

TESIS

**ANALISA RIVIEW PELANGGAN SHOPEE di TWITTER
MENGUNAKAN NAIVE BAYES**



Disusun oleh:

Nama : Aditya Hastami Ruger
NIM : 18.51.1132
Konsentrasi : Business Intelligence

PROGRAM STUDI S2 TEKNIK INFORMATIKA
PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2022

TESIS

**ANALISA RIVIEW PELANGGAN SHOPEE di TWITTER
MENGUNAKAN NAIVE BAYES**

**ANALYSIS OF SHOPEE CUSTOMER REVIEWS ON TWITTER USING
NAIVE BAYES**

Diajukan untuk memenuhi salah satu syarat memperoleh derajat Magister



Disusun oleh:

Nama : Adltya Hastami Ruger
NIM : 18.51.1132
Konsentrasi : Business Intelligence

**PROGRAM STUDI S2 TEKNIK INFORMATIKA
PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2022

HALAMAN PENGESAHAN

**ANALISA RIVIEW PELANGGAN SHOPEE di TWITTER MENGGUNAKAN
NAIVE BAYES**

**ANALYSIS OF SHOPEE CUSTOMER REVIEWS ON TWITTER USING NAIVE
BAYES**

Dipersiapkan dan Disusun oleh
Aditya Hastami Ruger
18.51.1132

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Teknik Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari jumat, 4 Februari 2022

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 11 Februari 2022
Rektor

Prof. Dr. M. Suyanto, M.M.
NIK. 190302001

HALAMAN PERSETUJUAN

ANALISA RIEW PELANGGAN SHOPEE di TWITTER MENGGUNAKAN NAIVE BAYES

ANALYSIS OF SHOPEE CUSTOMER REVIEWS ON TWITTER USING NAIVE BAYES

Dipersiapkan dan Disusun oleh

Aditya Hastami Ruger

18,51.1132

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Teknik Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari jumat, 4 Februari 2022

Pembimbing Utama

Anggota Tim Penguji

Prof. Dr.M. Suyanto,MM
NIK. 190302001

Prof. Dr. Ema Utami, S.Si.,M.Kom.
NIK. 190302037

Pembimbing Pendamping

Prof. Dr. Kusriani, M.Kom.
NIK. 190302106

Mei P Kurniawan, S.Kom. M.Kom.
NIK. 1903022187

Prof. Dr.M. Suyanto,MM
NIK. 190302001

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 11 Februari 2022
Direktur Program Pascasarjana

Prof. Dr. Kusriani, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN TESIS

HALAMAN PERNYATAAN KEASLIAN TESIS

Yang bertandatangan di bawah ini,

Nama mahasiswa : Aditya Hastami Ruger
NIM : 18.51.1132
Konsentrasi : Business Intelligence

Menyatakan bahwa Tesis dengan judul berikut,
Analisa Riwiew Pelanggan Shopee di Twitter Menggunakan Naive Bayes

Dosen Pembimbing Utama : Prof. Dr.M.Suyanto,M.M.
Dosen Pembimbing Pendamping : Mei,P Kurniawati, S.Kom. M.Kom.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Tim Dosen Pembimbing
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi

Yogyakarta, 10 Februari 2022

Yang Menyatakan,



Aditya Hastami Ruger

HALAMAN PERSEMBAHAN

Pertama saya ucapkan puji dan syukur kehadiran Allah SWT yang telah memberikan kelancaran serta kemudahan pada proses pembuatan laporan tesis.

Tesis ini saya persembahkan untuk :

1. Kedua orang tua saya serta adik-adik saya yang selalu memberi suport dan bantuannya kepada saya, sehingga saya dapat menyelesaikan tesis ini dengan lancar. Semoga kedua orang tua saya serta adik-adik diberikan kesehatan selalu oleh Allah SWT
2. Bapak Prof.Dr.M. Suyanto, M.M dan Bapak Mei P Kurniawan, S.Kom. M.Kom, yang dengan sabar dalam membimbing saya dalam penyelesaian penelitian saya, semoga mendapatkan banyak keberkahan dan dilancarkan segala urusannya.
3. Teman-teman saya yang selalu membantu saya serta mendoakan saya dalam proses pembuatan Tesis ini.

Keluarga besar yang selalu mendukung dan memberikan semangat tanpa henti kepada saya dan selalu ada ketika saya membutuhkan bantuan. Serta semua pihak yang tidak bisa penulis sebutkan satu persatu.

HALAMAN MOTTO

Tetap berjuang dalam menghadapi berbagai masalah tanpa henti dan tidak menyerah, karena permasalahan itu membantu kita lebih dewasa untuk menghadapi cobaan dalam hidup*Aditya Hastami Ruger*****



KATA PENGANTAR

Puji dan syukur penulis panjatkan kehadirat Allah SWT karena atas berkat hidayah serta rahmat dan inayahNYA, penulis masih diberi kesempatan, kemudahan dan kesehatan untuk menyelesaikan tesis ini.

Tesis ini disusun untuk memenuhi salah satu syarat untuk kelulusan perguruan tinggi Program Studi Strata-2 Magister Teknik Informatika di Universitas Amikom Yogyakarta dan meraih gelar Magister Komputer (M.Kom). Selain itu tesis ini bertujuan untuk menambah ilmu dan wawasan dalam mempelajari sentimen analisis grup Shopee di Twitter dengan metode *Naive Bayes*.

Penulis juga mengucapkan terimakasih yang sebesar-besarnya kepada :

1. Bapak Prof. Dr. M. Suyanto, MM, selaku ketua Amikom Yogyakarta.
2. Bapak Prof. Dr. M. Suyanto, MM dan Bapak Mei P Kurniawan, S.Kom. M.Kom, selaku dosen pembimbing yang dengan sabar dalam memberikan bantuan, saran, masukan, serta bimbingan dalam menyelesaikan naskah tesis ini.
3. Ibu Prof. Dr. Kusriani, M.Kom selaku Direktur Program Pascasarjana Universitas Amikom Yogyakarta.
4. Kedua orang tua dan keluarga saya yang tanpa lelah mendoakan saya maupun membantu saya dalam proses pembuatan Tesis ini.
5. Para Dosen Amikom Yogyakarta yang sudah memberikan berbagai ilmunya kepada saya.

Semua pihak yang tentunya sangat berharga bagi saya, karena selalu memberi suport maupun bantuan kepada saya dalam menyelesaikan tesis ini. Penulis juga meminta maaf apabila ada penulisan atau kata-kata yang salah atau tidak cocok, karena penulisan ini masih jauh dari kata sempurna. Penulis juga menerima kritik maupun sarannya guna memperbaiki penulisan ini. Semoga tesis ini dapat memberi ilmu maupun manfaat bagi penulis maupun bagi mahasiswa jurusan Ilmu Teknologi, serta memberi refrensi untuk penelitian yang akan datang.

Yogyakarta, 11 Februari 2022

Aditya Hastami Ruger

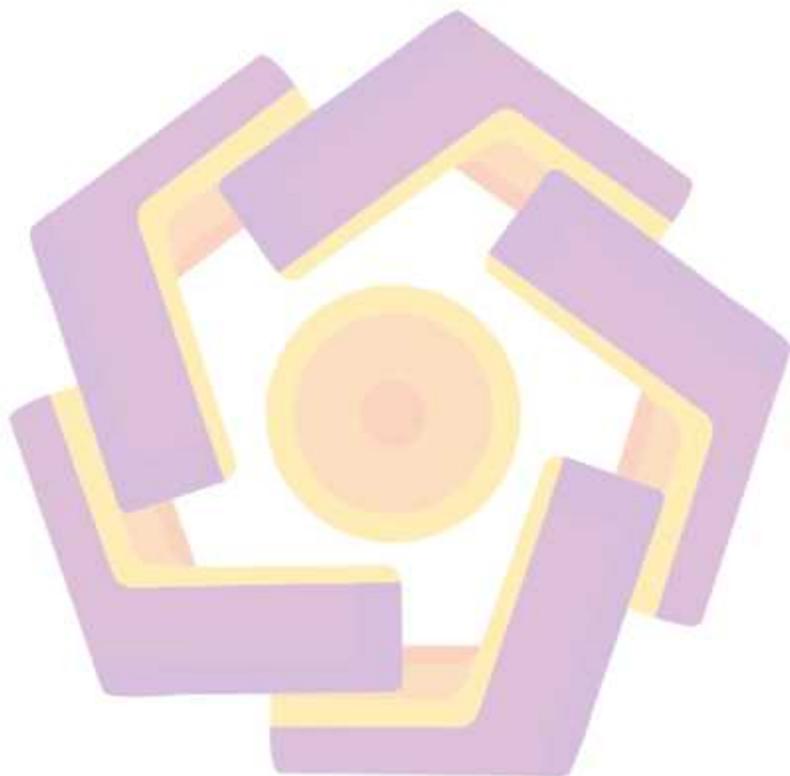
DAFTAR ISI

TESIS	ii
HALAMAN PENGESAHAN.....	iii
HALAMAN PERSETUJUAN.....	iv
HALAMAN PERNYATAAN KEASLIAN TESIS	v
HALAMAN PERSEMBAHAN	vi
HALAMAN MOTTO.....	vii
KATA PENGANTAR.....	viii
DAFTAR ISI.....	x
DAFTAR TABEL.....	xii
DAFTAR GAMBAR	xiii
ABSTRACT.....	xvi
BAB I PENDAHULUAN.....	1
1.1. Latar Belakang Masalah.....	1
1.2. Rumusan Masalah.....	2
1.3. Batasan Masalah.....	3
1.4. Tujuan Penelitian	3
1.5. Manfaat Penelitian	3
BAB II TINJAUAN PUSTAKA.....	5
2.1. Tinjauan Pustaka.....	5
2.2. Keaslian Penelitian.....	8
2.3. Landasan Teori.....	12

2.4. Twitter.....	12
2.5. <i>Sentimen Analysis</i>	17
2.6. Naive Bayes	19
2.7. RStudio.....	21
2.8. <i>PreProcessing</i>	24
2.9. Shopee.....	27
2.10. Convusion Matriks.....	30
2.11. Sublime Text.....	32
BAB III METODE PENELITIAN	33
3.1. Jenis, Sifat, dan Pendekatan Penelitian.....	33
3.2. Metode Pengumpulan Data.....	33
3.3. Alur Penelitian.....	34
BAB IV HASIL PENELITIAN DAN PEMBAHASAN	38
4.1. Tampilan RStudio.....	38
4.2. Pengambilan Data.....	38
4.3. Proses <i>Cleaning Data</i>	41
4.4. Aplikasi <i>Naive Bayes</i>	44
BAB V PENUTUP	50
5.1. Kesimpulan.....	50
5.2. Saran.....	51
DAFTAR PUSTAKA	52
LAMPIRAN	54

DAFTAR TABEL

Tabel 2. 1 Matriks Literatur Riview dan Posisi Penelitian	8
---	---



DAFTAR GAMBAR

Gambar 2.1 Menulis Tweet.....	13
Gambar 2.2 Hashtag.....	14
Gambar 2.3 Re Tweet	15
Gambar 2.2 Peningkat URL.....	15
Gambar 2.3 Trending Topics	16
Gambar 2.4 Tampilan Merubah Warna	17
Gambar 2.5 Rumus Naive Bayes.....	20
Gambar 2.6 Tampilan Platform Shopee.....	28
Gambar 2.7 Tabel <i>Convulsion Matriks</i>	30
Gambar 3.1 Diagram Alur Penelitian.....	35
Gambar 3.2 <i>Normalisasi Fitur</i>	37
Gambar 3.3 Proses <i>Preprocessing</i>	37
Gambar 4.1 Tampilan awal RStudio.....	38
Gambar 4.2 <i>Source Code Directory</i>	38
Gambar 4.3 <i>Source Code install package</i>	39
Gambar 4.4 Perintah <i>Library</i>	39
Gambar 4.5 Download Sertifikat <i>Curl</i>	39
Gambar 4.6 <i>Key Twitter</i>	40
Gambar 4.7 Mengakses <i>Consumer key</i>	40
Gambar 4.8 Kode Ambil Data Di Twitter.....	40
Gambar 4.9 Tampilan Data Twitter	41

Gambar 4.10 Perintah <i>Cleanning</i> Data	41
Gambar 4.10 Perintah <i>Cleanning</i> Data (Lanjutan)	42
Gambar 4.11 Perintah <i>skoring</i> data Positif dan Negatif.....	42
Gambar 4.12 <i>Skoring</i> data.....	43
Gambar 4.13 Memisahkan Data Positif dan negatif	43
Gambar 4.14 Hasil Data Positif dan Data Negatif.....	44
Gambar 4.15 Tampilan Aplikasi <i>Naive Bayes</i>	45
Gambar 4.16 Data Training.....	46
Gambar 4.17 Tampilan <i>Stemming</i> Data Twitter	47
Gambar 4.18 Hasil dari Proses <i>Naive Bayes</i>	47
Gambar 4.19 Rumus <i>Convusion Matriks</i>	48
Gambar 4.20 Hasil <i>Convusion Matriks</i>	49

INTISARI

sentimen analisis merupakan suatu bidang studi yang menganalisis tentang opini-opini seseorang, mengevaluasi, menilai, perilaku, dan emosi seperti produk-produk, layanan, kejadian, serta topik-topik. Sentimen analisis dalam dunia bisnis yang biasanya dipakai untuk menganalisis kebutuhan masyarakat dan kebutuhan pasar, yang diharapkan mampu menyusun strategi pemasaran yang dapat meningkatkan suatu pendapatan dari perusahaan mereka. Penelitian ini mengambil data-data dari grup Shopee yang ada di media sosial Twitter. Shopee merupakan salah satu market place yang sering digunakan oleh berbagai masyarakat maupun berbagai kalangan yang ada didunia ini. Market place Shopee yang menjual berbagai jasa maupun menjual barang-barang untuk kebutuhan sehari-hari. Tujuan dari penelitian ini yaitu menganalisis sentimen pada komentar maupun *review* dari grup Shopee di Twitter. Sentimen yang dianalisis adalah layanan dari Shopee seperti kuis-kuis. Data diambil dari grup Shopee di Twitter, data tersebut yaitu *review-review* dari komentar positif maupun komentar negatif yang diberikan oleh pelanggan Shopee di grup Shopee yang ada di media sosial Twitter, pengambilan datanya menggunakan *software* Rstudio, Rstudio merupakan suatu aplikasi yang dipakai untuk menulis program-program dengan menggunakan bahasa R yang dapat dijalankan pada operasi sistem Windows, Linux, maupun Apple. Serta untuk proses menghitungnya menggunakan metode Naive Bayes yang dimasukkan kedalam sentimen analisis yang pengambilan data-datanya di ambil dari grup Shopee yang ada di Twitter. Dalam penghitungan tingkat keakurasian menggunakan convusion matriks. Data yang digunakan yaitu data pada tanggal 4 bulan januari 2022. Hasil dari penelitian ini yaitu dapat diketahui bahwa jumlah komentar positif 250 data sedangkan negatif 100 data sisanya netral yang berjumlah 50 data, karena data yang diambil adalah 400 data dan tingkat keakurasiannya adalah sebesar 97%.

Kata kunci: RStudio, Convusion Matriks, Naive Bayes, Shopee

ABSTRACT

Sentiment analysis is a field of study that analyzes people's opinions, evaluates, judges, behaviors, and emotions such as products, services, events, and topics. Sentiment analysis in the business world is usually used to analyze community needs and market needs, which are expected to be able to develop marketing strategies that can increase an income from their company. This study takes data from the Shopee group on Twitter social media. Shopee is a market place that is often used by various people and groups in the world. Shopee market place that sells various services and sells goods for daily needs. The purpose of this study is to analyze sentiment in comments and reviews from the Shopee group on Twitter. The sentiments analyzed are services from Shopee such as quizzes. The data is taken from the Shopee group on Twitter, the data are reviews of positive and negative comments given by Shopee customers in the Shopee group on Twitter social media, the data collection uses Rstudio software, Rstudio is an application used to write programs - a program using the R language that can be run on Windows, Linux, and Apple operating systems. As well as for the process of calculating it using the Naive Bayes method which is included in the sentiment analysis whose data is taken from the Shopee group on Twitter. In calculating the level of accuracy using the confusion matrix. The data used is data on January 4, 2022. The results of this study are that it can be seen that the number of positive comments is 250 data while the negative 100 data is the remaining neutral, totaling 50 data, because the data taken is 400 data and the accuracy rate is 97%.

Keyword: RStudio, Confusion Matrix, Naive Bayes, Shopee

BAB I PENDAHULUAN

1.1. Latar Belakang Masalah

Pada saat ini banyak orang-orang yang menggunakan sosial media Twitter karena Twitter sangat mudah digunakan oleh berbagai kalangan dari remaja hingga orang tua. Media sosial Twitter mudah digunakan oleh semua orang, dan memperoleh berbagai informasi dari dalam maupun luar negeri dengan mudah. Serta kemudahan dalam berkomunikasi dengan keluarga serta teman-teman. Negara Indonesia merupakan salah satu negara terbesar kelima dalam penggunaan media sosial Twitter dalam memberikan tweet-tweetnya setiap harinya.

Shopee pun membuat suatu grup di Twitter untuk para konsumennya agar dapat saling memberi masukan-masukannya sehingga Shopee dapat memperbaiki kekurangannya dalam memberikan layanan jasanya untuk para konsumen-konsumennya. Serta memberikan informasi-informasi tentang bonus-bonus, potongan harga atau diskon maupun kuis-kuis yang dibuat oleh Shopee untuk para pengguna Shopee. Shopee merupakan salah satu marketplace yang sering digunakan oleh berbagai kalangan di Indonesia. Karena Shopee memberikan suatu kemudahan untuk para pengguna maupun konsumennya. Berdasarkan dari latar belakang yang telah dipaparkan di atas, maka peneliti membuat suatu penelitian yang berjudul "Analisa Riview Pelanggan Shopee di Twitter Menggunakan Naive Bayes". Penelitian ini menganalisis pada riview-riview atau komentar para pelanggan ataupun pengguna Shopee yang ada di Twitter dan mensentimen pada layanan yang diberikan oleh Shopee, yaitu seperti

kuis-kuis maupun diskon. Setelah meneliti grup Shopee di Twitter, sebagian besar pelanggan Shopee membahas atau *review* kuis yang diberikan oleh pihak Shopee. Penelitian sebelumnya dari Elly Indrayuni dan Mochamad Wahyudi (2015) yang melakukan suatu penelitian yang berjudul “Penerapan Character N-Gram untuk Sentimen Analisis pada Review Hotel dengan Algoritma Naive Bayes”. Merupakan penelitian sebelumnya yang menjadi referensi utama untuk penelitian ini. Pada penelitian ini menggunakan penerapan *N-Gram* serta *Naive Bayes*. Penelitian ini mengambil data dari *review-review* hotel. *Review* tersebut dibagi menjadi 2 data yaitu *review* positif maupun *review* negatif. Perbedaan dari penelitian sebelumnya yaitu terletak pada cara pengambilan data nya serta penghitungannya. Pada penelitian ini menggunakan aplikasi Rstudio untuk pengambilan data-datanya dari Twitter, serta untuk menghitung tingkat keakurasiannya menggunakan *Confusion Matrix*.

1.2. Rumusan Masalah

Berdasarkan latar belakang masalah yang telah dijelaskan, maka rumusan masalah dalam penelitian ini yaitu :

1. Bagaimana cara penerapan sentimen analisis pada *review* pelanggan atau pengguna di grup Shopee yang ada di media sosial Twitter menggunakan algoritma *Naive Bayes*?
2. Berapa tingkat keakurasiannya dengan algoritma *Naive Bayes* dalam sentimen analisis pada komentar-komentar grup Shopee yang ada di Twitter dengan *Confusion Matrix*?
3. Bagaimana cara mengambil data-data dari Twitter menggunakan RStudio ?

1.3. Batasan Masalah

Adapun beberapa batasan masalah dalam penelitian ini adalah sebagai berikut :

1. Tweet ini menggunakan tweet berbahasa indonesia, Dengan jumlah data yang digunakan sebanyak 400 data yang terdiri dari 200 data, yaitu 100 data untuk komentar positif dan 100 data untuk komentar negatif serta untuk data latihnya terdiri dari 200 buah komentar, yang terbagi sama rata antara komentar positif dan negatifnya untuk data ujinya.
2. Penelitian ini menggunakan metode algoritma *Naive bayes* untuk sentimen analisisnya dalam menganalisis *review* dari penjual maupun pelanggan yang memakai jasa *market place* Shopee di Twitter.
3. Dalam penelitian ini pengujian untuk tingkat keakurasiannya menggunakan metode *Confusion Matrix*.
4. Dalam mengolah maupun menambang data-data dari grup Shopee di Twitter menggunakan aplikasi RStudio.

1.4. Tujuan Penelitian

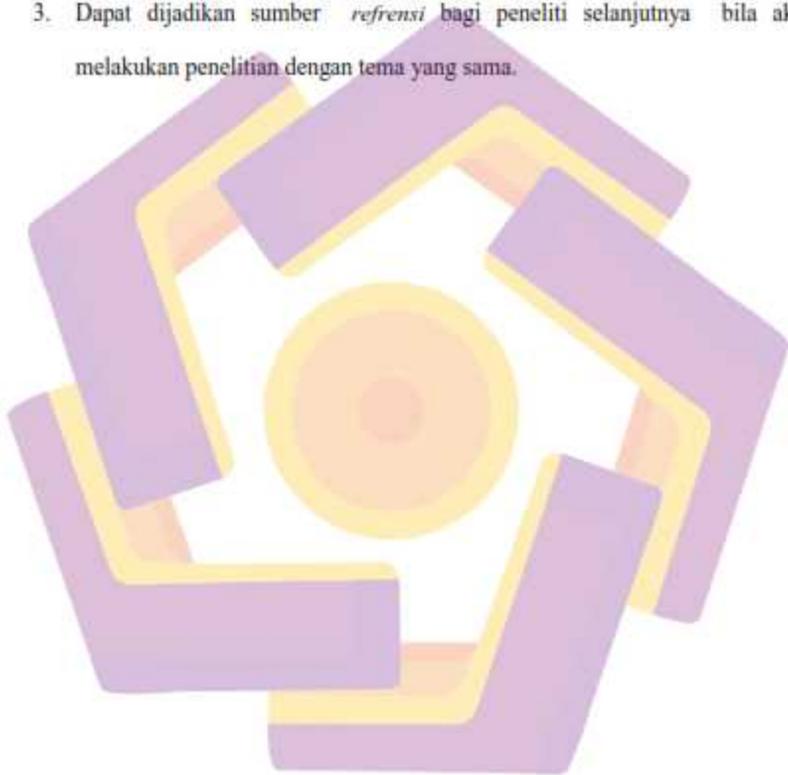
Adapun tujuan dari penelitian ini adalah :

1. Mengetahui berapa tingkat akurasi dalam menggunakan metode *Naive bayes* dalam sentimen analisis *review-review* pelanggan maupun pengguna Shopee .
2. Menganalisis *review-review* pelanggan maupun pengguna Shopee di grup Shopee di media sosial Twitter.

1.5. Manfaat Penelitian

Manfaat yang dapat diperoleh dari penelitian ini adalah :

1. Bagi peneliti mengetahui berapa tingkat akurasi menggunakan metode *Naive Bayes* untuk sentimen analisis *review* pelanggan pada grup Shopee di Twitter.
2. Dapat memberikan beberapa informasi pada *review-review* atau komentar-komentar dari pelanggan Shopee di Twitter.
3. Dapat dijadikan sumber *refrensi* bagi peneliti selanjutnya bila akan melakukan penelitian dengan tema yang sama.



BAB II TINJAUAN PUSTAKA

2.1. Tinjauan Pustaka

Ai Nurhayatul Kamilah (2017) melakukan pengujian dalam meriview pelanggan Tokopedia. Metode yang digunakan yaitu algoritma *Naive Bayes*. Pengumpulan datanya dari komentar-komentar pelanggan Tokopedia. Tahap *Pre Processing* menggunakan Rapid Minner dalam memining data dari Tokopedia. Mengolah data yang tidak terstruktur menggunakan *Pre Processing* yaitu *tokenization*, *filtering* dan *steeming*. Hasil kerja *Naive Bayes* pada pada klasifikasi teks adalah sebesar 77%.

Lingga Aji andika, Pratiwi Amalia Nur Azizah, dan Respatiwan (2019) melakukan pengujian menggunakan metode *Naive Bayes* terhadap hasil *quick count* pemilihan presiden 2019 . Langkah yang dilakukan untuk perancangan sentimen analisis menggunakan *Pre Processing* data seperti *tokenizer*, *stemming* serta *stopwordse*. Dalam penelitian ini menggunakan data sebanyak 2000 data berkurang menjadi 1322 data karena terindikasi sama atau hanya menggunakan retweet pengguna.

Fauziah Afshoh (2017) melakukan pengujian dalam persepsi masyarakat terhadap kenaikan harga jual rokok. Data diambil dari Twitter dan metode yang digunakan adalah algoritma *Naive Bayes*. Penelitian ini melakukan seleksi fitur menggunakan *fitur Chi Square* dan data *Pre Processing* dengan *tokenizer*, *cleansing*, *normalization*, dan *case folding*. Pengujian yang telah dilakukan menggunakan data latih sebanyak 350 buah dari opini masyarakat mengenai topik

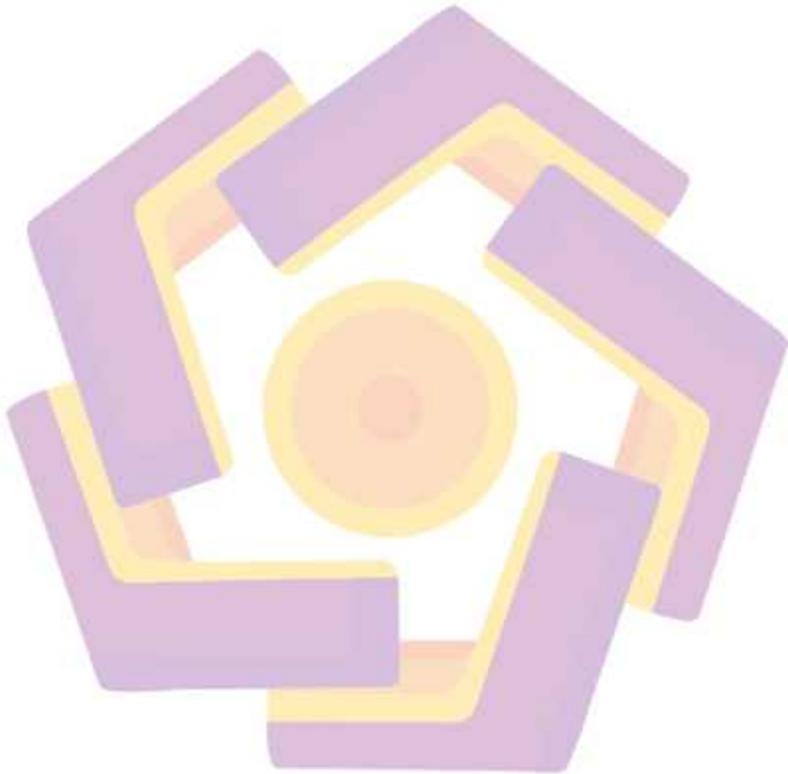
kenaikan harga jual rokok menggunakan metode *Lexicon Based* menghasilkan nilai *precision* 53% pada data opini negatif dan *recall* sebesar 57% pada data opini negatif, kemudian untuk nilai *accuracy* menggunakan *Lexicon Based* menghasilkan nilai persentase sebesar 81% pada data opini netral.

Fajar Ratnawati (2018) melakukan pengujian opini film pada sosial media Twitter. Metode yang digunakan adalah *Naive Bayes*. Penelitian ini menggunakan 200 tweet positif dan 200 tweet negatif untuk melatih *classifier* dan disimpan di file.txt dengan nama file juga untuk data tes sehingga dapat dinilai ketepatan dari *classifier* terlatih. Akurasi tertinggi didapat pada *fold* kedua yaitu 90%, *precision* 92%, *Recall* 90% dan *f-measure* 90%.

Fransiska Vina Sari dan Arief Wibowo (2019) melakukan pengujian menganalisis pelanggan toko onlin JD.ID berbasis konversi ikon emosi. Metode penelitian menggunakan algoritma *Naive Bayes*. Pengolahan data dalam penelitian ini menggunakan *preprocessing* yaitu *case folding*, *convert emoticon*, *cleansing*, *text transformation*, *stopwords removal*, *tokenizing*, pembobotan *tf-idf* dan *stemming* kemudian dilakukan menggunakan text klasifikasi menggunakan *Naive Bayes*.

Penelitian tentang menganalisis perbandingan akurasi deteksi serangan pada jaringan komputer dengan metode *Naive bayes* dan *Support Vector Machine* yang dilakukan oleh Mercury Fluorida Fibrianda, Adhitya Bhawiyuga (2018) data ini mengambil survey dari lab kapersky 2017, 33% organisasi mengalami serangan DDoS mpada tahun 2017. Penelitian ini mempunyai tujuan melihat

tingkat akurasi dari metode naive Bayes, SVM Linier, SVM Polynomial, dan SVM Sigmoid dengan menggunakan dataset ISCX 2012.



2.2. Keaslian Penelitian

Tabel 2. 1 Matriks Literatur Riview dan Posisi Penelitian

No	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
1	Rizqiyah (2018)	menganalisis penerimaan dan penolakan masyarakat terhadap pengesahan UUMD3 berdasarkan hasil akurasi dari kedua algoritma.	Pada <i>Cross-validation</i> data dipisahkan menjadi dua subset yaitu data prosed dan pembelajaran dan data validasi.	Pengembangan selanjutnya dapat dikombinasikan dengan metode klasifikasi lainnya selain <i>K-NN</i> dan <i>Naive Bayes</i>	Penelitian menganalisis sentimen tweet di Tweeter pada pelanggan Shopee dengan metode <i>Naive Bayes</i>
2	Kamilah (2017)	memperoleh informasi kualitas produk yang berdasarkan <i>review</i> pelanggan. Mendapatkan informasi dari <i>review</i> teks yang tidak terstruktur dengan proses teks <i>mining</i> .	Dengan adanya klasifikasi <i>review</i> menggunakan metode <i>Naive Bayes</i> , terbukti data sentimen dapat di kategorikan kedalam <i>class review</i> negatif dan positif. Sehingga informasi dapat sampai pada konsumen	Ada beberapa metode selain metode tersebut, dengan tujuan memperoleh nilai akurasi yang tinggi.	Penelitian menganalisis sentimen tweet di Tweeter pada pelanggan Shopee dengan metode <i>Naive Bayes</i>

Tabel 2.1 Matriks Literatur Riwiew dan Posisi Penelitian (Lanjutan)

3	Andika, Azizah, dan Respawatiwulan (2019)	Menganalisis sentimen Hasil <i>Quick Count</i> Pemilihan presiden 2019	Model Naive Bayes yang terbaik memiliki tingkat akurasi sebesar 82,90% dengan $\alpha = 0,05\%$. Klasifikasi yang diperoleh masing-masing sebesar 34,5% <i>tweet</i> positif dan 65,5% <i>tweet</i> negatif.	Kalau bisa perbandingan jangan metode <i>Naive Bayes</i> saja, agar hasil lebih maksimal	Penelitian menganalisis sentimen tweet di Tweeter pada pelanggan Shopee dengan metode <i>Naive Bayes</i>
4	Ratnawati (2018)	Menganalisis sentimen opini film di Twitter menggunakan <i>Naive Bayes</i>	Pengklasifikasian data opini film berbahasa Indonesia berdasarkan sentimennya dapat dilakukan dengan algoritma <i>Naive Bayes Classifier</i> dengan pembagian datasetnya menggunakan <i>5-fold cross validation</i> .	Kendala yang sering ditemui dalam memahami kalimat untuk diolah dalam proses klasifikasi sentimen	Penelitian menganalisis sentimen tweet di Tweeter pada pelanggan Shopee dengan metode <i>Naive Bayes</i>

Tabel 2.1 Matriks Literatur Riview dan Posisi Penelitian (Lanjutan)

5	Afsboh (2017)	<p>Memprediksi apakah masyarakat lebih banyak yang beranggapan positif, negatif, ataupun netral, sehingga dari hasil prediksi tersebut pemerintah dapat menentukan kebijakannya terhadap wacana kenaikan harga jual rokok.</p>	<p>Penelitian yang telah dilakukan menggunakan data opini sebanyak 350 buah dapat diketahui bahwa nilai sentimen positif yang paling banyak terbentuk dalam menanggapi topik kenaikan harga jual rokok dengan persentase tweets sebesar 53%. Sedangkan untuk evaluasi kinerja sistem menggunakan <i>Naive Bayes Classifier</i> menunjukkan hasil klasifikasi dokumen yang lebih baik daripada metode <i>Lexicon Based</i>.</p>	<p>Data yang tidak seimbang mengakibatkan proses klasifikasi tidak berjalan secara maksimal, untuk penelitian lebih lanjut dapat menggunakan data yang seimbang agar klasifikasi dokumen mampu bekerja dengan baik.</p>	<p>Penelitian menganalisis sentimen tweet di Tweeter pada pelanggan Shopee dengan metode <i>Naive Bayes</i></p>
---	-----------------	--	--	---	---

Tabel 2.1 Matriks Literatur Riwiew dan Posisi Penelitian (Lanjutan)

6	Fibrianda dan Bhawiyuga (2018)	Melakukan perbandingan akurasi deteksi serangan jaringan pada komputer dengan SVM	Tahapan klasifikasi serangan menggunakan metode behavior based membutuhkan sebuah dataset dan metode. Dengan melakukan perbandingan pola atau aktivitas yang ada pada sebuah data, kemudian dilakukan klasifikasi dengan sebuah metode dan menghasilkan sebuah model	Untuk meningkatkan nilai akurasi dari sebuah metode dapat dilakukan dengan beberapa teknik diantaranya teknik bagging dan boosting.	Penelitian menganalisis sentimen tweet di Tweeter pada pelanggan Shopee dengan metode <i>Naive Bayes</i>
---	----------------------------------	---	--	---	--

2.3. Landasan Teori

Penulis melakukan suatu penelitian dengan mengkaitkan beberapa teori-teori yang dapat mendukung penerapannya serta sebagai acuan referensi adalah :

2.4. Twitter

Twitter adalah suatu layanan jejaring sosial yang memungkinkan penggunanya untuk mengirim dan membaca pesan berbasis teks. Twitter didirikan pada tanggal 21 Maret 2006, oleh Jack Dorsey, Noah Glass, Biz Stone, dan Evan Williams. Tempatnya yaitu di San Francisco, California, Amerika Serikat. Tempat Twitter merupakan peringkat ke dua sebagai situs jejaring sosial yang paling sering dikunjungi di dunia. Tingginya popularitas Twitter menyebabkan layanan ini dimanfaatkan untuk keperluan dalam berbagai aspek, contohnya sebagai sarana protes, politik, sarana pembelajaran, dan sebagai media komunikasi. Twitter dimiliki oleh Twitter, Inc, yang berbasis di San Francisco, dengan kantor terdapat di New York City, Boston, dan San Antonio. Mei 2015, Twitter telah memiliki lebih dari 500 juta pengguna, 302 juta diantaranya adalah pengguna aktif.

Sedangkan pertama kali Twitter diperkenalkan oleh Jack Dorsey pada acara diskusi yang diselenggarakan oleh Odeo (sebuah perusahaan *podcast*). Jack Dorsey merupakan mahasiswa dari Universitas di New York yang memberikan gagasannya dalam penggunaan layanan pesan singkat untuk berkomunikasi maupun interaksi. Jack Dorsey merupakan karyawan dari Odeo, sedangkan Evan Williams dan Biz Stone dari Google. Pada tahun 2006, Biz Stone dan Evan Williams, mendirikan sebuah perusahaan yang bernama Obvious Corp dengan

mengakuisisi Odeo beserta semua asetnya termasuk Odeo.com dan Twitter.com dari para pemegang saham dan investor. Twitter mulai populer pada tahun 2007 bertepatan saat diselenggarakan sebuah festival musik yang populer yaitu South by Southwest (SXSW). Pada saat penyelenggaraan festival tersebut penggunaan Twitter meningkat yaitu sebanyak 40.000 tweets per harinya. Yang pada awalnya hanya sebanyak 20.000 tweets per hari menjadi 60.000 tweets per hari. Sedangkan mempubliknya pada tanggal 15 Juli 2006.

Fitur-fitur yang ada di Twitter yaitu :

1. Tweet, merupakan salah satu fitur yang dapat digunakan oleh para pengguna untuk membagikan tulisan, video, foto dan gif atau gambar gerak ke publik. Secara umum, tweet yang dibagikan dapat dilihat oleh semua orang atau semua pengguna Twitter yang terdaftar ataupun tidak terdaftar, tetapi pengguna juga bisa mengatur privasi tweetnya sehingga hanya bisa dilihat oleh orang-orang tertentu. Contoh gambar dibawah



Gambar 2.1 Menulis Tweet

2. *Follow*, *Followers*, *Unfollowing*, pengguna Twitter pasti menemukan istilah pengikut (*follow*), menjadi pengikut (*followers*), dan membatalkan pertemanan (*unfollowing*). fitur tersebut berguna untuk para pengguna dapat

3. saling berklomunikasi dengan para pengguna lain serta mendapatkan tweet-tweet dari pengguna lainnya di beranda Twitter. Pengguna juga bisa melakukan pemblokiran pada akun-akun yang dirasa pengguna mengganggu.
4. *Hashtag* (#), simbol ini sangat penting ketika digunakan untuk mengawali sebuah topik yang ada di Twitter. Contohnya #hidupindonesia, #Covid-19, #Kemerdekaan, dan lainnya. Sebuah topik dengan penggunaan hashtag yang tinggi atau banyak dapat memungkinkan menjadi trending topik di Twitter. Contoh *hashtag* :



Gambar 2.2 Hashtag

5. Nama Pengguna, ketika kita membuat akun Twitter, pengguna diharuskan untuk membuat nama pengguna (*username*). Nama pengguna ini mengidentifikasi sebuah akun yang fungsinya penting untuk membedakan pengguna satu dengan lainnya. Penggunaan "@" diikuti dengan nama pengguna, contohnya @aditya, @rere12, dan lainnya.
6. *Retweet*, adalah membagikan ulang kiriman yang pernah dibuat. Pengguna dapat membagikan kembali kiriman sendiri ataupun kiriman orang lain. Pengguna juga dapat me-*retweet* balasan (*reply*) untuk dibagikan ke pengikut. Contoh gambar di bawah.



Gambar 2.3 Re Tweet

7. Penyingkat URL, ketika kita membagikan tautan, secara otomatis Twitter memperpendek menjadi tautan (<https://t.co/>), yang diikuti dengan URL Shortener. Dengan fungsi ini pengguna dapat membagikan tulisan maksimal 240 karakter tanpa gangguan dari panjangnya tautan yang ini ditambahkan. Lihat gambar di bawah

Warning: this link may be unsafe

<https://goo.gl/2R9WMM>

The link you are trying to access has been identified by Twitter or our partners as being potentially harmful or associated with a violation of Twitter's Terms of Service. The link could lead to a site that:

- steals your password or other personal information
- installs malicious software programs on your computer
- collects your personal information for spious purposes
- has been associated with a violation of Twitter's Terms of Service

[Back to previous page](#)

[Learn more about unsafe links](#)

[Ignore this warning and continue](#)

Gambar 2.2 Penyingkat URL

8. Topik hangat (*Trending Topics*), Twitter merupakan media sosial yang pertama kali memperkenalkan topik hangat (*Trending Topics*). Topik hangat

9. yaitu tautan atau topik yang banyak di perbincangkan para pengguna Twitter berdasarkan tagar, kata, atau frasa. Trending topics berguna untuk para pengguna Twitter agar tau apa yang sedang terjadi di dunia ini serta letak geografisnya. Contoh gambar di bawah



Gambar 2.3 Trending Topics

10. **Tampilan**, merupakan salah satu fitur yang terdapat di Twitter, yakni merubah tampilan dengan tiga pilihan warna yaitu bawaan, temaram, dan lampu mati. Pengguna juga dapat merubah ukuran huruf. Dengan fitur ini menambah kesan pengguna agar lebih nyaman ketika memakai Twitter. Contoh gambar dibawah.



Gambar 2.4 Tampilan Merubah Warna

Twitter juga memiliki manfaat yaitu :

1. Media berbagi dan bertukar informasi
2. Mendapatkan informasi yang up to date.
3. Sebagai media hiburan.
4. Meningkatkan kemampuan dalam menulis.
5. Ajakan untuk menggerakkan atau mempengaruhi masa.

2.5. *Sentimen Analysis*

Sentimen Analysis adalah salah satu bidang dari NLP (*Natural Language Processing*) yang membangun suatu sistem untuk mengenali serta mengekstraksi opini-opini kedalam bentuk teks. Informasi berbentuk teks banyak ditemukan di forum, media sosial, blog, serta situs-situs yang berbasis *review-review*. Dengan adanya *sentimen analysis* informasi-informasi yang tidak terstruktur menjadi terstruktur. Dalam hal bisnis, *sentimen analysis* sangat berguna. Suatu perusahaan

mampu mengolah data masukan, kritikan, dan keluhan dari produk yang ditawarkan akhirnya bisa mereka gunakan untuk memperbaiki kualitas dari pelayanan atau barang yang mereka dagangkan atau tawarkan.

Menurut (Liu, 2011) *sentimen analysis* atau *opinion mining* mengacu pada bidang yang luas dari pengolahan bahasa alami, komputasi *linguistic* dan *text mining* yang memiliki tujuan menganalisa pendapat, sentimen, evaluasi, sikap, penilaian dan emosi seseorang apakah pembicara atau penulis berkenan dengan suatu topik, produk, layanan, organisasi, individu, ataupun kegiatan tertentu.

Analisis sentimen sangatlah berguna sebagai pemroses penganalisis komentar seperti pendapat kemudian di proses menjadi sesuatu yang lebih bermakna (Palucoumputer, 2015) seperti diolah menjadi bentuk ranting.

Ada beragam jenis dari *sentimen analysis*, yaitu *Fine-Grained Sentimen Analysis* yang berfokus pada tingkat polaritas pendapat maksudnya mengelompokkan pendapat-pendapat atau respon-respon kedalam beberapa kategori yaitu positif, netral dan negatif. Ke dua yaitu *Intent Sentiment Analysis* tujuan dari *sentimen analysis* ini yaitu mengidentifikasi atau menggali lebih dalam di balik pesan pengguna untuk melihat apakah itu termasuk pesan keluhan, saran pertanyaan, pendapat atau suatu penghargaan terhadap suatu produk atau layananmu. Ke tiga yaitu *Aspect-Based Sentiment Analysis*, *analysis* ini memungkinkan untuk menghubungkan sentimen spesifik dengan berbagai aspek dari produk atau layanan pengguna. Cara kerja *sentimen analysis* dibagi menjadi tiga yaitu evaluasi, klasifikasi dan visualisasi hasil.

2.6. Naive Bayes

Naive Bayes merupakan metode yang dimanfaatkan untuk perhitungan probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris yang bernama Thomas Bayes, yaitu memprediksi probabilitas di masa depan berdasarkan pengalaman sebelumnya. *Naive Bayes* merupakan metode yang sangat baik dibandingkan dengan model *classifier* lainnya. Hal ini dibuktikan oleh Daniela, Chris J. Hinde, dan Roger G. Stone. “*Naive Bayes vs decision trees vs neural network in the classification of training web pages.*” 2009 pada jurnal Xhemali, mengatakan bahwa *Naive Bayes Classifier* memiliki tingkat akurasi yang lebih baik dibandingkan dengan *Classifier* yang lainnya.

Keuntungan dari *Naive Bayes* yaitu metode ini hanya membutuhkan jumlah data latih yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian. Karena yang diasumsikan sebagai *variable independent*, maka hanya *varians* dari suatu *variable* dalam sebuah kelas yang dibutuhkan untuk menentukan klasifikasi, bukan keseluruhan dari *matriks kovarian*. Kegunaan dari metode *Naive Bayes* yaitu :

1. Mengklasifikasikan dokumen seperti teks berita atau teks akademis.
2. Untuk membuat diagnosis medis secara otomatis.
3. Mendeteksi atau menyaring spam.
4. Sebagai *machine learning* yang menggunakan probabilitas.

Sedangkan kelebihan dari metode ini yaitu :

1. Metode ini dapat digunakan untuk data kuantitatif maupun kualitatif.
2. Tidak perlu data yang banyak.
3. Perhitungannya cepat.
4. Tidak memerlukan training yang banyak.
5. Mudah dipahami.

Metode juga memiliki suatu kekurangan yaitu :

1. Keakuratannya tidak bisa diukur memakai satu probabilitas saja, butuh bukti lain untuk membuktikannya.
2. Apabila probabilitasnya 0 maka prediksinya juga 0.

3. Untuk membuat suatu keputusan, diperlukan pengetahuan sebelumnya. Keberhasilannya sangat bergantung pada pengetahuan awal, maka banyak celah yang dapat mengurangi efektivitasnya.
4. Dirancang untuk mendeteksi kata-kata saja, tidak bisa mendeteksi gambar.

Rumus *Teorema bayes* menurut (Bustami, 2013) adalah

Keterangan

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

X	: Data dengan class yang belum diketahui
H	: Hipotesis data merupakan suatu class spesifik
P(H X)	: Probabilitas hipotesis H berdasarkan kondisi X (posteriori probabilitas)
P(H)	: Probabilitas hipotesis H (prior probabilitas)
P(X H)	: Probabilitas X berdasarkan kondisi hipotesis H
P(X)	: Probabilitas X

Teorema Bayes sering pula dikembangkan menjadi :

$$P(H|X) = \frac{P(X|H)}{\sum_{i=1}^n P(H_i|X)} \cdot P(H)$$

Gambar 2.5 Rumus Naive Bayes

Keterangan

- I** : 1,2,3,..., n jumlah data Hipotesis (prior probabilitas) Dimana
 : $H_1 \cup H_2 \cup H_3 \dots \cup H_n = S$
- S** : Probabilitas total H

2.7. RStudio

R adalah bahasa pemrograman dan sistem perangkat lunak yang dirancang khusus untuk mengerjakan segala hal terkait komputasi statistik. R dikembangkan oleh dua pakar statistik yaitu Ross Ihaka dan Robert Gentleman di Auckland University, New Zealand, pada tahun 1993. R dan Python menjadi pilihan utama bagi para peneliti di bidang *science* untuk mengolah dan menganalisa data baik untuk kepentingan penelitian maupun bisnis. R Studio tersedia dalam dua edisi yaitu RStudio *Dekstop*, RStudio ini dijalankan secara lokal sebagai aplikasi *dekstop* biasa sedangkan RStudio Server yang dapat mengakses RStudio menggunakan browser saat sedang berjalan di server Linux jarak jauh. RStudio tersedia untuk OS Windows, Linux, dan MacOS.

Berdasarkan publikasi yang dirilis *IEE Spectrum ranks languages* di tahun 2017, R termasuk kedalam 10 bahasa pemrograman terpopuler. R di tingkat posisi ke 6 diatas pemrograman *Javascript* dan *PHP*. Sedangkan diperingkat pertama dipegang oleh bahasa pemrograman *Python*, yang juga diminati oleh data *scientist*. R dikembangkan secara *open source* sehingga dapat dipergunakan secara legal tanpa perlu membayar lisensi apapun. Pengerjaan RStudio dimulai pada bulan Desember 2010 sedangkan versi betanya (v0.92) secara resmi dipublik pada bulan Februari 2011. Versi 1.0 dirilis pada tanggal 1 November 2016, dan versi 1.1 dirilis pada tanggal 9 Oktober 2017.

Pada bulan April 2018 RStudio mengumumkan akan menyediakan dukungan operasional serta infrastruktur untuk Ursa Labs. Ursa Labs merfokus dalam membangun runtime sains data baru yang didukung oleh *Apache Arrow*,

RStudio beserta timnya telah berkontribusi banyak dalam membangun paket R, selain paket R terdapat juga

1. *Tidyverse*- paket R untuk data sains (*ggplot2*, *dplyr*, *tidyr* dan *purrr*).
2. *Shiny*- paket R untuk teknologi web interaktif.
3. *Knitr*- paket R untuk laporan dinamis yang menggabungkan R, *Markdown*, HTML, dan TeX.
4. *Packrat*- paket R untuk alat ketergantungan paket.
5. *Devtools*-paket R untuk alat pengembangan.

Disini peneliti belum menjelaskan perbedaan R dengan RStudio. R merupakan mesin dari RStudio, sedangkan Rstudio merupakan *interfacenya* atau tampilannya. RStudio memudahkan kita dalam berkomunikasi dengan R, sehingga kita dapat menjalankan fungsi-fungsi statistika dan data sains. Rstudio memiliki keunggulan atau kelebihan, yaitu:

1. **Gratis**, maksud dari gratis yaitu aplikasi ini dapat dimanfaatkan atau digunakan secara bebas dan gratis. Kita dapat dengan bebas menginstall, memperbarui, memodifikasi, mendistribusikan ulang, bahkan bisa menjualnya kembali. Keunggulan ini merupakan hal terpenting, jika dibandingkan dengan mahalnya software-software statistika lainnya seperti SPSS yang harganya kisaran 1,4 juta perbulan.
2. **Populer**, banyaknya para pengguna R. Setelah mengalami penurunan pada tahun 2018-2019, R kembali menanjak popularitasnya pada tahun 2020, kemungkinan karena adanya Covid-19. Menurut Tiobe Index per Januari 2021, R menempati urutan ke 8, naik dari urutan ke 18 pada bulan januari

3. 2020. R populer di berbagai bidang, yaitu bidang analisis *intelligent*, perbankan, *e-commerce*, *finance*, media sosial, dan bisnis. R dipakai oleh perusahaan-perusahaan besar seperti Facebook, Google, Twitter, Foursquare, Mozilla, dan New York Times.
4. *Powerful*, R memiliki fungsi yang *powerful*, seperti *package* “*caret*” yang digunakan oleh para data sains untuk menjalankan *machine learning*.
5. Lengkap, R memiliki fungsi yang lengkap mulai dari proses data cleaning dan data *manipulation*, data *visualization*, analisis dan permodelan statistika, analisis spasial, teks NLP, dan pembuatan laporan *website*.
6. *Reproducible*, Skrip R dapat kita simpan, skrip R juga dapat digunakan untuk data yang berbeda, artinya dengan sekali klik, kita dapat menghasilkan analisis dan visualisasi yang sama.
7. Dukungan Komunitas, dengan adanya dukungan komunitas ketika kita mengalami suatu kendala, kita tahu dimana kita harus bertanya. Bahkan di Google banyak bermacam-macam alternatif jawaban terhadap pertanyaan kita yang muncul.
8. Dibutuhkan, para perusahaan pencari kerja sering memasukkan skill R sebagai syarat saat melamar kerja. Robert A. Meuchen melakukan analisis terhadap jumlah lowongan pekerjaan berdasarkan *software* analisis populer di indeed.com. Hasilnya, R berada di urutan kelima setelah *Java*, *SAS*, *Python*, dan *C*. Bukan hanya banyak dibutuhkan, pekerja dengan skill R juga mendapatkan gaji yang tinggi. Peluang mendapatkan karir yang lebih baik dan gaji yang lebih baik membuat R semakin populer.

2.8. PreProcessing

Merupakan tahapan yang utama dalam data mining untuk mengubah data mentah atau bisa dikenal *raw* data yang telah dikumpulkan dari berbagai sumber menjadi informasi yang bersih dan bisa digunakan untuk pengolahan data selanjutnya. Pada tahapan ini teks akan dinormalisasi agar mempermudah dalam pembobotan. *Preprocessing* melibatkan validasi suatu data dan imputasi data. Tujuan dari validasi data adalah untuk menilai suatu data apakah data tersebut sudah lengkap atau akurat. Sedangkan tujuan dari imputasi data yaitu untuk memperbaiki kesalahan dan memasukan nilai yang hilang, baik secara manual atau otomatis melalui program *bussines process automation* (BPA).

Data *preprocessing* dipakai dalam aplikasi berbasis *database* dan berbasis aturan. Dalam proses *machine learning*, data *preprocessing* sangat penting untuk memastikan kumpulan data besar di format sedemikian rupa sehingga data yang didalamnya dapat diartikan atau diuraikan oleh algoritma *machine learning*.

Terdapat 2 jenis fitur untuk mendeskripsikan data yaitu :

1. Fitur kategoris, fitur yang penjelasan atau nilainya diambil dari berbagai serangkaian kemungkinan penjelasan atau nilai yang ditentukan. Nilai kategoris bisa berupa jenis hewan, warna rumah, benar atau salah, positif dan negatif.
2. Fitur numerik, fitur dengan nilai yang berkelanjutan pada skala, statistik, atau yang berkaitan dengan bilangan bulat. Nilai numerik diwakili oleh bilangan bulat, presentase, dan pecahan. Fitur numerik juga dapat berupa

3. harga motor, jumlah data dari dokumen, waktu yang diperlukan untuk berpindah ke satu tempat ke tempat lainnya.

Langkah dari data *preprocessing* yaitu :

1. *Data Cleaning*, proses ini melakukan pembersihan data melalui suatu proses seperti mengisi nilai-nilai yang hilang atau menghapus baris dengan data yang hilang, menyelesaikan inkonsistensi dalam data. Menghaluskan data kasar sangat penting untuk set data *machine learning*, karena mesin tidak dapat menggunakan data yang tidak dapat mereka uraikan. Data dapat dibersihkan dengan membagi menjadi segmen-segmen dengan ukuran yang sama kemudian akan dihaluskan, dengan menyesuaikan fungsi regresi linier atau berganda, atau dengan mengelompokkan ke dalam kelompok-kelompok data yang serupa (*clustering*).
2. *Data Integration*, Data dengan representasi yang berbeda disatukan dan konflik didalam data dapat diselesaikan.
3. *Data Transformasi*, Data dinormalisasi dan digeneralisasikan. Normalisasi adalah suatu proses yang memastikan bahwa tidak ada data yang berlebihan, semuanya disimpan di suatu tempat, dan semua data dependensinya logis.
4. *Data Reduction*, Jika volume data sangat besar, maka basis datanya akan lebih lambat, mahal untuk diakses, dan sulit untuk disimpan dengan benar. Data reduction bertujuan untuk menyajikan representasi data yang direduksi dalam sebuah gudang data. Ada beberapa metode untuk mereduksi data yaitu setelah subset atribut yang relevan dipilih untuk signifikansinya, dibawah level tertentu apapun akan dibuang.

5. *Data Discretization*, langkah ini melibatkan pengurangan jumlah nilai suatu atribut kontinu dengan membagi rentang interval atribut.
6. *Data Sampling*, Terkadang karena keterbatasan oleh waktu, penyimpanan, kumpulan data terlalu besar, teknik pengambilan sampel dapat digunakan untuk memilih dan bekerja hanya dengan subset dari dataset.

Preprocessing data memiliki manfaat, yaitu diantaranya:

1. Membuat data lebih mudah untuk dibaca.
2. Mengurangi durasi data mining secara signifikan.
3. Mempermudah proses analisis data dalam machine learning.
4. Mengurangi beban representasi dalam data.

Dalam penelitian ini mempreprocessing datanya dengan cara sebagai berikut:

1. Hapus URL

Tahap ini peneliti melakukan *cleanning* data yaitu dengan menghapus semua url yang ada di komentar.

2. Hapus *Mention*

Tahap ini melakukan *cleanning* data yaitu menghapus semua mention di komentar.

3. *Stemming*

Tahapan ini mengubah semua kata kedalam bentuk kata dasar.

4. Hapus *Hashtag*

Tahapan ini melakukan *cleanning* data dengan menghapus semua tanda hashtag di komentar.

5. Hapus Koma

Tahapan ini melakukan *cleaning* data dengan menghapus semua tanda koma yang ada di komentar.

6. Hapus Titik

Tahapan ini melakukan *cleaning* data dengan menghapus tanda titik yang ada di komentar.

2.9. Shopee

Shopee merupakan *e-commerce* yang berkantor pusat di Singapura yang dimiliki oleh Sea Limited. Shopee pertama kali diluncurkan di Singapura pada tahun 2015, serta memperluas jangkauannya ke Malaysia, Thailand, Taiwan, Indonesia, Vietnam, dan Filipina. Pada tahun 2019 Shopee sudah aktif di negara Brazil, menjadikan Brazil negara pertama di Amerika Selatan dan luar Asia yang dikunjungi Shopee. Shopee dipimpin oleh Chris Feng, mantan karyawan Rocket Internet yang dipimpin oleh Zalora dan Lazada. Pertama kali Shopee meluncur sebagai *marketplace consumer (C2C)*.

Sekarang Shopee beralih ke model hibrid C2C dan business-to consumer (B2C) semenjak meluncurkan Shopee Mall yang merupakan *platform* toko daring yang ternama. Pada tahun 2017, platform ini mencatat 80 juta unduhan aplikasi dengan lebih dari empat juta penjual dan lebih dari 180 juta produk. Pada kuartal tahun keempat tahun 2017, Shopee melaporkan nilai perdagangan bruto (GMV) sebanyak US\$1,6 miliar, naik 206 persen dari tahun sebelumnya. Shopee memiliki nilai total GMV pada tahun 2018 sebesar US\$2,7 miliar, naik 153 persen dari tahun 2017. Di negara Malaysia, Shopee menjadi portal perdagangan

elektronik ke 3 yang paling banyak dikunjungi di Q4 2017, yang menggantikan Lelong dan melampaui peringkat dari market place Lazada sebagai aplikasi yang terbaik di Google Play dan iOS App store. Di bawah ini adalah tampilan dari platform Shopee.



Gambar 2.6 Tampilan Platform Shopee

Shopee merupakan aplikasi *marketplace* online yang digunakan untuk jual beli di ponsel maupun di website sehingga digunakan dengan mudah dan cepat. Shopee menawarkan berbagai macam produk-produk dari produk *fashion* hingga kebutuhan sehari-hari maupun menawarkan atau menjual jasa-jasa untuk para pelanggan atau *customer*. Shopee juga hadir aplikasi *mobile* sehingga memudahkan penggunaanya dalam melakukan belanja online tanpa harus membuka *website* di komputer maupun laptop.

Sasaran bagi Shopee yaitu para kalangan muda yang selalu melakukan kegiatan dengan bantuan gadgetnya terutama untuk berbelanja online. Shopee

hadir dalam aplikasi *mobile* untuk menunjang kegiatan berbelanja dengan mudah dan cepat. Produk-produk yang ditawarkan oleh Shopee lebih mengarah ke produk fahion dan perlengkapan alat rumah tangga.

Di halaman awal pengguna akan disambut dengan 21 kategori di Shopee, yaitu seperti pakaian wanita, pakaian pria, sepatu wanita, sepatu pria, fashion muslim, aksesoris fashion, kecantikan & kesehatan, jam tangan, gadget, perlengkapan olahraga, hobi & mainan, perlengkapan bayi & anak, perlengkapan rumah, elektronik, komputer & aksesoris, makanan & minuman, voucher, fotografi, otomotif, dan serba serbi yang di dalamnya terdapat camilan serta dekorasi rumah.

Segmentasi

CEO Shopee Chris Feng menyatakan basis konsumen Shopee di Indonesia didominasi oleh kelompok pada usia 15-25 tahun. Tidak hanya konsumen saja, mitra pedagang di platform kebanyakan berada pada usia 25-30 tahun. Dengan para milenial sebagai pembawa perubahan bagi dunia digital serta merupakan salah satu penggerak utama di industri *e-commerce*, hal ini sangat penting serta efektif untuk terlibat bersama mereka, untuk menumbuhkan *e-commerce* di negara Indonesia.

Prestasi Shopee diantaranya yaitu :

1. Netizen brand *Choice Award*, Shopee menerima “ The Indonesian netizen *Brand Choice Award 2017*” dalam kategori Belanja Online pada Maret 2017.
2. *Marketing Award*, Shopee menerima penghargaan sebagai salah satu pemenang”*The Best in Marketing Campaign*” di penghargaan *Marketing Award 2017* di Majalah *Marketing* pada September 2017.

3. *Bright Awards* Indonesia 2017, Shopee menerima penghargaan "*Bright Awards* Indonesia 2017" kategori iklan paling berkesan.

Kontroversi dalam Promosi

1. Shopee menggaet Blackpink sebagai brand ambassador mereka pada tahun 2018.
2. Shopee menggaet Syahrini sebagai brand ambassador pada tahun 2019.
3. Shopee menggaet Didi Kempot sebagai brand ambassador pada tahun akhir 2019.
4. Serta pelawak dan presenter Tukul Arwana ditunjuk sebagai bintang iklan di Shopee untuk keperluan promo Shopee 11.11 big Sale.

2.10. Confusion Matriks

Confusion Matrix merupakan alat yang digunakan untuk mengetahui suatu nilai dari tingkat keakurasian. *Confusion Matrix* berbentuk tabel matriks yang digambarkan kinerja dari model klasifikasi pada serangkaian data uji yang nilai sebenarnya diketahui. *Confusion Matrix* mempunyai 4 nilai yaitu *True Positif* (TP), *False Positif* (FP), *False Negative* (FN), dan *True Negative* (TN). Contoh tabel *Confusion Matrix* dapat dilihat pada gambar 2. 1

		Actual Values	
		a (Positif)	b (Negatif)
Predicted Values	A (Positif)	TP (True Positive)	FP (False Positive)
	B (Negatif)	FN (False Negative)	TN (True Negative)

Gambar 2.7 Tabel *Confusion Matriks*

Keterangan :

1. *True Positive* (TP) : Merupakan data positif yang telah diprediksi benar. Contohnya, pasien menderita penyakit jantung (*class 1*) dan dari model yang dibuat telah diprediksi pasien tersebut menderita jantung (*class 1*).
2. *True Negative* (TN) : Merupakan data negatif yang sudah diprediksi benar. Contohnya pasien itu tidak menderita penyakit jantung (*class 2*) dan dari model yang dibuat telah diprediksi pasien tidak menderita penyakit jantung (*class 2*).
3. *False Positif* (FP) : Merupakan data positif tetapi sudah diprediksi hasilnya data negatif, Contohnya pasien menderita penyakit jantung (*class 2*) tetapi dari model yang dibuat diprediksi pasien tersebut tidak menderita penyakit jantung (*class 2*).
4. *False Negative* (FN) : Merupakan data negatif tetapi setelah diprediksi hasilnya data positif (*class 1*), Contohnya pasien itu menderita penyakit jantung, tetapi dari model yang dibuat diprediksi pasien itu tidak menderita penyakit jantung (*class 1*).

Beberapa kasus "*False Negative*" lebih berbahaya, ketika kita menghubungkan pernyataan tersebut dengan contoh prediksi jantung. Jika pasien tidak menderita jantung tetapi diprediksi menderita jantung (FP), maka diagnosa selanjutnya pasien tersebut dapat mengetahui keadaan sebenarnya bahwa pasien tersebut benar tidak menderita jantung. Tetapi jika ada pasien yang sebenarnya menderita jantung tetapi diprediksi tidak menderita penyakit jantung (FN), maka pasien tersebut akan mengetahui keadaan sebenarnya dengan sangat terlambat dan

pasien tersebut tidak dengan segera mengambil tindakan pencegahan medis untuk penyakit jantung. Sehingga dapat memperburuk keadaan pasien tersebut. Jadi dapat dikatakan bahwa “*False Negative*” sangat berbahaya.

Cara mudah mengingatnya adalah :

1. Jika diawal *True* maka prediksinya benar, walaupun diprediksi benar atau salah.
2. Jika diawal *false* maka prediksinya salah.
3. Positif dan negatif adalah hasil prediksi dari model.

Cara Penghitungannya :

$$Accuracy = (TP + TN) / (TP + FP + FN + TN)$$

Hasilnya = _____

2.11. Sublime Text

Sublime text adalah salah satu text editor yang banyak diminati atau digunakan, karena aplikasi ini dapat digunakan di berbagai macam *platform* OS. Sublime text diresmikan pada tahun 2008 dan sudah memiliki beberapa versi. Sublime text sangat berguna bagi para *web developet* atau *programer* seperti memudahkan *programer* dalam membedakan *syntak* dengan warna yang dimilikinya.

BAB III METODE PENELITIAN

3.1. Jenis, Sifat, dan Pendekatan Penelitian

Penelitian ini termasuk ke dalam jenis penelitian kuantitatif, yaitu dengan melakukan perhitungan secara matematis terhadap data-data yang diuji ataupun data yang di *training*.

3.2. Metode Pengumpulan Data

Penelitian ini dilakukan pada tweet berbahasa Indonesia, data yang diambil dari komentar-komentar atau *review* dari pengguna maupun penjual toko online yang memakai jasa *market place* Shopee di grup Twitter menggunakan *software* RStudio. Data komentar akan dibagi menjadi dua data yaitu data komentar positif dan negatif. Jumlah data yang digunakan sebanyak 400 data, yang terdiri dari 200 data yaitu 100 data untuk komentar positif dan 100 data untuk komentar negatif dan untuk data latihnya berjumlah 200 buah komentar yang terbagi sama rata antara komentar positif dan negatif untuk data uji. Data yang dipakai adalah data yang diambil pada tanggal 4 Januari 2021. Proses pelabelan pada komentar-komentar yang ada di grup Shopee di Twitter menggunakan aplikasi RStudio, dengan *coding-coding* yang telah disediakan di aplikasi RStudio. Proses *mentraining* data menggunakan algoritma *Naive Bayes*, dan penghitungan tingkat keakurasiannya menggunakan *Convusion Matriks*.

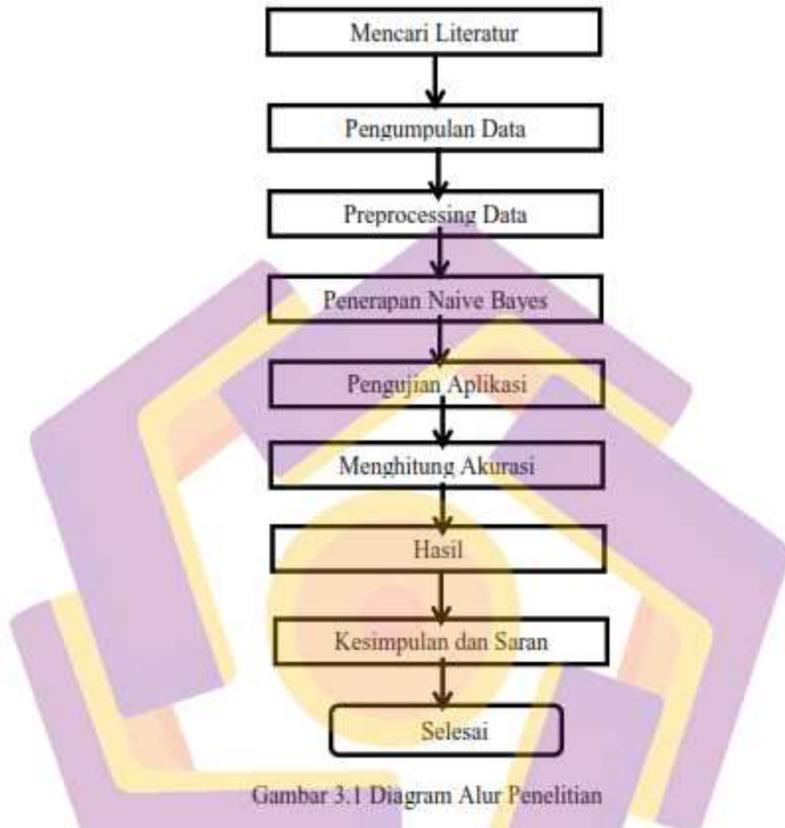
3.3. Alur Penelitian

Adapun tahapan yang dilakukan pada penelitian ini yaitu :

1. Peneliti melakukan pengamatan pada *review* atau komentar pelanggan yang ada di grup Shopee yang ada di Twitter.
2. Mengumpulkan beberapa *literatur* yang mengacu pada penelitian yang sedang dikerjakan atau dilakukan.
3. Mengambil data-data *review* atau komentar pelanggan dari grup Shopee di Twitter.
4. Setelah data-data terkumpul, melakukan preprocessing data untuk menghapus *hashtag*, URL, *Retweet*, dan *mention*.
5. Setelah selesai *preprocessing* data selanjutnya memasukkan data ke aplikasi *Naive bayes* untuk di *training*, dan proses *stemming* data.
6. Menghitung tingkat akurasinya menggunakan *convusion matriks*.

Berikut merupakan diagram alur penelitian





Penjelasan dari alur penelitian

Pada gambar 3.1 diagram alur penelitian yang digunakan dalam penelitian hanya 7 tahapan yaitu pendahuluan, latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, pembahasan dan pengumpulan data, dan kesimpulan dan saran.

Penelitian ini membangun sistem terdiri dari empat proses utama yaitu pengumpulan data, *preprocessing*, pelabelan, serta klasifikasi data. Data yang

dipakai adalah data dari grup Shopee di Twitter berbahasa Indonesia. Data tweet ini diperoleh dari *scraping* menggunakan aplikasi RStudio yaitu *library* (tm), *library* (twitterR). Pertama dari penelitian ini yaitu mengumpulkan komentar-komentar atau *review* dari para pelanggan Shopee di grup Shopee di Twitter. Pengumpulan data komentar menggunakan atau memanfaatkan *API Search* Twitter dari RStudio. Data tweet yang telah terkumpul otomatis tersimpan ke dalam penyimpanan komputer yang kita inginkan. Tahap selanjutnya yaitu melakukan pelabelan komentar Twitter melalui aplikasi RStudio sesuai dengan sentimen yang sudah ditentukan sebelumnya yaitu sentimen positif untuk komentar positif dan sentimen negatif untuk komentar negatif. Setelah seluruh data komentar memiliki kelas masing-masing kemudian dilakukan pemisahan data menjadi dua bagian yaitu data *testing* dan data *training*.

Data latih digunakan untuk mengklasifikasikan tweet pada kelas sentimennya dan jumlah datanya lebih banyak. Dalam penelitian menggunakan data latih sebesar 80% dari total data keseluruhannya. Data uji digunakan untuk mengukur atau menilai sejauh mana *classifier* berhasil melakukan klasifikasi dengan benar. Jumlah data yang digunakan dalam penelitian ini yaitu sebanyak 20%. Selanjutnya kedua bagian tersebut dilakukan proses *preprocessing*. Adapun tahap *preprocessingnya* antara lain :

1. Normalisasi Fitur

Biasanya di dalam komentar-komentar tweet terdapat *username*, URL, dan karakter khusus seperti *hashtag* (#), *retweet* (RT), tanda baca (.,&,%,\$, dan lainnya). karena URL dan *username* merupakan karakter khusus, tanda baca dan

karakter lainnya tidak akan berpengaruh pada nilai sentimen, maka komponen tersebut akan dibuang. Seperti gambar 3.2 *normalisasi fitur*.

Data Input	Data Output
@ShopeeID Tahun baru, hp baru, Rejeki aku amin	Tahun baru hp baru rejekiaku amin

Gambar 3.2 *Normalisasi Fitur*

Proses *preprocessing* dapat dilihat pada gambar 3.3 *preprocessing*



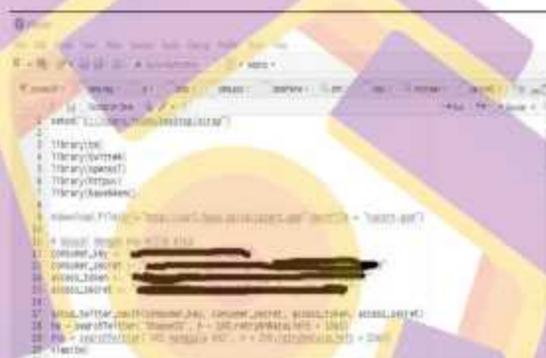
Gambar 3.3 *Proses Preprocessing*

Setelah tahap *preprocessing* selesai kemudian dilakukan *implementasi Naive Bayes* yaitu proses *training* data. Data *testing* di penelitian ini yaitu pelabelan pada komentar-komentar yang ada di grup Shopee di Twitter. Data *testing* yang sudah dilabeli data negatif maupun data positif dimasukkan ke aplikasi *Naive Bayes* untuk melakukan *training* data.

BAB IV HASIL PENELITIAN DAN PEMBAHASAN

4.1. Tampilan RStudio

Untuk tahap awal, kita harus sudah menginstal *Software* RStudio terlebih dahulu, setelah sudah terinstal kita bisa mulai untuk menambang data dari grup Shopee di Twitter. Berikut tampilan dari RStudio



```
1 setwd("D:/TUGAS/Penting/scrap")
2 library(twit)
3 library(twitapi)
4 library(twitnamer)
5 library(twitpic)
6 library(twittext)
7 library(twiturl)
8 library(twituser)
9 library(twitvideo)
10 library(twitgeo)
11 library(twitmedia)
12 library(twitdm)
13 library(twitdmapi)
14 library(twitdmapi2)
15 library(twitdmapi3)
16 library(twitdmapi4)
17 library(twitdmapi5)
18 library(twitdmapi6)
19 library(twitdmapi7)
20 library(twitdmapi8)
21 library(twitdmapi9)
22 library(twitdmapi10)
23 library(twitdmapi11)
24 library(twitdmapi12)
25 library(twitdmapi13)
26 library(twitdmapi14)
27 library(twitdmapi15)
28 library(twitdmapi16)
29 library(twitdmapi17)
30 library(twitdmapi18)
31 library(twitdmapi19)
32 library(twitdmapi20)
33 library(twitdmapi21)
34 library(twitdmapi22)
35 library(twitdmapi23)
36 library(twitdmapi24)
37 library(twitdmapi25)
38 library(twitdmapi26)
39 library(twitdmapi27)
40 library(twitdmapi28)
41 library(twitdmapi29)
42 library(twitdmapi30)
43 library(twitdmapi31)
44 library(twitdmapi32)
45 library(twitdmapi33)
46 library(twitdmapi34)
47 library(twitdmapi35)
48 library(twitdmapi36)
49 library(twitdmapi37)
50 library(twitdmapi38)
51 library(twitdmapi39)
52 library(twitdmapi40)
53 library(twitdmapi41)
54 library(twitdmapi42)
55 library(twitdmapi43)
56 library(twitdmapi44)
57 library(twitdmapi45)
58 library(twitdmapi46)
59 library(twitdmapi47)
60 library(twitdmapi48)
61 library(twitdmapi49)
62 library(twitdmapi50)
63 library(twitdmapi51)
64 library(twitdmapi52)
65 library(twitdmapi53)
66 library(twitdmapi54)
67 library(twitdmapi55)
68 library(twitdmapi56)
69 library(twitdmapi57)
70 library(twitdmapi58)
71 library(twitdmapi59)
72 library(twitdmapi60)
73 library(twitdmapi61)
74 library(twitdmapi62)
75 library(twitdmapi63)
76 library(twitdmapi64)
77 library(twitdmapi65)
78 library(twitdmapi66)
79 library(twitdmapi67)
80 library(twitdmapi68)
81 library(twitdmapi69)
82 library(twitdmapi70)
83 library(twitdmapi71)
84 library(twitdmapi72)
85 library(twitdmapi73)
86 library(twitdmapi74)
87 library(twitdmapi75)
88 library(twitdmapi76)
89 library(twitdmapi77)
90 library(twitdmapi78)
91 library(twitdmapi79)
92 library(twitdmapi80)
93 library(twitdmapi81)
94 library(twitdmapi82)
95 library(twitdmapi83)
96 library(twitdmapi84)
97 library(twitdmapi85)
98 library(twitdmapi86)
99 library(twitdmapi87)
100 library(twitdmapi88)
101 library(twitdmapi89)
102 library(twitdmapi90)
103 library(twitdmapi91)
104 library(twitdmapi92)
105 library(twitdmapi93)
106 library(twitdmapi94)
107 library(twitdmapi95)
108 library(twitdmapi96)
109 library(twitdmapi97)
110 library(twitdmapi98)
111 library(twitdmapi99)
112 library(twitdmapi100)
```

Gambar 4.1 Tampilan awal RStudio

4.2. Pengambilan Data

Pada tahap selanjutnya pertama-tama kita harus membuat perintah *setting working directory* terlebih dahulu, perintah ini untuk menentukan dimanakah kita mau menyimpan data-data yang telah kita ambil dari RStudio. Di gambar di bawah ini saya menyimpan data nya di *directory* D/di folder PENTING/folde scrap.

```
1 setwd("D://TUGAS/Penting/scrap")
```

Gambar 4.2 Source Code Directory

Setelah kita telah menentukan mau di mana data tersebut disimpan, selanjutnya kita harus menginstal *package* yang diperlukan terlebih dahulu agar kita bisa menambang data dari Twitter serta agar Twitter bisa mendeteksi RStudio kita, dengan perintah di bawah ini agar kita dapat menjalankan RStudio nya

```
install.packages("RCurl")
install.packages("tm")
install.packages("twitter")
```

Gambar 4.3 Source Code install package

Setelah kita menginstal code yang kita butuhkan maka selanjutnya kita gunakan perintah *library* untuk menjalankan perintah package yang telah kita install tadi contoh gambarnya seperti dibawah ini

```
library(RCurl)
library(tm)
library(twitter)
```

Gambar 4.4 Perintah Library

Dan jangan lupa untuk mendownload *sertifikat Curl* nya, agar kita bisa *mencrawling* data-data yang ada di twitter seperti gambar dibawah ini

```
download.file(url="http://curl.haxx.se/ca/cacert.pem",destfile = "cacert.pem")
```

Gambar 4.5 Download Sertifikat Curl

Setelah kita mendownload *curl* nya kita dapat melanjutkan prosesnya yaitu meminta izin ke Twitter dengan cara memasukan *key* tweet kita, agar kita bisa mengakses twitter kita untuk menambang data-datanya, lihat gambar dibawah ini

```

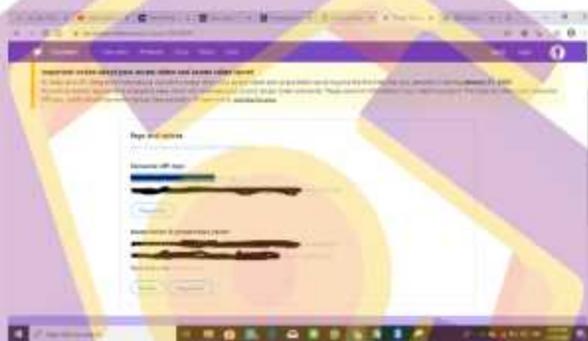
consumer_key <- "rEyb656LzWatTAHARcZYEQj7m"
consumer_secret <- "rvRj0svDQRWssYSN3EtVngFA7QpTfa1xHu55c2u0sZd1Y42DsH"
access_token <- "1357132242355593218-BFNBCYxagaRyRVIZQ25N0UkTZ1QMP7"
access_secret <- "BnKWG4rEQXxa5RXNk6aRn4hfk2k8ZT6DuspX72uTzdIHr"

setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)

```

Gambar 4.6 Key Twitter

Kalian bisa mengakses *consumer key* serta *consumer secret* nya melalui *Twitter Developer*, berikut contoh gambarnya

Gambar 4.7 Mengakses *Consumer key*

Setelah kita tahu *consumer key*nya kita lanjut ke tahap selanjutnya kita bisa menambang data dari twitter dengan *code*

```
tw = searchTwitter("ShopeeID", n = 400, retryOnRateLimit = 10e3)
```

Gambar 4.8 Kode Ambil Data Di Twitter

Di kode gambar di atas saya mencari data dari grup Shopee yang ada di Twitter dengan jumlah data 400 data. Setelah kita dapat datanya, jangan lupa kita untuk menyimpan dulu datanya dengan perintah (`saveRDS(tw, file = 'shopee.rds')`). Maksud yang tulisan yang di tandai warna merah itu adalah nama


```
#Hapus Koma
replacecomma <- function(y) gsub(",", "", y)
twitclean <- tm_map(twitclean, replacecomma)

#Hapus Titik
removetitik2 <- function(y) gsub(":", "", y)
twitclean <- tm_map(twitclean, removetitik2)
```

Gambar 4.10 Perintah *Cleaning* Data (Lanjutan)

Setelah selesai *cleaning* data yang telah kita ambil selanjutnya kita melakukan *skoring* data positif dan negatif seperti gambar 4.11

```
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)
}
```

Gambar 4.11 Perintah *skoring* data Positif dan Negatif

Contoh hasil *skoring* yang telah kita panggil, bisa kita lihat pada gambar 4.12, pada gambar dibawah data masih tercampur antara positif, negatif, dan netral.

score	text	klasifikasi
1	0 semangat jika ini good luck !!	Positif
2	1 acma milih gatel dan ini koo kuat menang baru atiger aku per dia ngak keawalan info mari mlttstn agom	Positif
3	0 dika kaku	Positif
4	1 acm mentu banget ini kaku an	Positif
5	1 ita ada koo mentu banget jika dia dlm ga jika jagan remeh ini ga	Positif
6	0 mri koo dm dang	Positif
7	1 semih jika semangat itu ya kuat kudu konyonya godudu koo ya	Positif
8	0 acm mlttstn agom	Positif
9	0 in koo koo jagan omg kung hwa koo kuat banget mari ini ya koo klttstn ur	Positif
10	0 jagan omg dang koo koo koo koo koo	Positif
11	0 acm jga koo semangatttt an	Positif
12	0 kaku koo mri	Positif

Gambar 4.12 Skoring data

Setelah melakukan skoring data, selanjutnya membagi atau memisahkan data ke data positif dan data negatif dengan seperti gambar 4.14

```
data.pos <- hasil[hasil$score>0,]
View(data.pos)
write.csv(data.pos, file = "data-pos.csv")

data.neg <- hasil[hasil$score<0,]
View(data.neg)
write.csv(data.neg, file = "data-neg.csv")
```

Gambar 4.13 Memisahkan Data Positif dan negatif

Setelah berhasil memanggil *code* diatas akan menghasilkan data seperti gambar

4.14

score	text	klasifikasi
19	2 Bantlah semoga kali ini aku bisa menang ya Allah Butuh banget hp baru http://susi14a76wT	Positif
57	2 Bantlah semoga kali ini aku bisa menang ya Allah Butuh banget hp baru http://suzbaQbc5Dz	Positif
95	2 Bantlah semoga kali ini aku bisa menang ya Allah Butuh banget hp baru http://weflBemfemr	Positif
140	1 baru pulang	Positif
152	2 Bantlah semoga kali ini aku bisa menang ya Allah Butuh banget hp baru http://suzbaieqdi1Zz	Positif
154	2 Bantlah semoga kali ini aku bisa menang ya Allah Butuh banget hp baru http://suzbaieqdi1Zz	Positif
235	2 Bantlah semoga kali ini aku bisa menang ya Allah Butuh banget hp baru http://suzbaieqdi1Zz	Positif
273	2 Bantlah semoga kali ini aku bisa menang ya Allah Butuh banget hp baru http://suzbaieqdi1Zz	Positif
312	2 Bantlah semoga kali ini aku bisa menang ya Allah Butuh banget hp baru http://suzbaieqdi1Zz	Positif
354	2 Bantlah semoga kali ini aku bisa menang ya Allah Butuh banget hp baru http://suzbaieqdi1Zz	Positif
388	2 Bantlah semoga kali ini aku bisa menang ya Allah Butuh banget hp baru http://suzbaieqdi1Zz	Positif

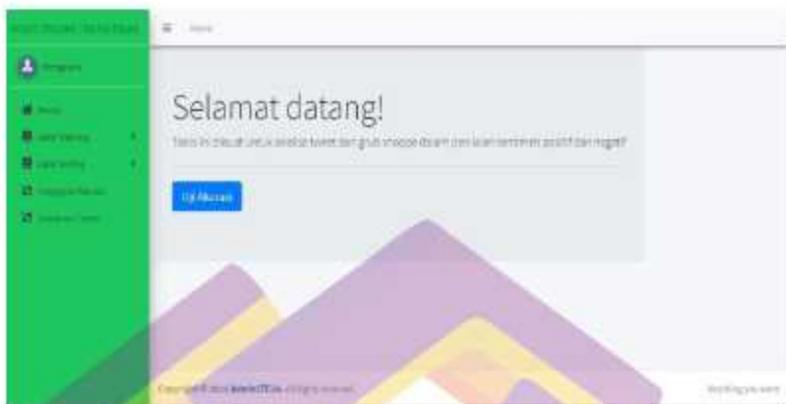
score	test	Klasifikasi
5	-1. Bantilan ateh nyoba bekal ngikut GA seroga kali in beruntung antr BELAJA Di Sini Httstoz5uCYGQ	Negatif
9	-1. Ya ampun bisa deh kak cek di Shopee ada namanya ShopeeFood nah bisa tuh kak jajan gitu Htt bke dan ...	Negatif
15	-1. Bantilan udh bekal kal kut to ga pernah menang seroga bisa kal in menang A HttstozGjAPRZOf	Negatif
21	-1. Bantilan ateh nyoba bekal ngikut GA seroga kali in beruntung antr BELAJA Di Sini Httstoz5uCYGQ...	Negatif
26	-1. Bantilan udh bekal kal kut to ga pernah menang seroga bisa kal in menang A Httstoz5uCYGQ	Negatif
45	-1. Bantilan ateh nyoba bekal ngikut GA seroga kali in beruntung antr BELAJA Di Sini HttstozGGR6au...	Negatif
65	-1. Bantilan ateh nyoba bekal ngikut GA seroga kali in beruntung antr BELAJA Di Sini Httstoz5uCYGQ...	Negatif
81	-1. Bantilan udh bekal kal kut to ga pernah menang seroga bisa kal in menang A Httstoz5uCYGQ	Negatif
98	-1. Bantilan udh bekal kal kut to ga pernah menang seroga bisa kal in menang A Httstoz5uCYGQ	Negatif
109	-1. Bantilan ateh nyoba bekal ngikut GA seroga kali in beruntung antr BELAJA Di Sini Httstoz5uCYGQ...	Negatif
114	-1. Bantilan udh bekal kal kut to ga pernah menang seroga bisa kal in menang A Httstoz5uCYGQ	Negatif
118	-1. Bantilan udh bekal kal kut to ga pernah menang seroga bisa kal in menang A Httstoz5uCYGQ	Negatif

Gambar 4.14 Hasil Data Positif dan Data Negatif

Gambar di atas merupakan hasil data yang sudah kita bagi datanya yaitu data positif dan negatif, dengan cara tersebut kita dengan mudah melihat data positif dan negatifnya.

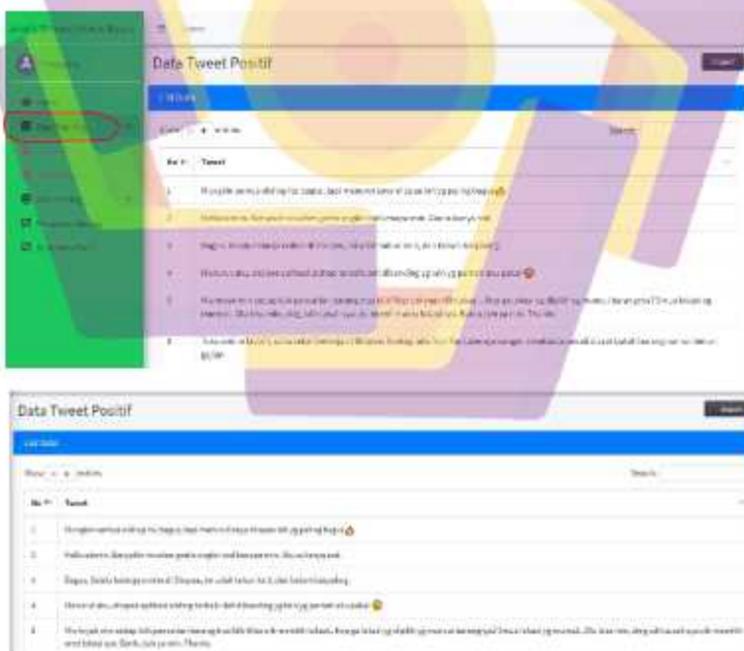
4.4. Aplikasi *Naive Bayes*

Setelah data-data selesai pemrosesan *cleanning*, *scoring* data negatif serta positif, selanjutnya data-data tersebut kita masukan kedalam aplikasi *Naive Bayes* untuk melakukan *training* data. Di aplikasi ini kita melakukan proses *stemming* data dari twitter, serta *training* data, dan proses penghitungan tingkat akurasi. Di bawah ini adalah tampilan dari aplikasi *Naive Bayes*



Gambar 4.15 Tampilan Aplikasi *Naive Bayes*

Fitur-fitur dari aplikasi *Naive Bayes*:



Hasil Pengujian					
List Data					
No	Uraian	Benar	Salah	Real	Benar
1	Siapa yang bilang dia bagus, ya memang ya memang dia yang bilang	siapa yang bilang dia bagus ya memang ya memang dia yang bilang	siapa yang bilang dia bagus ya memang ya memang dia yang bilang	positif	positif
2	Itu adalah... Bawalah... dan jangan lupa... dan jangan lupa...	itu adalah... bawalah... dan jangan lupa... dan jangan lupa...	itu adalah... bawalah... dan jangan lupa... dan jangan lupa...	positif	positif
3	Siapa... Siapa... Siapa... Siapa... Siapa... Siapa... Siapa... Siapa...	siapa... siapa... siapa... siapa... siapa... siapa... siapa... siapa...	siapa... siapa... siapa... siapa... siapa... siapa... siapa... siapa...	positif	positif
4	Siapa... Siapa... Siapa... Siapa... Siapa... Siapa... Siapa... Siapa...	siapa... siapa... siapa... siapa... siapa... siapa... siapa... siapa...	siapa... siapa... siapa... siapa... siapa... siapa... siapa... siapa...	positif	positif
5	Siapa... Siapa... Siapa... Siapa... Siapa... Siapa... Siapa... Siapa...	siapa... siapa... siapa... siapa... siapa... siapa... siapa... siapa...	siapa... siapa... siapa... siapa... siapa... siapa... siapa... siapa...	positif	positif
6	Siapa... Siapa... Siapa... Siapa... Siapa... Siapa... Siapa... Siapa...	siapa... siapa... siapa... siapa... siapa... siapa... siapa... siapa...	siapa... siapa... siapa... siapa... siapa... siapa... siapa... siapa...	positif	positif

Gambar 4.17 Tampilan *Stemming* Data Twitter

Dari gambar 4.17, merupakan gambar hasil *stemming* dari komentar-komentar Twitter. Disana terdapat hasil sebelum di *stemming* serta telah di *stemming*, isi Twitter kita masukan dari tombol *import* yang terdapat di aplikasi Naive Bayes. Setelah proses *stemming* selesai dapat dilihat hasil dari *Naive Bayesnya* seperti gambar 4.18

Confusion Matrix		
	Actual	Predicted
Actual		
(Positive)	15	1
(Negative)	0	14

Gambar 4.18 Hasil dari Proses *Naive Bayes*

Maksud dari gambar 4.18 yaitu aplikasi mengklasifikasikan dengan benar sebesar 14 data sebagai negatif dan 15 data positifnya. Selain itu aplikasi salah dalam memprediksi 1 data ke dalam negatif yang seharusnya data positif, serta salah dalam memprediksi 0 data ke dalam positif yang seharusnya negatif. Setelah diketahui hasilnya kita lanjutkan untuk menghitung tingkat akurasi dengan rumus dibawah ini

$$\begin{aligned}
 \text{Akurasi} &= (TP+TN) / (TP+FP+FN+TN) \\
 &= (15+14) / (15+19+14) \\
 &= 29/58 \\
 &= 0,507 * 100\% \\
 &= 50,7\% (51\%) \\
 \\
 \text{Precision} &= TP / (TP+FP) \\
 &= 15 / (15+1) \\
 &= 15 / 16 \\
 &= 0,937 * 100\% \\
 &= 93,7\% (94\%) \\
 \\
 \text{Recall} &= TP / (TP+FN) \\
 &= 15 / (15+0) \\
 &= 15 / 15 \\
 &= 1 * 100\% \\
 &= 100\% \\
 \\
 \text{F1 Score} &= 2 * \text{Accuracy} * \text{Precision} / (\text{Accuracy} + \text{Precision}) \\
 &= 2 * 0,507 * 0,937 / (0,507 + 0,937) \\
 &= 1,024 / 1,444 \\
 &= 0,709 * 100\% \\
 &= 70,9\% (71\%)
 \end{aligned}$$

Gambar 4.19 Rumus *Convusion Matriks*

Penjelasan dari rumus di atas :

TP : Pernyataan positif yang benar-benar terjadi

TN : Pernyataan negatif yang benar-benar terjadi

FP : Suatu prediksi positif, tetapi kebenarannya bersifat negatif

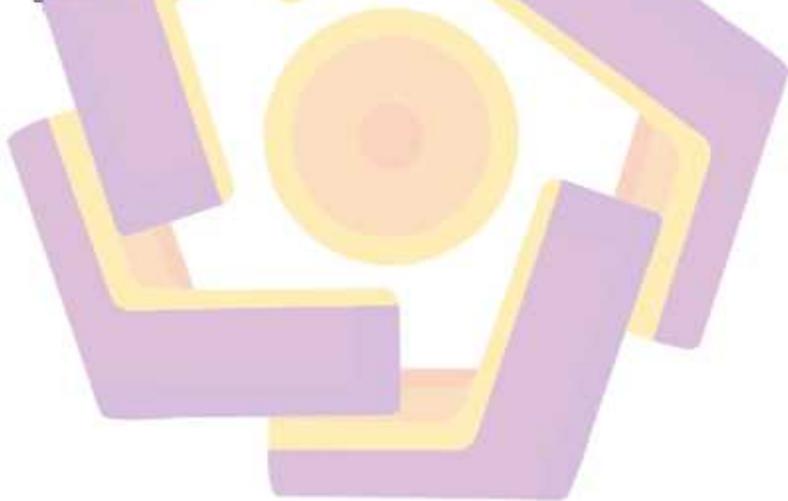
FN : Suatu prediksi negatif, tetapi kebenarannya bersifat positif

1. *Accuracy* menggambarkan seberapa akurat model dalam pengklasifikasian dengan benar.
2. *Precision* menggambarkan akurasi antara data yang diminta dengan prediksi yang diberikan oleh model
3. *Recall* merupakan keberhasilan model dalam menemukan kembali sebuah informasi
4. *F1 Score* merupakan perbandingan rata-rata antara *precision* dengan *recall*.

Hasil Akurasi	
Akasi	97%
Presi	94%
Recall	100%

Gambar 4.20 Hasil *Convusion Matriks*.

Hasil dari tingkat akurasinya dalam sentimen pada komentar-komentar grup Shopee yang ada di Twitter yaitu sebesar 97%, *precision* nya sebesar 94 %, sedangkan *recall* nya sebesar 100% dan F1 score nya sebesar 97 % ,seperti gambar 4.20.



BAB V PENUTUP

5.1. Kesimpulan

Berdasarkan hasil analisis yang telah dilakukan, diperoleh beberapa kesimpulan sebagai berikut :

1. Menghitung sentimen analisis pada komentar grup Shopee di Twitter dengan metode *Naive Bayes* sudah berhasil.
2. Tingkat keakurasian yang dihasilkan dalam sentimen analisis pada komentar-komentar pelanggan atau pengguna Shopee di grup Shopee sebesar 97% dengan menggunakan *Convusion Matriks*.
3. Dari hasil pengambilan data dari grup Shopee di Twitter menggunakan software RStudio menghasilkan 250 data positif dan 100 data yang negatif, sisanya 50 data netral. Dengan memanggil 400 data dari grup ShopeeID di Twitter dengan code "tw = searchTwitter('ShopeeID', n = 400, retryonRateLimit = 10e3)" Setelah itu melakukan pelabelan data untuk dibagi menjadi dua data yaitu data positif dan negatif.
4. Penerapan *Naive Bayes* dalam *sentimen* pada komentar di grup Shopee dengan cara mengambil data komentar di Twitter menggunakan RStudio, serta melakukan pelabelan pada komentar yang dibagi menjadi 2 data yaitu komentar positif dan komentar negatif, selanjutnya komentar yang telah dilabeli dimasukkan kedalam aplikasi *Naive Bayes* untuk di training datanya, maka akan ditemukan hasilnya yaitu aplikasi mengklasifikasikan dengan benar sebesar 14 data sebagai data negatif dan 15 data untuk data

5. positifnya. Selain itu aplikasi salah dalam memprediksi 1 data ke dalam data negatif yang seharusnya data positif, serta salah dalam memprediksi 0 data ke dalam data positif yang seharusnya data negatif.

5.2.Saran

Berdasarkan kesimpulan di atas, dapat diberikan beberapa saran sebagai berikut :

1. Untuk penelitian selanjutnya diharapkan dalam memasukan data dari Twitter ke aplikasi *Naive Bayesnya* secara otomatis.
2. Perlunya ditambahkan perbandingan antara metode satu dengan metode lainnya agar hasil yang didapat lebih baik lagi.
3. Pada penelitian selanjutnya diharapkan hasilnya tidak hanya tahu bahwa data itu adalah positif atau negatif saja, tetapi apa manfaatnya kalau sudah diketahui bahwa data itu adalah positif atau negatif.

DAFTAR PUSTAKA

PUSTAKA BUKU

- E. Kouloumpis, T. Wilson, and J. Moore, 2011, "Twitter sentiment analysis: The good the bad and the omg!" Proc. Fifth Int. AAAI Conf. Weblogs Soc. Media (ICWSM 11)
- Liu, B., Blasch, E., Chen, Y., Shen, D. & Chen, G. 2013, Scalable sentiment classification for Big Data analysis using Naive Bayes Classifier. 2013 IEEE International Conference on Big Data. 2013 IEEE International Conference on Big Data. hal.99-104
- Liu, B., 2012, Sentiment Analysis and Opinion Mining. In: Chicago: Morgan & Claypool
- Taheri, S. & Mammadov, M., 2013, Learning The Naive Bayes Classifier With Optimization Models. In International Journal of Applied Mathematics and Computer Science. pp. 787-795

PUSTAKA MAJALAH, JURNAL ILMIAH ATAU PROSIDING

- Andris Faezal, Aziz Muslim, Aditya Hastami Ruger. 2020, Sentimen Analisis Pada Data Tweet Pengguna Twitter Terhadap Produk penjualan Toko Online Menggunakan Metode K-Means. Jurnal Matrik. Volume 19 No 2. ISSN : 2476-9843.
- Hartanto. 2017, Text Mining dan Sentimen Analisis Twitter Pada Gerakan LGBT. Jurnal Psikologi Ilmiah. Volume 9 No 1. ISSN : 2541-2965
- Kepin Sihotang, Rajib Ghaniy . 2019, Penerapan Metode Naive Bayes Classifier Untuk Penentuan Topik Tugas Akhir Pada Website Perpustakaan STIKOM Binaniaga. Jurnal Ilmiah Teknologi-Informasi dan Sains (TeknoIS). Volume 9 No 1. ISSN : 2597-8918
- Rifiana Arif, Karel Imanuel. 2019, Analisis Sentimen Topik Viral Desa Penari Pada Media Sosial Twitter Dengan Metode Lexicon Based. Jurnal Ilmiah Matrik. Vol. 21 No. 3. ISSN : 2621-8089
- Suwanda Aditya Saputra, Didi Rosiyadi, Windu Gata, Syepri Maulana Husain, 2019, Analisis Sentimen E-Wallet Pada Google Play Menggunakan Algoritma Naive Bayes Berbasis Particle Swarm Optimization. Jurnal RESTI. Vol.3 No.3. ISSN Media Elektronik: 2580-0760

PUSTAKA LAPORAN PENELITIAN

- Saraswati, Ni Wayan S., 2011, "Text Mining dengan Metode Naïve Bayes Classifier dan Support Vector Machine Untuk Sentimen Analysis", Tesis - Universitas Udayana, Tesis Magister Program Studi Teknik Elektro Program Pascasarjana, Denpasar
- Suryadi.2017, Text Mining dengan Metode Naive Bayes Classifier dan Support Vector Machines untuk Sentiment Analysis, Tesis, Program Pascasarjana, Universitas Udayana, Denpasar



LAMPIRAN

Code Menambang Data Komentar di Grup Shopee di Twitter dengan RStudio

```
setwd("D://TUGAS/Penting/scrap")
install.packages("RCurl")
install.packages("ROAuth")
install.packages("tm")
install.packages("twitteR")
library(RCurl)
library(tm)
library(twitteR)
library(ROAuth)
download.file(url="http://curl.haxx.se/ca/cacert.pem",destfile=
cacert.pem")
```

Key untuk masuk ke Twitter

consumer_key

```
<- "rEvb6S6LzWatTAHARcZYEQJ7m"consumer_secret<-
```

```
"rvRj0SvDqRwSsySN3tTvngFA7QpTfa1xHu5Sc2uoszd1Y42DsH"
```

access_token

```
<- "1357132242355593218-BFNBCYxagaRyRVIZQ2SNoUxTZ1QMP7"
```

access_secret

```
<- "BnKWG4rEQXXaSRXNk6aRm4hfk2kBZT6DUspX72uTzdlHr"
```

```
setup_twitter_oauth(consumer_key, consumer_secret, access_token,  
access_secret)
```

Carl data di Twitter

```
tw = searchTwitter('ShopeeID', n = 400, retryOnRateLimit = 10e3)
```

```
View(tw)
```

Save data

```
saveRDS(tw, file = 'njeh.rds')
```

```
tw <- readRDS('njeh.rds')
```

```
d = twListToDF(tw)
```

Code ambil data komen saja

```
komen <- d$text
```

```
komenc <- Corpus(VectorSource(komen))
```

Cleaning Data

Hapus URL

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

```
twitclean <- tm_map(komenc, removeURL)
```

Hapus Mention

```
removeMention <- function(x) gsub("@\\S+", "", x)
```

```
twitclean <- tm_map(komenc, removeMention)
```

Hapus Hashtag

```
removeHashtag <- function(x) gsub("#\\S+", "", x)
```

```
twitclean <- tm_map(komenc, removeHashtag)
```

Hapus RT

```
removeRT <- function(y) gsub("RT ", "", y)
```

```
twitclean <- tm_map(twitclean, removeRT)
```

Save data Jenis .csv

```
write.csv(dataframe,file = 'wait.csv')
```

Skoring Data

```
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)
hasil = score.sentiment(kalimat2$text, kata.positif, kata.negatif)
head(hasil)

```

Convert teks ke sentimen

```

hasil$klasifikasi<- ifelse(hasil$score<0, "Negatif",ifelse(hasil$score>0,
"Positif","Netral"))
hasil$klasifikasi
View(hasil)

```

Pelabelan data positif dan negatif

```

data.pos <- hasil[hasil$score>0,]
View(data.pos)
data.neg <- hasil[hasil$score<0,]
View(data.neg)

```

