TEXT MINING ANALYSIS DAN *SENTIMENT ANALYSIS* DENGAN MENGGUNAKAN METODE *NAÏVE BAYES CLASSIFIER*

(Studi Kasus: Data Tanggapan Mengenai Tokopedia Melalui Media Sosial *Twitter*)

TUGAS AKHIR

Khofiyya Mulia Rahmi

14611180

PROGRAM STUDI STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS ISLAM INDONESIA

YOGYAKARTA

2021

TEXT MINING ANALYSIS DAN *SENTIMENT ANALYSIS* DENGAN MENGGUNAKAN METODE *NAÏVE BAYES CLASSIFIER*

(Studi Kasus: Data Tanggapan Mengenai Tokopedia Melalui Media Sosial *Twitter*)

TUGAS AKHIR

Diajukan Sebagai Salah Satu Syarat Untuk Memperoleh Gelar Jurusan Statistika

Khofiyya Mulia Rahmi

14611180

PROGRAM STUDI STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS ISLAM INDONESIA

YOGYAKARTA

2021

HALAMAN PERSETUJUAN PEMBIMBING

TUGAS AKHIR

Judul	:	*Text Mining Analysis* dan *Sentiment Analysis* dengan Menggunakan Metode *Naive Bayes Classifier*. (Studi Kasus : Data Tanggapan Tokopedia melalui media sosial *Twitter*).
Nama Mahasiswa	:	Khofiyya Mulia Rahmi
Nomor Mahasiswa	:	14611180

TUGAS AKHIR INI TELAH DIPERIKSA DAN DISETUJUI UNTUK DIUJIKAN

Yogyakarta, 29 November 2020

Pembimbing

(Dr. Edy Widodo,S.Si.,M.Si)

HALAMAN PENGESAHAN

TUGAS AKHIR

***TEXT MINING ANALYSIS* DAN *SENTIMENT ANALYSIS* DENGAN MENGGUNAKAN METODE *NAÏVE BAYES CLASSIFIER* TERHADAP TOKOPEDIA**

Nama Mahasiswa : KHOFIYYA MULIA RAHMI

NIM : 14611180

TUGAS AKHIR INI TELAH DIUJIKAN
PADA TANGGAL: 8 MARET 2021

Nama Penguji:	**Tanda Tangan**
1. Tuti Purwaningsih, S.Stat., M.Si	……
2. Dina Tri Utari, S.Si., M.Si	……
3. Dr. Edy Widodo,S.Si.,M.Si	……

Mengetahui,

Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

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

MOTTO DAN PERSEMBAHAN

MOTTO:

" ketika telah melakukan yang terbaik yang kita bisa, maka kegagalan bukan sesutau yang harus disesalkan, tapi jadikanlah pelajaran dan motivasi diri "

PERSEMBAHAN :

1. Dengan segenap rasa di hati, kupersembahakan karya kecil yang telah berhasil ku selesaikan ini kepada bapak dan mama, terima kasih untuk segalanya yang telah diberikan selama ini. Terima kasih atas kasih sayang, doa yang tak pernah putus, dukungan moral maupun materil serta harapan kepadaku. Terima kasih juga untuk rasa percaya yang telah diberikan dan tidak bisa terbayarkan dengan apapun. *This litte achivement for you, Pak Ma.*
2. Kakaku tercinta, Irfan Widy Setyoko, Novita Windayuningtyas, si kecil keponakan Alisya, yang selalu mengingatkan dan menjadi alasan untuk segera menyelesaikan studi agar dapat kembali berkumpul dan bersama membahagiakan kedua orang tua.

KATA PENGANTAR

Assalamu'alaikum Warahmatullaahi wabarakaatuh

Puji syukur penulis panjatkan atas kehadirat Aah SWT karena atas berkat, rahmat, kesehatan dan kekuatan yang diberikan oleh-Nya tugas akhir ini dapat berjalan dengan lancar. Tugas akhir ini memberikan begitu banyak pembelajaran yang kemudian dapat dikembangkan pada penelitian selanjutnya.

Keberhasilan dalam pembuatan tugas akhir ini tentunya tidak lepas dari pihak-pihak yang memberikan semangat, dorongan atau motivasinya selama tugas akhir ini berlangsung. Ucapan terima kasih ini saya sampaikan kepada :

1. Bapak Prof. Riyanto, S.Pd.,M.Si.,Ph.D selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia.
2. Bapak Dr. Edy Widodo,S.Si.,M.Si selaku Ketua Jurusan Statistika dan dosen Pembimbing yang sudah membimbing dan memberikan dukungan serta motivasi kepada penulis dari awal hingga akhir ini tidak hanya skripsi ini tetapi juga untuk menjadi pribadi yang lebih baik.
3. Seluruh staf pengajar Program Studi Statistika Universitas Islam Indonesia yang telah memberikan bekal kepada penulis.
4. Seluruh civitas Fakultas Matematika dan Ilmu Pengetahuan Alam yang telah memberikan pelayanan yang baik kepada penulis, hingga bisa menyelesaikan studi.
5. Teruntuk sahabatku rumah kedua. Tempatku berbagi keluh kesah, tempat berbagi tawa bahagia hingga air mata. Hanny Cahya, Dwika Lucky, Meviana Rizki , Mega Ayu, Megi Ayu, Khafiya Nur, Ine Layna, Tyas Ira, Evi, Hawila Sonya, Satibi Mulyadi, Gita Evi, Afifah Mutaroh, Rakhil Khaeriyah, Anita Roikhatul, Husna Laela, Adhelia, Ardjun Wibowo, Rachel Ayuningtyas, Meimunah, Adim, Aysah Karla, Shofura. *Im blessed to having you guys.*
6. Teruntuk sahabat yang jauh di sana yang selalu memberikan semangat secara *virtual* Ridha Nur Izza, Riza Indriani, Eren Fajrila, Inayatus, Gustiara, Nur hidayah, Syarifah Rosita, Rina Sriwiji, Sri Hardianti, Ayu Renya, Nadhiroth,

Sekar Faika, Mia Rizky, Kia, Ellysa Lutfiana, Suci Insani, Salwa Yudanti, Selvina Selva, Eli, Enggar, Lazuardy, Dinda Septiani, Dini Ayu, Raja Ilman, Panji Satriok, Budi Ramdani, dll. Terima kasih telah menemani kehidupan perkuliahan ini, bersama manis pahitnya perjuangan. Terimakasih tawa, tangis dan segalanya. Sukses terus kedepannya untuk semuanya *and keep in touch.*

7. Teman-teman yang ada di Banjarnegara yang tidak dapat penulis sebutkan satu persatu dan teman-teman Komunitas Cinta Anak Yatim dan Duafa Banjarnegara. Terima kasih keluarga baru atas semangatnya, selalu memberikan pembelajaran hidup untuk terus bersyukur.
8. Teman-teman seperbimbingan, salah satunya Defi yang selalu ada 24 jam.
9. Teman-teman Statistika angkatan 2014, teman-teman satu jurusan, fakultas, organisasi atau kepanitian yang tidak dapat penulis sebutkan satu per satu, terima kasih atas bantuan dan semangatnya selama proses perkuliahan.Terima kasih telah menjadi sahabat sekaligus keluarga baru bagi penulis.
10. Seluruh Anggota IKS (Ikatan Keluarga Statistika) yang selalu menjalin silaturahmi, saya ucapkan terima kasih.
11. Serta semua pihak yang telah membantu dalam penyelesaian Tugas Akhir ini yang tidak dapat penulis sebutkan satu per satu.

Demikianlah yang dapat disampaikan, semoga Allah SWT senantiasa melimpahkan rahmat dan ridho-Nya kepada semua pihak yang telah membantu penulis. Penulis menyadari bahwa Tugas Akhir ini masih jauh dari kata sempurna dan masih banyak kekurangan. Hal tersebut dikarenakan keterbatasan ilmu dan pengetahuan yang dimiliki penulis semata. Penulis berharap semoga penulis Laporan Tugas Akhir ini dapat bermanfaat bagi semua pihak.

Wassalamu'alaikum Warahmatullaahi wabarakaatuh

Yogyakarta, 01 Januari 2021

Khofiyya Mulia Rahmi

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam Tugas Akhir ini tidak terdapat 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, 01 Januari 2021

Khofiyya Mulia Rahmi

TEXT MINING ANALYSIS* DAN *SENTIMENT ANALYSIS* DENGAN MENGGUNAKAN METODE *NAÏVE BAYES CLASSIFIER
(Studi Kasus: Data Tanggapan Mengenai Tokopedia Melalui Media Sosial *Twitter*)

Khofiyya Mulia Rahmi
14611180

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

INTISARI

Perkembangan teknologi yang semakin canggih telah membawa perubahan. Hal ini kemudian mempengaruhi gaya hidup masyarakat, dengan adanya belanja *online* melalui internet. Situs ini biasa disebut pasar *e-commerce. E-commerce* adalah penjualan yang dilakukan melalui media elektronik. Salah satu *e-commerce* di Indonesia adalah Tokopedia. Dalam strategi pemasaran untuk menarik daya tarik konsumen untuk menggunakan jasa maupun produk dari perusahaan adalah melakukan strategi pemasaran modern melalui media digital dalam memasarkan aktivitas dan inovatif dalam meningkatkan *brand awareness*, *brand image* perusahaan melalui *digital marketing*, *advertising* dan *social media marketing*, banyaknya promo dan informasi-informasi menarik yang disajikan tentunya membutuhkan marketing yang kuat dari perusahaan tersebut. Dalam hal ini pihak perusahaan tentunya diharapkan mampu memahami jenis konten dalam sebuah informasi yang diberikan agar memperoleh respon positif sehingga dapat meningkatkan jumlah pengunjung, akan semakin kuat, semakin dikenal. Oleh karena itu akan dilakukan analisis menggunakan akun sosial media Twitter dengan NBC (*Naive Bayes Classifier*). Data yang digunakan adalah *tweet* dan *retweet* para pengguna *Twitter* mulai dari tanggal 25 September 2020 sampai 28 September 2020. Diperoleh hasil *Sentiment Analysis* dari data *Twitter* yakni sebesar 5055 data terklasifikasi, 69,8% atau sebanyak 3530 tweet masuk dalam kelas sentimen negatif dan 30,2% atau 1525 masuk dalam kelas sentimen positif.

Kata Kunci : *Text Mining*, Tokopedia, *Sentiment Analysis*, *Naïve Bayes Classifier*.

TEXT MINING ANALYSIS AND SENTIMENT ANALYSIS USING METODE NAÏVE BAYES CLASSIFIER

(Case study: Data from Tokopedia in Social Media Twitter)

Khofiyya Mulia Rahmi

14611180

ABSTARK

Department of Statistics, Faculty of Mathematics and Natural Science

Islamic University of Indonesia

The sophistication of development technology has brought changes. Today shopping through the internet is affecting of people's lifestyles. This is commonly called the e-commerce marketplace. E-commerce is a sale made through electronic media. One of the e-commerce in Indonesia is Tokopedia. For the marketing strategy to attract consumers to use the services and products of the company, modern marketing strategies are applied through digital media in marketing activities and to be innovative in increasing the company's brand awareness and brand image through digital marketing, advertising and social media marketing, lots of attractive promotions and informations which is presented by the company certainly requires strong marketing from itself. In this case, the company is certainly expected to be able to understand the type of content in the information provided in order to get a positive response so that it could increase the number of visitors to make it firm and known as well. Therefore an analysis will be carried out using a Twitter social media account with NBC (Naive Bayes Classifier). The data used are tweets and retweets of Twitter users from September 25 to September 28, 2020. Data obtained from Twitter as many as 5055 classified as a result of Sentiment Analysis with a percentage of 69.8% or as many as 3530 Tweets into the negative sentiment class and 30.2% or as many as 1525 into positive sentiment.

Keywords: Text Mining, Tokopedia, Sentiment Analysis, Naïve Bayes Classifier.

BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi yang semakin canggih telah membawa perubahan pada masyarakat. Dengan adanya internet ini, memungkinkan bagi seseorang untuk melakukan komunikasi dengan pihak lain tanpa dibatasi waktu dan jarak. Hal ini juga yang kemudian memengaruhi gaya hidup masyarakat, dengan adanya belanja *online* melalui internet. (Ishak, 2012).

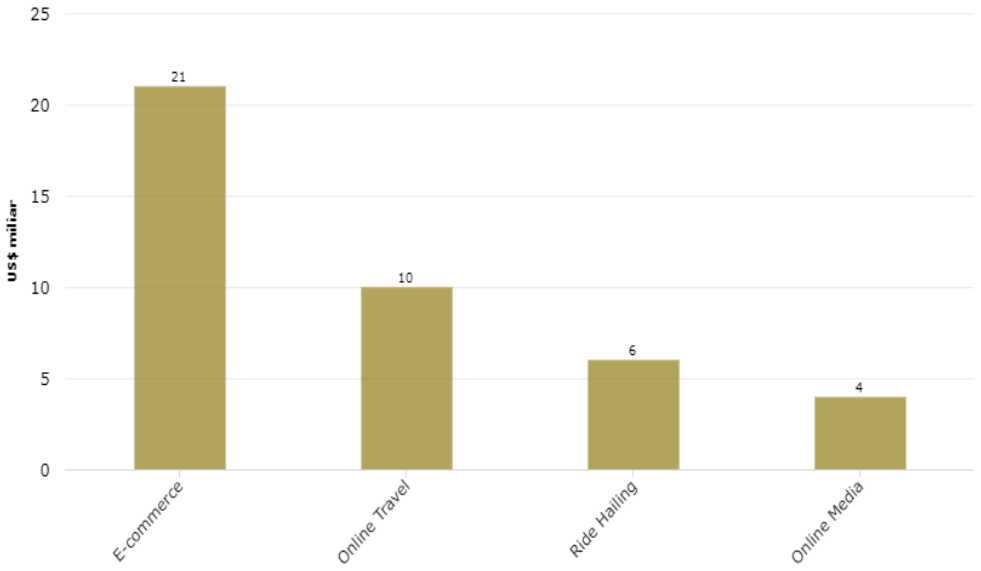
Situs belanja *online* ini disebut pasar *e-commerce*. Para pengguna jasa jual beli online ini dapat dengan mudah bertransaksi. *E-commerce* adalah penjualan yang dilakukan melalui media elektronik. *E-commerce* sebagai salah satu jenis dari mekanisme bisnis secara elektronik yang berfokus pada transaksi bisnis dengan menggunakan internet sebagai media pertukaran barang atau jasa sebagai penunjang sistem penjualan dan sebagai pemasaran pengembangan bisnis. (Ayu, 2018).

Menurut data yang diperoleh dari Databoks pada **Gambar 1.1**, yang paling tinggi terhadap kontribusi sektor ekonomi internet ASEAN di 2019 adalah kontribusi *e-commerce*, yang mencapai 21 miliar US$. Hal ini dikarenakan semakin banyaknya pengguna internet yang melakukan belanja *online* melalui *platform digital*, selanjutnya sektor *online travel* tercatat memiliki kontribusi sebesar 10 miliar US$. Lalu, *ride hailing* dan *online media* masing-masing berkontribusi sebesar 6 miliar US$ dan 4 milliar US$. (Databoks, 2020).

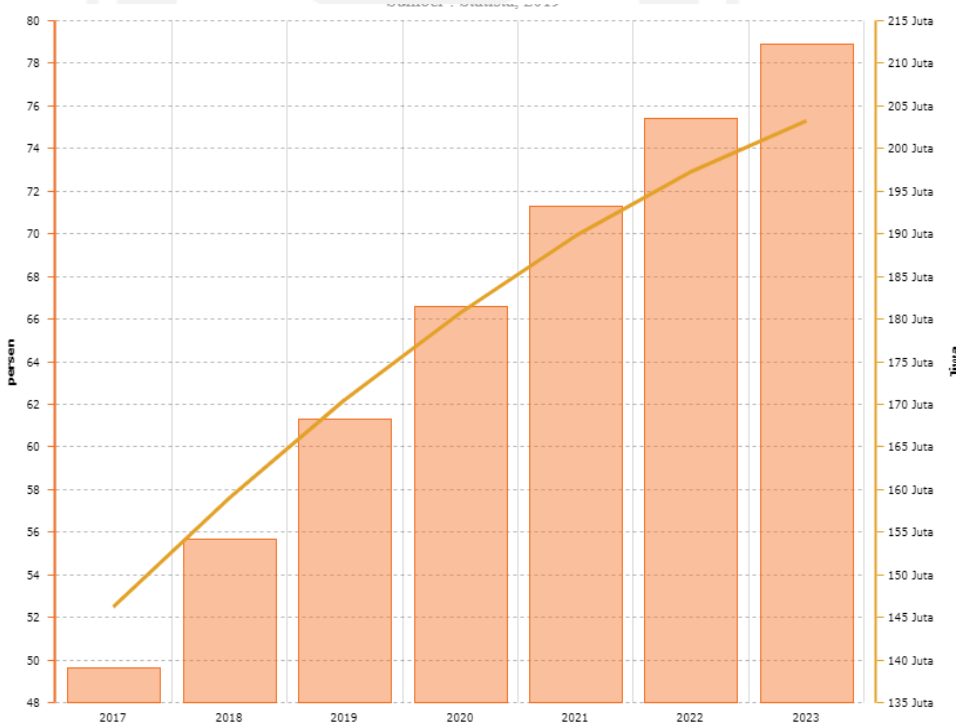
Perkembangan *e-commerce* di Indonesia juga bisa dikatakan cukup pesat. Tren pengguna *e-commerce* di Indonesia tumbuh cukup besar dari beberapa tahun terakhir. Prediksinya, pertumbuhan masih akan terus terjadi dalam beberapa tahun kedepan. (Databoks, 2019).

Menurut data yang diperoleh dari Databoks pada **Gambar 1.2**. Statistika mencatat jumlah pengguna *e-commerce* di Indonesia pada 2017 mecapai 139 juta penggnuna, kemudian naik 10,8 % menjadi 154,1 juta pengguna di tahun 2018.

Tahun 2019 mencapai 168.3 juta pengguna dan akan selalu meningkat hingga 212,2 juta pada Tahun 2023. (Databoks, 2019).



Gambar 1.1 Kontribusi Sektor Ekonomi Internet ASEAN 2019
(Sumber : Databoks, 2020)

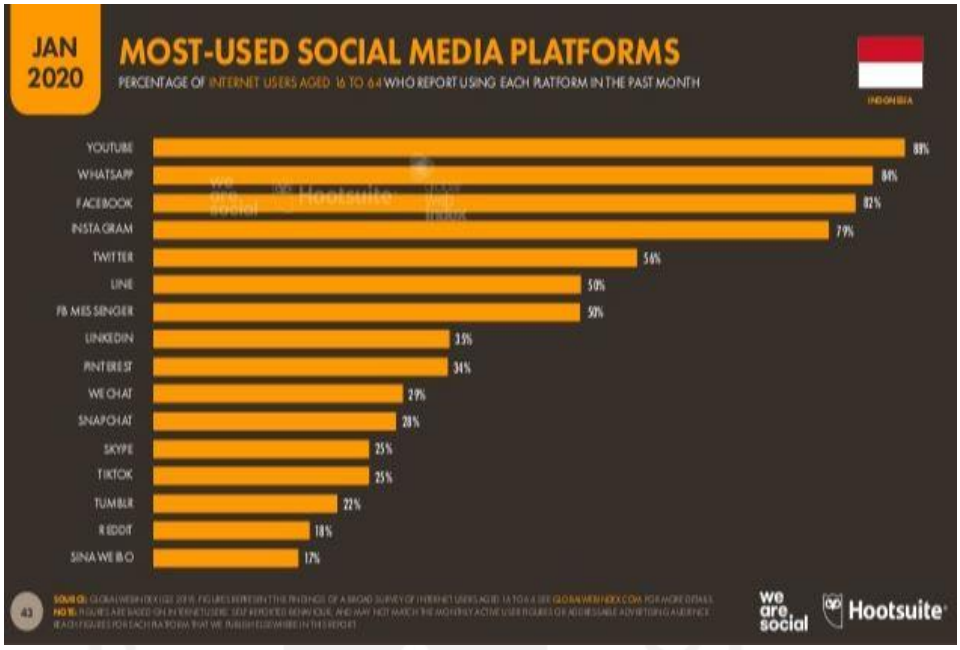


Gambar 1.2 Penggunaan dan Tingkat Penetrasi *E-Commerce* di Indonesia 2017-2023
(Sumber : Databoks, 2020)

Salah satu pelaku usaha *e-commerce* ada di Indonesia adalah Tokopedia. Tokopedia adalah toko pertama dalam *website* di Indonesia yang menyediakan peluang bisnis *online*. Tokopedia berdiri pada 6 Februari 2009 dan *website*-nya diperkenalkan pada 17 Agustus 2009 oleh William Tanuwijaya. (Retzen & Nurdin, 2016).

Menurut data yang diperoleh dari I Price Group. Tokopedia dalam jumlah pengunjung *website* pada kuarter ke-4 di tahun 2018 menduduki peringkat pertama hingga kuarter ke-3 di tahun 2019, namun mengalami penurunan dimulai dari kuarter ke-4 di tahun 2019 hingga kuarter ke-2 tahun 2020. Aplikasi Tokopedia menduduki peringkat kedua di AppStore, tetapi hanya peringkat keempat di PlayStore (I Price Group, 2020). Dalam strategi pemasaran untuk menarik daya tarik konsumen untuk menggunakan jasa maupun produk dari perusahaan adalah melakukan strategi pemasaran modern melalui media digital dalam memasarkan aktivitas dan inovatif dalam meningkatkan *brand awareness*, *brand image* perusahaan melalui *digital marketing*, *advertising* dan *social media marketing*, banyaknya promo dan informasi-informasi menarik yang disajikan tentunya membutuhkan marketing yang kuat dari perusahaan tersebut. (Defrianto & Loisa, 2019). Dalam hal ini pihak perusahaan tentunya diharapkan mampu memahami jenis konten atau topik dalam sebuah informasi yang diberikan agar memperoleh respon positif sehingga dapat meningkatkan jumlah pengunjung, supaya semakin kuat, semakin dikenal, dan penggunanya tetap menggunakan layanannya. (Helmi, 2019).

Saat ini, banyak orang menggunakan situs media sosial untuk menyampaikan pendapat atau untuk melihat topik yang sedang dibincangkan. Salah satunya adalah media sosial *Twitter*. Menurut data yang diperoleh dari wearesocial.com per Januari 2020 pada **Gambar 1.3**, pengguna media sosial di Indonesia *Twitter* menduduki peringkat ke lima dengan jumlah sebanyak 56%.



Gambar 1.3 Jumlah Pengguna Media Sosial di Indonesia tahun 2020
(Sumber : WeAreSocial, 2020)

Akun media sosial *Twitter* memiliki beberapa keungulan. Hal itu didukung oleh adanya fitur *Trends for you*. Pada fitur ini akan ditampilkan serangkaian kata yang *trends* dalam arti banyak yang membuat postingan kata tersebut. Semakin ramai sebuah postingan maka semakin banyak juga pengguna yang terlibat dalam postingan tersebut. Dengan begitu kemudahan sebuah informasi diterima oleh pengguna *Twitter* melalui *trending.* (Faesal, dkk., 2020). Hal ini dibuktikan dengan ramainya pengguna *Twitter* yang berkicau mengenai Tokopedia dalam perayaan festival belanja bulanan Waktu Indonesia Belanja. Tokopedia yang mengadakan acara TV Show mengundang idol Korea yang tembus *trending* pada tanggal 25 September 2020. (Tokopedia, 2020).

Untuk mengetahui informasi dari *tweet* dan *retweet* di *Twitter* mengenai topik atau konten tentang Tokopedia tersebut didapatkan data dan informasi dengan *text mining* digunakan untuk mengumpulkan data dari opini yang ada dari suatu peristiwa yang berupa kata-kata yang dapat menyimpulkan dari suatu pembahasan kalimat dari beberapa respon dan *sentiment analysis* merupakan proses memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam suatu kalimat apakah beopini positif

atau negatif setelah itu dapat digunakan analisis klasifikasi *naive bayes. Naive bayes* adalah klasifikasi statistik yang bisa memprediksi probabilitas sebuah kelas, dan kelebihan dari metode ini adalah tingkat akurasi yang tinggi juga waktu komputasi yang lebih cepat. (Sihotang, & Ghaniy, 2019). *Naïve Bayes Classifier* dirasa cocok digunakan pada analisis sentimen dikarenakan algoritma ini bertujuan sebagai metode klasifikasi kedalam kategori positif dan negatif (Jati, 2019).

Maka pada penelitian ini peniliti akan menggunakan metode *text mining* dan *sentiment analysis* meggunakan metode *Naïve Bayes Classifier* untuk mengetahui topik/konten atau kata-kata yang sering muncul dari masyarakat melalui jejaring sosial *Twitter* tentang Tokopedia, yang nantinya bisa dijadikan rekomendasi upaya meningkatkan jumlah pengunjung, penggunanya dll.

1.2 Rumusan Masalah

Berdasarkan pada latar belakang yang telah di paparkan pada poin 1.1, maka terdapat beberapa rumusan masalah yaitu:

1. Bagaimana hasil analisis *text mining* dan *sentiment Analysis* tentang Tokopedia di tahun 2020?
2. Bagaimana hasil tingkat akurasi yang didapatkan dari pengklasifikasian *sentiment Analysis* dengan menggunakan metode *Naïve Bayes Classifier* ?

1.3 Batasan Masalah

Berikut ini batasan masalah yang digunakan dalam penelitian ini:

1. Pada penelitian ini menggunakan data ulasan mengenai Tokopedia pada media sosial *Twitter*.
2. Metode yang digunakan yaitu *Text mining* dan *Naive Bayes Classifier*.
3. *Software* yang digunakan untuk membantu penelitian adalah *Microsoft Excel* dan *R 3.4.1, R 3.6.1*.

1.4 Tujuan Penelitian

Berikut ini merupakan tujuan dari penelitian yang akan dilakukan:

1. Mengetahui hasil *text mining* dan *sentiment Analysis* tentang pendapat masyarakat terhadap Tokopedia.

2. Mengetahui hasil tingkat akurasi yang didapatkan dari pengklasifikasian dengan metode *Naïve Bayes Classifier*.

1.5 Manfaat Penelitian

Manfaat yang didapatkan dari penelitian ini yaitu,

1. Memberikan pengetahuan mengenai analisis *text mining, sentimen analisis* menggunakan metode *naive bayes classifier* terhadap Tokopedia.
2. Memberikan rekomendasi kepada tokopedia dari hasil *text mining, sentimen analisis.*

BAB II
TINJAUAN PUSTAKA

Tinjauan pustaka ini dilakukan sebagai kajian litetatur untuk mengetahui keterkaitan antara penelitian yang sudah pernah dilakukan dengan penelitian yang akan dilakukan oleh peneliti. Berikut ini merupakan beberapa penelitian terdahulu yang dijadikan sebagai acuan oleh peneliti.

Tabel 2 Penelitian Terdahulu

No	Nama (Tahun)	Judul	Metode	Hasil Penelitian
1.	Afshoh (2017).	Analisa Sentimen menggunakan *Naive Bayes* untuk Melihat Persepsi Masyarakat terhadap Kenaikan Harga Jual Rokok pada Media Sosial *Twitter*.	*Analisa Sentimen* dan *Naive Bayes Classifier*.	Sentimen positif paling banyak terbentuk untuk menanggapi wacana kenaikan harga jual rokok dan hasil pengujian kinerja sistem menggunakan data *training* 150 positif, 150 negatif, dan 50 netral yang menunjukan hasil klasifikasi dokumen yang lebih baik.
2.	Faishal Nuruz Zuhri (2017)	Analisis Sentimen Masyarakat terhadap Bran Smartfren Menggunakan *Naive Bayes Classifier* di Forum Kaskus.	*Naive Bayes Classifier*	Dari hasil pengujian untuk kasus pada penelitian ini didapatkan bahwa NBC dapat diimplementasikan dengan nilai akurasi 98.40%. Dari 6338 data uji, 4049 berhasil terklasifikasi kedalam sentimen positif dan 3233 data bersentimen negatif. Komentar terbanyak pada bulan januari yaitu sebanyak 1472 data. Dari proses wordcloud, dapat disimpulkan bahwa kata-kata dalam komentar user yang mendominasi dari hasil analisis sentimen brand Smartfren, yang bersentimen positif, antara lain smartfren, paket, true unlimited, speed, lancar dan kencang dan bersentimen negatif, antara lain paket, smartfren, beli, true unlimited, lamban, masalah dan gangguan.

No	Nama (Tahun)	Judul	Metode	Hasil Penelitian
4.	Adhi Viky Sudiantoro, Eri Zuiarso (2018).	Analisis Sentimen *Twitter* Menggunakan *Text Minig* dengan *Algoritma Naive Bayes Classifer*.	*Naive Bayes.*	Algoritma *Naïve Bayes Classifier* sangat efektif untuk digunakan sebagai proses klasifikasi *tweet* yang dibutuhkan dalam sistem analisis sentiment ini dimana nilai yang di dapatkan dalam pengujian sampai 84%. Metode Naïve Bayes *Classifier* dapat digunakan untuk melakukan klasifikasi *tweets* dengan cukup baik pada sistem analisis sentimen. Hasil dari 100 data uji yang klasifikasi menghasilkan 32 data bersentimen positif dan sebanyak 68 data sentimen negatif. Bahwa 100 data uji yang diklasifikasi masuk dalam kategori bersentimen negatif dikarenakan data positif lebih kecil daripada data yang bersentimen negatif.
5.	Ghulam Asrofi Buntoro (2018)	*Sentiment Analysis for Governor of East Java 2018 in Twitter*	*Text mining & Naïve Bayes Classifier*	Didapatkan hasil akurasi sebesar 77% untuk Khofifah dan 76% untuk Gus Ipul.
6.	Sigit suryono (2018).	Klasifikasi Sentimen pada *Twitter* dengan *Naive Bayes Classifier.*	*Naive Bayes Classifier.*	Didapatkan hasil uji coba sebanyak 3 kali, nilai akurasi yang pertama sebesar 64,95%, yang kedua sebesar 66,36% dan yang ketiga sebesar 66,79%.Dari proses klasifikasi didapatkan opini positif sebesar 28%, opini negative sebesar 20%, dan opini netral sebesar 52%. Hasil yang diperoleh opini netral yang dikaitkan dengan topic Presiden Joko Widodo.
7.	Diba Farach & Jaka Nugraha. (2020).	Implementasi *Metode Naive Bayes Classifier* dalam Analisis Sentimen Pada Opini Masyarakat Tentang RUU KUHP.	- *Metode Naive Bayes Classifier*	Hasil analisis sentimen mengenai demo mahasiswa terkait pasal-pasal kontroversial dalam RUU KUHP didapatkan sebesar 3561 data *tweet* dengan jumlah 2483 data *tweet* negatif dan 1078 data *tweet* positif. Berdasarkan hasil klasifikasi menggunakan metode *naive bayes* didapatkan nilai akurasi sebesar 93,1%, untuk nilai *recall* sebesar 78,9% dan untuk nilai *precision* sebesar 97,6%. Berdasarkan nilai

No	Nama (Tahun)	Judul	Metode	Hasil Penelitian
				akurasi, *recall* dan *precision* yang tinggi, maka dapat dikatakan bahwa klasifikasi sudah tepat. Dapat juga dilihat dari nilai lain seperti *spesificity*, *false positive rate* (FPR) dan *area under curve* (AUC) juga menghasilkan nilai yang besar. Untuk nilai AUC didapatkan hasil sebesar 0,89 yang artinya bahwa nilai tersebut sudah baik atau klasifikasi baik.
8.	Eko Budi Santoso dan Aryo Nugroho (2019).	Analisis Sentimen Calon Preiden Indonesia 2019 Berdasarkan Komentar Publik di *Facebook*.	- *Text Mining* - *Sentiment Analysis* - *Naive Bayes Classifier*.	Hasil dari penelitian berdasarkan data yang telah dikumpulkan AJoko Widodo lebih unggul polaritas sentimen positif dari data sebanyak 5.000 komentar yang dipilih secara acak pada masing-masing calon presiden dan melalui tahap preprocessing yang menghasilkan polaritas sentimen. Joko Widodo diperoleh 85% sentimen positif, dan 15% untuk sentimen negatif. Sedangkan Prabowo Subianto diperoleh 76% sentimen positif, dan 24% untuk sentimen negatif. Metode klasifikasi Naïve Bayes Classifier terhadap penelitian ini memperoleh hasil akurasi 86,4%.
9.	Enggar Prima Jati (2019).	Implementasi *Text Mining* dan *Sentiment Analysis* pada Jejaring Sosial *Twitter* dengan Menggunakan Metode *Naive Bayes Classifier*.	- *Text Mining* - *Sentiment Analysis* - *Naive Bayes Classifier*.	Hasil *Sentiment Analysis* dari data *Twitter* yakni sebesar 6011 data terklasifikasi, 71% atau sebanyak 4276 tweet masuk dalam kelas sentimen negatif dan 29% atau 1735 masuk dalam kelas sentimen positif. Untuk kasus banjir di DKI Jakarta 5527 data terklasifikasi, yaitu 44% masuk kedalam kelas positif dan 56% masuk kedalam kelas negatif.
10.	Kepin Sihotang dan Rajib Ghaniy (2019)	Penerapan Metode *Naive Bayes Classifier* Untuk Penentuan Topik Tugas Akhir pada *Website* Perpustakaan STIKOM Binaniaga.	- *Naive Bayes Classifier*	Mahasiswa jurusan TI dapat melihat referensi juduljudul yang telah di tetapkan pada sistem website perpusatakaan stikom, dan penelitian menerapkan metode naive bayes classifier untuk pengklasifikasian dimana judul yang ada akan terklasifikasi dengan pernyataan topik yang berhubungan, demikian data judul yang didapat sebelumnya hanya mahasiswa jurusan TI, karena data yang dijadikan untuk uji coba

No	Nama (Tahun)	Judul	Metode	Hasil Penelitian
				pembelajaran adalah hanya judul alumni mahasiswa jurusan TI stikom pada tahun 2015-2017. Dengan data judul-judul TI tersebut metode naive bayes classifier sudah dapat membuktikannya. Hasil dari kuesioner yang disebarkan kepada dosen dan ahli sistem menunjukkan hasil sangat layak pada kedua responden.

Berdasarkan **Tabel 2.1** yang dilakukan oleh peneliti dalam melakukan penelitian yang sekarang. Objek penelitian sekarang adalah terkait Tokopedia di tahun 2020. Sedangkan metode analisis yang digunakan pada penelitian adalah Metode Klasifikasi *Naive Bayes*.

BAB III
LANDASAN TEORI

3.1 *E-commerce*

Salah satu penerapan bidang bisnis dan perdagangan menggunakan jaringan internet dalam adalah *electronic commerce* atau disingkat dengan *e-commerce*. *E-commerce* merupakan penggunaan aplikasi seluler untuk bertaransaksi bisnis dalam melakukan transaksinya. *E-commerce* adalah suatu jenis dari mekanisme bisnis secara elektronik yang memfokuskan diri pada transaksi bisnis berbasis individu dengan menggunakan internet (teknologi berbasis jaringan digital) sebagai medium pertukaran barang atau jasa baik antara dua buah institusi (*business to business*), institusi dengan konsumen langsung (*business to Consumer*). Dengan aplikasi *e-commerce*, hubungan antar perusahaan dengan entitas eksternal lainnya (pemasok, distributor, rekanan, konsumen) dapat dilakukan secara lebih cepat, lebih intensif, dan lebih murah daripada aplikasi prinsip manajemen secara konvensional (*door to door*, *one-to-one relationship*). (Pujastuti, dkk., 2014).

3.2 Tokopedia

PT. Tokopedia merupakan salah satu *mall online* di Indonesia. Wujud sebuah mall online yang mempertemukan penjual dan pembeli dan memungkinkan untuk terjadinya transaksi jual beli online dengan aman dan nyaman. Bergabung untuk menggunakan Tokopedia sangatlah mudah dan tidak dipungut biaya. Setelah beroperasi www.tokopedia.com telah menjadi salah satu *online marketplace* dengan tingkat pertumbuhan yang sangat pesat di Indonesia walaupun usianya masih seumur jagung, baik dalam jumlah anggota, *took*, *online* aktif, jumlah produk hingga jumlah transaksi pembelian dan penjualan setiap harinya. Tokopedia sudah mampu bersaing di pasar Indonesia, selain mempunyai metode yang berbeda dari pesaingnya Tokopedia mampu terus maju dalm persaingan bisnis *e-commerce.* (Pratiwi,dkk, 2020).

Tokopedia sejatinya tidak mempunyai cabang perusahaan. Tokopedia hanya memiliki kantor pusat yang berlokasi di Jakarta namun memiliki berbagai pengguna

(penjual) diseluruh penjuru Indonesia. Tokopedia.com resmi diluncurkan ke publik pada 17 Agustus 2009 dibawah naungan PT. Tokopedia yang didirikan oleh William Tanuwijaya dan Leontinus Alpha Edison. (Pratiwi,dkk, 2020).

PT. Tokopedia mendapatkan *seed fundig* (pendanaan awal) dari PT Indonusa Dwitama pada tahun 2009. Kemudian pada tahun-tahun berikutnya, Tokopedia kembali mendapatkan suntikan dana dari pemodal *ventura global* seperti *East Ventures* (2010), *Cyber Agent Ventures* (2011), *Netprice* (2012) dan *Soft Bank Ventures* (2013). Lalu pada Oktober 2014, Tokopedia berhasil mencetak sejarah sebagai perusahaan teknologi pertama di Asia Tenggara, yang menerima investasi sebesar USD 100 juta atau sekitar RP. 1,2 triliun dari *Squoia Capital* dan *Softbank Internet* dan *Media Inc* (SIMI). Pada tanggal April, Tokopedia kembali dikabarkan mendapatkan investasi sebesar USD 147 juta atau sekitar RP 1,9 triliun. Sejauh ini PT. Tokopedia telah beberapa kali di anugerahi penghargaan antara lain : *Marketeers of the Year* 2014 untuk sektor *E-Commerce* pada acara *Markplus Conference* 2015 yang di gelar oleh *Markplus Inc* tanggal 11 Desember 2014. Pada tanggal 12 Mei 2016, Tokopedia terpilih sebagai *Best Company in Consumer Industry* dari Indonesia *Digital Economy Award* 2016. (Pratiwi,dkk., 2020).

3.3 *Twitter*

Twitter adalah layanan jejaring sosial *microblogging* gratis yang memungkinkan penggguna yang terdaftar menyiarkan pesan singkat yang disebut *tweet*s. Anggota *Twitter* dapat menyiarkan *tweet* dan mengikuti *tweet* pengguna lain dengan menggunakan beberapa *platform* dan perangkat. Pencipta *Twitter* ialah Jack Dorsey pada tahun 2006 dengan *link URL* http://www.*Twitter*.com. *Tweet*s bisa dilihat secara publik, namun pengirim dapat membatasi pengiriman pesan ke daftar teman-teman tertentu saja. Pengguna dapat melihat *Tweet*s pengguna lain yang dikenal dengan sebutan pengikut (*follower*). *Twitter* API *(Application Programming Interface)* merupakan sejumlah fungsi yang dapat digunakan pengembang perangkat lunak untuk mengolah data saat membangun perangkat lunak. *Twitter* API menyediakan beberapa fungsi untuk melakukan suatu tugas tertentu, sehingga pengembang perangkat lunak hanya memanggil fungsi tersebut di dalam perangkat lunak yang dibangun. *Twitter* API menggunakan arsitektur

REST *(Representational State Transfer)* sehingga *Twitter* API dapat digunakan pada format data yang beragam seperti XML maupun JSON. *Twitter* API terdiri atas *Twitter Search* API dan *Twitter Streaming* API. Perbedaan keduanya yaitu, *Twitter Search* API menitikberatkan fungsi pencarian ke masa lampau sedangkan *Twitter Streaming* API menitikberatkan fungsi pencarian ke masa yang akan datang (Rustiana,dkk., 2017).

Beberapa terminologi yang ada pada *Twitter* adalah sebagai berikut:

1. *Tweet*

   Tweet adalah pesan standar yang ada pada *Twitter* yang terdiri dari 140 karakter atau kurang.

2. *Retweet*

   Tweet yang telah di bagikan ulang oleh pengikut seseorang.

3. *Hashtag*

   Simbol # digunakan untuk menandai kata kunci atau topik pada *tweet* dengan tujuan untuk mempermudah pencarian.

4. *Mention*

   Tweet dapat berisi balasan dan mention dari pengguna lainnya dengan memberi awalan @ pada *username* mereka.

5. *Handle*

   Fitur ini menandakan *username* dan URL pada

6. *Feed*

   Kumpulan *tweet* pada beranda *Twitter* pengguna yang berisi akun – akun yang diikuti oleh pengguna tersebut.

7. *Lists*

   Mekanisme pembagian pengguna yang diikuti ke dalam beberapa urutan atau kelompok dan memunculkan *tweet* dari pengguna pada urutan tersebut.

8. *Direct Message*

   Juga dikenal dengan istilah DM, fitur ini memberikan dukungan untuk pemberian pesan secara langsung antar pengguna.

3.4 *Data Mining*

Suatu proses dengan menggunakan teknik statistik, matematika, *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan dari berbagai *database* yang besar disebut *data mining*. Istilah *data mining* mempunyai arti sebagai suatu disiplin ilmu yang bertujuan untuk menemukan, menggali, atau menambang pengetahuan dari data atau informasi. Data mining disebut sebagai *Knowledge Discovery in Database* atau KDD. Suatu kegiatan yang meliputi pengumpulan, pemakaian data historis dalam menemukan keteraturan, hubungan atau pola dalam dataset yang berukuran besar disebut KDD. (Nurhafidzah, 2019).

3.5 *Machine Learning*

Machine learning atau pembelajaran mesin adalah pendekatan dalam *Artificial Intelligence* (AI) yang banyak digunakan untuk menggantikan atau menirukan perilaku manusia untuk menyelesaikan masalah atau melakukan otomatisasi. *Machine Learning* mencoba menirukan bagaimana proses manusia atau makhluk cerdas belajar dan mengeneralisasi. Dua aplikasi utama dalam *machine mearning* yaitu, klasifikasi dan prediksi. Ciri khas dari *machine learning* adalah adanya proses pelatihan, pembelajaran, atau *training*. Oleh karena itu, *machine mearning* membutuhkan data untuk dipelajari yang disebut sebagai data *training.* Tujuan dari *text mining* adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen. Jadi, sumber data yang digunakan dalam *text mining* adalah sekumpulan teks yang memiliki format yang tidak terstruktur atau minimal semi terstruktur. Adapun tugas khusus dari *text mining* antara lain yaitu pengkategorisasian teks dan pengelompokkan teks. (Nurhafidzah, 2019).

3.6 *Text Mining*

Text mining merupakan ilmu yanng bertujuan untuk memproses teks agar menjadi informasi, menambang suatu data yang berupa teks yang bersumber dari data tersebut (Luqyana,dkk. 2018). Data yang biasanya diperoleh dari dokumen dan digunakan untuk mencari kata-kata yang dapat mewakili isi dari dokumen tersebut. (Farach & Nugraha, 2020).

3.7 *Text Preprocessing*

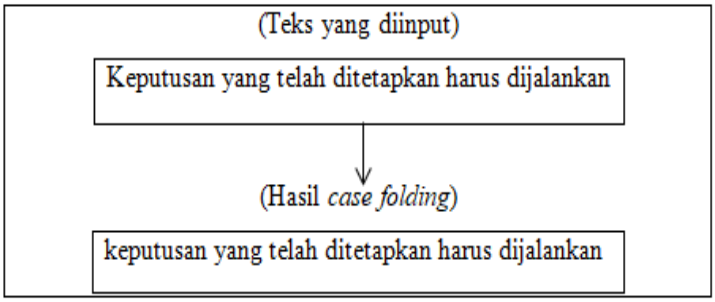
Text Preprocessing untuk merupakan tahapan awal pada *Text Mining* yang merupakan suatu proses yang mempersiapkan teks agar dapat diolah lebih lanjut. *Preprocessing* secara umum bertujuan untuk mengubah informasi dari tiap – tiap sumber data ke dalam bentuk atau format yang baku sebelum menerapkan berbagai metode – metode pengambilan data terhadap dokumen yang akan diproses (Fitria, 2018). Adapun tahapan *pre-processing* adalah :

1. *Cleaning Data*

   Cleaning Data tahapan yang bertujuan untuk membersihkan *tweet* dari tanda baca, *mention, hashtag, link* dan karakter lainnya dalam dokumen dengan seperti menghilangkan tanda baca seperti koma(,), titik(.), titik koma (;), titik dua (:),*mention*, RT, *hastagh* dan lainnya yang kurang penting yang bertujuan untuk mengurangi *noise*.

2. *Case Folding*

   Case folding adalah tahap mengubah semua huruf kapital dalam dokumen menjadi huruf kecil.

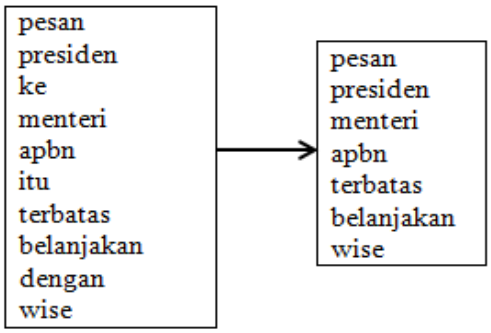


Gambar 3.1 Proses *Case Folding*

(sumber: Farach & Nugraha, 2019)

3. *Filtering*

   Filtering adalah tahapan proses penghapusan kata yang kurang penting, seperti kata ganti, kata hubung, kata keterangan, dan lain-lain dengan menggunakan *stopword* yakni daftar kata yang akan dihapus pada dokumen, pada *filtering* juga dilakukan penghapusan terhadap spasi yang berlebih akibat dari penghapusan beberapa kata.

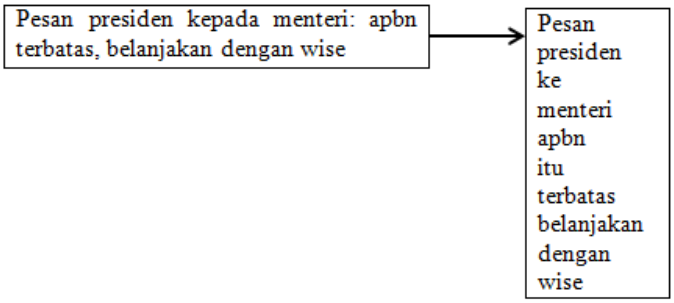


Gambar 3.2 Proses *Filtering*

(sumber: Farach & Nugraha, 2019)

4. *Tokenizing*

Tokenizing adalah proses memotong suatu kalimat menjadi beberapa bagian berdasarkan kata perkata. Potongan kata perkata tersebut disebut dengan token.



Gambar 3.3 Proses *Tokenizing*

(sumber: Farach & Nugraha, 2019)

3.8 Pembobotan Kata *(Term Weighting)*

Term Frequency-Inverse Document Frequency (TF-IDF) adalah metode yang digunakan untuk menghitung bobotan setiap kata yang telah diekstrak. Pemberian bobot hubungan pada suatu kata (*term*) terhadap dokumen sering dikenal dengan metode *Term Frequency-Inverse Document Frequency* (TF-IDF). Dengan menggunakan TF-IDF dapat menjadi ukuran statistik yang digunakan untuk mengetahui seberapa penting sebuah kata dalam sebuah dokumen. Model pembobotan TF-IDF merupakan metode yang mengintegrasikan model *term frequency* (tf) dan *inverse document frequency* (idf). *Term frequency (tf)* merupakan proses untuk menghitung jumlah kemunculan term dalam satu dokumen dan

inverse document frequency (idf) digunakan untuk menghitung *term* yang muncul diberbagai dokumen (komentar). (Farach&Nugraha, 2020).

Tahapan pembobotan dengan TF-IDF adalah (Farach&Nugraha, 2020).

1. Hitung *term frequency* $tf_{t,j}$

   $$tf_{t,j} = \frac{n_{i,j}}{\Sigma_k n_{i,j}} \quad (3.1)$$

2. Hitung *document frequency* (df)
3. Hitung bobot *inverse document frequency* (idf)

   $$idf = \log \frac{N}{df_t} \quad (3.2)$$

4. Hitung nilai bobot TF-IDF

   $$W_{i,j} = tf_{t,j} \times \text{idf} \quad (3.3)$$

   Ketarangan:

   $tf_{t,j}$ = frekuensi term

   $n_{i,j}$ = banyak kata i dalam dokumen j

   df = jumlah frekuensi dokumen yang mengandung term

   N = jumlah total dokumen

   $W_{i,j}$ = bobot TF-IDF

Contoh Perhitungan TF-IDF. (Nurhafizah, 2019).

Kalimat
Seharusnya janji bapak menjadi beban masa lalu

Langkah pertama, hitung TF adalah frekuensi kata pada suatu dokumen. Perhitungan TF dapat dilihat pada Tabel 3.1.

Tabel 3.1 Contoh Perhitungan TF

Kata/ Dokumen	**1**	···	**23**	···	**8261**
Seharusnya	0	···	1/7= 0.143	···	···
Janji	0	···	1/7= 0.143	···	···

Kata/ Dokumen	1	···	23	···	8261
Bapak	0	···	1/7= 0.143	···	···
Menjadi	1	···	1/7= 0.143	···	···
Beban	0	···	1/7= 0.143	···	···
Masa	0	···	1/7= 0.143	···	···
Lalu	0	···	1/7= 0.143	···	···
⋮	⋮	⋮	⋮	⋮	⋮
Jumlah	···	···	**7**	···	···

Berdasarkan **Tabel 3.1** dapat diketahui bahwa kata "seharusnya" pada dokumen 23 berjumlah 1 dan jumlah keseluruhan kata dalam dokumen 23 tersebut adalah 7 kata.

Sehingga didapatkan nilai TF untuk kata "seharusnya" yaitu 1/7 = 0,143. Perhitungan yang sama juga dilakukan untuk kata yang lain dalam dokumen. Setelah menghitung nilai TF, dilanjutkan dengan menghitung nilai IDF.

Document frequency (DF) adalah banyaknya dokumen dimana suatu *term* (t) muncul. Perhitungan IDF yaitu *log* dari jumlah seluruh dokumen dibagi dengan jumlah kata yang muncul dalam keseluruhan dokumen. Contoh perhitungan *Inverse Document Frequency (*IDF) dapat dilihat pada Tabel 3.2 dibawah ini.

Tabel 3.2 Contoh Perhitungan IDF

Kata/ Dokumen	1	...	23	...	8261	Jumlah Kata	IDF= log $\frac{N}{df}$
Seharusnya	0	···	1	···	···	5	$\log(\frac{8261}{5}) = 3.218$
Janji	0	···	1	···	···	24	$\log(\frac{8261}{24}) = 2{,}524$
Bapak	0	···	1	···	···	29	$\log(\frac{8261}{29}) = 2{,}454$

Kata/ Dokumen	1	...	23	...	...	8261	IDF= log $\frac{N}{df}$
Menjadi	0	...	1	...	...	32	$\log(\frac{8261}{32}) = 2{,}411$
Beban	0	...	0	...	...	14	$\log(\frac{8261}{14}) = 2{,}77$
Masa	0	...	1	...	...	26	$\log(\frac{8261}{26}) = 2{,}502$
Lalu	0	...	1	...	...	27	$\log(\frac{8261}{27}) = 2{,}485$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Jumlah	...	...	7	...	...	...	...

Berdasarkan **Tabel 3.2** dapat diketahui bahwa jumlah kata "janji" pada keseluruhan dokumen adalah 24 kata. Sehingga didapatkan nilai IDF untuk kata "janji" yaitu 2.524. Perhitungan yang sama juga dilakukan untuk kata yang lain dalam dokumen.

Setelah didapatkan nilai TF dan nilai IDF, maka kemudian akan menghitung nilai TF-IDF yang digunakan pembobotan kata dalam analisis klasifikasi. Nilai TF-IDF diperoleh dengan mengalikan nilai TF dengan nilai IDF. Perhitungan TF-IDF dapat dilihat pada Tabel 3.3.

Tabel 3.3 Contoh Perhitungan TF-IDF

Kata/ dokumen	TF					IDF	TF-IDF				
	1	...	23	...	8261		1	...	23	...	8261
Seharusnya	0	...	0.143	...	...	3,218	0	...	0,465	...	...
Janji	0	...	0.143	...	...	2,524	0	...	0,361	...	...
Bapak	0	...	0.143	...	...	2,454	0	...	0,351	...	...

⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Menjadi	0	...	0.143	...	...	2,411	0	...	0,344	...	...
Beban	0	...	0.143	...	...	2,77	0	...	0.396	...	...
Kata/ dokumen	**TF**					**IDF**	**TF-IDF**				
	1	...	**23**	...	**8261**		**1**	...	**23**	...	**8261**
Masa	0	...	0.143	...	...	2,502	0	...	0,360	...	...
Lalu	0	...	0.143	...	...	2,485	0	...	0,355	...	...

Berdasarkan **Tabel 3.3** dapat diketahui bahwa nilai TF pada kata “bapak” adalah 0,143 sedangkan nilai IDF pada kata “bapak” yaitu 2,454. Sehingga didapatkan nilai TF-IDF dengan mengalikan nilai TF dan IDF didapatkan hasil 0,351. Sehingga, dengan nilai yang didapatkan diartikan bahwa semakin tinggi nilai, semakin penting kata tersebut.

3.9 Wordcloud

Wordcloud merupakan sebuah sistem yang memunculkan visualisasi kata-kata dengan memberikan penekanan pada frekuensi kemunculan kata terkait dalam wacana tertulis (Fitria, 2018).

Gambar 3.4 Tampilan *Wordcloud*

(Sumber : Adiyana & Hakim, 2015)

3.10 Asosiasi Kata

Asosiasi kata dapat digunakan untuk mengetahui kata apa saja yang sering muncul pada sebuah dokumen. Asosiasi kata juga dapat mengetahui keterkaitan dan hubungan antar kata, misalnya antar dua kata atau lebih digunakan secara bersamaan dalam sebuah dokumen. Dalam asosiasi kata dapat juga dilihat dari nilai korelasi antar kata, dimana nilai korelasi berkisar antara -1 sampai 1. Jika nilai mendekati 1 atau -1 maka hubungan antar kata tersebut semakin kuat, sedangkan jika nilai mendekati 0 maka hubungan antar kata semakin lemah. Ada beberapa kategori nilai korelasi yang digunakan sebagai berikut**.** (Farach dan Nugraha, 2019).

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

$$r_{xy} = \frac{n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{\sqrt{(n\Sigma x_i^2 - (\Sigma x_i)^2\}\{n\Sigma y_i^2 - (\Sigma y_i)^2\}}} \quad (3.4)$$

Dengan

r_{xy} = nilai korelasi antar variabel x dan variabel y

n = banyaknya pasangan data x dan y

Σx_i = jumlah nilai pada variabel , i = 1,2,3,....., n

Σy_i = jumalah nilai pada variabel y

Σx_i^2 = kuadrat dari total nilai variabel x

Σy_i^2 = kuadrat dari total nilai variabel y

$\Sigma x_i \, \Sigma y_i$ = jumlah dari hasil perkalian antara nilai variabel x dan variabel y

Simulasi perhitungan Asosiasi kata Proses perhitungan asosiasi ini menggunakan pendekatan korelasi. Sebelumnya, kata akan diubah (transformasi) menjadi document term matrix. Berikut ini merupakan contoh perhitungan asosiasi kata dengan menggunakan 6 dokumen dan 6 kumpulan kata (Adawiyah, 2018).

1. Dokumen 1 aplikasi
   Dokumen 2 aplikasi gagal
   Dokumen 3 aplikasi gagal mendaftarkan
   Dokumen 4 aplikasi gagal mendaftarkan rekening
   Dokumen 5 aplikasi gagal mendaftarkan rekening tujuan
   Dokumen 6 aplikasi gagal mendaftarkan rekening tujuan transfer.

2. Kemudian kumpulan kata di atas dibentuk ke dalam *document term matrix*

Dok	aplikasi	gagal	mendaftarkan	rekening	tujuan	transfer
1	1	0	0	0	0	0
2	1	1	0	0	0	0
3	1	1	1	0	0	0
4	1	1	1	1	0	0
5	1	1	1	1	1	0
6	1	1	1	1	1	1

3. Selanjutnya akan dilakukan perhitungan dengan rumus korelasi mendapatkan asosiasi kata. Pada contoh ini akan dilakukan perhitungan untuk asosiasi kata gagal dan rekening.

Dok	Gagal	rekening	gagal^2	rekening^2	gagal*rekening
1	0	0	0	0	0
2	1	0	1	0	0
3	1	0	1	0	0
4	1	1	1	1	1
5	1	1	1	1	1
6	1	1	1	1	1
total	5	3	5	3	3

$$r = \frac{n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{\sqrt{(n\Sigma x_i^2 - (\Sigma x_i)^2\}\{n\Sigma y_i^2 - (\Sigma y_i)^2\}}} \quad (3.5)$$

$$r = \frac{(6*3) - 5*3}{\sqrt{\{(6*5) - (5^2)\}\{6*3) - (3^2) - (3)^2\}}} = \frac{3}{\sqrt{45}} = 0.447 \quad (3.6)$$

Jadi, berdasarkan perhitungan di atas, diketahui bahwa besar asosiasi kata antara gagal dengan rekening yaitu 0.447, yang artinya korelasi cukup.

```
> findAssocs(myTdm, 'kpk', 0.3)
       kpk
polri 0.50
lapor 0.35
```

Gambar 3.5 Tampilan Asosiasi Kata

(Sumber : Adiyana & Hakim, 2015)

3.11 *Sentiment Analysis*

Sentiment Analysis dilakukan untuk melihat pendapat atau kecenderungan pendapat terhadap suatu masalah atau objek oleh seseorang, apakah cenderung berpendapat negatif atau positif. *Sentiment Analysis* biasa dilakukan untuk memantau perkembangan pasar atau menanggapi suatu permasalahan, salah satu contoh penggunaannya di dunia nyata adalah indentifikasi kecenderungan pasar

atau pendapat terhadap suatu objek. Analisis sentimen juga menganalisis sebagian data untuk mengetahui emosi manusia. Analisis sentimen dapat dikategorikan kedalam tiga *task*, yaitu *informative text detection, information extraction* dan *sentiment interestingness classification* (*emotional, polarity indentification*). *Sentiment classification* (negatif atau positif) digunakan untuk memprediksi *sentiment polarity* berdasarkan data sentimen dari pengguna. (Fitria, 2018).

Contoh perhitungan skor sentimen sebagai berikut. (Santoso & Nugroho, 2019). Berdasarkan teks komentar "Jalan tol yang sudah dikerjakan sangat bagus dan indah, tapi untuk pembayaran tolnya sangat mahal", terdapat 1 kata negatif dan 2 kata positif yang terdeteksi pada kamus lexicon, yaitu "bagus" dan "indah" sebagai kata positif, untuk kata negatif yaitu "mahal" sebagai kata negatif.

Adapun rumus yang digunakan dalam proses perhitungan skor sentimen adalah sebagai berikut:

$$\text{Skor} = (\sum \text{kata positif}) - (\sum \text{kata negatif}) \tag{3.7}$$

Tabel 3.4 Perhitungan Skor

	Jumlah Kata Positif	**Jumlah Kata Negatif**
Jalan tol yang sudah dikerjakan sangat bagus dan indah, tapi untuk pembayaran tolnya sangat mahal	Bagus	Mahal
	Indah	
Total Jumlah Kata	**2**	**1**

Maka, diperoleh hasil dari perhitungan: Skor = (Jumlah kata positif) – (Jumlah kata negatif) Skor = 2 – 1 = 1 Nilai akhir yang diperoleh dari perhitungan menghasilkan skor 1 atau > 0, maka diidentifikasi kata positif.

3.12 Klasifikasi

Klasifikasi merupakan suatu proses penemuan model (atau fungsi) yang membedakan kelas data atau konsep yang bertujuan agar dapat digunakan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui. Model

ditemukan berdasarkan analisis data training (objek data yang kelasnya diketahui). Klasifikasi juga merupakan suatu proses yang memiliki tujuan untuk mempelajari fitur dari sebuah atau sekelompok data yang sudah diketahui kelasnya dan menghasilkan sebuah model klasifikasi yang dapat digunakan untuk memprediksi kelas atau label data yang baru. Pada klasifikasi terdapat beberapa Algoritma yang sering digunakan, yaitu *K-Nearest Neighbor*, *Rough Set*, Algoritma Genetika, Metode *Rule Based*, C4.5, *Naive Bayes Classifier* (NBC), *Memory Based Reasoning*, dan *Support Vector Machines* (SVM). (Nurhafidzah, 2019).

Pada proses klasifikasi data terdiri dari 2 langkah proses. Pertama adalah *learning* (*fase training*), dimana algoritma klasifikasi dibuat untuk menganalisa data *training* lalu direpresentasikan dalam bentuk aturan klasifikasi. Proses kedua dimana data tes digunakan untuk memperkirakan akurasi dari aturan klasifikasi.

Berdasarkan hal itu, data yang terdapat pada data *testing* semestinya tidak terdapat pada data *training* sehingga dapat diketahui model klasifikasi dapat melakukan klasifikasi dengan baik dan benar. Pada proses pembagian antara data *training* dan data *testing* tidak mengikat akan tetapi agar variasi dalam model tidak terlalu besar maka dianjurkan data *training* memiliki perbandingan lebih besar dari pada data *testing.* Pada umumnya 2/3 dari total data dijadikan sebagai data *training* sementara sisanya dijadikan sebagai data *testing.* Penelitian yang menghasilkan keakuratan model klasifikasi optimum dengan proporsi 80:20 untuk data *training* dan data *testing.* (Nurhafidzah, 2019).

3.13 *Naïve Bayes Classifier*

Metode *Naïve Bayes Classifier* (NBC) didasari oleh *teorema bayes* yang ditemukan oleh Thomas Bayes pada abad ke 18. *Naïve Bayes Classifier* (NBC) merupakan salah satu teknik klasifikasi dalam statistik dimana pengklasifikasian ini dapat memprediksi probabilitas keanggotaan kelas suatu data. *Naïve Bayes Classifier* (NBC) mengasumsikan bahwa setiap atribut dalam data bersifat saling terpisah atau independen dan juga metode ini mengklasifikasikan kelas berdasarkan pada probabilitas sederhana. Klasifikasi *bayes* sederhana yang lebih dikenal dengan Naïve Bayes Classifier dapat diasumsikan bahwa efek dari atribut suatu kelas yang diberikan adalah bebas dari atribut lain yang disebut dengan *class conditional*

independence. Class conditional independence ini dianggap sebagai "*naive*' dan dibuat untuk memudahkan perhitungan – perhitungan yang artinya bahwa kemunculan suatu *term* kata dalam suatu kalimat tidak dipengaruhi oleh kata – kata lain, namun dalam kenyataannya bahwa kemungkinan kata dalam kalimat sangat dipengaruhi oleh kemungkinan keberadaan kata – kata yang ada dalam kalimat (Meimunah, 2019).

Persamaan yang digunakan dalam metode *Naïve Bayes Classifier* (NBC) adalah sebagai berikut: (Meimunah, 2019).

$$P\ (V|X) = \frac{P(X|V).P(V)}{P(X)} \tag{3.8}$$

Keterangan :

X : Data dengan kelas yang belum diketahui

V : Hipotesis data merupakan suatu kelas spesifik

P(V|X) :Probabilitas hipotesis V berdasar kondisi X (*posterior* probabilitas)

P(V) : Probabilitas hipotesis V (*prior* probabilitas)

P(X|V) : Probabilitas hipotesis X berdasarkan kondisi V

P(X) : Probabilitas X

Proses klasifikasi dalam metode *Naive Bayes* memerlukan sejumlah petunjuk untuk menentukan kelas apa yang cocok untuk sampel yang dianalisi tersebut. Oleh karena itu, persamaan yang digunakan dalam metode Naïve Bayes Classification (NBC) adalah sebagai berikut:

$$P(V|X_1 \ldots . X_n) = \frac{P(V)P(X_1 \ldots X_n|V)}{P(V|X_1 \ldots X_n)} \tag{3.9}$$

Variabel V pada persamaan 3.9 merepresentasikan kelas, sedangkan variabel $X_1 \ \ldots \ . X_n$ menjelaskan karakterisik petunjuk yang dibutuhkan untuk melakukan klasifikasi. Rumus pada persamaan 3.9. Menjelaskan bahwa peluang masuknya sampel karakteristik tertentu dalam kelas V (*Posterior*) adalah peluang munculnya kelas V (sebelum masuknya sampel tersebut, sering disebut *prior*), dikali dengan peluang kemunculan karakterisik – karakteristik sampel pada kelas V (disebut dengan *likelihood*), kemudian dibagi dengan peluang kemunculan

karakteristikkarakteristik sampel secara global (disebut *evidence*). Oleh sebab itu persamaan 3.9. Dapat disederhanakan menjadi persamaan berikut:

$$Posterior = \frac{Prior \times Likelihood}{Evidance} \tag{3.10}$$

Nilai *evidence* pada persamaan selalu tetap untuk setiap kelas pada satu sampel. Nilai yang didapatkan dari posterior tersebut nantinya akan dibandingkan dengan nilai – nilai posterior kelas lainnya untuk menentukan ke dalam kelas apa suau sampel akan diklasifikasikan. Dari persamaan akan dijabarkan lebih lanjut dengan menjabarkan $(V|X_1 \ldots . X_n)$ menggunakan aturan perkalian seperti pada persamaan berikut :

$$P(V|X_1,....X_n) = P(V)P(X_1,...X_n|V) \tag{3.11}$$

$$= P(V)P(X_1|V)P(X_2,...,X_n|V,X_1)$$

$$= P(V)P(X_1|V)P(X_2|V,X_1)P(X_3,...,X_n|V,X_1,X_2)$$

$$= P(V)P((X_1|V)P(X_2|V,X_1)P(X_3|V,X_1,X_2)P(X_4,...,X_n|V,X_1,X_2,X_3)$$

$$= P(V)P((X_1|C)P(X_2|C,X_1)P(X_3|C,X_1,X_2)...P(X_n|C,X_1,X_2,X_3,....,X_{n-1})$$

Hasil penjabaran dari persamaan menyebabkan semakin banyak dan semakin kompleksnya faktor – faktor yang hampir mustahil untuk dianalisa satu persatu. Oleh sebab itu persamaan tersebut menjadi sulit untuk dilakukan. Pada metode inilah digunakan asumsi independensi yang sangat tinggi (*naive*) bahwa masing – masing petunjuk $(X_1, X_2 \ldots X_n)$ independen (saling bebas) satu sama lain. Berdasarkan asumsi tersebut maka berlaku suatu kesamaan sebagai berikut :

$$P(X_i|X_j) = \frac{P(X_i \cap X_j)}{P(X_j)} = \frac{P(F_i)\,P(F_j)}{P(F_j)} = P(F_i) \tag{3.12}$$

dimana i≠j, sehingga:

$$P(F_i|C, F_j) = P(F_i|C) \tag{3.13}$$

Salah satu kelebihan metode Naïve Bayes adalah metode ini hanya membutuhkan jumlah data training yang sedikit untuk menentukan estimasi parameter yanng diperlukan dalam proses klasifikasi dan menghasilkan nilai akurasi yang tinggi sehingga proses klasifikasi menjadi cepat dan sederhana. Dalam

algoritma naïve bayes classifier setiap dokumen dipresentasikan dengan pasangan atribut "$x_1, x_2, x_3 \dots x_n$" dimana x_1 merupakan kata pertama, x_2 merupakan kata kedua dan seterusnya, sedangkan himpuan kelas disimbolkan dengan "V". Pada waktu klasifikasi algoritma akan mencari probabilitas tertinggi dari semua kategori kategori yang diujikan yang disimbolkan dengan "VMAP" dan dituliskan dengan persamaan dari *naïve bayes classifier* berikut (Meimunah,2019):

$$V_{MAP} = argmax_{vj\epsilon v}\left(\frac{P(x_1, x_2, x_3 \dots x_n|Vj)P(V_j)}{P(x_1,x_2,x_3 \dots x_n)}\right) \quad (3.14)$$

$P(x_1, x_2, x_3 \dots x_n)$ memiliki nilai konstan untuk semua kategori (V_j) sehingga persamaan 3.14 menjadi 3.15

$$V_{MAP} = argmax_{vj\epsilon v}P\left(x_1, x_2, x_3 \dots x_n|V_j\right)P(V_j) \quad (3.15)$$

Menghitung $P(x_1, x_2, x_3 \dots x_n)$ akan memiliki tingkat kesulitan yang tinggi karena jumlah *term*/kata $P(x_1, x_2, x_3 \dots x_n)$ tergantung jumlah kombinasi posisi kata dikalikan jumlah kelas, sehingga persamaan 3.15 dapat disederhanakan menjadi persamaan 3.16.

$$V_{MAP} = argmax_{vj\epsilon v} \prod_{i=1}^{n}\left(P(x_1|V_j)P(V_j)\right) \quad (3.16)$$

Keterangan :

V_j = kategori sentimen j=1, 2 dimana pada penelitian ini j_1 = kategori positif, j_2 = kategori negatif

$P\left(x_i\middle|V_j\right)$ = probabilitas kata x_i pada kategori V_j

$P(V_j)$ = probabilitas dari V_j

Untuk $P(V_j)$ dan $P\left(x_1\middle|V_j\right)$ dihitung pada saat pelatihan dimana menggunakan

$$P(V_j) = \frac{n\ docs\ j}{n\ contoh} \quad (3.17)$$

$$P(x_1|Vj) = \frac{n_k+1}{n + |n\ kosakata|} \quad (3.18)$$

Keterangan :

$n\ docs\ j$ = jumlah dokumen setiap kategori j

$n\ contoh$ = jumlah dokumen dari semua kategori

n_k = jumlah frekuensi kemunculan setiap kata

n = jumlah frekuensi kemunculan kata dari setiap kategori

$|n\ kosakata|$ = jumlah semua kata dari semua kategori

Simulasi Contoh Soal dengan Perhitungan Metode *Naive Bayes Classifier***.** Misalkan terdapat empat buah dokumen yang telah melalui tahapan *preprocessing*, dua dokumen diambil dari kelas positif dan dua dokumen diambil dari kelas negatif. Dokumen tersebut adalah sebagi berikut. (Fitria, 2018).

Dokumen 1 : beli tiket pesawat murah situs

Dokumen 2 : beli tiket periode promo maret

Dokumen 3 : server tiket habis

Dokumen 4 : habis beli tiket buka Twitter

Selanjutnya akan dihitung frekuensi kemunculan kata pada setiap dokumen seperti pada **tabel 3.5**.

Tabel 3.5 Frekuensi Kemunculan Kata

Query	**Dokumen**			
	1	**2**	**3**	**4**
Beli	1	1	0	1
Tiket	1	1	1	1
Pesawat	1	0	0	0
Murah	1	0	0	0
Situs	1	0	0	0
periode	0	1	0	0
Promo	0	1	0	0
Maret	0	1	0	0
server	0	0	1	0
Habis	0	0	1	1

Buka	0	0	0	1
Twitter	0	0	0	1
Kelas	**Positif**	**Positif**	**Negatif**	**Negatif**

Berdasarkan **tabel 3.5** diketahui bahwa kelas positif terdiri dari 2 dokumen dengan jumlah kata sebanyak 8 kata dari 12 kosakata yang ada, sedangkan 2 dokumen kelas negatif terdiri dari 6 kata dari 12 kosakata yang ada. Berdasarkan jumlah kata tersebut, dapat dihitung nilai probabilitas untuk setiap kelasnya.

1. Probabilitas kata kelas positif

   Contoh perhitungan probabilitas untuk kata "beli" yang terdapat dalam kelas positif.

$$P\ \alpha_i\ v_j = \frac{n_i+1}{n + kosakata} \quad (3.19)$$

$$P\ \ beli\ positif = \frac{2+1}{8+\ 12} = 0{,}15 \quad (3.20)$$

dimana,

n_i : jumlah kata "beli" dalam dokumen kelas positif

n : jumlah seluruh kata pada dokumen kelas positif

$kosakata$: jumlah kata dalam fase *training*

Nilai probabilitas untuk kata yang lain dalam kelas positif ditunjukkan oleh

Tabel 3.6 Probabilitas Kata Kelas Positif

Query	Probabilitas	Query	Probabilitas
beli	0.15	Promo	0.1
tiket	0.15	Maret	0.1
pesawat	0.1	Situs	0.1
murah	0.1	Periode	0.1

2. Probabilitas kata kelas negatif

   Contoh perhitungan kata "beli" pada kelas negatif.

$$P\ \alpha_i\ v_j = \frac{n_i+1}{n + kosakata} \quad (3.21)$$

$$P\ \ beli\ negatif = \frac{2+1}{6+\ 12} = 0{,}11 \quad (3.22)$$

dimana,

n_i : jumlah kata "beli" dalam dokumen kelas negatif

n : jumlah seluruh kata pada dokumen kelas negatif

$kosakata$: jumlah kata dalam fase *training*

Nilai probabilitas untuk kata yang lain dalam kelas negatif ditunjukkan oleh.

Tabel 3.7 Probabilitas Kata Kelas Negatif

Query	Probabilitas
Beli	0.11
Tiket	0.16
Server	0.11
Habis	0.16
Buka	0.11
Twitter	0.11

Nilai probabilitas kata pada masing-masing kelas tersebut, kemudian disimpan pada *database* yang nantinya akan digunakan untuk menguji data baru. Misal ingin diketahui kelas data dari tanggapan baru "tiket pesawat promo murah". Langkah pertama yang dilakukan untuk melakukan klasifikasi adalah memecah kalimat dalam tanggapan tersebut menjadi kata per kata kemudian menghitung nilai probabilitas dari kata pada masing-masing kelas dengan menggunakan tabel probabilitas kata yang telah diperoleh sebelumnya, sedangkan probabilitas masing-masing kelas ditentukan dengan menggunakan.

$$P\ v_j = \frac{doc_j}{training} \quad (3.23)$$

$$P\ \ beli\ positif = \frac{2}{4} = 0{,}5 \quad (3.24)$$

$$P\ \ beli\ negatif = \frac{2}{4} = 0{,}5 \quad (3.25)$$

Probabilitas untuk tanggapan baru yang ingin diklasifikasikan ditentukan

dengan menggunakan. Nilai hasil perhitungan ditunjukan oleh.

Tabel 3.8 Nilai Probabilitas Tanggapan Baru

Kelas	tiket	pesawat	Promo	Murah	Nilai Probabilitas
Positif (P=0,5)	0.15	0.1	0.1	0.1	0.95
Negatif (P=0,5)	0.16	0.05	0.05	0.05	0.81

$$P \quad positif = 0.5 + 0.15 + 0.1 + 0.1 + 0.1 = 0.95$$

$$P \quad negatif = 0.5 + 0.16 + 0.05 + 0.05 + 0.5 = 0.81$$

Hasil klasifikasi dari kelas tanggapan baru tersebut adalah kelas atau kategori yang memiliki probabilitas tertinggi. Berdasarkan **tabel 3.8** diketahui bahwa nilai probabilitas tertinggi adalah probabilitas kelas positif, maka tanggapan baru "tiket pesawat promo murah" masuk ke dalam kelas positif.

3.14 *Confusion Matrix*

Confusion matrix adalah suatu metode yang biasanya digunakan untuk melakukan perhitungan akurasi pada konsep *data mining*. *Confusion matrix* digambarkan dengan tabel yang menyatakan jumlah data uji yang benar diklasifikasikan dan jumlah data uji yang salah diklasifikasikan. (Fitria, 2018).

Pengukuran efektif dapat dilakukan dengan perhitungan perolehan atau *recall*, nilai ketepatan atau presisi, nilai akurasi, dan nilai *spesificity*.

Tabel 3.9 Confusion Matrix

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

1. *True Negative* (TN) merupakan kelas yang dihasilkan dari prediksi pada klasifikasi negatif padahal kelas sebenarnya negatif.

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

Dari tabel diatas, didapatkan perhitungan *recall*, presisi, akurasi, dan perhitungan lainnya dalam rumus sebagai berikut: (Khairani, 2019).

Recall merupakan proporsi jumlah yang dapat ditemukan kembali dalam proses pencarian.

$$Recall = \frac{TP}{FP + FN} x100 \quad (3.26)$$

Presisi merupakan proporsi jumlah dokumen yang ditemukan dan dianggap relevan untuk kebutuhan suatu informasi.

$$\text{Presisi} = \frac{TP}{FP + TP} x100 \quad (3.27)$$

Akurasi adalah nilai ketepatan suatu klasifikasi dalam bentuk persen.

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} x100 \quad (3.28)$$

Spesificity digunakan untuk mengukur proporsi negatif yang benar diidentifikasi.

$$Spesificity = \frac{TN}{TN + FP} x100 \quad (3.29)$$

False Positive Rate (FPR) = 1-spesificity (3.30)

$$Area\ Under\ Curve = \frac{1 + Recall - FPR}{2} \quad (3.31)$$

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

Tabel 3.10 Nilai *Area Under Curve* (AUC)

Nilai AUC	Keterangan
0.91 - 1.00	Klasifikasi sangat baik
0.81 – 0.90	Klasifikasi baik
0.71 – 0.80	Klasifikasi cukup
0.61 – 0.70	Klasifikasi buruk
≤ 0.60	Klasifikasi salah

(Sumber : Fitria, 2018)

BAB IV
METODE PENENELITIAN

4.1 Populasi dan Sampel

Populasi dalam penelitian ini adalah *tweet* dan *reetweet* dari media sosia Twitter. Sampel penelitian ini yaitu selama 4 hari pada tanggal 25 September 2020-28 September 2020. Sedangkan total sampel yang digunakan dalam penelitian sebesar 14243 *tweet* dalam *Twitter*.

4.2 Variabel Penelitian dan Definisi Operasional Variabel

Variabel penelitian yang digunakan dalam penelitian ini dapat dilihat pada tabel berikut:

Tabel 4 Variabel Penelitian

Variabel	Definisi
Tweet	Kicauan atau status yang diposting pada *Twitter*
Retweet	Mengulang atau memposting kembali sebuah *tweet* dalam *Twitter*

4.3 Metode Pengambilan Data

Dalam melakukan pengambilan data dari *Twitter* dengan metode scrapping menggunakan *Twitter* API. Menggunakan kata kunci "tokopedia", "tokped". Pada penelitian ini, menggunakan metode dengan teknik *web scrapping*. Alat yang digunakan untuk *crawling* data adalah *R Studio*.

4.4 Metode Analisis Data

Dalam melakukan analisis data terdapat beberapa bantuan menggunakan *API Twitter*, *software Rstudio*, dan *Microsoft Excel*. Beberapa metode yang digunakan dalam penelitian ini yaitu:

1. *Text Mining*

   Digunakan dalam melakukan analisis data yang berupa teks yang tidak terstruktur.

2. *Wordcloud*

   Digunakan untuk menampilkan visualisasi data yang paling banyak digunakan.

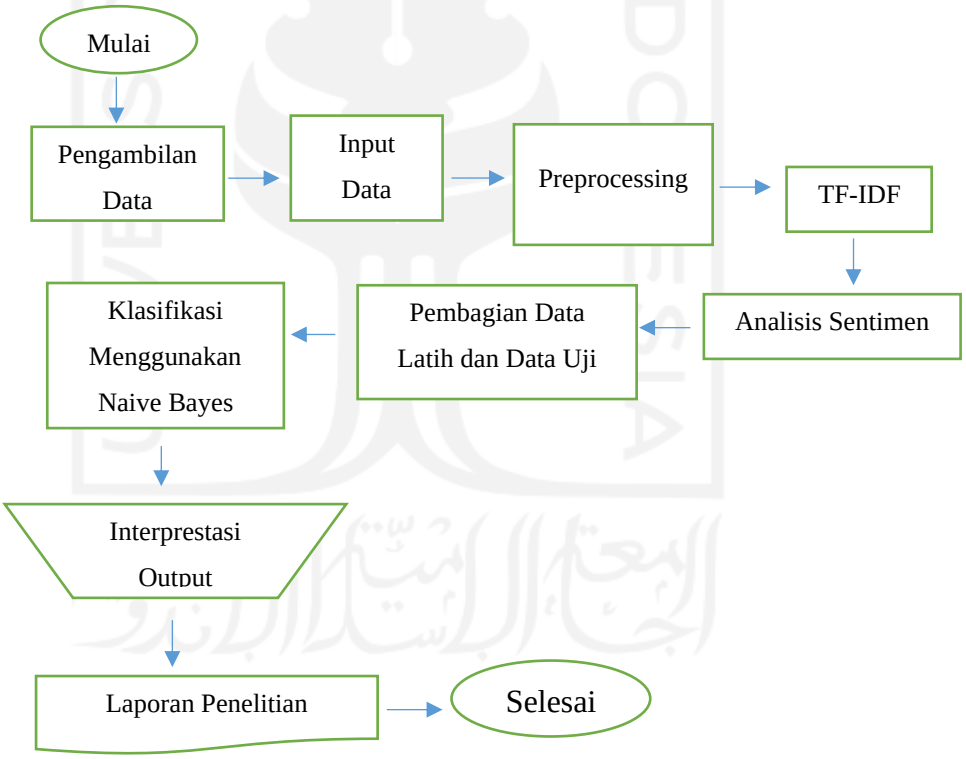
3. Analisis Sentimen

   Analisis sentimen bertujuan untuk pelabelan kata-kata yang terbagi menjadi 2 yaitu positif, dan negatif.

4. Klasifikasi *Naive Bayes*

   Digunakan untuk mengklasifikan *tweet* berdasarkan *tweets* positif, negatif. Selain itu *Naive Bayes* juga digunakan untuk melihat tingkat akurasi dalam klasifikasi data.

4.5 Tahapan Penelitian



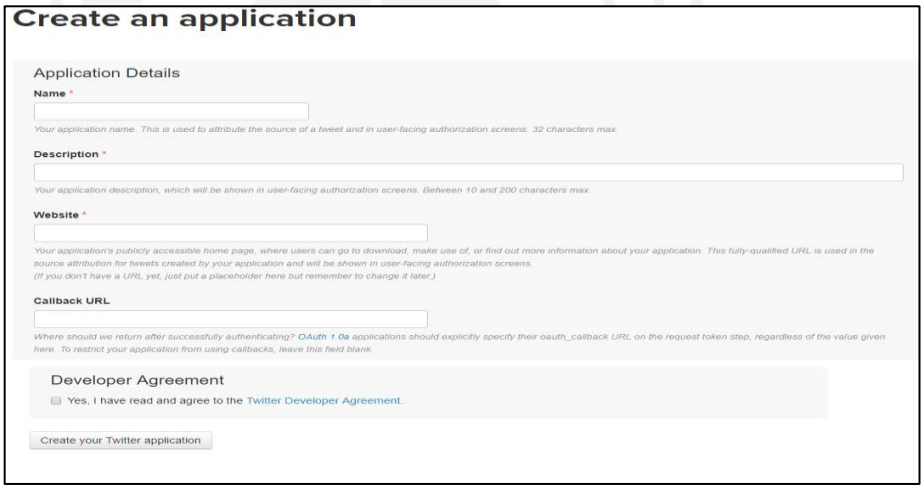
Gambar 4.1 Tahapan Penelitian

BAB V
PEMBAHASAN

Pada bab ini, akan menjelaskan mengenai hasil dan pembahasan dari penelitian yang telah dilakukan oleh peneliti.

5.1 *Authentication*

Untuk memulai pengambilan data dari *Twitter*, hal pertama yang di lakukan adalah membuat kode API pada *Twitter*. Kode API didapatkan dengan mengakses *Twitter* API, yaitu sebuah aplikasi yang bertujuan untuk mempermudah *developer* untuk mengakses informasi *web Twitter*. Kemudian, peneliti diharuskan untuk melakukan registrasi terlebih dahulu.



Gambar 5.1 Form Registrasi

Gambar 5.1 merupakan *form* yang harus diisi pada saat melakukan registrasi.

Pendaftaran ini bertujuan untuk menginformasikan kepada pihak *Twitter* tentang tujuan peneliti dalam mengakses atau mengambil data pada *Twitter*. Setelah berhasil melakukan registrasi maka didapatkan beberapa kode, yaitu *consumer key*, *consumer secret*, *access token*, dan *access key* dari *Twitter* terdapat pada lampiran 1. Kode-kode tersebut merupakan penghubung antara *Twitter* dan aplikasi lainnya, dalam hal ini peneliti menggunakan aplikasi *R Studio*.

5.2 *Scrapping* Data *Twitter*

Pada proses scraping atau pengambilan data pada *Twitter* kata kunci yang digunakan adalah “tokopedia”, dan “tokped ”.

Tabel 5.1 Data *Tweet*

No.	Text	Created	Retweet Count
1	RT @tokopedia: <U+2728><U+2728> MENANGKAN ALBUM/MERCHANDISE ITZY UNTUK 8 PEMENANG ITZY	25/10/20120	3797
2	@tokopedia Terimakasih TOKOPEDIA karena telah mengundang TREASURE ke WIB TV SHOW<U+2764>? https://t.co/n6Etl6VZtc	25/10/2020	0
3	@bhecex90 @tokopedia @tokopediacare terima kasih kembali dan sukses selalu<f0><U+009F><U+0098><U+008A>	25/10/2020	0

Tabel 5.1 adalah beberapa hasil dari *scraping* data pada *Twitter* mengenai Tokopedia, hasil yang didapatkan berjumlah 14243 *tweets* mulai tanggal 25 sept 2020 sampe tgl 28 sept 2020. Pada hasil tersebut terdapat data *text, created¸* dan *retweet count*. Data *text* berisikan kicauan atau *tweet* yang dituliskan oleh pengguna *Twitter*. *Created* adalah data tanggal kapan *tweet* dibuat, sedangkan *retweet count* merupakan data yang berisikan banyaknya pengguna *Twitter* lain ikut menyebarkan ulang *tweet* aslinya.

5.3 *Preprocessing*

Penelitian ini menggunakan data komentar Tokopedia tahun 2020 yang diambil dari media sosial *Twitter*. Berikut merupakan sebagian data yang digunakan dalam penelitian.

Tabel 5.2 Data Awal Penelitian

Data Awal Komentar
RT @tokopedia: <U+2728><U+2728> MENANGKAN ALBUM/MERCHANDISE ITZY UNTUK 8 PEMENANG ITZY
@tokopedia Terimakasih TOKOPEDIA karena telah mengundang TREASURE ke WIB TV SHOW<U+2764>? https://t.co/n6Etl6VZtc
@bhecex90 @tokopedia @tokopediacare terima kasih kembali dan sukses selalu<f0><U+009F><U+0098><U+008A>

Pada **Tabel 5.2** menunjukkan sebagian data komentar Tokopedia tahun 2020. Data komentar akan melalui beberapa tahapan *preprocessing* yang meliputi *cleaning, case folding, filtering,tokenizing,*

1. Cleaning Data

Proses *cleaning* data adalah menghilangkan URL, *retweet, username,hastag,* dan tanda baca seperti titik koma, titik, spasi.

Tabel 5.3 Proses *Cleaning* Data

No	Sebelum *Cleaning*	Sesudah *Cleaning*
1	RT @tokopedia: <U+2728><U+2728> MENANGKAN ALBUM/MERCHANDISE ITZY UNTUK 8 PEMENANG ITZY	MENANGKAN ALBUM/MERCHANDISE ITZY UNTUK 8 PEMENANG ITZY
2	@tokopedia Terimakasih TOKOPEDIA karena telah mengundang TREASURE ke WIB TV SHOW<U+2764>? https://t.co/n6Etl6VZtc	Terimakasih TOKOPEDIA karena telah mengundang TREASURE ke WIB TV SHOW
3	@bhecex90 @tokopedia @tokopediacare terima kasih kembali dan sukses selalu<f0><U+009F><U+0098><U+008A>	terima kasih kembali dan sukses selalu

2. *Case Folding*

Case Folding adalah tahapan untuk mengubah semua huruf capital menjadi huruf non capital atau kecil.

Tabel 5.4 Proses *Case Folding*

No	Sebelum *Case Folding*	Sesudah *Case Folding*
1	MENANGKAN ALBUM/MERCHANDISE ITZY UNTUK 8 PEMENANG ITZY	menangkan albummerchandise itzy untuk 8 pemenang
2	Terimakasih TOKOPEDIA karena telah mengundang TREASURE ke WIB TV SHOW	Terimakasih tokopedia karena mengundang treasureke wib tv show
3	terima kasih kembali dan sukses selalu	terima kasih kembali dan sukses selalu

3. *Filtering*

Proses *filtering* adalah proses penghapusan yang tidak perlu dengan menggunakan stopword yang ada, seperti kata “yang”, “dan”,”ke”,”dari”,”oleh”,”agak” dan lain-lain.

Tabel 5.5 Proses Filtering

No	Sebelum *Filtering*	Sesudah *Filtering*
1	menangkan albummerchandise itzy untuk 8 pemenang itzy	menangkan album marchandise pemenang itzy
2	Terimakasih tokopedia karena mengundang treasure wib tv show	terimakasih mengundang treasure tv show
3	terima kasih kembali dan sukses selalu	terima kasih sukses selalu

4. *Tokenizing*

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

Tabel 5.6 Proses *Tokenizing*

No	Sebelum *Tokenizing*	Sesudah *Tokenizing*
1	menangkan album marchandise pemenang itzy	"menangkan" "albummarchindes" "pemenang" "treasure"
2	terimakasih mengundang treasure tv show	""terimakasih" "mengundang" "treasure" "treasure" "tv" "show"
3	terima kasih sukses selalu	"terima kasih" "sukses" "selalu"

5.4 *Statistical Term Frequency-Invers Document Frequency* (TF-IDF)

Data yang telah melalui tahapan *preprocessing* kemudian diubah ke bentuk numerik sebelum di analisis. Metode pembobotan TF-IDF digunakan untuk mengubah data teks menjadi numerik. Metode TF-IDF adalah metode yang digunakan untuk menentukan keterhubungan kata terhdapat dokumen dengan memberikan bobot stiap kata. Nilai TF-IDF sangat dipengaruhi oleh kemunculan kata dalam satu dokumen dan jumlah kata atau frekuensi kata yang muncul secara keseluruhan. Tahapan awal yang dilakukan adalah mencari *Term Frequency* (TF) yaitu jumlah kata pada dokumen dibagi dengan jumlah kata dalam dokumen. Contoh yang akan dihitung sebagai berikut.

Tabel 5.7 Contoh Kalimat

terimakasih mengundang treasure tv show
terima kasih sukses selalu

Tabel 5.8 Perhitungan TF

Kata	**Dokumen**				
	1	**...**	**6**	**20**	**14243**
Terimakasih	0	...	1/5 = 0.2	0	0
Mengundang	0	...	1/5 = 0.2	0	0
Treasure	0	...	1/5 = 0.2	0	0
Tv	0	...	1/5 = 0.2	0	0
show	0		1/5 = 0.2	0	0
⋮	⋮	⋮	⋮	⋮	⋮
terima	0	...	0	1/4 = 0.25	0
kasih	0	...	0	1/4 = 0.25	0
sukses	0	...	0	1/4 = 0.25	0
selalu	0	...	0	1/4 = 0.25	0
Jumlah		...	0	...	0

Berdasarkan **Tabel 5.8** diketahui bahwa kata "terimakasih" pada dokumen 6 berjumlah 1 dan jumlah keseluruhan kata pada dokumen 6 berjumlah 5 kata. Sehingga didapatkan nilai TF untuk kata "terimakasiht" yaitu 1/5 = 0.2. Untuk kata yang lain, nilai TF juga didapatkan dengan perhitungan yang sama. Setelah mendapatkan nilai TF adalah mencari *Invers Document Frequency* (IDF) yaitu l*og* dari jumlah dokumen dibagi dengan jumlah kata yang muncul., Perhitungan TF-IDF dapat dililihat pada **Tabel 5.9**

Tabel 5.9 Nilai IDF

Kata	**Dokumen**				**Jumlah**	**IDF = log** $(\frac{N}{df})$
	1	**6**	**20**	**14243**		
terimakasih	0	1	0	0	22	$\log\left(\frac{14243}{22}\right) = 2.81$
mengundang	0	1	0	0	9	$\log\left(\frac{14243}{9}\right) = 3.19$
treasure	0	1	0	0	1874	$\log\left(\frac{14243}{1874}\right) = 0.88$
tv	0	1	0	0	3	$\log\left(\frac{14243}{3}\right) = 3.67$
show	0	1	0	0	674	$\log\left(\frac{14243}{674}\right) = 1.32$

Kata	Dokumen				Jumlah	IDF = log $(\frac{N}{df})$
	1	6	20	14243		
terima	0	1	0	0	23	$\log\left(\frac{14243}{23}\right) = 2.79$
kasih	0	1	0	0	113	$\log\left(\frac{14243}{113}\right) = 2.1$
sukses	0	1	0	0	12	$\log\left(\frac{14243}{12}\right) = 3.07$
selalu	0	1	0	0	2	$\log\left(\frac{14243}{2}\right) = 3.85$

Berdasarkan **Tabel 5.9** dapat diketahui nilai IDF, nilai IDF untuk kata "terimakasih" pada seluruh dokumen adalah sebesar 2.81. Perhitungan yang sama juga dilakukan untuk kata-kata yang lainnya sehingga didapatkan nilai IDF dari masing masing kata. Setelah mendapatkan nilai IDF masing-masing kata, selanjutnya melakukan perhitungan untuk TF-IDF. Nilai TF-IDF diperoleh dengan mengalikan nilai TF dengan nilai IDF.

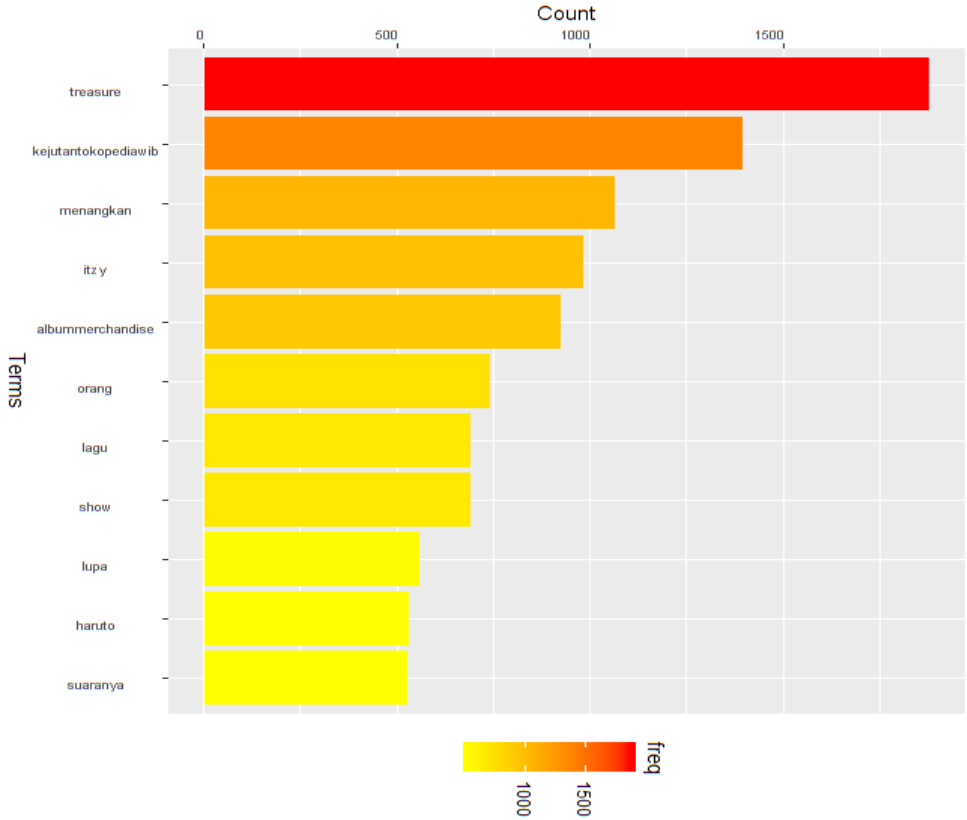
Tabel 5.10 Perhitungan TF-IDF

Kata	Dokumen (TF)					IDF	TF-IDF				
	1	...	6	20	14243		1	...	6	20	14243
terimakasih	0	...	0.2	0	0	2.81	0	...	0.562	...	0
mengundang	0		0.2	0	0	3.19			0.638		
treasure	0		0.2	0	0	0.88			0.176		
tv	0		0.2	0	0	3.67			0.734		
show	0	...	0.2	0	0	1.32	0	...	0.264	...	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
terima	0	....	0	0.25	0	2.79	0	....	0	0.697	0
kasih	0		0	0.25	0	2.1				0.525	
sukses	0		0	0.25	0	3.07				0.768	
Selalu	0	....	0	0.25	0	3.85	0	....	0	0.963	0

Berdasarkan Tabel 5.10 dapat diketahui bahwa nilai TF pada kata "terimakasih" adalah 0.2, sedangkan nilai IDF pada kata "terimakasih" yaitu 2.81.

Sehingga didapatkan nilai TF-IDF dengan mengalikan nilai TF dan IDF yaitu 0.562.

Setelah mendapatkan TF-IDF dapat diketahui bahwa semakin besar nilai TF-IDF dari suatu kata maka semakin besar pula pentingnya kata tersebut pada sebuah dokumen. Selanjutnya untuk mengetahui frekuensi kata yang sering muncul pada data *tweet* dapat dilihat seperti pada **Gambar 5.2**.

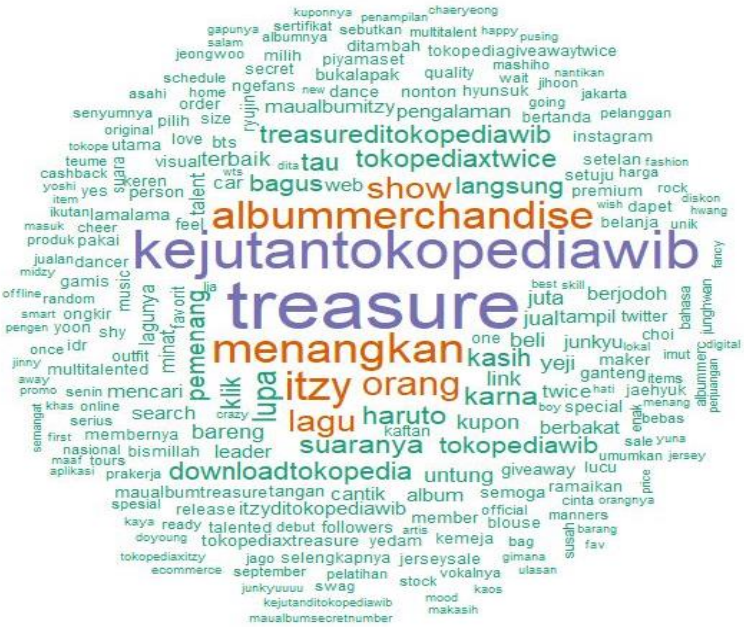


Gambar 5.2 Tampilan *Statistical Term Frequency*

Pada **Gambar 5.2** dapat dilihat hasil dari *Statistical Term Frequency* diperoleh 11 besar kata topik/ konten yang paling banyak disebutkan dalam topik Tokopedia adalah kata “treasure” , “ kejutantokopediawib”, “menangkan”,”itzy” dan diikuti oleh kata-kata lainnya, **gambar 5.2** menunjukan juga semakin kata tersebut *frequency*nya semakin tinggi, warna pada grafik semakin berwarna pekat.

5.5 *Wordcloud*

Wordcloud adalah visualisasi data teks yang paling umum ditampilkan dalam bentuk yang menarik namun tetap mudah dipahami. Berikut akan ditampilkan *wordcloud.*



Gambar 5.3 Tampilan *Wordcloud* Tokopedia yang ada di *Twitter*.

Dapat dilihat pada **Gambar 5.3** di bawah, kata “treasure”, ”kejutantokopediawib”, ”menangkan” dan diikuti oleh kata-kata yang lainnya menjadi kata yang paling banyak diucapkan oleh pengguna *Twitter*. Hal ini ditandai dengan kata yang memiliki ukuran yang besar pada *wordcloud*. Selain menarik *wordcloud* juga memudahkan pembaca dalam mencari informasi terkait.

5.6 *Sentiment Analysis*

Memasuki tahap analisis sentiment, pada tahap ini setiap *tweet* akan diberikan label yaitu, sentimen positif dan sentimen negatif. Penilaian kelas sentiment didasarkan pada kumpulan kata Bahasa Indonesia yang berisikan kumpulan kata positif dan kata negatif. Kemudian dilakukan pelabelan otomatis dengan menggunakan aplikasi *R* dengan cara menghitung selisih jumlah kata positif atau kata negatif yang ada pada sebuah *tweet*. Jika *tweet* memiliki nilai >0 maka akan terklasifikasi dalam kelas sentimen positif, sedangkan untuk *tweet* yang memiliki nilai <0 maka akan terklasifikasi dalam kelas sentimen negatif. Dari 14243 data yang didapatkan, hanya digunakan 5055 data yang masuk kedalam kelas sentiman positif atau negatif.

Tabel 5.11 Perhitungan *Score* Sentimen Manual

Data Komentar	Kata Positif	Kata Negatif
Menangkan albummarchandise itzy pemenang itzy	Menangkan Pemenang	
Jumlah	2	
Perhitungan *Score*	*Score* = 2-0 = 2	

Contoh perhitungan *score* untuk melakukan pelabelan data adalah seperti pada **Tabel 5.11** dengan data komenter *Twitter* "menangkan albummarchandise itzy pememenang itzy". Data tersebut mengandung sentimen positif karena terdapat 2 kata yang mengandung kata positif yaitu menangkan dan pemenang, sedangkan untuk kata negatif tidak ada dalam data tersebut yang berarti 0. Dalam melakukan perhitungan score pelabelan yaitu dengan mengurangi jumlah kata positif dikurangi dengan jumlah kata negatif sehingga didapatkan hasil sebesar 2 yang berarti masuk dalam kategori sentimen positif.

Gambar 5.4 *Sentiment Analysis*

Pada **Gambar 5.4** hasil klasifikasi *sentiment Analysis* didapatkan sebanyak 69,8% atau sebanyak 3530 *tweet* masuk dalam kelas sentimen negatif dan 30,2% atau 1525 masuk dalam kelas sentimen positif.

Untuk melihat gambaran kata yang sering muncul dalam *tweet* kelas positif dan kelas negatif dengan wordcloud dan asosiasi kata dapat dilihat dibawah ini.

Gambar 5.5 Tampilan *Wordcloud* Kelas Positif

Pada **Gambar 5.5** dapat dilihat kata "itzy" menjadi kata yang sering muncul dalam *tweet* kelas positif terkait dengan topik Tokopedia dan diikuti oleh kata-kata lain. Itzy merupakan *girlgroup* Idol asal Korea yang berkolaborasi dengan Tokopedia.

Selanjutnya, melihat asosiasi kata digunakan untuk melihat keterkaitan kata dengan kata lainnya sehingga dapat dihubungkan satu dengan yang lainnya. Berdasarkan hasil *wordcloud* positif . Dipilihlah kata "itzy" sebagai kata yang paling sering muncul pada kelas positif karena kata "itzy" terbanyak atau terbesar *visualisasi wordcloud*nya. Berikut hasil asosiasi dari kata "itzy".

Tabel 5.12 Asosiasi Kata “itzy”

Kata	Nilai
Albummerchandise	0.91
Pemenang	0.90
Menangkan	0.70
Nantikan	0.19
Penampilan	0.18
Tokopediawib	0.18
Show	0.16

Berdasarkan **tabel 5.12** menunjukan asosiasi kata yang terbentuk pada kelas positif terhadap “itzy”, diperoleh kata-kata yang memiliki hubungan dengan kata “itzy” pada komentar positif sebagai berikut :

Kata “itzy” berasosiasi dengan kata “albummerchandise”,”pemenang”, “menangkan”,“nantikan”,”penampilan”,”tokopediawib”, dan “show”. Informasi topik atau konten yang membuat masyarakat berkomentar positif pada Tokopedia adalah mengambarkan masyarakat antusias sebagai pemenang memenangkan albummarchandise, masyarakat sangat menantikan penampilan “itzy” menjadikan poin positif yang ada pada tanggapan tentang Tokopedia.

Gambar 5.6 Tampilan *Wordcloud* Kelas Negatif

Pada **Gambar 5.6** dapat dilihat kata “treasure” menjadi kata yang sering muncul dalam *tweet* kelas negatif terkait dengan topik Tokopedia dan diikuti oleh kata-kata lain. Treasure merupakan *boygroup* Idol asal Korea yang berkolaborasi dengan Tokopedia.

Selanjutnya, melihat asosiasi kata digunakan untuk melihat keterkaitan kata dengan kata lainnya sehingga dapat dihubungkan satu dengan yang lainnya. Berdasarkan hasil *wordcloud* negatif. Dipilihlah kata “treasure” sebagai kata yang paling sering muncul pada kelas negatif karena kata “treasure” terbanyak atau terbesar *visualisasi wordcloud*nya. Berikut hasil asosiasi dari kata “treasure”.

Tabel 5.13 Asosiasi Kata “treasure”

Kata	**Nilai**
Albummerchandise	0.88
Menangkan	0.88
Orang	0.82
Milih	0.21
Gapunya	0.21
Pusing	0.20

Berdasarkan **tabel 5.13** diperoleh kata-kata yang memiliki hubungan dengan kata “treasure” pada komentar negatif memberikan informasi topik perbincangan terhadapat Tokopedia antara lain albummerchandise, menangkan, orang, milih, gapunya, pusing. Informasi topik atau konten yang membuat masyarakat berkomentar negatif pada Tokopedia adalah menggambarkan masyarakat

Berdasarkan **tabel 5.13** menunjukan asosiasi kata yang terbentuk pada kelas negatif terhadap “treasure”, diperoleh kata-kata yang memiliki hubungan dengan kata “treasure” pada komentar negatif sebagai berikut :

Kata “treasure” berasosiasi dengan kata “albummerchandise”,”menangkan”, “orang”,”milih”,”gapunya”, dan “pusing”. Informasi topik atau konten yang membuat masyarakat berkomentar negatif pada Tokopedia adalah mengambarkan masyarakat untuk memenangkan albummerchandise treasure dibuat pusing dalam hal syarat cara memenangkannya.

5.7 *Naïve Bayes Classifier*

Setelah mengetahui klasifikasi sentimen pada topik, hal berikutnya yang dilakukan oleh peneliti adalah mengetahui tingkat akurasi dari klasifikasi sentimen tersebut. Dalam hal ini peneliti menggunakan metode *naïve bayes classifier*.

Naïve bayes classifier memerlukan data latih dan data uji. Data latih dapat mempengaruhi tingkat akurasi yang dihasilkan. Data uji merupakan data yang digunakan untuk menguji akurasi dari model yang telah dibuat oleh data latih. Dari total 5055 data keseluruhan digunakan 80% sebagai data latih dan 20% digunakan untuk data uji.

Data Latih = 80% x 5055
= 4044
Data Uji = 20% x 5055
= 1011

Tabel 5.14 Pembagian Data *Training* dan *Testing*

	Data Training	***Data Testing***
Ratio	80%	20%
Jumlah	4044	1011

Dari perhitungan proporsi untuk menentukan jumlah data latih dan data uji, diperoleh jumlah data yang digunakan untuk data latih sebesar 4044 data untuk data uji yang diperoleh sebesar 1011 data.

Tabel 5.15 *Confusion Matrix*

Prediksi	**Aktual**	
	Negatif	**Positif**
Negatif	649	26
Positif	58	278

Dari **Tabel 5.15** dapat diketahui data masuk dalam *True Negative* artinya yang terprediksi tepat negatif 649, *False Positive* artinya yang terprediksi positif sebenarnya negatif terdapat 58. *True Positive* artinya yang terprediksi positif tepat

positif sebanyak 278, dan *False Negative* yang artinya terprediksi negatif sebenarnya positif 26. Setelah mendapatkan *confusion matrix* langkah selanjutnya adalah menentukan nilai akurasi yang didapatkan.

Dengan menggunakan aplikasi *R Studio* didapat nilai akurasi *Naïve bayes classifier* sebesar 0.91. Dengan nilai akurasi yang cukup tinggi maka dapat dikatakan bahwa klasifikasi sudah baik.

Berdasarkan **tabel 5.15**, didapatkan bahwa untuk nilai prediksi dan nilai-nilai lainnya bisa dihitung secara manual sebagai berikut :

Recall $= \frac{TP}{TP+FN} 100 = \frac{278}{278+26} = 0{,}9144 = 91\%$

Didapatkan bahwa dari nilai *recall* dengan menggunakan *naive bayes classifier* sebesar 0,9144 atau 91%. Hal ini menunjukan bahwa hasil ketepatan 91% jumlah data kelas positif yang terprediksi benar.

Precision $= \frac{TP}{TP+FP} 100 = \frac{278}{278+58} = 0{,}827 = 82\%$

Didapatkan bahwa dari nilai *precision* dengan menggunakan *naive bayes classifier* sebesar 0,82 atau 82%. Hal ini menunjukan bahwa hasil ketepatan prediksi yang benar berdasarkan kelas positif sebesar 82%.

Accuracy $= \frac{TP+TN}{TP+FP+TN+FN} 100 = \frac{278+649}{278+58+649+26} = 0{,}916 = 91\%$

Didapatkan bahwa dari nilai *accuracy* dengan menggunakan *naive bayes classifier* sebesar 0,91 atau 91%. Hal ini menunjukan bahwa hasil ketepatan menggunakan *naive bayes classifier* sebesar 91%.

Spesificity $= \frac{TN}{TN+FP} 100 = \frac{649}{649+58} = 0{,}91 = 91\%$

Didapatkan bahwa dari nilai *spesifiticity* dengan menggunakan *naive bayes classifier* sebesar 0,91 atau 91%. Hal ini menunjukan bahwa hasil ketepatan prediksi yang benar berdasarkan kelas negatif sebesar 91%.

FPR = 1- Spesificity

= 1 – 0,91 = 0,09

Didapatkan bahwa dari nilai FPR dengan menggunakan *naive bayes classifier* sebesar 0,09 atau 9%. Hal ini menunjukan bahwa hasil rating pada kesalahan prediksi positif sebesar 9%.

AUC $= \frac{1+recall-FPR}{2} = \frac{1+0,91-0,09}{2} = 0,91 = 91\%$

Didapatkan bahwa dari nilai *AUC* dengan menggunakan *naive bayes classifier* sebesar 0,91 atau 91%. Hal ini menunjukan bahwa hasil ketepatan menggunakan *naive bayes classifier* sebesar 91%.

Berdasarkan perhitungan diatas didapatkan bahwa nilai accuracy 91%, nilai recall 91%, nilai precision 82%. Berdasarkan nilai *accuracy*, *recall*, dan *precision* yang tinggi makan dapat dikatakan bahwa klasifikasi sudah baik. Pengukuran lain seperti *spesificity*, FPR, dan AUC juga menghasilkan nilai yang besar juga. Dapat dilihat bahwa nilai AUC yang didapatkan sebesar 0,91 yang artinya bahwa nilai tersebut sudah sangat baik atau klasifikasi dikatakan sudah sangat baik.

BAB VI
KESIMPULAN

6.1 Kesimpulan

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

1. Pada proses pengumpulan data, diperoleh data sebanyak 14243 kicauan/*tweets* pada tanggal 25 September 2020 sampai 28 September 2020 mengenai Tokopedia. Dari hasil pembahasan diatas mengenai aplikasi *teks mining* untuk penanganan data besar hasil pencarian topik-topik terkait studi kasus Tokopedia pada *Twitter* maka dapat ditarik kesimpulan yang menghasilkan sebuah informasi bahwa yaitu "treasure", "kejutantokopediawib", "menangkan", dan dikuti kata-kata lainnya. Hasil *Sentiment Analysis* dari data *Twitter* yakni sebesar 5055 data terklasifikasi, 69,8% atau sebanyak 3530 *tweet* masuk dalam kelas sentimen negatif dan 30,2% atau 1525 masuk dalam kelas sentimen positif.
2. Berdasarkan hasil klasifikasi untuk tanggapan tentang Tokopedia menggunakan *Naïve Bayes Classifier* diperoleh hasil akurasi sebesar 91%, nilai *recall* 91%, dan nilai *precision* 82%. Dari hasi akurasi, *recall,* dan *precision* yang tinggi maka dapat dikatakan bahwa klasifikasi sudah tepat. Pengukuran lain seperti *specificity, false positive rate,* dan *Area Under Rate* (AUC) juga menghasilkan nilai yang tepat. Dapat dilihat nilai AUC yang didapatkan sebesar 0.91 yang artinya bahwa nilai tersebut sudah sangat baik atau klasifikasi dikatakan sudah baik.

6.2 Saran

1. Memperbanyak atau memperluas rentang waktu pengambilan data.
2. Mencoba membandingkan beberapa metode klasifikasi lainnya agar dapat melihat hasil yang lebih spesifik untuk penelitian selanjutnya.

3. Melakukan analisis sentimen dengan melihat dokumen/ kalimat secara utuh dengan bantuan ahli bahasa dan pembalajaran mesin yang lebih bagus.

DAFTAR PUSTAKA

Adawiyah, R. (2018). Analisis Sentimen pada Aplikasi Mobile Banking Menggunakan Metode Naive Bayes Classifier dan Asosiasi. Skripsi Jurusan Statsitika FMIPA UII.

Adiyana, I & Hakim, F. (2015). Implementasi *Text Mining* Pada Mesin Pencarian *Twitter* Untuk Menganalisis Topik-topik Terkait KPK dan Jokowi. *Jurnal Prosiding Seminar Nasional Matematika dan Pendidikan Matematika UMS 2015. ISBN :978.602.361.002.0.*

Afshoh. (2017). Analisa Sentimne Menggunakan Naive Bayes untuk Melihat Persepsi Masyarakat terhadap Kenaikkan Harga Jual Rokok pada Media Sosial Twitter.

Ayu, F. (2018). Analisis Pengaruh Kualitas Pelayanan E-commerce Terhadap Konsumen Shopee Indonesia Pada Mahasiswa Fakultas Ekonomi UII Pengguna Shopee. Skripsi Jurusan Manajemen FE UII.

Buntor, A. G (2018). Analisis Sentimen Calon Gubernur Jawa Timur 2018 di *Twitter. Computer Science & Informatics Journal*. Vol1, No. 8.

Databooks. (2020). Kontribusi Sektor Ekonomi Internet ASEAN Tahun 2019. https://databoks.katadata.co.id/datapublish/2020/10/20/e-commerce-kontributor-ekonomi-internet-terbesar-di-asean. Diakses 1 Oktober 2020.

Databooks. (2019). Pengguna dan Tingkat Penetrasi *E-commerce* di Indonesia Tahun 2017-2023. https://databoks.katadata.co.id/datapublish/2019/10/10/tren-pengguna-e-commerce-2017-2023. Diakses 1 Oktober 2019.

Defrianto & Loisa, R.(2019). *Efektivitas Social Media Marketing E-commerce* dalam Meningkatkan Brand Image Perusahaan. Fakultas Ilmu Komunikasi Universitas Tarumanegara.

Farach, D & Nugraha, J. (2019). Implementasi Metode *Naïve Bayes Classifier* dalam Analisis Sentimen Pada Opini Masyarakat Tentang RUU KUHP. *Jurnal Advance in Social, Education and Humanities Research. Vol, 474.*

Faesal, A., Muslim, A., Ruger, A. H & Kusrini. (2020). Sentimen Analisis pada Data Tweet Pengguna Twitter Terhadap Produk Penjualan Online Menggunaka Metode K-Means. Jurnal Matrik, Vol.19 No.2.

Fitria, U. E.(2018). Perbandingan Kinerja *Machine Learning Berbasis Algoritma Sepport Vectore Machine* dan *Naive Bayes*. Skripsi Jurusan Statistika FMIPA UII.

Helmi, B. (2019). *Topic Modeling* Menggunakan Metode *Latent Dirichet Allocation* (LDA) Pada Start-Up PT. Global Tiket Network. Skripsi Jurusan Statistika. FMIPA.UII.

I Price Group. (2020). Persaingan Toko Online di Indonesia. https://iprice.co.id/insights/mapofecommerce/. Diakses 2 Oktober 2020.

Ishak, A. (2012). Analisis Kepuasan Pelanggan dalam Belanja Online: Sebuah Studi Tentang Penyebab (*Antecedents)* dan (*Consequents*). *Jurnal Siasat Bisnis, Vol. 16 No.* 2, 141-154.

Jati, P. E. (2019). Implementasi Text Mining dan Sentiment Analysis pada Jejaring Sosial Twitter dengan Menggunakan Metode Naive Bayes Classifier. Skripsi Jurusan Statistika FMIPA UII.

Khairani, F.(2019). Analisis Topic Modelling dan Klasifikasi Ujaran Kebencian Menggunakan *Algoritma Support Vector Machine* (SVM). Skripsi Jurusan Statistika FMIPA UII.

Liu, Bing, Hu, Minqing, and Cheng, Junsheng (2005). *"Opinion Observer: Analyzing and Comparing Opinions on the Web." Proceedings of the 14th International World Wide Web Conference (WWW-2005), May 10-14, Chiba, Japan.*

Luqyana, W. A., Cholissodin, I., & Perdana, R. S. (2018). Analisis Sentimen Cyrberbullyong pada Komentar Instangram dengan Metode Klasifikasi Support Vector Machine. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, Vol. 2 No 11. 4704-4713.

Meimunah.(2019). Perspektif Warga Global Terhadap Islam Dalam Analisis Sentimen Twitter Menggunakan Perbandingan *Metode Support Vector* Machine dan *Naive Bayes Classifier*. Skripsi Jurusan Statistika FMIPA UII.

Nurhafidzah, N. (2019). Analisis *Topic Modeling* dan Klasifikasi Ujaran Kebencian Menggunakan *Algotirma Support Vector Machiene.* Skripsi Jurusan Statistika FMIPA UII.

Pramiti, D., Saptono, R. & Anggrainingsih, R. (2018). Academic Articles Classification dengan Metode *Naïve Bayes Classifier*. Jurnal Ilmiah Teknologi dan Informasi. Vol. 7, No. 2.

Pratiwi, A., Wahab, Z. & Widiyanti, M. (2020). Pengaruh Consumer Onine *Rating* dan *Review* Terhadap Keputusan Pembelian Pada Pengguna Tokopedia. Jurnal Bisnis dan Manajemen, Vol. 7 No.1, 25-33.

Pujastuti, E., Winarno, W. & Sudarmawan. (2014). Pengaruh *E-commerce* Toko *Online* Terhadap Kepercayaan Konsumen. *Citec Journal. Vol. 1, No.2.*

Retzen, F., & Nurdin. (2016). Analisis Strategi Pemasaran dan Penjualan E-commerace pada Tokopedia. *Jurnal Elektronik Sistem Informasi dan Komputer STMIK Bina Mulia, Vol. 2 No.1.*

Ridwan, M., Suyono, H., & Sarosa, M. (2013). Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier. *Jurnal EECCIS Vol.7, No. 1*.

Rustiana, Deden., Rahayu, Nina. (2017) . Analisis Sentimen Pasar Otomotif Mobil : *Tweet Twitter* Menggunakan Naive Bayes. *Jurnal SIMETRIS, Vol 8 No 1 April 2017. ISSN: 2252-4983.*

Santoso, E. B., & Nugroho, A. (2019). Analisis Sentimen Calon Presiden Indonesia 2019 Berdasarkan Komentar Publik di Facebook. *Jurnal Eksplora Informatika. Vol. 9, No. 1, September 2019.*

Sigit Suryono, E. U. (2018). Klasifikasi Sentimen pada Twitter dengan Naive Bayes Classifier. *Jurnal Ilmiah Bidang Teknologi*, 89-86.

Sihotang, K., & Ghaniy, R. (2019). Penerapan Metode Naïve Bayes Classifier Untuk Penentuan Topik Tugas Akhir pada Website Perpustakaan STIKOM Binaniaga. *Jurnal TeknoIS, Vol. 9, No. 1. 63-72.*

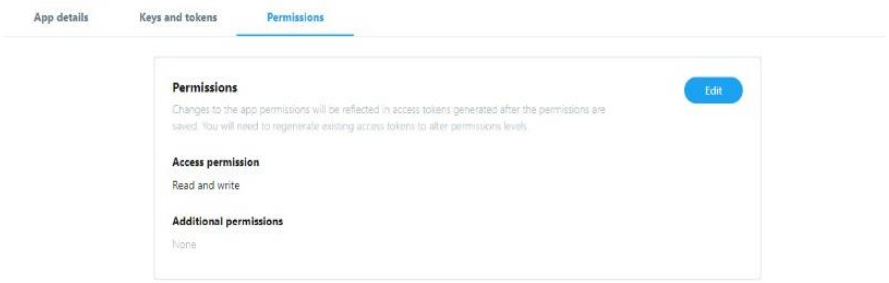
Sudiantoro. V. A & Zuiarso. E. (2018). Analisis Sentimen *Twitter* Menggunakan *Text Mining* dengan *Algoritma Naive Bayes Classifier*. *Jurnal Informatika Vol. 10, No. 2, Oktober 2018.*

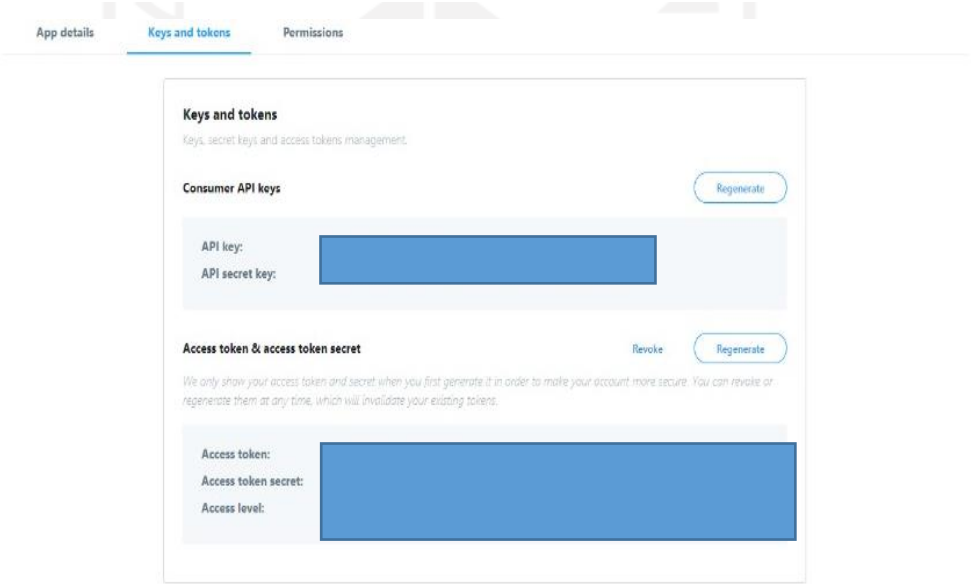
Tokopedia. (2020). *Trending Worldwide*. https://Twitter.com/tokopedia. Diakses 25 September 2020.

Wearesocial. (2020). Jumlah Penggunaan Sosial di Indoensia. https://wearesocial.com/digital-2020. Diakses 29 Sept 2020.

Zuhri. F, N. (2017). Analisis Sentimen Masyarakat terhadap Bran Smartfren Menggunakan Naive Bayes Classifier di Forum Kaskus. *e-Proceeding of Management*, 242-251.

Lampiran 1: Kode *API*





Lampiran 2: Data *Twitter*

Text	Created	screenname	retweet Count
RT @MarcellaWibowoo: Hi siapa tau ada yang minat dan berjodoh. Tokopedia sedang mencari talent berikut ini, bisa langsung search di web car…	25/09/2020 17:09	___renjana	0
@tokopedia Bias ak di itzy itu ryujin min soalnya dia cantik dan suka dance nya #KejutanTokopediaWIB #MauAlbumITZY… https://t.co/Pd8pDTXAEt	25/09/2020 17:06	pandelia__	440
@tokopedia @treasuremembers Mengenalnya melalui suara tanpa melihat rupa meskipun dgn part yg sedikit, suara unikny… https://t.co/o2JpFutaEz	25/09/2020 16:43	Ritamelsayaaaaa	142
@tokopedia @treasuremembers HARUTO Kenapa HARUTO ya karna dia masih muda udah berbakat paling utama ditambah visual… https://t.co/s6Jz0VZDVc	25/09/2020 16:40	broxkennn	0
@tokopedia @treasuremembers Bias aku Mashiho. Karna dia kawaii bgt. Suaranya khas bgt dong,favorit bgt wktu cio nya… https://t.co/a4T2CArLME	25/09/2020 16:38	WMawardani	0
Yang butuh buah bit /beetroot ready tiap hari kak..... Fast Respon Call/WhatsApp :083101475050 Beli BUAH BIT/BEET… https://t.co/EK6zOQF1CA	25/09/2020 16:33	frenchfrys	0
@tokopedia @treasuremembers Bias aku tu doyoung tapi sering oleng ke member lain:v.Alasan aku suka doyoung karna di… https://t.co/evqlNpVMRw	25/09/2020 16:05	_dedyfirmansyah	0
RT @tokopedia: <U+2728>Giveaway Time<U+2728> MENANGKAN 10 ALBUM BERTANDA TANGAN @JYPETWICE! <U+2764> Pertanyaan: Sebutkan lagu TWICE favorit kamu dan jelaskan…	25/09/2020 15:45	1MenujuDamai	2

RT @tokopedia: <U+2728>Giveaway Time<U+2728> MENANGKAN 10 ALBUM BERTANDA TANGAN @JYPETWICE! <U+2764> Pertanyaan: Sebutkan lagu TWICE favorit kamu dan jelaskan…	25/09/2020 15:21	dinodyy	0
RT @tokopedia: Buat MIDZY INDONESIA, yang udah ikutan giveaway ini mana suaranya??? Coba kasih tau mimin pakai #KejutanTokopediaWIB #Tokope…	25/09/2020 15:20	pemalasygaktif	1
@tokopedia @JYPETWICE More & More. .. aku suka banget sama lagu ini, lagunya enak, gak bosen dengernya, pokoknya ni… https://t.co/gyRQYKDow2	25/09/2020 15:15	sklet0n	0
@tokopedia @JYPETWICE More & More. .. aku suka banget sama lagu ini, lagunya enak, gak bosen dengernya, pokoknya ni… https://t.co/gyRQYKDow2	25/09/2020 15:15	albanahasan_	2

Lampiran 3: Syntax Sentiment Analysis

```
library(NLP)
library(tm)
library(wordcloud2)
library(Twitter)
library(rtweet)
library(base64enc)
library(ROAuth)
library(devtools)
library(memoise)
library(whisker)
library(rstudioapi)
library(git2r)
library(withr)
library(rjson)
library(glue)
library(bit64)
library(httr)
```

```
library(httpuv)
library(ggplot2)
library(SnowballC)
library(graph)
library(Rgraphviz)
library("openssl")
library("httpuv")
devtools::install_github("jrowen/Twitter", ref = "oauth_httr_1_0")

# Ganti Sesuai dengan Key Milik Kita
options(httr_oauth_cache=T)
consumer_key <- “xxxxxxxxx”
consumer_secret <- "xxxxxxx"
access_token <- "xxxxxxxxx"
access_secret <- "xxxxxxx"

##REtrieve Tweets
setup_Twitter_oauth(consumer_key,  consumer_secret,  access_token,
access_secret)
tweets = searchTwitter("tokopedia",
                       n = 10000,
                       retryOnRateLimit = 1000)
tweets.df <- twListToDF(tweets)

dataframe=data.frame(tweets.df)

write.csv(dataframe,file                                          =
'F:\\SKRIPSIFEA\\DATA\\DATATOKOPEDIA.csv')

tweets[1:5]
dataframe=data.frame(tweets.df)

#cleaning data
tweets.df <- twListToDF(tweets)
dim(tweets.df)

#data open
```

```
tweets.df <- read.csv('F:\\SKRIPSI MBAK FEA\\DATA\\DATATOKOPEDIA2.csv')

tweets.df <- read.csv('F:\\SKRIPSI MBAK FEA\\DATA\\DATA SENTIMENT TERLABELI.csv')

#RUNNING STELAH DATA UTAMA SELESAI ANALISIS
#RUNNING POSITIF
tweets.df <- read.csv("F:\\SKRIPSI MBAK FEA\\DATA\\data-pos4.csv")

#RUNNING NEGATIF
tweets.df <- read.csv("F:\\SKRIPSI MBAK FEA\\DATA\\data-neg4.csv")

myCorpus <- Corpus(VectorSource(tweets.df$text))
myCorpus <- sapply(myCorpus,function(row) iconv(row, "latin1", "ASCII", sub="byte"))

myCorpus<- Corpus(VectorSource(myCorpus))
removeURL<-function(x) gsub ("http[^[[:space:]]*","",x)
myCorpus<-tm_map(myCorpus,content_transformer(removeURL))

removeRT <- function(x) gsub("RT ", "", x)

myCorpus <- tm_map(myCorpus, removeRT)
removetitik2 <- function(x) gsub(":", "", x)

myCorpus <- tm_map(myCorpus, removetitik2)
removetitikkoma <- function(x) gsub(";", " ", x)

myCorpus <- tm_map(myCorpus, removetitikkoma)
removeamp <- function(x) gsub("&amp;", "", x)

myCorpus <- tm_map(myCorpus, removeamp)
removeUN <- function(x) gsub("@\\w+", "", x)
myCorpus <- tm_map(myCorpus, removeUN)


removeNumFuct<-function(x) gsub("[^[:alpha:][:space:]]*","",x)
```

```
myCorpus<-tm_map(myCorpus,content_transformer(removeNumFuct))
myCorpus = tm_map(myCorpus, PlainTextDocument)
myCorpus<-tm_map(myCorpus,content_transformer(tolower))

#stop word
file_Stop<-file("sword.csv", open="r")
id_stopwords<-readLines(file_Stop)
close(file_Stop)
id_stopwords=c(id_stopwords,"tokopedia","tokped")
myCorpus<-tm_map(myCorpus,removeWords,id_stopwords)

#menghapus ekstraksi angka
myCorpus<-tm_map(myCorpus, stripWhitespace)
myCorpus <- tm_map(myCorpus, removeNumbers)
myCorpusCopy<-myCorpus
dtm = DocumentTermMatrix(myCorpus)
tdm = TermDocumentMatrix(myCorpus)
tdm

##Frekuensi data dan asosiasi(hubungan)
idx <- which(dimnames(tdm)$Terms == "tokopedia")
inspect(tdm[idx + (1:5), 91:100])
(freq.terms <- findFreqTerms(tdm, lowfreq = 15))
term.freq <- rowSums(as.matrix(tdm))
term.freq <- subset(term.freq, term.freq >= 15)
df <- data.frame(term = names(term.freq), freq = term.freq)
View(df)
library(wordcloud)
library(RColorBrewer)
m<-as.matrix(tdm)

#tfidf
View(m)
tf<-data.frame(m)
write.csv(tf,file = 'F:\\SKRIPSI MBAK FEA\\DATA\\datatfidf2.csv')


word.freq<-sort(rowSums(m),decreasing = T)
```

```
pal<-brewer.pal(5, "BuGn")[-(1:4)]
windows()
wordcloud(words = names(word.freq),freq = word.freq, min.freq = 50,
scale=c(3.5,0.50),random.order = F, colors = brewer.pal(2,
"Dark2"))

library(ggplot2)
term.freq <- rowSums(as.matrix(tdm))
term.freq<-subset(term.freq,term.freq>=500)
df<-data.frame(term=names(term.freq),freq=term.freq)
View(df)

#diurut dari tertinggi ke terkecil
ggplot(df, aes(x=reorder(term,-freq), y=freq))+
  geom_col((aes(fill=freq)))+
  xlab("Terms")+
  ylab("Count")+scale_fill_gradient(low='yellow', high='red')+
  theme(axis.text=element_text(size=7, angle=90))

#hubungan kata
source("https://bioconductor.org/biocLite.R")
biocLite()
library(graphTweets)
biocLite("Rgraphviz")

#langsung sini
library(Rgraphviz)freq.terms<-findFreqTerms(tdm, lowfreq = 400)
windows()
plot(tdm, term = freq.terms, corThreshold = 0.01, weighting = T)

vee<-as.list(findAssocs(tdm, terms =c("treasure"), corlimit =
c(0.15,0.15,0.15,0.15,0.15,0.15)))
vee

library(graph)


#terakhir untuk tampilan positf negatif
```

```
## save data
dataframe<-data.frame(text=unlist(sapply(myCorpusCopy,      `[`)),
stringsAsFactors=F)
View(dataframe)

#data sudah dicleaning
write.csv(dataframe,file       =       'F:\\SKRIPSIFEA\\DATA\\DATA
SENTIMENT.csv')
dataframe[110,]


#proses negatif dan positif sentimen
kalimat2<-read.csv("F:\\SKRIPSIFEA\\DATA\\DATA
SENTIMENT.csv",header=TRUE)

#ambil kata kata untuk skoring
positif <- scan("s-pos.txt",what="character",comment.char=";")
negatif <- scan("s-neg.txt",what="character",comment.char=";")
kata.positif = c(positif)
kata.negatif = c(negatif)
score.sentiment  =  function(kalimat2,  kata.positif,  kata.negatif,
.progress='none')
{
  require(plyr)
  require(stringr)
  scores   =   laply(kalimat2,   function(kalimat,   kata.positif,
kata.negatif) {
    kalimat = gsub('[[:punct:]]', '', kalimat)
    kalimat = gsub('[[:cntrl:]]', '', kalimat)
    kalimat = gsub('\\d+', '', kalimat)
    kalimat = tolower(kalimat)

    list.kata = str_split(kalimat, '\\s+')
    kata2 = unlist(list.kata)
    positif.matches = match(kata2, kata.positif)
    negatif.matches = match(kata2, kata.negatif)
    positif.matches = !is.na(positif.matches)
```

```
    negatif.matches = !is.na(negatif.matches)
    score = sum(positif.matches) - (sum(negatif.matches))
    return(score)
  }, kata.positif, kata.negatif, .progress=.progress )
  scores.df = data.frame(score=scores, text=kalimat2)
  return(scores.df)
}

#melakukan skoring text
hasil = score.sentiment(kalimat2$text, kata.positif, kata.negatif)
head(hasil)

#CONVERT SCORE TO SENTIMENT
hasil$klasifikasi<- ifelse(hasil$score<0, "Negatif","Positif")
hasil$klasifikasi
View(hasil)

#Tukar Row
data <- hasil[c(3,1,2)]
View(data)
write.csv(data, file = "F:\\SKRIPSIFEA\\DATA\\DATA SENTIMENT
TERLABELI4.csv")

#Memisahkan twit
data.pos <- hasil[hasil$score>0,]
View(data.pos)
write.csv(data.pos, file = "F:\\SKRIPSIFEA\\DATA\\data-pos4.csv")

data.neg <- hasil[hasil$score<0,]
View(data.neg)
write.csv(data.neg, file = "F:\\SKRIPSI MBAK FEA\\DATA\\data-
neg4.csv")

#proses naive bayer harus dgnti r 6
#r.6
library(tm)
library(RTextTools)
library(e1071)
```

```
library(dplyr)
library(caret)
library(maxent)
library(NLP)
install.packages("pbkrtest")

df<-  read.csv("F:\\SKRIPSI  MBAK  FEA\\DATA\\DATA  NAIVE3.csv",
stringsAsFactors = FALSE)
glimpse(df)

set.seed(1)
df <- df[sample(nrow(df)), ]
df <- df[sample(nrow(df)), ]
glimpse(df)

df$class <- as.factor(df$class)

corpus <- Corpus(VectorSource(df$text))

corpus

inspect(corpus[1:3])


dtm <- DocumentTermMatrix(corpus)
inspect(dtm[40:50, 10:15])
```

#setting sesuai jumlah data pembagian data train 80% dan testing 20%

```
df.train <- df[1:4044,]
df.test <- df[4045:5055,]

dtm.train <- dtm[1:4044,]
dtm.test <- dtm[4045:5055,]

corpus.train <- corpus[1:4044]
corpus.test <- corpus[4045:5055]
```

```
dim(dtm.train)

fivefreq <- findFreqTerms(dtm.train, 10)
length((fivefreq))
fivefreq
## [1] 12144

# Use only 5 most frequent words (fivefreq) to build the DTM

dtm.train.nb <- DocumentTermMatrix(corpus.train,
control=list(dictionary = fivefreq))

dim(dtm.train.nb)
## [1]  1500 12144

dtm.test.nb <- DocumentTermMatrix(corpus.test,
control=list(dictionary = fivefreq))

dim(dtm.train.nb)

# Function to convert the word frequencies to yes (presence) and no
(absence) labels
convert_count <- function(x) {
  y <- ifelse(x > 0, 1,0)
  y <- factor(y, levels=c(0,1), labels=c("No", "Yes"))
  y
}

# Apply the convert_count function to get final training and testing
DTMs
trainNB <- apply(dtm.train.nb, 2, convert_count)
testNB <- apply(dtm.test.nb, 2, convert_count)

library(naivebayes)
# Train the classifier
naive=system.time( classifier <- naiveBayes(trainNB,
df.train$class, laplace = 0) )
```

```
print(naive)
# Use the NB classifier we built to make predictions on the test set.
system.time( pred <- predict(classifier, newdata=testNB) )

# Create a truth table by tabulating the predicted class labels with the actual class labels
table("Predictions"= pred,  "Actual" = df.test$class )

# Prepare the confusion matrix
conf.mat <- confusionMatrix(pred, df.test$class)

conf.mat

conf.mat$byClass

conf.mat$overall
conf.mat$overall['Accuracy']
```

Lampiran 5. Output Runing

```
> conf.mat
Confusion Matrix and Statistics

          Reference
Prediction neg pos
       neg 649  26
       pos  58 278

               Accuracy : 0.9169
                 95% CI : (0.8982, 0.9332)
    No Information Rate : 0.6993
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8082

 Mcnemar's Test P-Value : 0.0007186

            Sensitivity : 0.9180
            Specificity : 0.9145
         Pos Pred Value : 0.9615
         Neg Pred Value : 0.8274
             Prevalence : 0.6993
         Detection Rate : 0.6419
   Detection Prevalence : 0.6677
      Balanced Accuracy : 0.9162

       'Positive' Class : neg
```

```
> conf.mat$byClass
         Sensitivity          Specificity       Pos Pred Value       Neg Pred Value
           0.9179632            0.9144737            0.9614815            0.8273810
              Recall                   F1           Prevalence       Detection Rate Detecti
           0.9179632            0.9392185            0.6993076            0.6419387
    Balanced Accuracy
           0.9162185
>
> conf.mat$overall
      Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull AccuracyPValue
  9.169139e-01   8.081906e-01   8.981655e-01   9.331884e-01   6.993076e-01   4.075999e-64
> conf.mat$overall['Accuracy']
 Accuracy
0.9169139
```

```
> vee<-as.list(findAssocs(tdm, terms =c("treasure"), corlimit = c(0.15,0.15,0.15,0.15,0.15,0.15)))
> vee
$treasure
                orang        albummerchandise             menangkan                 maker
                 0.49                    0.42                  0.38                  0.29
               tampil                  effect                 perfom          tokopediawib
                 0.25                    0.23                  0.23                  0.21
```

```
> vee<-as.list(findAssocs(tdm, terms =c("itzy"), corlimit = c(0.15,0.15,0.15,0.15,0.15,0.15)))
> vee
$itzy
albummerchandise        pemenang        menangkan         nantikan       penampilan             shy      tokope
diawib
            0.91            0.90            0.70             0.19             0.18            0.18
  0.18
            show
            0.16
```

```
> vee<-as.list(findAssocs(tdm, terms =c("treasure"), corlimit = c(0.15,0.15,0.15,0.15,0.15,0.15)))
> vee
$treasure
albummerchandise       menangkan          orang          milih        gapunya         pusing
            0.88            0.88           0.82           0.21           0.21           0.20
```