri.*

IndiHome, 2020. [Online] Available at: https://indihome.co.id/news/targetkan-83-juta-pelanggan-indihome-tingkatkan-pelayanan-pelanggan-hingga-inovasi-produk-digital-di-tahun-2020 [Diakses 27 Desember 2020].

Jihaderajad, 2017. *Profil Perusahaan PT. Telkom Indonesia (Indihome).* [Online]
Available at: https://jihadyakhir.wordpress.com/2017/10/04/profil-perusahaan-pt-telkom-indonesia-indihome/ [Diakses 27 Juli 2021].

Josi, A., Abdillah, L. A. & S., 2014. Penerapan Teknik Web Scrapingpada Mesin Pencari Artikel Ilmiah. *Jurnal Sistem Informasi.*

Josina, 2020. *detikInet.* [Online] Available at: https://inet.detik.com/business/d-5106784/jumlah-pengguna-twitter-naik-tapi-pendapatannya-turun [Diakses 31 Desember 2020].

Karmayasa, O., 2012. Implementasi Vector Space Model dan Beberapa Notasi Metode Term Frequency Inverse Document Frequency (TF-IDF) ada Sistem Temu Kembali Informasi. *JELIKU (Jurnal Elektronik Ilmu Komputer Udayana).*

Katz, G., Ofek, N. & Shapira, B., 2015. ConSent: Context-based sentiment analysis. *Knowledge-Based Systems,* p. 162–178.

Kemp, S., 2021. *DATAREPORTAL*. [Online] Available at: https://datareportal.com/reports/digital-2021-indonesia [Diakses 3 Februari 2021].

Larose, D. T., 2005. *Discovering Knowledge in Data An Introduction to Data Mining.* Canada: John Wiley & Sons, Inc..

Luqyana, W. A., Cholissodin, I. & Perdana, R. S., 2018. Analisis Sentimen Cyberbullying pada Komentar Instagram dengan MetodeKlasifikasi Support Vector Machine. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer* .

Maclean, F., Jones, D., Carin-Levy, G. & Hunter, H., 2013. Understanding twitter. *British Journal of Occupational Therapy,* pp. 295-298.

Mailo, F. F. & Lazuardi, L., 2019. Analisis Sentimen Data Twitter Menggunakan Metode Text Mining Tentang Masalah Obesitas di Indonesia. *Journal of Information Systems for Public Health.*

Medhat, W., Hassan, A. & Korashy, H., 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal.*

Muis, I. A. & Affandes, M., 2015. Penerapan Metode Support Vector Machine (SVM) Menggunakan Kernel Radial Basis Function (RBF) Pada Klasifikasi Tweet. *Jurnal Sains, Teknologi dan Industri.*

Nasution, L. M., 2017. Statistik Deskriptif. *Jurnal Hikmah,* Volume 14.

Nugroho, A. S., Witarto, A. B. & Handoko, D., 2003. *Support Vector Machine: Teori dan Aplikasinya dalam Bioinformatika.* s.l.:s.n.

Pozzi, F. A., Fersini, E., Messina, E. & Liu, B., 2017. *Sentiment Analysis in Social Networks.* India: Todd Green.

Prayoginingsih, S. & Kusumawardani, R. P., 2018. Klasifikasi Data Twitter Pelanggan Berdasarkan Kategori myTelkomsel Menggunakan Metode Support Vector Machine (SVM) Studi Kasus: Telekomunikasi Selular. *Jurnal Sisfo.*

Pristiyanti, R. I., Fauzi, M. A. & Muflikhah, L., 2018. Sentiment Analysis Peringkasan Review Film Menggunakan Metode Information Gain dan K-Nearest Neighbor. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya,* pp. 1179-1186.

Putri, E. K. & Setiadi, T., 2014. Penerapan Text Mining pada Sistem Klasifikasi Email Spam Menggunakan Naive Bayes. *Jurnal Sarjana Teknik Informatika.*

Rizkia, S., Setiawan, E. B. & Puspandari, D., 2019. Analisis Sentimen Kepuasan Pelanggan Terhadap Internet Provider Indihome di Twitter Menggunakan Metode Decision Tree dan Pembobotan TF-IDF. *e-Proceeding of Engineering*, pp. 9683-9693.

Sicotte, X. B., 2015. *Support Vector Machine: calculate coefficients manually*. [Online]
Available at: https://xavierbourretsicotte.github.io/SVM_by_hand.html [Diakses 5 Agustus 2021].

Simon, A., Deo, M. S., Venkatesan, S. & Babu, D. R., 2015. An Overview of Machine Learning and its Applications. *International Journal of Electrical Science & Engineering (IJESE).*

Sitefanus, H., 2020. Analisis Kinerja Metode Cross Validation dan K-Nearest Neighbor dalam Klasifikasi Data. *Tesis.*

Tempola, F., Muhammad, M. & Khairan, A., 2018. Perbandingan Klasifikasi Antara KNN dan Naive Bayes Pada Penentuan Status Gunung Berapi Dengan K-Fold Cross Validation. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK).*

Turland, M., 2010. *php|architect's Guide to Web Scraping with PHP.* Canada: Marco Tabini & Associates, Inc..

Waloeyo, Y. J., 2010. *Twitter Best Social Networking.* Yogyakarta: Elcom.

Wibowo, A., 2015. Pengaruh Elektronik Word Of Mouth dan Brand Image Terhadap Purchase Intention pada Konsumen Smartphone Samsung yang Berbasis Android. *Jurnal Ilmu Manajemen,* pp. 71-88.

Yan, B. et al., 2014. Beam Structure Damage Identification Based onBP Neural Network and Support Vector Machine. *Hindawi Publishing Corporation.*

Zafikri, A., 2008. Implementasi Metode Term Frequency Inverse Document Frequency (Tf-Idf) pada Sistem Temu Kembali Informasi. *Tugas Akhir Program Studi Ilmu Komputer Fakultas Matematika Dan Ilmu Pengetahuan Alam Universitas Sumatera Utara: Medan.*

LAMPIRAN

Lampiran 1 Tampilan Data

	Datetime	Text	Username
0	2021-03-31 23:59:55+00:00	Gue semalem abis kena april mop ama PLN 1 peru...	chavikook
1	2021-03-31 23:59:01+00:00	@IndiHome tolong cek DM yaa	azamazn
2	2021-03-31 23:58:01+00:00	@sunadios iya anjir menghibur jugaa soalnya gu...	ackermeong
3	2021-03-31 23:57:46+00:00	@IndiHome kenapaa lagi ini lemot banget dah si...	ghemaans
4	2021-03-31 23:57:04+00:00	@IndiHome Woyy telorr nge april mop jugak kele...	frankiesayrelay
...	**...**	**...**	**...**
58306	2021-03-01 00:06:29+00:00	@Kurniasari3yyn @IndiHome Sama kak aku juga😊 S...	mey_ayam
58307	2021-03-01 00:02:24+00:00	Kalo wangi kemenyan bisa memanggil hantu,\nApa...	firdausdah
58308	2021-03-01 00:02:03+00:00	@Kurniasari2yyn @IndiHome Aku juga ngerasain s...	SenjaaMaharani
58309	2021-03-01 00:00:32+00:00	Indihomee kenapaa sih haruss di hari senin err...	nadila_mawarni
58310	2021-03-01 00:00:15+00:00	indihome bngstt	widhyourbae

Lampiran 2 *Script Python Web Scraping*

```
import snscrape.modules.twitter as sntwitter
import pandas as pd

# Creating list to append tweet data to
tweets_list = []

# Using TwitterSearchScraper to scrape data and append tweets to
list
for i,tweet in enumerate(sntwitter.TwitterSearchScraper('indihome
lang:id since:2021-03-01 until:2021-05-01').get_items()):
    if i>i:
        break
    tweets_list.append([tweet.date,tweet.content,
tweet.user.username])

# Creating a dataframe from the tweets list above
tweets_df = pd.DataFrame(tweets_list, columns=['Datetime', 'Text',
'Username'])

tweets_df

# Export dataframe into a CSV
tweets_df.to_csv('tweets-march21.csv', sep=';', index=False)
```

Lampiran 3 *Script Python Preprocessing Data*

```
# -------- Input Data -------- #
import pandas as pd
data = pd.read_csv("tweets-march21.csv", sep=";")
data

# -------- Case Folding -------- #
data['Text'] = data['Text'].str.lower()
print('Case Folding Result : \n')
print(data['Text'])
print('\n\n\n')

# -------- Tokenizing -------- #
import string
import re

# import word_tokenize & FreqDist from NLTK
from nltk.tokenize import word_tokenize
from nltk.probability import FreqDist
def remove_tweet_special(text):
    # remove tab, new line, ans back slice
    text      =      text.replace('\\t',"      ").replace('\\n',"
").replace('\\u'," ").replace('\\',"")
    # remove non ASCII (emoticon, chinese word, .etc)
    text = text.encode('ascii', 'replace').decode('ascii')
    # remove mention, link, hashtag
    text = ' '.join(re.sub("([@#][A-Za-z0-9]+)|(\w+:\/\/\S+)"," ",
text).split())
    # remove incomplete URL
    return text.replace("http://", " ").replace("https://", " ")
data['Text'] = data['Text'].apply(remove_tweet_special)

#remove number
def remove_number(text):
    return  re.sub(r"\d+", "", text)
data['Text'] = data['Text'].apply(remove_number)

#remove punctuation(tanda baca)
def remove_punctuation(text):
    return text.translate(str.maketrans("","",string.punctuation))
data['Text'] = data['Text'].apply(remove_punctuation)

#remove whitespace(menghapus spasi di awal dan akhir) leading &
trailing
def remove_whitespace_LT(text):
    return text.strip()
data['Text'] = data['Text'].apply(remove_whitespace_LT)

#remove multiple whitespace into single whitespace
def remove_whitespace_multiple(text):
```

```
    return re.sub('\s+',' ',text)
data['Text'] = data['Text'].apply(remove_whitespace_multiple)

# remove single char
def remove_singl_char(text):
    return re.sub(r"\b[a-zA-Z]\b", "", text)
data['Text'] = data['Text'].apply(remove_singl_char)

# remove elongation word
def remove_consecutive_dups(text):
    return re.sub(r'(?i)(.)\1+', r'\1', text)
data['Text'] = data['Text'].apply(remove_consecutive_dups)

# NLTK word tokenize
def word_tokenize_wrapper(text):
    return word_tokenize(text)
data['Text_tokens'] = data['Text'].apply(word_tokenize_wrapper)
print('Tokenizing Result : \n')
print(data['Text_tokens'])
print('\n\n\n')

# -------- Filtering -------- #
from nltk.corpus import stopwords

# add stopword from txt file
stopwords = open("stopword.txt","r").read().split()

# manualy add stopword
# append additional stopword
stopwords.extend(['dm','cek','via','kak','kakak','admin','sih','te
rima','kasih','nih','hai','terimakasih','nya','halo','trims','thx'
,'tks','in','amp','mots','youre','welcome','dear','an','sayang','a
rrk','the','april','mop','it','slot','dah','min','mimin','adm','in
di','home','mutual','mutualan','woy','telor','nge','papale','aq','
mengkren','deh','cp','btw','helo'])

# remove stopword
def stopwords_removal(words):
    return [word for word in words if word not in stopwords]
data['tweets'] = data['Text_tokens'].apply(stopwords_removal)
print(data['tweets'])

# -------- Normalization -------- #
normalizad_word = pd.read_csv("normalisasi.csv", sep=';')
normalizad_word_dict = {}
for index, row in normalizad_word.iterrows():
    if row[0] not in normalizad_word_dict:
        normalizad_word_dict[row[0]] = row[1]
def normalized_term(document):
    return     [normalizad_word_dict[term]     if     term     in
normalizad_word_dict else term for term in document]
```

```
data['tweet_normalized'] = data['tweets'].apply(normalized_term)
data['tweet_normalized']

# -------- Stemming -------- #
# import Sastrawi package
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
import swifter

# create stemmer
factory = StemmerFactory()
stemmer = factory.create_stemmer()

# stemmed
def stemmed_wrapper(term):
    return stemmer.stem(term)
term_dict = {}
for document in data['tweet_normalized']:
    for term in document:
        if term not in term_dict:
            term_dict[term] = ' '
print(len(term_dict))
print("------------------------")
for term in term_dict:
    term_dict[term] = stemmed_wrapper(term)
    print(term,":" ,term_dict[term])
print(term_dict)
print("------------------------")

# apply stemmed term to dataframe
def get_stemmed_term(document):
    return [term_dict[term] for term in document]
data['text']                                                    =
data['tweet_normalized'].swifter.apply(get_stemmed_term)
print(data['text'])
data['text'].to_csv("Pre_processing.csv")
```

Lampiran 4 *Script R* Pelabelan Sentimen

```
# Load packages #
library(tm)
setwd("C:\\Users\\desyrr\\Documents\\KULIAH\\SKRIPSI")
tweet <- read.csv("Pre_processing.csv", sep=";")
View(tweet)

# Scoring #
library(stringr)
library(plyr)
kata.positif<-
scan("positif.txt",what="character",comment.char=";")
kata.negatif<-
scan("negatif.txt",what="character",comment.char=";")
score.sentiment = function(tweet, kata.positif, kata.negatif,
.progress='none')
{
  require(plyr)
  require(stringr)
  scores = laply(tweet, 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) - (1*sum(negatif.matches))
    return(score)
  }, kata.positif, kata.negatif, .progress=.progress )
  scores.df = data.frame(score=scores, text=tweet)
  return(scores.df)
}
hasil = score.sentiment(tweet$text, kata.positif, kata.negatif)
View(hasil)

# Convert Score to Sentiment #
hasil$klasifikasi <- ifelse(hasil$score<0, "Negatif", "Positif")
View(hasil)

# Exchange Row Sequence #
docs <- hasil[c(3,1,2)]
View(docs)
write.csv(docs, file = "hasil_pelabelan.csv")
```

Lampiran 5 *Script R* Klasifikasi dengan Algoritma SVM

```
# Klasifikasi SVM #
setwd("C:\\Users\\desyrr\\Documents\\KULIAH\\SKRIPSI")
library(RTextTools)
library(e1071)
library(caTools)

# Split Data #
set.seed(1234)
split <- sample.split(label$klasifikasi, SplitRatio = 0.8)
train <- subset(label, split == TRUE)
write.csv(train[,-1], file = "training_maret_3.csv")
test <- subset(label, split == FALSE)
write.csv(test[,-1], file = "testing_maret_3.csv")

# Train Data #
train <- read.csv("training_maret_3.csv")
train <- train[,-1]
View(train)

# Test Data #
test <- read.csv("testing_maret_3.csv")
test <- test[,-1]
View(test)

# Join #
train.data <- rbind(train,test)

# Cleaning #
train.data$text = tolower(train.data$text)
text = train.data$text
text = removePunctuation(text)
text = removeNumbers(text)
text = stripWhitespace(text)
stopwords = unique(gsub("'"," ",text))

cor = Corpus(VectorSource(text))
dtm = DocumentTermMatrix(cor,
                         control = list(weighting =
```

```
                                        function(x)
                                        weightTfIdf(x,
                                        normalize = F)))
training_codes = train.data$klasifikasi

container <- create_container(dtm,
                              t(training_codes),
                              trainSize = 1:nrow(train),
                              testSize=
                              (nrow(train)+1):nrow(train.data),
                              virgin = FALSE)

models <- train_models(container, "SVM", kernel="linear")
models2 <- train_models(container, "SVM", kernel="polynomial")
models3 <- train_models(container, "SVM", kernel="radial")
models4 <- train_models(container, "SVM", kernel="sigmoid")

results <- classify_models(container, models)
results2 <- classify_models(container, models2)
results3 <- classify_models(container, models3)
results4 <- classify_models(container, models4)

# Accuracy #
recall_accuracy(as.numeric(as.factor(test$klasifikasi)),
results[,"SVM_LABEL"])
recall_accuracy(as.numeric(as.factor(test$klasifikasi)),
results2[,"SVM_LABEL"])
recall_accuracy(as.numeric(as.factor(test$klasifikasi)),
results3[,"SVM_LABEL"])
recall_accuracy(as.numeric(as.factor(test$klasifikasi)),
results4[,"SVM_LABEL"])

# Confusion Matrix #
library(caret)
conf.matSVM <-
confusionMatrix(results[,"SVM_LABEL"],train.data$klasifikasi[(nrow
(train)+1):nrow(train.data)])
conf.matSVM2 <-
```

```
confusionMatrix(results2[,"SVM_LABEL"],train.data$klasifikasi[(nro
w(train)+1):nrow(train.data)])
conf.matSVM3 <-
confusionMatrix(results3[,"SVM_LABEL"],train.data$klasifikasi[(nro
w(train)+1):nrow(train.data)])
conf.matSVM4 <-
confusionMatrix(results4[,"SVM_LABEL"],train.data$klasifikasi[(nro
w(train)+1):nrow(train.data)])
```

Lampiran 6 *Visualisasi Sentimen*

```
# Pie Chart #
library(plotrix)
slices <- c(16218, 16619, 24110)
lbls <- c("Netral", "Positif", "Negatif")
pct <- round((slices/sum(slices))*100, 2)
label <- paste(lbls, pct)
lbl <- paste(label,"%",sep="")
pie3D(slices,labels=lbl,explode=0.1,
col=c("skyblue","green","red"))

# Barplot #
hasil <- read.csv("data gabung.csv", sep = ";")
hasil <- hasil[,-2]
Maret <- subset(hasil, bulan %in% c("Maret"))
April <- subset(hasil, bulan %in% c("April"))

neutral <- length(which(Maret$score == 0))
positive <- length(which(Maret$score > 0))
negative <- length(which(Maret$score < 0))
neutral2 <- length(which(April$score == 0))
positive2 <- length(which(April$score > 0))
negative2 <- length(which(April$score < 0))

Month <- c(" March"," March"," March","April","April","April")
Count <- c(neutral,positive,negative,neutral2,positive2,negative2)
Sentiment                                                    <-
c("Neutral","Positive","Negative","Neutral","Positive","Negative")
output <- cbind.data.frame(Month,Sentiment,Count)

ggplot(output, aes(x=Month, y=Count, fill=Sentiment)) +
  geom_bar(stat="identity", position=position_dodge()) +
  geom_text(aes(label=Count), vjust=1.6, color="black",
            position = position_dodge(0.9), size=3.5) +
  theme_minimal()

# Word Cloud #
library(wordcloud2)
dt <- read.csv("hasil_pelabelan_maret2.csv",sep=";")
```

```
data <- subset(dt, klasifikasi %in% c("Positif"))
data2 <- subset(dt, klasifikasi %in% c("Negatif"))
data3 <- subset(dt, klasifikasi %in% c("Netral"))

library(tm)
docs <- Corpus(VectorSource(data$text))
docs2 <- Corpus(VectorSource(data2$text))
docs3 <- Corpus(VectorSource(data3$text))
dtm <- TermDocumentMatrix(docs)
dtm <- TermDocumentMatrix(docs2)
dtm <- TermDocumentMatrix(docs3)
m <- as.matrix(dtm)
words <- sort(rowSums(m),decreasing=TRUE)
df <- data.frame(word=names(words), freq=words)
wordcloud2(data=df, size=1.3, color='random-dark')

# Asosiasi Kata #
dt <- read.csv("hasil_pelabelan_april2.csv",sep=";")
data <- subset(dt, klasifikasi %in% c("Positif"))
doc <- Corpus(VectorSource(data$text))
dtm <- TermDocumentMatrix(doc)
a <- as.list(findAssocs(dtm, terms
                        =c("nomor","internet","bantu"),
                        corlimit = c(0.15,0.15,0.15,0.15)))
a
```