ANALISIS SENTIMEN MENGENAI KOMISI PEMBERANTASAN KORUPSI (KPK) PADA MEDIA SOSIAL TWITTER DENGAN MENERAPKAN ALGORITMA *NAÏVE BAYES CLASSIFIER*

(Skripsi)

Oleh

HENDY SYUHADA

FAKULTAS TEKNIK
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2022

ANALISIS SENTIMEN MENGENAI KOMISI PEMBERANTASAN KORUPSI (KPK) PADA MEDIA SOSIAL TWITTER DENGAN MENERAPKAN ALGORITMA *NAIVE BAYES CLASSIFIER*

Oleh

HENDY SYUHADA

Skripsi

Sebagai Salah Satu Syarat untuk Mencapai Gelar
SARJANA TEKNIK

Pada

Program Studi Teknik Informatika
Jurusan Teknik Elektro
Fakultas Teknik Universitas Lampung

FAKULTAS TEKNIK
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2022

ABSTRAK

ANALISIS SENTIMEN MENGENAI KOMISI PEMBERANTASAN KORUPSI (KPK) PADA MEDIA SOSIAL TWITTER DENGAN MENERAPKAN ALGORITMA *NAÏVE BAYES CLASSIFIER*

Oleh

HENDY SYUHADA

Komisi Pemberantasan Korupsi (KPK) merupakan lembaga resmi di Indonesia yang ditugaskan secara khusus untuk menangani kasus korupsi. Seiring dengan maraknya kasus korupsi di Indonesia, publik turut menyampaikan opini terhadap kinerja KPK yang disampaikan salah satunya melalui Twitter. Namun opini tersebut masih samar bernada positif ataupun negatif. Oleh karena itu, pada penelitian ini dilakukan analisis sentimen terhadap opini tersebut dengan menggunakan algoritma *naïve bayes classifier* berbasis *machine learning*.

Data bersumber dari Twitter yang diambil dengan teknik *crawling* melalui API (*Application Programming Interface*). Data tersebut diolah melalui beberapa tahapan yaitu *preprocessing* yang meliputi menghilangkan tanda baca, menghapus kata yang berulang dan kata yang sering muncul tetapi tidak terlalu memiliki makna dalam kalimat. Tahap selanjutnya adalah *labeling* data yang dilakukan secara manual dengan memberikan label atau *class* pada data tersebut. Berikutnya adalah proses *modeling* yaitu proses untuk membangun sebuah model yang tepat untuk memprediksi probabilitas data yang akan masuk serta mengelompokkannya sesuai dengan perhitungan probabilitas sebelumnya. Data yang digunakan dalam proses *modeling* yaitu sebanyak 2055 data *tweet* yang dibagi menjadi *training set* dan *testing set* dengan perbandingan 80:20. Selanjutnya dilakukan *deployment system* dengan model yang dipilih untuk menganalisis sentimen terhadap KPK di Twitter.

Hasil dari penelitian ini menunjukkan dengan menggunakan *Naïve Bayes Classifier* model *multinomial* didapatkan nilai *precision* 0.69, *recall* 0.89, *F-1 Score* 0.74, dan akurasi sebesar 64%. Pada penelitian ini juga dikembangkan sebuah *website* untuk mengambil data baru yang kemudian secara otomatis mengklasifikasikannya ke dalam label positif, negatif, atau netral. *Website* ini juga menampilkan hasil dalam bentuk tabel dan grafik.

Kata kunci: Analisis Sentimen, KPK, *Multinomial model*, *Naïve Bayes Classifier*, Twitter

ABSTRACT

SENTIMENT ANALYSIS OF THE KOMISI PEMBERANTASAN KORUPSI (KPK) ON SOCIAL MEDIA TWITTER BY APPLYING THE ALGORITHM NAÏVE BAYES CLASSIFIER

By

HENDY SYUHADA

Komisi Pemberantasan Korupsi (KPK) is an official institution in Indonesia specifically assigned to handle corruption cases. Along with the rise of corruption cases in Indonesia, the public also expressed opinions on the performance of the KPK, which was conveyed through Twitter. However, this opinion is still vaguely positive or negative. Therefore, in this study, sentiment analysis was carried out on this opinion using the machine learning-based naïve bayes classifier algorithm.

Data comes from Twitter which is taken by crawling technique through API (Application Programming Interface). The data is processed through several stages, namely preprocessing which includes removing punctuation marks, removing repetitive words and words that often appear but do not really have meaning in sentences. The next stage is data labeling which is done manually by assigning a label or class to the data. Next is the modeling process, which is the process of building an appropriate model to predict the probability of incoming data and classifying them according to the previous probability calculations. The data used in the modeling process is 2055 tweet data which is divided into training sets and testing sets with a ratio of 80:20. Next, a system deployment with the chosen model was carried out to analyze sentiment towards the KPK on Twitter.

The results of this study indicate that using the multinomial Naïve Bayes Classifier model, the precision value is 0.69, the recall is 0.89, the F-1 Score is 0.74, and the accuracy is 64%. In this study, a website was also developed to retrieve new data which then automatically classified it into positive, or neutral labels. This website also displays the results in the form of tables and graphs.

Keywords: Sentiment Analysis, KPK, Multinomial model, Naïve Bayes Classifier, Twitter

Judul Skripsi	:	**ANALISIS SENTIMEN MENGENAI KOMISI PEMBERANTASAN KORUPSI (KPK) PADA MEDIA SOSIAL TWITTER DENGAN MENERAPKAN ALGORITMA *NAÏVE BAYES CLASSIFIER***
Nama Mahasiswa	:	*Hendy Syuhada*
Nomor Pokok Mahasiswa	:	**1715061018**
Program Studi	:	**Teknik Informatika**
Jurusan	:	**Teknik Elektro**
Fakultas	:	**Teknik**

MENYETUJUI

1. Komisi Pembimbing

Ir. Gigih Forda Nama, S.T., M.T.I., IPM.
NIP 19830712 200812 1 003

Titin Yulianti, S.T., M.Eng.
NIP 19880709 201903 2 015

2. Mengetahui

Ketua Jurusan
Teknik Elektro

Herlinawati, S.T., M.T.
NIP 19710314 199903 2 001

Ketua Program Studi
Teknik Informatika

Mona Arif Muda, S.T., M.T.
NIP 19711112 200003 1 002

MENGESAHKAN

1. Tim Penguji

Ketua : **Ir. Gigih Forda Nama, S.T., M.T.I., IPM.**

Sekretaris : **Titin Yulianti, S.T., M.Eng.**

Penguji : **Yessi Mulyani, S.T., M.T.**

2. Dekan Fakultas Teknik

Dr. Eng. Ir. Helmy Fitriawan, S.T., M.Sc.
NIP 19750928 200112 1 002

Tanggal Lulus Ujian Skripsi : **30 September 2022**

SURAT PERNYATAAN

Saya yang bertandatangan dibawah ini, menyatakan bahwa skripsi saya yang berjudul "Analisis Sentimen Mengenai Komisi Pemberantasan Korupsi (KPK) Pada Media Sosial Twitter Dengan Menerapkan Algoritma *Naive Bayes Classifier*" dengan ini menyatakan bahwa skripsi saya dibuat oleh saya sendiri. Semua hasil yang tertuang dalam skripsi ini telah mengikuti kaidah penulisan karya ilmiah Universitas Lampung.

Apabila di kemudian hari terbukti bahwa skripsi ini merupakan Salinan atau dibuat oleh orang lain, maka saya bersedia menerima sanksi sesuai dengan ketentuan hukum atau akademik yang berlaku.

Bandar Lampung, 30 September 2022

Pembuat Pernyataan,

Hendy Syuhada

NPM. 1715061018

RIWAYAT HIDUP

Penulis dilahirkan di Kota Metro pada tanggal 26 Februari 1999. Penulis merupakan anak pertama dari dua bersaudara, dari pasangan Bapak Edi Hariyanto dan Ibu Eliya Mustafa.

Penulis memulai jenjang pendidikan di TK Aisyiyah Kalirejo pada tahun 2004, dilanjutkan ke SD Negeri 1 Kalirejo pada tahun 2005 dan lulus pada tahun 2011, kemudian penulis melanjutkan ke SMP Negeri 1 Kalirejo dan lulus pada tahun 2014, selanjutnya penulis menyelesaikan pendidikan tingkat SMA di MA Negeri 1 Bandar Lampung pada tahun 2014 dan lulus pada tahun 2017 dan ditahun yang sama penulis diterima sebagai mahasiswa pada Program Studi Teknik Informatika Universitas Lampung melalui jalur SBMPTN.

Selama menjadi mahasiswa, penulis aktif di Organisasi Himpunan Mahasiswa Teknik Elektro (HIMATRO) sebagai anggota Divisi Minat dan Bakat pada Periode 2018 dan menjadi anggota Divisi Kewirausahaan pada Periode 2019. Selain itu, penulis juga pernah bergabung dengan Koperasi Mahasiswa Universitas Lampung (KOPMA UNILA) pada tahun 2019 sebagai anggota bidang Pengembangan Sumber Daya Anggota. Ditengah proses perkuliahan, pada bulan Juli hingga Agustus tahun 2020 penulis melaksanakan Praktek Kerja Lapangan di Dinas Komunikasi, Informatika, dan Statistik (DISKOMINFOTIK) Provinsi Lampung yang bertempat di Kota Bandar Lampung dan masuk dalam bidang atau divisi Pengelolaan dan Layanan Informasi Publik. Kemudian pada bulan September 2021 hingga September 2022 penulis berhasil menyelesaikan penelitian Tugas Akhir dengan judul penelitian **"Analisis Sentimen Mengenai Komisi Pemberantasan Korupsi (KPK) Pada Media Sosial Twitter Dengan Menerapkan Algoritma *Naive Bayes Classifier*".**

Dengan menyebut nama Allah Yang Maha Pengasih lagi Maha Penyayang. *Alhamdulillahirobbilalamin* dengan segala rahmat dan karunia-Nya, skripsi ini dapat diselesaikan dengan baik dan lancar. Mudah-mudahan beserta keberhasilan yang telah digapai ini, saya mampu menuju masa depan yang lebih baik lagi serta mampu meraih cita-cita yang diharapkan serta selalu berada di jalan-Mu. Aamiin.

KUPERSEMBAHKAN SKRIPSI INI TERUNTUK:

"Kedua orang tua terkasih, Ibu Eliya Mustafa dan Bapak Edi Hariyanto atas kasih sayang, doa dan dukungannya yang senantiasa diberikan kepada saya sehingga skripsi ini mampu terselesaikan dengan baik. Tak lupa terima kasih pula atas kebahagiaan-kebahagian yang diberikan selama ini. Semoga ini menjadi tahap awal dalam meraih kesuksesan sehingga dapat membahagiakan Ibu dan Bapak"

"Adikku satu-satunya Khalisa Sahda yang selalu menjadi alasanku untuk pulang ke rumah. Terima kasih karena selalu menemani ketika berada di rumah. Mudah-mudahan dirimu kelak akan menjadi pribadi yang lebih sukses dari abangmu ini"

"Diri ini. Maaf karena telah merusak jam tidurmu selama ini, maaf sudah memaksamu melakukan hal-hal diluar batas kemampuanmu, maaf telah mengkritikmu bahkan menyalahkanmu atas segala kesalahan, dan maaf karena selalu saja lupa mengapresiasi pada tiap keberhasilan yang telah dicapai. Terima kasih telah berjuang dan menemani hingga detik ini. Kamu Luar Biasa!"

"Rekan-rekan Teknik Informatika 2017 yang telah mendampingi serta menolong saya selama menempuh pendidikan di kampus tercinta Universitas Lampung ini. Terima kasih atas segala kenangan indah yang diberikan selama masa perkuliahan ini. Terima kasih atas begitu banyaknya pembelajaran dan pengalaman hidup yang kalian berikan mulai dari awal bersama hingga akhir saat ini. Semoga kelak kita kedepannya akan menjadi orang yang sukses dijalannya masing-masing."

“…dan bersabarlah. Sungguh, Allah beserta orang-orang yang sabar.”

(Q.S. Al-Anfal Ayat 46)

“Allah tidak membebani seseorang melainkan sesuai dengan kesanggupannya…”

(Q.S. Al-Baqarah Ayat 286)

“Tidak masalah jika kamu berjalan dengan lambat, asalkan kamu tidak pernah berhenti berusaha.”

(Confucius)

“Melihat ke atas sebagai motivasi bukan menjadi rendah diri, melihat ke bawah agar lebih bersyukur bukan untuk menjadi sombong”

“The One Piece Is Real”

(Edward Newgate)

SANWACANA

Puji syukur penulis panjatkan kepada Allah سُبْحَانَهُ وَ تَعَالَى, yang telah memberikan karunia serta ridho-Nya sehingga penulis dapat melaksanakan dan menyelesaikan penelitian ini yang berjudul "Analisis Sentimen Mengenai Komisi Pemberantasan Korupsi (KPK) Pada Media Sosial Twitter Dengan Menerapkan Algoritma *Naive Bayes Classifier*". Penelitian ini merupakan salah satu syarat untuk menyelesaikan kurikulum mata kuliah penelitian skripsi pada Program Studi Teknik Informatika, Jurusan Teknik Elektro, Fakultas Teknik, Universitas Lampung.

Pada penelitian ini penulis mendapatkan bantuan, bimbingan serta pengarahan dari berbagai pihak. Maka dari itu, penulis mengucapkan terima kasih sebanyak-banyaknya kepada:

1. Allah سُبْحَانَهُ وَ تَعَالَى yang senantiasa memberikan kemudahan dan kelancaran kepada penulis serta Nabi Muhammad صَلَّى اللهُ عَلَيْهِ وَسَلَّمَ yang telah menjadi suri tauladan selama penelitian berlangsung;
2. Ibu dan Bapak serta keluarga penulis yang selalu memberikan motivasi dan dukungan kepada penulis;
3. Bapak Dr. Eng. Ir. Helmy Fitriawan, S.T., M.Sc. selaku Dekan Fakultas Teknik Universitas Lampung;
4. Ibu Herlinawati, S.T., M.T. selaku Ketua Jurusan Teknik Elektro Universitas Lampung;
5. Bapak Mona Arif Muda, S.T., M.T. selaku Ketua Program Studi Teknik Informatika Universitas Lampung;
6. Bapak Ir. Gigih Forda Nama, S.T., M.T.I., IPM selaku Pembimbing Utama penelitian yang selalu meluangkan waktunya untuk memberikan bimbingan dan dukungan;

7. Ibu Titin Yulianti, S.T., M.Eng. selaku Pembimbing Pendamping penelitian yang selalu memberikan motivasi dan memberikan bimbingan kepada penulis untuk menjadi lebih baik;
8. Ibu Yessi Mulyani, S.T., M.T., selaku Penguji penelitian yang telah banyak memberikan saran dan masukan;
9. Bapak Ir. Meizano Ardhi Muhammad, S.T., M.T., selaku Pembimbing Akademik yang telah memberikan bimbingan selama perkuliahan dan selalu memberikan motivasi;
10. Mbak Rika selaku Admin Program Studi Teknik Informatika yang telah memberikan bantuan dalam proses administrasi penelitian;
11. Rekan-rekan KKN (Agung, Anam, Belia, Kris, Resti, Tina) yang sudah memberikan pengalaman yang berharga selama masa KKN, semoga kita semua menjadi orang yang sukses kelak;
12. Teman-teman Program Studi Teknik Informatika 2017 atas kebersamaan dan kerjasamanya selama ini;
13. Semua pihak yang turut serta dalam membantu menyelesaikan penelitian dan tidak mungkin penulis sebutkan satu persatu.

Penulis menyadari bahwa dalam penulisan laporan penelitian ini masih bisa disempurnakan kembali. Oleh karena itu, penulis mengharapkan saran dan kritik yang bersifat membangun dari para pembaca. Penulis berharap laporan skripsi ini dapat bermanfaat bagi banyak pihak.

Bandar Lampung, 30 September 2022

Penulis,

Hendy Syuhada

DAFTAR ISI

DAFTAR TABEL

Halaman

DAFTAR GAMBAR

Halaman

I. PENDAHULUAN

1.1 Latar Belakang

Korupsi dikategorikan sebagai salah satu kejahatan luar biasa (*extra ordinary crime*) dikarenakan korupsi menyebabkan kerugian proses demokrasi serta hak-hak sosial dan ekonomi masyarakat luas. Dalam perkembangannya, korupsi di Indonesia telah terjadi secara sistematis, meluas, serta terjadi di mana-mana, baik di lembaga pemerintahan maupun non pemerintahan. Maka dari itu diperlukan pencegahan dan usaha penanganan yang luar biasa juga. Lembaga yang berwenang melakukan itu adalah Komisi Pemberantasan Korupsi atau KPK yang sejak awal memang dibentuk dengan kewenangan luar biasa agar mampu mengungkap praktik licik dan kotor serta menembus benteng pertahanan koruptor yang paling kuat sekalipun. Komisi Pemberantasan Korupsi adalah lembaga yang dibentuk berdasarkan Undang-Undang Nomor 30 Tahun 2002. Secara harfiah, Komisi Pemberantasan Korupsi adalah lembaga yang bergerak dalam pemberantasan tindak pidana korupsi. Namun berdasarkan Pasal 6 UU No. 30 tahun 2002 tentang Komisi Pemberantasan Tindak Pidana Korupsi, tugas KPK tidak hanya dalam hal pemberantasan saja, tetapi juga melakukan koordinasi dengan instansi yang berwenang melakukan pemberantasan tindak pidana korupsi, melakukan pengawasan, penyelidikan, penyidikan, dan penuntutan terhadap tindak pidana korupsi, dan melakukan monitor terhadap penyelenggaraan pemerintahan Negara [1].

Akhir-akhir ini cukup banyak masyarakat yang memberikan pendapatnya mengenai KPK, contohnya seperti kinerja perorangan maupun lembaganya dianggap sudah tidak sesuai dengan tugasnya yang tercantum dalam Pasal 6 UU No. 30 tahun 2002.

Berita mengenai KPK banyak ditampilkan dalam portal berita online seperti kompas dan detikcom, akan tetapi portal berita tersebut tidak menyediakan API agar dapat mengakses lebih tentang berita yang ada di dalamnya. Maka dari itu, dibutuhkan media lain yang menampung banyak berita mengenai KPK dan memiliki API untuk mengakses berita tersebut. Salah satu media sosial yang banyak digunakan oleh masyarakat serta menyediakan API adalah Twitter. Twitter merupakan media yang digunakan untuk memfasilitasi pengguna dalam berkomunikasi dan mendapatkan informasi dengan berbagai topik. Banyak masyarakat yang memberikan opini mereka terhadap KPK melalui Twitter, akan tetapi opini tersebut masih samar apakah bernada positif ataupun negatif. Berdasarkan masalah tersebut, maka perlu dilakukan analisis sentimen terhadap opini masyarakat agar diketahui opini tersebut bernada positif atau negatif sehingga pada akhirnya kinerja KPK dapat dinilai baik atau buruk.

1.2 Rumusan Masalah

Rumusan masalah pada penelitian ini adalah bagaimana membangun sebuah sistem yang melakukan analisis terhadap sentimen masyarakat di Twitter berkenaan dengan kinerja Komisi Pemberantasan Korupsi.

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut :

1. Membuat sebuah sistem analisis sentimen berdasarkan *tweet* masyarakat di media sosial Twitter mengenai kinerja dari Komisi Pemberantasan Korupsi.
2. Melakukan analisis terhadap kinerja metode *Naïve Bayes* dalam mengklasifikasikan *tweet* terkait Komisi Pemberantasan Korupsi.
3. Membuat *website* untuk visualisasi data sentimen analisis terkait Komisi Pemberantasan Korupsi.

1.4 Manfaat Penelitian

Penelitian ini diharapkan memberikan manfaat untuk mengetahui sentimen berdasarkan *tweet* masyarakat di media sosial Twitter berkenaan dengan kinerja Komisi Pemberantasan Korupsi.

1.5 Batasan Masalah

Penelitian ini memiliki batasan masalah sebagai berikut :

1. Dataset yang digunakan adalah *tweet* pengguna Twitter mengenai kinerja Komisi Pemberantasan Korupsi.
2. *Website* yang dikembangkan menggunakan server lokal.

1.6 Sistematika Penulisan

Sistematika penulisan yang digunakan dalam penulisan skripsi ini terdiri dari beberapa bab, antara lain:

BAB I : PENDAHULUAN

Bab ini menjelaskan secara umum mengenai latar belakang, rumusan masalah, tujuan dan manfaat penelitian, serta batasan dari penelitian yang akan dilakukan terkait sentimen masyarakat terhadap KPK pada media sosial Twitter.

BAB II : TINJAUAN PUSTAKA

Bab ini berisi teori-teori yang mendasari penelitian ini serta berbagai referensi dari penelitian-penelitian sebelumnya yang berfungsi sebagai sumber dalam memahami permasalahan mengenai analisis sentimen.

BAB III : METODOLOGI PENELITIAN

Bab ini membahas mengenai metode penelitian yang digunakan dalam

menganalisis sentimen masyarakat pada media sosial Twitter.

BAB IV : PEMBAHASAN

Bab ini berisi tentang pembahasan serta hasil yang diperoleh dalam penelitian secara keseluruhan.

BAB V : SIMPULAN DAN SARAN

Bab ini berisi kesimpulan dari hasil penelitian dan saran yang digunakan sebagai masukan untuk penelitian lebih lanjut di masa mendatang.

DAFTAR PUSTAKA

LAMPIRAN

II. TINJAUAN PUSTAKA

2.1 Penelitian Terkait

Terdapat beberapa penelitian terkait yang dijadikan sebagai perbandingan dan rujukan mengenai metode yang digunakan pada penelitian ini.

Penelitian oleh Alec Go et al. dari *Stanford University* bertujuan untuk mengklasifikasikan sentimen pesan Twitter sebagai pesan positif dan pesan negatif. Sistem ini akan berguna bagi konsumen yang ingin mencari sentimen dari produk sebelum membeli atau bagi perusahaan yang ingin memantau sentimen publik dari produk yang dimiliki oleh perusahaan tersebut. Penelitian ini menggunakan tiga algoritma dari *machine learning*, yaitu *naïve bayes*, *maximum entropy* dan *support vector machine*. Hasil dari penelitian ini memiliki akurasi diatas 80% ketika dilatih menggunakan *emoticon data* untuk ketiga algoritma tersebut [2]. Penelitian ini membahas tentang sentimen pesan Twitter berkenaan tentang produk yang dimiliki oleh perusahaan yang berguna bagi konsumen sebelum membeli atau sebagai bahan evaluasi bagi perusahaan tersebut. Pada penelitian yang akan dilakukan ini juga membahas sentimen pesan Twitter, akan tetapi topik yang dibahas adalah mengenai Komisi Pemberantasan Korupsi. Serta dibuat juga sebuah *website* untuk visualisasi dataset yang dapat melakukan penarikan data secara *real time* dan data baru tersebut sudah diklasifikasikan berdasarkan *class* atau labelnya, yaitu positif, netral, dan negatif.

Pada penelitian tahun 2018 dilakukan sebuah survei opini publik terhadap suatu peristiwa, objek, ataupun sebuah lokasi. Kemudian sentimen tersebut dapat digunakan wisatawan lain untuk memutuskan pergi ke tempat itu atau tidak [3].

Penelitian ini dilakukan untuk mengetahui tentang sentimen atau opini mengenai suatu peristiwa, objek, ataupun sebuah lokasi agar berguna bagi wisatawan lain. Sedangkan pada penelitian yang akan dilakukan adalah membahas sentimen pesan Twitter mengenai Komisi Pemberantasan Korupsi. Serta dibuat juga sebuah *website* untuk visualisasi dataset yang dapat melakukan penarikan data secara *real time* dan data baru tersebut sudah di klasifikasikan berdasarkan *class* atau labelnya, yaitu positif, netral, dan negatif.

Pada tahun 2009 dilakukan penelitian untuk mencari perbandingan algoritma dalam data mining oleh Daniela Xhemali et al. dari *Loughborough University.* Penelitian ini berfokus pada perbandingan *naïve bayes, decision tree,* dan *neural network* untuk menganalisis otomatis dan mengklasifikasi data pada *web* kursus pelatihan. Penelitian ini menggunakan tiga tahapan, yaitu *crawling* data, mencari *data training* dan mengklasifikasi data untuk perbandingannya. Hasil dari penelitian ini membuktikan bahwa Algoritma *naïve bayes* lebih mengungguli dalam sistem mengklasifikasi data [4]. Penelitian ini dilakukan untuk mencari perbandingan algoritma *naïve bayes, decision tree,* dan *neural network* untuk menganalisis otomatis dan mengklasifikasi data pada *web* kursus pelatihan. Sedangkan pada penelitian yang akan dilakukan hanya menggunakan algoritma *naïve bayes* untuk mengetahui sentimen masyarakat mengenai Komisi Pemberantasan Korupsi.

Pada penelitian lainnya dilakukan pemodelan dan peramalan *time series* dengan Python dan Jupyter Notebook sebagai software *open source* untuk menjalankan sistem pada saham PT. Bank Negara Indonesia pada tahun 2020. Dalam makalah ini diselidiki pula faktor seasonal atau musiman, sehingga pada identifikasi model digunakan analisis *time series Seasonal Autoregressive Integrated Moving Average* (SARIMA) berbantuan software Jupyter Notebook berbahasa Python. Dalam analisis ini Python membantu untuk melakukan perhitungan dan visualisasi data agar lebih mudah dan efisien. Pada hasil peramalan didapatkan kesimpulan bahwa harga saham PT BNI dalam 3 tahun kedepan memiliki tren naik. Dalam peramalan ini terdapat kemungkinan bahwa harga saham tidak hanya dipengaruhi oleh waktu tetapi juga dapat dipengaruhi oleh faktor lainnya [5]. Pada penelitian ini

menggunakan pemodelan dan peramalan *Time Series* terhadap saham PT. Bank Negara Indonesia dengan menggunakan Python dan Jupyter Notebook sebagai software *open source* untuk menjalankan sistem. Sedangkan pada penelitian yang akan dilakukan ini menggunakan *naïve bayes* untuk mengetahui sentimen masyarakat mengenai Komisi Pemberantasan Korupsi menggunakan Python dan Jupyter Notebook sebagai software *open source* untuk menjalankan sistemnya.

Pada tahun 2019 lalu, dilakukan juga sebuah analisis sentimen terhadap kinerja Dewan Perwakilan Rakyat (DPR) yang diungkapkan masyarakat melalui media sosial Twitter. Penelitian ini menggunakan metode *naïve bayes classifier* dan menggunakan sebanyak 1546 data *tweet*. Hasil dari penelitian ini didapatkan bahwa DPR mendapatkan 95 *tweet* positif dengan polaritas 0.75 atau 75% sentimen positif, 693 *tweet* netral dengan polaritas 0.79 atau 79% sentimen netral dan 758 *tweet* negatif dengan polaritas 0.82 atau 82% sentimen negatif dengan accuracy score 0.8 atau 80% berdasarkan data *testing* sebanyak 20% [6]. Pada penelitian ini dilakukan analisis sentimen terhadap kinerja Dewan Perwakilan Rakyat (DPR) yang diungkapkan masyarakat melalui media sosial Twitter menggunakan metode *naïve bayes classifier*. Pada penelitian yang akan dilakukan ini juga menggunakan algoritma *naïve bayes* untuk mengetahui sentimen masyarakat mengenai Komisi Pemberantasan Korupsi. Akan tetapi dibuat juga sebuah *website* untuk visualisasi dataset yang dapat melakukan penarikan data secara *real time* dan data baru tersebut sudah di klasifikasikan berdasarkan *class* atau labelnya, yaitu positif, netral, dan negatif.

Analisis *text mining* dari Twitter mengenai infrastruktur di Indonesia dengan metode klasifikasi *naïve bayes* pada tahun 2019 mendapatkan hasil bahwa proporsi sentimen negatif lebih besar dibandingkan dengan sentimen positif. Selain itu, hasil pengklasifikasian dengan menggunakan metode *naïve bayes* diperoleh model yang paling baik pada model bandara dengan akurasi sebesar 82%, presisi sebesar 0,84 dan recall sebesar 0,48 [7]. Pada penelitian ini dilakukan analisis *text mining* dari Twitter mengenai infrastruktur di Indonesia dengan metode klasifikasi *naïve bayes.* Penelitian yang akan dilakukan juga menggunakan algoritma *naïve bayes* untuk

mengetahui sentimen masyarakat mengenai Komisi Pemberantasan Korupsi. Akan tetapi dibuat juga sebuah *website* untuk visualisasi dataset yang dapat melakukan penarikan data secara *real time* dan data baru tersebut sudah di klasifikasikan berdasarkan *class* atau labelnya, yaitu positif, netral, dan negatif.

Algoritma *naïve bayes* juga digunakan dalam penelitian terhadap analisis sentimen opini film pada Twitter menggunakan sekitar 500 data pengujian yang dibagi menjadi 400 data *training* dan 100 data *testing* dengan memakai metode evaluasi *K-Fold Cross Validation.* Pengujian menggunakan data yang sudah dipartisi akan diulang sebanyak 5 kali (k=5) dengan posisi data tes berbeda di setiap iterasinya. Berdasarkan hasil eksperimen tersebut, analisis sentimen yang dapat dilakukan oleh sistem dengan akurasi yang didapat adalah 90 % dengan rincian nilai precision 92%, recall 90% dan f-measure 90% [8]. Pada penelitian ini dilakukan analisis sentimen opini film pada Twitter dengan metode evaluasi *K-Fold Cross Validation.* Sedangkan penelitian yang akan dilakukan, untuk mengetahui sentimen masyarakat mengenai Komisi Pemberantasan Korupsi dengan metode evaluasi *splitting* data.

Pada penelitian lain menggunakan algoritma KNN (*K-Nearest Neighbor*) dalam analisis sentimen pengguna Twitter tentang topik Pilkada DKI 2017. Data *tweet* yang digunakan adalah sebanyak 2000 data *tweet* berbahasa Indonesia yang dikumpulkan selama bulan Januari 2017 menggunakan package Python bernama Twitterscraper. Menggunakan algoritma KNN dengan pembobotan kata TF-IDF dan fungsi Cosine Similarity, dan dilakukan pengklasifikasian nilai sentimen ke dalam dua kelas: positif dan negatif. Hasil pengujian diketahui bahwa nilai akurasi terbesar adalah 67,2% ketika k=5, presisi tertinggi 56,94% ketika k=5, dan recall 78,24% dengan k=15 [9]. Penelitian ini dilakukan menggunakan algoritma KNN *(K-Nearest Neighbor)* untuk analisis sentimen pengguna Twitter tentang topik Pilkada DKI 2017. Sedangkan penelitian yang akan dilakukan menggunakan algoritma *naïve bayes* untuk mengetahui sentimen masyarakat mengenai Komisi Pemberantasan Korupsi. Serta dibuat juga sebuah *website* untuk visualisasi dataset yang dapat melakukan penarikan data secara *real time* dan data baru tersebut sudah di klasifikasikan berdasarkan *class* atau labelnya, yaitu positif, netral, dan negatif.

2.2 Twitter

Twitter adalah media jejaring sosial yang mengizinkan penggunanya untuk mengirim dan membaca pesan berbasis teks hingga 280 karakter yang disebut sebagai kicauan atau *tweet*. Twitter didirikan pada bulan Maret 2006 oleh Jack Dorsey dan diluncurkan pada bulan Juli. Twitter menjadi salah satu dari sepuluh situs yang paling sering dikunjungi di internet. Pada Januari 2013, terdapat lebih dari 500 juta pengguna terdaftar di media sosial Twitter, dan Twitter menangani lebih dari 1,6 miliar permintaan pencarian per hari. Tingginya popularitas Twitter menyebabkan layanan ini telah dimanfaatkan untuk berbagai keperluan dalam berbagai aspek, misalnya sebagai sarana protes, kampanye politik, sarana pembelajaran, dan sebagai media komunikasi darurat. Data teks Twitter yang begitu beragam bentuk dan kandungan isinya, memiliki banyak arti jika diproses lebih lanjut, dalam konteks tersebut maka teknik Data Mining memiliki peran yang signifikan selama data Twitter tersebut bisa diperoleh dalam jumlah besar, ratusan hingga ribuan bahkan jutaan *tweet*. Kelebihan pada media sosial Twitter ini salah satunya adalah menyediakan API (*Application Programming Interface*) yang sangat baik, sehingga memudahkan setiap orang untuk mengambil data dari Twitter. Pengumpulan data dari Twitter dapat digunakan untuk berbagai kebutuhan seperti, mengetahui popularitas kandidat pilkada atau pemilu, mendapat informasi mengenai popularitas suatu produk, mengetahui pendapat masyarakat terhadap topik tertentu atau untuk yang sederhana dapat digunakan untuk melihat semua *mention, retweet* atas suatu akun Twitter tertentu [10].

Twitter API menyediakan akses untuk data *tweet* dari rentang waktu tertentu, dari pengguna tertentu, dengan kata kunci tertentu, atau dari suatu wilayah geografis tertentu, namun tidak memberikan fitur untuk mengekstrak struktur dari *tweet*. API merupakan cara program komputer "berbicara" satu sama lain agar mereka dapat meminta dan menyajikan informasi. Ini dilakukan dengan mengizinkan aplikasi perangkat lunak memanggil apa yang disebut sebagai *endpoint*: alamat yang terkait dengan informasi jenis tertentu yang disediakan (*endpoint* umumnya unik seperti nomor telepon). Twitter mengizinkan akses ke bagian dari layanan melalui API untuk memungkinkan orang-orang membangun perangkat lunak yang terintegrasi

dengan Twitter seperti solusi yang membantu sebuah perusahaan menjawab umpan balik pelanggan di Twitter. Data Twitter berbeda dari data yang dibagi oleh kebanyakan platform sosial lain karena data tersebut mencerminkan informasi yang dipilih pengguna untuk dibagikan ke publik. *Platform* API Twitter menyediakan akses luas ke data Twitter publik yang telah dipilih pengguna untuk dibagikan ke dunia. Twitter juga mendukung API yang memungkinkan pengguna mengelola informasi Twitter pengguna yang non-publik dan memberikan informasi ini ke pengembang yang telah diizinkan pengguna untuk melakukannya [11].

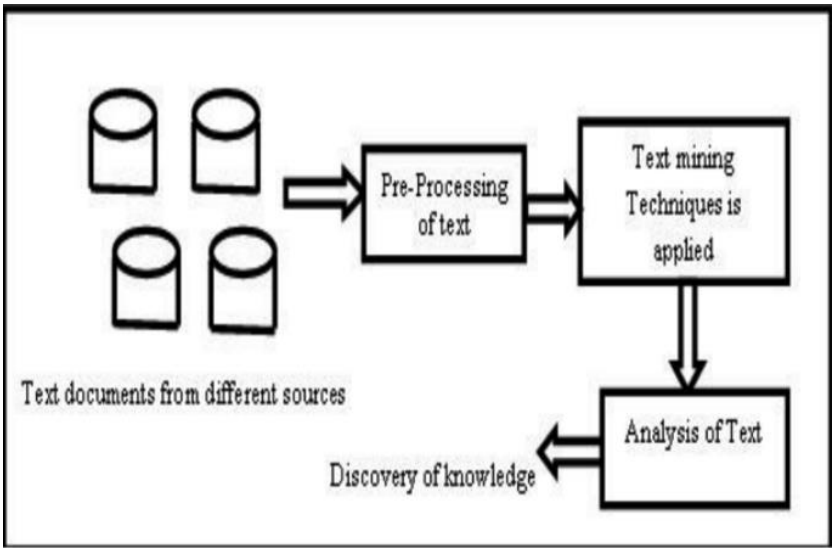
2.3 *Sentiment Analysis*

Sentiment analysis merupakan suatu proses dalam data mining yang digunakan untuk mengidentifikasi dan mengekstrak suatu informasi dari suatu teks yang bertujuan untuk memahami *social sentiment* pada teks tersebut. *Sentiment analysis* juga digunakan untuk mendapatkan informasi mengenai sikap, pendapat dan emosi yang ada pada teks informasi tersebut. Tugas dari *sentiment analysis* adalah mengelompokkan polaritas dari teks yang ada dalam teks, tergolong teks positif atau teks negatif. *Sentiment Analysis* difokuskan untuk riview klasifikasi berdasarkan polaritas. Berdasarkan klasifikasi, analisis sentimen dibagi menjadi dua yaitu klasifikasi subjektivitas dan klasifikasi ke dalam positif atau negatif [12].

2.4 *Text Mining*

Text Mining adalah salah satu penambangan informasi yang berguna dari data-data yang berupa tulisan, dokumen atau *text* dalam bentuk klasifikasi maupun *clustering*. Dengan *Text Mining* maka kita akan melakukan proses mencari atau penggalian informasi yang berguna dari data tekstual. Pada *Text Mining*, pertama diperlukan pengambilan data kemudian data tersebut perlu di *pre-processing* sebelum proses klasifikasi. *Pre-processing* adalah bagian dimana saat data sudah didapat dan selanjutnya diproses informasi yang terkandung di dalamnya. Proses *Text Mining* adalah sama dengan data mining, kecuali, beberapa metode dan data yang dikelola

nya seperti data teks yang tidak terstruktur, terstruktur sebagian maupun terstruktur seperti teks email, teks HTML, maupun teks komentar serta dari berbagai sumber [13].



Gambar 2.1 Proses t*ext mining*

2.5 Python

Python adalah bahasa pemrograman tingkat tinggi yang dibuat oleh Guido van Rossum dan pertama kali dirilis pada tahun 1991. Python merupakan salah satu bahasa pemrograman interpretatif, interaktif, berbasis objek dan bisa dijalankan di semua platform. Program pada Python dijalankan melalui interpreter sehingga program-program yang ada pada Python dapat diuji secara langsung. Python banyak diaplikasikan pada berbagai sistem operasi seperti *Linux, Microsoft Windows, Mac OS, Android, Symbian OS, Amiga, Palm* dan lain sebagainya. Bahasa pemrograman Python mendukung konsep pemrograman berorientasi objek sehingga pada Python terdapat berbagai *library* dan *framework* yang digunakan untuk menganalisis data [14]. Berikut adalah *library* yang digunakan dalam penelitian ini:

1. Tweepy

   Tweepy adalah *library* Python yang berfungsi untuk mengakses API milik

Twitter sehingga dapat mengambil informasi dari Twitter dengan skrip Python. Tweepy digunakan untuk menjembatani Twitter dengan bahasa pemrograman Python dalam penelitian ini. Dengan menggunakan *library* Tweepy, data mengenai KPK yang ada di Twitter dapat diambil untuk bahan penelitian.

2. NLTK (*Natural Language Toolkit)*

Library NLTK pada penelitian ini digunakan pada tahap *tokenization*, yang berfungsi untuk memisahkan kata-kata pada tiap kalimat yang didapat untuk kemudian diolah.

3. Sastrawi

Library Sastrawi pada penelitian ini digunakan dalam tahapan *stop removal* atau *stemming* untuk menghapus kata-kata yang tidak berhubungan dan menyederhanakan kata menjadi kata dasar.

4. *Scikit-learn*

Scikit-learn adalah library machine learning open source berbasis Python yang umumnya digunakan dalam *data science. Library* ini juga memberikan fitur algoritma *naïve bayes* untuk keperluan *data science.*

2.6 *Naïve Bayes*

Naïve bayes adalah salah satu algoritma klasifikasi yang populer dan memiliki performa yang kompetitif dalam proses klasifikasi yang dikemukakan oleh Thomas Bayes. Algoritma *naïve bayes* memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai *teorema bayes*. Keuntungan menggunakan metode *naïve bayes* ini adalah hanya memerlukan jumlah data pelatihan atau data *training* yang kecil untuk menentukan parameter yang dibutuhkan pada proses pengklasifikasian. Metode *naïve bayes classifier* ini memiliki kelebihan dibanding dengan metode yang lainnya yaitu implementasi yang sederhana, bekerja dengan cepat, serta memberikan hasil yang baik. Selain itu, metode *naïve bayes* memiliki tingkat akurasi yang tinggi dengan perhitungan yang sederhana [15]. *Naïve bayes classifier* menunjukkan akurasi dan kecepatan yang tinggi bila diterapkan pada database yang besar. Metode ini sering digunakan dalam menyelesaikan masalah dalam bidang mesin pembelajaran karena metode ini

dikenal memiliki tingkat akurasi yang tinggi dengan perhitungan sederhana [16].

Metode algoritma *naïve bayes* digolongkan menjadi beberapa tipe berdasarkan fungsinya yaitu *Multinomial NB, Bernoulli NB,* dan *Gaussian NB.*

2.6.1. *Multinomial Naïve Bayes*

Multinomial Naïve Bayes merupakan suatu kondisi probabilitas yang dilakukan tanpa memperhitungkan urutan pada kata dan informasi yang telah ada pada dokumen atau kalimat pada umumnya. Dalam algoritma tersebut juga menghitung jumlah kata yang muncul pada dokumen. Model *multinomial naïve bayes* menggunakan rumus seperti yang ditunjukkan pada persamaan (1) berikut:

$$P(c \mid \text{term dok d}) = P(c) \times P(t1 \mid c) \times P(t2 \mid c) \times P(tn \mid c) \qquad (1)$$

Dimana :

P(c|term dok d)	= Probabilitas suatu dokumen dalam kelas c
P(c)	= Probabilitas prior dari kelas c
P(tn|c)	= Probabilitas kata ke-n pada kelas c
tn	= kata ke n pada dokumen

2.6.2. *Bernoulli Naïve Bayes*

Algoritma *Bernoulli Naïve Bayes* mengimplementasikan klasifikasi untuk data yang didistribusikan sesuai dengan distribusi Bernoulli multivariat; yaitu, mungkin terdapat beberapa fitur tetapi masing-masing dianggap sebagai variabel bernilai biner (Bernoulli, boolean). Oleh karena itu, kelas ini membutuhkan sampel untuk direpresentasikan sebagai vektor fitur bernilai biner. Aturan keputusan untuk algoritma *bernoulli naïve bayes* diberikan pada persamaan (2) berikut:

$$P(xi \mid y) = P(i \mid y)\, x\_i + (1 - P(i \mid y))(1 - x\_i\,) \qquad (2)$$

2.6.3. *Gaussian Naïve Bayes*

Untuk fitur bertipe numerik (kontinu), distribusi Gauss biasanya dipilih untuk merepresentasikan probabilitas bersyarat dari fitur kontinu pada sebuah kelas. Distribusi Gauss dikarakteristikkan dengan dua parameter: *mean* (μ) dan standar deviasi (σ^2). Berikut adalah persamaan dari algoritma *gaussian naïve bayes:*

$$P(Xi = xi|Y = yj) = \frac{1}{\sqrt{2\pi\sigma ij}} e^{-\frac{(Xi-\mu ij)^2}{2\sigma 2ij}} \quad (3)$$

Dimana :
P = Peluang
Xi = Atribut ke i
xi = Nilai atribut ke i
Y = Kelas yang dicari
μ = *Mean*, menyatakan rata-rata dari seluruh atribut.
σ = Standar deviasi, menyatakan varian dari seluruh atribut [17].

Naïve bayes classifier menggunakan *prior probability* (yaitu nilai probabilitas yang diyakini benar sebelum melakukan eksperimen) dari setiap label yang merupakan frekuensi masing-masing label pada *training set* dan kontribusi dari masing-masing fitur. Berdasarkan ciri alami dari sebuah model probabilitas, klasifikasi *naïve bayes* bisa dibuat lebih efisien dalam bentuk pembelajaran *supervised* atau terawasi. Dalam beberapa bentuk praktiknya, parameter untuk perhitungan model *naïve bayes* menggunakan metode *maximum likelihood* atau kemiripan tertinggi. Untuk ranah klasifikasinya yang dihitung adalah P(H|X) yaitu peluang bahwa hipotesa benar untuk data sampel X yang diamati, dapat diterapkan pada persamaan (4).

$$P(H|X) = \frac{P(H|X)P(H)}{P(X)} \quad (4)$$

Keterangan:
X = Data sampel dengan label yang belum diketahui
H = Hipotesa bahwa X adalah data dengan label C
P(H|X) = Peluang bahwa hipotesa benar untuk data sampel X yang diamati

P(X|H) = Peluang data sampel X, bila diasumsikan bahwa hipotesa benar.

P(H) = Peluang dari hipotesa H

P(X) = Peluang data sampel yang diamati [8].

Berikut ini adalah contoh perhitungan manual menggunakan metode klasifikasi *naïve bayes.*

Tabel 2.1 Contoh *tweet* yang sudah diberi label

No	**Isi *tweet***	**Label**
1	kamu pintar sekali	Positif
2	ayah sedang mengendarai motor	Netral
3	ibu marah karena roni nakal	Negatif

1. Pembobotan menggunakan tf (*term frequency*).

Tabel 2.2 Pembobotan

No	**Kosa kata**	**Tf (Positif)**	**Tf (Netral)**	**Tf (Negatif)**
1	kamu	1	0	0
2	pintar	1	0	0
3	sekali	1	0	0
4	ayah	0	1	0
5	sedang	0	1	0
6	mengendarai	0	1	0
7	motor	0	1	0
8	ibu	0	0	1
9	marah	0	0	1
10	karena	0	0	1
11	roni	0	0	1
12	nakal	0	0	1
Jumlah *term*		**3**	**4**	**5**

Dari proses pembobotan mengunakan *term frequency* ini didapatkan *count* positif berjumlah 3, *count* netral berjumlah 4, dan *count* negatif berjumlah 5, dengan total kata yaitu sebanyak 12 kata.

2. Menghitung probabilitas *prior*

Terdapat 3 label yaitu positif, netral, dan negatif. Selanjutnya adalah menghitung probabilitas *prior* pada setiap label dengan menggunakan persamaan (5) sebagai berikut:

$$P\ (\text{label}) = \frac{\frac{\text{positif}}{\text{netral}}}{|C|} \quad (5)$$

Berdasarkan persamaan (5) maka didapatkan nilai 0,33 pada probabilitas *prior* untuk setiap labelnya.

- P (positif) $= \frac{\text{fx(positif)}}{|C|} = \frac{1}{3} = 0{,}3333$
- P (netral) $= \frac{\text{fx(netral)}}{|C|} = \frac{1}{3} = 0{,}3333$
- P (negatif) $= \frac{\text{fx(negatif)}}{|C|} = \frac{1}{3} = 0{,}3333$

3. Menghitung probabilitas *likelihood*

Berdasarkan pembobotan kata sebelumnya, didapatkan *term* sebanyak 12 *term* dengan 3 *term* dari label positif, 4 *term* dari label netral, dan 5 *term* dari label negatif. Kemudian dilanjutkan dengan menghitung probabilitas *likelihood* setiap *term* dari seluruh dokumen menggunakan persamaan (6) sebagai berikut:

$$P\ (w \mid \text{label}) = \frac{\text{nk(label)} + 1}{term(\text{label}) + \text{total } term} \quad (6)$$

1) Probabilitas kata “kamu”

- P (kamu | positif) = $\frac{1+1}{3+12} = \frac{2}{15}$ = 0,1333
- P (kamu | netral) = $\frac{0+1}{4+12} = \frac{1}{16}$ = 0,0625
- P (kamu | negatif) = $\frac{0+1}{5+12} = \frac{1}{17}$ = 0,0588

2) Probabilitas kata “pintar”

- P (pintar | positif) = $\frac{1+1}{3+12} = \frac{2}{15}$ = 0,1333
- P (pintar | netral) = $\frac{0+1}{4+12} = \frac{1}{16}$ = 0,0625
- P (pintar | negatif) = $\frac{0+1}{5+12} = \frac{1}{17}$ = 0,0588

3) Probabilitas kata “sekali”

- P (sekali | positif) = $\frac{1+1}{3+12} = \frac{2}{15}$ = 0,1333
- P (sekali | netral) = $\frac{0+1}{4+12} = \frac{1}{16}$ = 0,0625
- P (sekali | negatif) = $\frac{0+1}{5+12} = \frac{1}{17}$ = 0,0588

4) Probabilitas kata “ayah”

- P (ayah | positif) = $\frac{0+1}{3+12} = \frac{1}{15}$ = 0,066
- P (ayah | netral) = $\frac{1+1}{4+12} = \frac{2}{16}$ = 0,125
- P (ayah | negatif) = $\frac{0+1}{5+12} = \frac{1}{17}$ = 0,058

5) Probabilitas kata “sedang”

- P (sedang | positif) = $\frac{0+1}{3+12} = \frac{1}{15}$ = 0,067
- P (sedang | netral) = $\frac{1+1}{4+12} = \frac{2}{16}$ = 0,125
- P (sedang | negatif) = $\frac{0+1}{5+12} = \frac{1}{17}$ = 0,058

6) Probabilitas kata “mengendarai”

- P (mengendarai | positif) = $\frac{0+1}{3+12} = \frac{1}{15}$ = 0,067

- P (mengendarai | netral) = $\frac{1+1}{4+12} = \frac{2}{16} = 0{,}125$
- P (mengendarai | negatif) = $\frac{0+1}{5+12} = \frac{1}{17} = 0{,}058$

7) Probabilitas kata “motor”

- P (motor | positif) = $\frac{0+1}{3+12} = \frac{1}{15} = 0{,}067$
- P (motor | netral) = $\frac{1+1}{4+12} = \frac{2}{16} = 0{,}125$
- P (motor | negatif) = $\frac{0+1}{5+12} = \frac{1}{17} = 0{,}058$

8) Probabilitas kata “ibu”

- P (ibu | positif) = $\frac{0+1}{3+12} = \frac{1}{15} = 0{,}0667$
- P (ibu | netral) = $\frac{0+1}{4+12} = \frac{1}{16} = 0{,}0625$
- P (ibu | negatif) = $\frac{1+1}{5+12} = \frac{2}{17} = 0{,}1176$

9) Probabilitas kata “marah”

- P (marah | positif) = $\frac{0+1}{3+12} = \frac{1}{15} = 0{,}0667$
- P (marah | netral) = $\frac{0+1}{4+12} = \frac{1}{16} = 0{,}0625$
- P (marah | negatif) = $\frac{1+1}{5+12} = \frac{2}{17} = 0{,}1176$

10) Probabilitas kata “karena”

- P (karena | positif) = $\frac{0+1}{3+12} = \frac{1}{15} = 0{,}0667$
- P (karena | netral) = $\frac{0+1}{4+12} = \frac{1}{16} = 0{,}0625$
- P (karena | negatif) = $\frac{1+1}{5+12} = \frac{2}{17} = 0{,}1176$

11) Probabilitas kata “roni”

- P (roni | positif) = $\frac{0+1}{3+12} = \frac{1}{15} = 0{,}0667$
- P (roni | netral) = $\frac{0+1}{4+12} = \frac{1}{16} = 0{,}0625$
- P (roni | negatif) = $\frac{1+1}{5+12} = \frac{2}{17} = 0{,}1176$

12) Probabilitas kata “nakal”

- P (nakal | positif) = $\frac{0+1}{3+12} = \frac{1}{15}$ = 0,0667
- P (nakal | netral) = $\frac{0+1}{4+12} = \frac{1}{16}$ = 0,0625
- P (nakal | negatif) = $\frac{1+1}{5+12} = \frac{2}{17}$ = 0,1176

4. Pengujian data baru

Setelah sistem mempelajari seluruh data yang ada, selanjutnya adalah menguji sistem yang telah dibuat untuk mengetahui performa dari sistem tersebut.

Tabel 2.3 Data *tweet* baru yang belum diketahui labelnya

No	**Isi *tweet***	**Label**
1	pintar sekali anakmu	?
2	bima sangat nakal hingga kakak marah	?

Proses klasifikasi pada pengujian data baru adalah dengan mengalikan semua nilai peluang, seperti yang ditunjukkan pada persamaan (7). Nilai yang paling tinggi ialah label dari data tersebut.

$$P(\text{uji} \mid \text{label}) = P(\text{label}) \times P(w \mid \text{label}) \qquad (7)$$

1) Pada data baru “pintar sekali anakmu”, yang termasuk ke dalam data *training* adalah kata “pintar” dan “sekali”.

- **P (uji | positif)**
  = P (positif) × P (pintar | positif) × P (sekali | positif)
  = 0,3333 × 0,1333 × 0,1333
  = **0,0059**
- **P (uji | netral)**
  = P (netral) × P (pintar | netral) × P (sekali | netral)

= 0,3333 × 0,0625 × 0,0625

= **0,0013**

- **P (uji | negatif)**

  = P (negatif) × P (pintar | negatif) × P (sekali | negatif)

  = 0,3333 × 0,058 × 0,0588

  = **0,0011**

Nilai probabilitas tertinggi terdapat pada P (uji | positif) yaitu sebesar 0,0059. Jadi data *tweet* baru "pintar sekali anakmu" dapat diklasifikasikan ke dalam kelas atau label positif.

2) Pada data baru "bima sangat nakal hingga kakak marah", yang termasuk ke dalam data *training* adalah kata "nakal" dan "marah".

- **P (uji | positif)**

  = P (positif) × P (nakal | positif) × P (marah | positif)

  = 0,3333 × 0,0667 × 0,0667

  = **0,0014**

- **P (uji | netral)**

  = P (netral) × P (nakal | netral) × P (marah | netral)

  = 0,3333 × 0,0625 × 0,0625

  = **0,0013**

- **P (uji | negatif)**

  = P (negatif) × P (nakal | negatif) × P (marah | negatif)

  = 0,3333 × 0,1176 × 0,1176

  = **0,0046**

Nilai probabilitas tertinggi terdapat pada P (uji | negatif) yaitu sebesar 0,0046. Jadi data *tweet* baru "bima sangat nakal hingga kakak marah" dapat diklasifikasikan ke dalam kelas atau label negatif.

Pengujian performa algoritma *naïve bayes classifier* dilakukan dengan menguji tingkat akurasi, presisi dan recall. Dalam melakukan penghitungan tingkatakurasi algoritma *naïve bayes classifier* dilakukan dengan persamaan (8) sebagai berikut:

$$Akurasi = \frac{\sum data\ benar}{n\ dokumen} x\ 100\ \% \quad (8)$$

Rumus untuk melakukan penghitungan tingkat presisi dapat dilihat pada persamaan (9) sebagai berikut:

$$Presisi = \frac{\sum data\ positif\ atau\ negatif}{n\ dokumen\ positif\ atau\ negatif} x\ 100\ \% \quad (9)$$

Rumus untuk melakukan penghitungan *recall* dapat dilihat pada persamaan (10) sebagai berikut:

$$recall = \frac{\sum doc\ relevan\ dan\ terambil}{\sum seluruh\ dokumen\ relevan}\ x\ 100\% \quad (10)$$

Proses perhitungan prioritas bantuan dilakukan untuk menyamakan data dan melihat data mana yang lebih tinggi dalam suatu dataset sehingga dapat ditampilkan prioritas dari bantuan yang dibutuhkan.

Perhitungan prioritas bantuan per kategori dapat dilihat pada persamaan (11) sebagai berikut:

$$Jumlah = \frac{\sum seluruh\ dataset}{\sum dataset\ perkategori} \quad (11)$$

Rumus untuk menghitung rata-rata dari dataset per kategori dapat dilihat pada persamaan (12) sebagai berikut :

$$JumlahKategori = \frac{\sum dataset\ perkategori}{\sum dataset\ positif\ atau\ negatif} \quad (12)$$

Selanjutnya menghitung keseluruhan dari hasil rata-rata perhitungan dengan persamaan (13) sebagai berikut [18] :

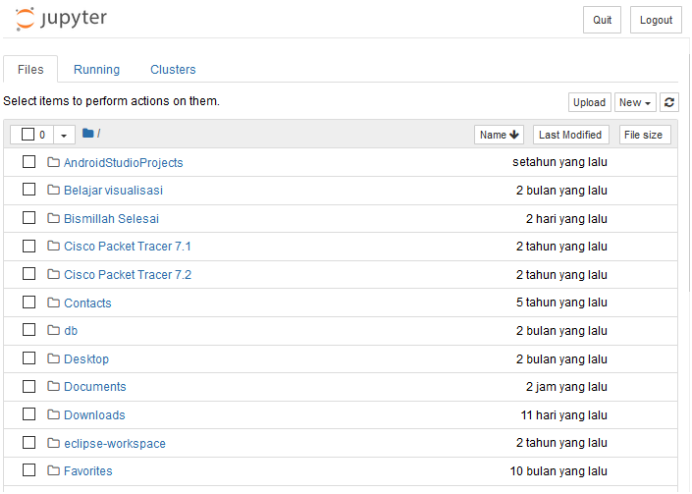
$$Hasil = Jumlah\ x\ JumlahKategori \quad (13)$$

2.7 Jupyter Notebook

Jupyter Notebook adalah aplikasi klien server yang memungkinkan pengeditan dan menjalankan dokumen notebook melalui *browser web*. Jupyter Notebook dapat dijalankan pada desktop lokal yang tidak memerlukan akses internet atau dapat diinstal pada server jarak jauh dan diakses melalui internet. Selain menampilkan, mengedit, dan menjalankan dokumen notebook, Jupyter Notebook App memiliki Dashboard Notebook yaitu sebuah panel kontrol yang menampilkan file lokal dan memungkinkan untuk membuka dokumen notebook atau mematikan kernelnya [19]. *Jupyter* mendukung beberapa komputasi yang dapat digunakan :

1. Jupyter Notebook

Jupyter Notebook digunakan untuk membuat dokumen Notebook Jupyter. Dokumen Notebook Jupyter disimpan dengan format ".ipynb".



Gambar 2.2 Jupyter Notebook

2. JupyterHub

JupyterHub merupakan server *multi-user* yang digunakan untuk menjalankan Jupyter Notebook. JupyterHub dirancangan untuk memperlancar pengguna dalam menggunakan Jupyter Notebook.

3. JupyterLab

JupyterLab adalah antarmuka pengguna yang baru untuk *Project Jupyter*. *Software* memiliki struktur modular, di mana dapat membuka beberapa buku catatan atau file (HTML, Markdown dll) sebagai tab di jendela yang sama.

2.8 PostgreSQL

PostgreSQL adalah adalah sistem manajemen basis data *open source* kelas perusahaan tercanggih di dunia yang dikembangkan oleh *PostgreSQL Global Development Group*. Ini adalah sistem database SQL (*Structured Query Language*) objek-relasional yang kuat dan sangat dapat dikembangkan serta populer karena keandalannya, ketahanan fiturnya, dan kinerjanya yang tinggi. PostgreSQL menggunakan model *client-server* di mana klien dan server dapat berada di host yang berbeda dalam lingkungan jaringan. PostgreSQL mendukung beberapa tipe data termasuk primitif (seperti string, integer, numerik, dan boolean), terstruktur (seperti *time*, array, rentang, dan UUID), dokumen (JSON, JSONB, XML), dan geometri (titik, garis, lingkaran, dan poligon). Ini mendukung integritas data menggunakan fitur-fitur seperti UNIQUE dan NOT NULL [20].

2.9 *Website*

Website merupakan suatu cara yang cukup efektif dan efisien untuk publikasi suatu produk dari perusahaan. Cara ini adalah yang paling menguntungkan jika dibandingkan melalui media massa seperti koran, majalah, tv dan radio yang membutuhkan investasi besar. Web merupakan sumber daya Internet yang sangat populer dan dapat digunakan untuk memperoleh informasi atau melakukan transaksi pembelian barang atau jasa. Web juga merupakan sistem pengiriman dokumenter besar yang berjalan di Internet [21].

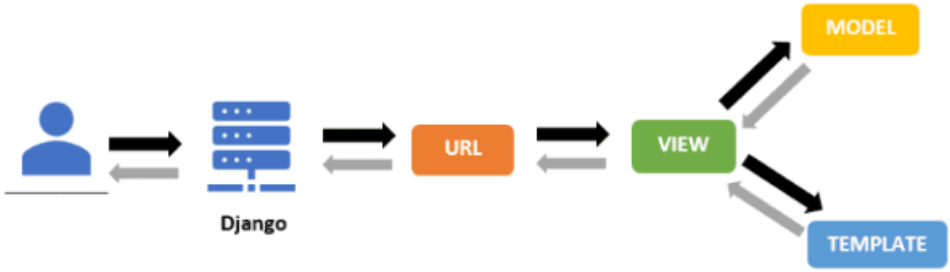
2.10 *Framework* Django

Django adalah kerangka kerja berbasis python yang populer, dan digunakan untuk pengembangan web. Ini adalah kerangka kerja web tingkat tinggi yang memungkinkan pembuatan situs web yang aman dan cepat. Django adalah kerangka kerja sumber terbuka dan gratis, yang berarti bebas untuk digunakan. Django juga mengikuti arsitektur Model-View-Controller (MVC), yang sekarang

menjadi standar untuk pengembangan aplikasi web.

Django mengikuti konvensinya sendiri dari arsitektur Model-View-Controller (MVC) bernama Model View Template (MVT). MVT adalah pola desain perangkat lunak yang terutama terdiri dari 3 komponen Model, View, dan Template.

- Model dalam arsitektur MVT adalah lapisan akses data yang digunakan untuk menangani data. Model menghubungkan seluruh arsitektur ke database. Setiap model ditautkan ke satu tabel *database* menggunakan file models.py.
- View dalam arsitektur MVT digunakan untuk mendefinisikan logika keseluruhan dari aliran data. Implementasinya menggunakan file view.py.
- Template dalam arsitektur MVT adalah lapisan presentasi yang menangani antarmuka pengguna.

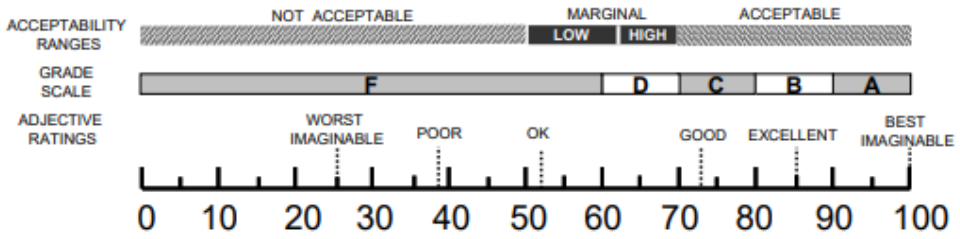


Gambar 2.3 Alur kerja Django

Setiap pengguna meminta beberapa sumber daya, maka Django bertindak sebagai pengontrol dan mencari sumber daya dalam file urls.py. Jika URL memetakan, maka view yang terkait dengan URL tersebut akan dipanggil. Setelah itu, view berinteraksi dengan model dan template. Pada akhirnya, Django merespons pengguna dan mengembalikan template sebagai respons [22].

2.11 *System Usability Score* (SUS)

System Usability Scale ialah salah satu survei yang dapat digunakan untuk menilai kegunaan dari berbagai produk atau jasa.



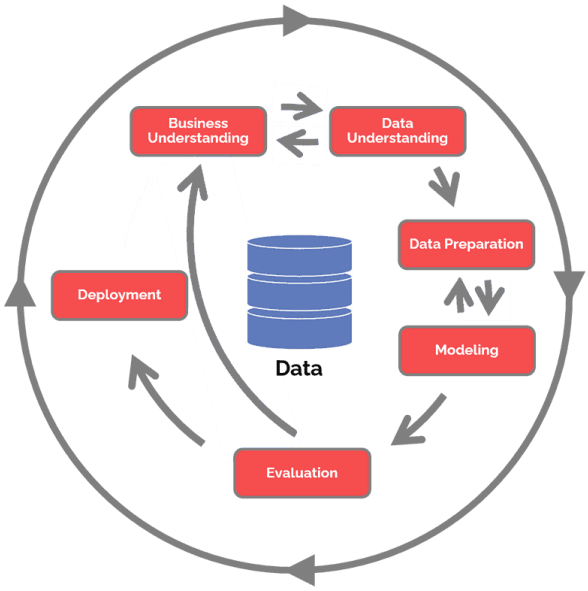
Gambar 2.4 Skor *System Usability Scale*

Hingga saat ini, SUS banyak digunakan untuk mengukur *usability* dan menunjukkan beberapa keunggulan, antara lain:

1. Karena hasil survei adalah skor tunggal, mulai dari 0 hingga 100, SUS relatif mudah dipahami oleh banyak orang dari disiplin ilmu lain.
2. Hanya terdiri dari sepuluh pernyataan, jadi relatif cepat dan mudah untuk menyelesaikan dan menilainya.
3. SUS sangat mudah digunakan, tidak membutuhkan perhitungan yang rumit.
4. SUS terbukti valid dan reliabel, walau dengan ukuran sampel yang kecil [23].

2.12 *Cross Industry Standard Process for Data Mining* (CRISP-DM)

Cross Industry Standard Process for Data Mining (CRISP-DM) dikembangkan pada tahun 1996 oleh analisis dari beberapa industri seperti standarisasi *Daimler Chrysler (Daimler-Benz)*, SPSS, NCR. *CRISP-DM* menyediakan standar proses data mining sebagai strategi pemecahan masalah secara umum dari bisnis atau unit penelitian [24]. Proses dalam *CRISP-DM* terdiri dari 6 fase kegiatan, yaitu:



Gambar 2.5 Tahapan metode penelitian CRISP-DM

2.12.1 *Business Understanding*

Pemahaman bisnis adalah tahapan yang pertama dilakukan dalam proses CRISP-DM. Pemahaman bisnis berfokus pada pemahaman tujuan dan kebutuhan dari sudut pandang bisnis, lalu selanjutnya diterjemahkan ke permasalahan pada data mining. Kemudian dibuat rencana yang spesifik untuk mencapai tujuan tersebut.

2.12.2 *Data Understanding*

Pada tahap pemahaman data ini dilakukan pemahaman terhadap kebutuhan data terkait dengan tujuan bisnis sebelumnya, serta menghimpun atau mengumpulkan dataset yang selanjutnya dilakukan pemahaman lebih lanjut tentang dataset tersebut.

2.12.3 *Data Preparation*

Tahap persiapan data dilakukan untuk membangun dataset akhir yang akan diproses pada tahap pemodelan berikutnya. Tahap ini juga dilakukan *preprocessing* dimana

proses memetakan data dari yang sebelumnya berupa data mentah menjadi data yang lebih sesuai untuk data akhir pemodelan. Tahap ini dapat dilakukan berulang kali.

2.12.4 *Modeling*

Dilakukannya tahap pemodelan ini adalah untuk membangun sebuah model yang tepat untuk memprediksi probabilitas data yang akan masuk serta mengelompokkannya sesuai dengan perhitungan probabilitas sebelumnya. Sebelum melakukan algoritma pemodelan, terlebih dahulu dilakukan pembagian dataset menjadi dua, yaitu data *training* dan data *testing*. Data *training* digunakan untuk menentukan *classifier* atau *labeling* pada data yang akan masuk selanjutnya dan data *testing* digunakan untuk mengetahui performa dari sistem yang telah dibuat berdasarkan data yang telah dilatih sebelumnya.

2.12.5 *Evaluation*

Tahap evaluasi ini adalah untuk memberikan penilaian pada model yang telah dibangun sebelumnya. Dari hasil evaluasi tersebut dapat ditentukan langkah selanjutnya. Jika pemodelan yang dibangun tersebut sudah cukup baik, maka dapat lanjut ke tahap *deployment* dan apabila model yang dibangun belum cukup baik, dapat dilakukan pengulangan kembali dalam pembuatan modelnya.

2.12.6 *Deployment*

Pada tahap penyebaran ini model yang telah selesai dibangun akan disampaikan kepada *customer* agar *customer* juga dapat menilai hasilnya. Pada tahap ini juga laporan hasil penelitian dibuat yang isinya adalah mencakup hasil akhir dari proses data mining tersebut.

III. METODOLOGI PENELITIAN

3.1 Waktu dan Tempat Penelitian

Waktu dan tempat pelaksanaan penelitian ini dilakukan pada :

Waktu : September 2021 sampai dengan Juli 2022

Tempat : Laboratorium Terpadu Jurusan Teknik Elektro Universitas Lampung

Tabel 3.1 Waktu Penelitian

No	Kegiatan	Waktu Penelitian										
		Sep	Okt	Nov	Des	Jan	Feb	Mar	Apr	Mei	Jun	Jul
1	Pemahaman Bisnis											
2	Pemahaman Data											
3	Persiapan Data											
4	Pemodelan											
5	Evaluasi											
6	Penyebaran											

3.2 Alat dan Bahan Penelitian

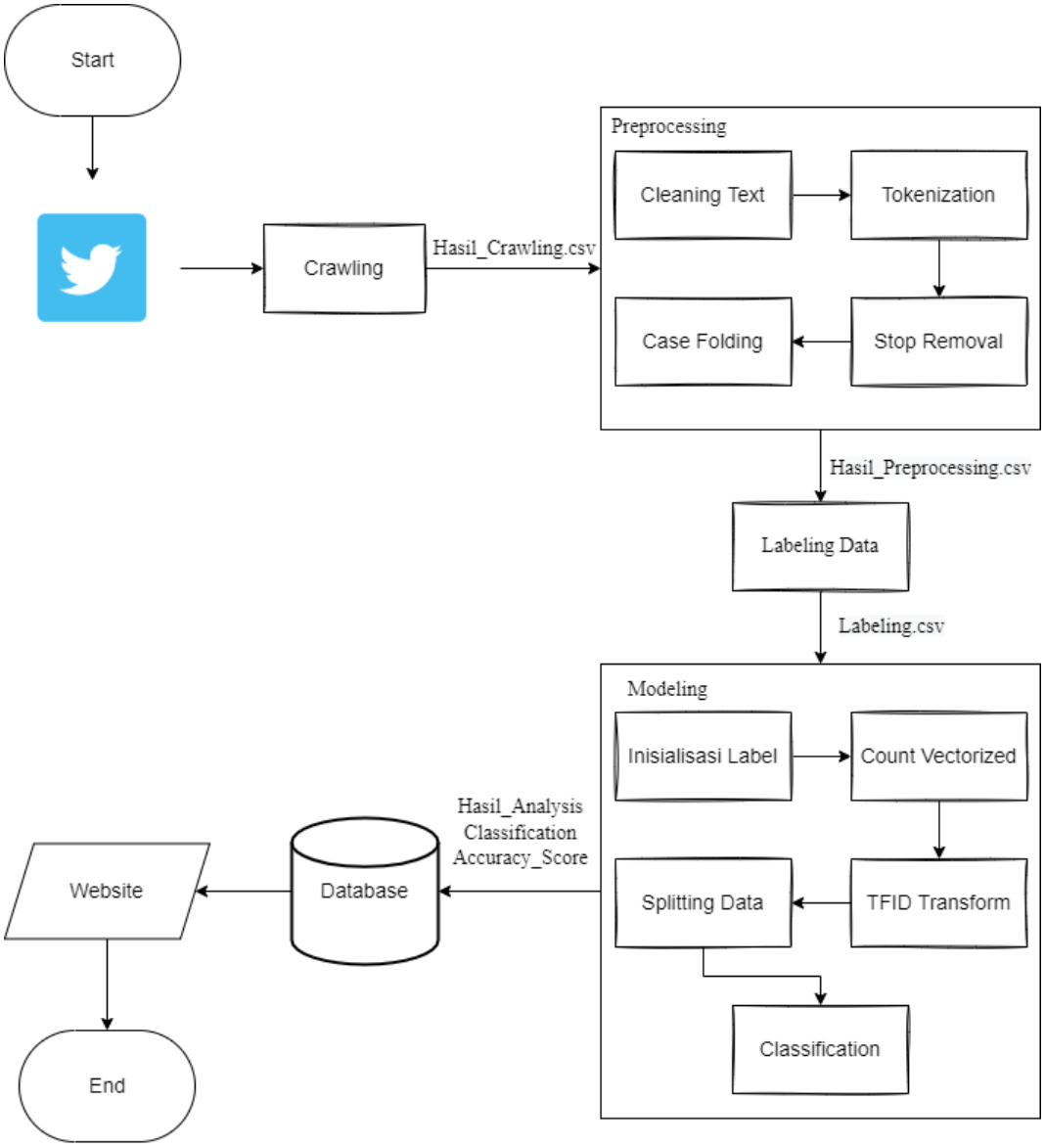
Dalam penelitian ini menggunakan alat dan bahan sebagai berikut :

Tabel 3.2 Alat dan Bahan Penelitian

No	Perangkat	Spesifikasi	Kegunaan
1	Laptop	Intel Pentium Gold, Ram 4 GB	Perangkat pengembangan dan pengujian aplikasi
2	Python	Jupyter *based on* Python 3.8.8	Bahasa pemrograman yang digunakan dalam pembuatan sistem
3	Visual Studio Code	Versi 1.49.2	Text Editor untuk pengembangan *website*
4	Windows	Windows 10	Sistem Operasi

3.3 Tahapan Penelitian

Pada penelitian ini, data didapatkan melalui Twitter API dan diolah menggunakan metode klasifikasi algoritma *naïve bayes*. Kemudian metode yang digunakan dalam penelitian ini adalah *Cross Industry Standard Process for Data Mining (CRISP-DM). CRISP-DM* menyediakan standar proses data mining sebagai strategi pemecahan masalah secara umum dari bisnis atau unit penelitian. Berdasarkan proses dari CRISP-DM, berikut adalah diagram alur dalam proses pengembangan sistem yang dilakukan dalam penelitian ini.



Gambar 3.1 *Flowchart* alur sistem yang dibuat

3.3.1 *Business Understanding*

Pemahaman bisnis merupakan tahapan pertama yang dilakukan dalam metode CRISP-DM. Pemahaman bisnis umumnya berfokus pada pemahaman tujuan serta kebutuhan dari sudut pandang bisnis, yang kemudian diterjemahkan ke dalam permasalahan pada data mining. Kemudian dibuat rencana yang spesifik untuk mencapai tujuan tersebut.

Tujuan pada penelitian ini adalah membuat sebuah model atau sistem analisis sentimen berdasarkan *tweet* masyarakat di media sosial Twitter mengenai kinerja dari Komisi Pemberantasan Korupsi. Kemudian menganalisis kinerja dari model *naïve bayes* dalam mengklasifikasikan *tweet-tweet* yang didapat, dan kemudian membuat sebuah *website* untuk visualisasi datanya.

3.3.2 *Data Understanding*

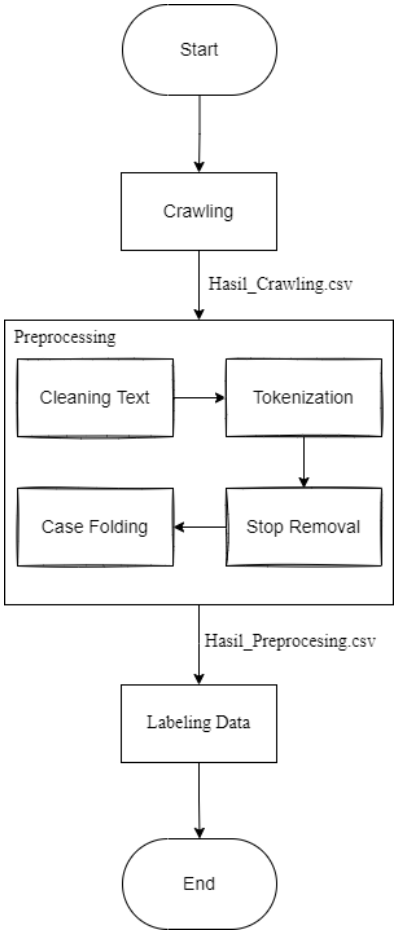
Pada tahap pemahaman data, akan dilakukan pemahaman terhadap kebutuhan data terkait dengan tujuan bisnis sebelumnya serta menghimpun dataset yang diperoleh. Pada penelitian ini, dataset didapatkan melalui media sosial Twitter menggunakan API *(Application Programming Interface)* berupa data *tweet* masyarakat Twitter. Pada proses ini dilakukan pengambilan data pada Twitter yang berisi atribut-atribut antara lain tanggal *tweet* dibuat, *username* pengguna, serta isi dari *tweet* dengan *query* KPK. Dataset yang didapat pada proses ini disimpan dengan ekstensi .csv. Kemudian data yang telah diperoleh tersebut akan dianalisa untuk mendapatkan informasi.

	tanggal	nama	tweet
1	2021-08-24 15:38:01	coffinyoursize_	@llrzm @mahendragilangg Tahanan KPK, antemi wae
2	2021-08-24 15:36:52	MRLaode	Kurang apa coba kebaikan @KPK_RI pada orang ini. Sudah tau tempat persembunyiannya tapi belum ditangkap alasa…
3	2021-08-24 15:34:32	Deny26241241	@detikcom Semakin kesini jadi KPK pro dengan para KORUPTOR.!!! TIDAK prok dengan menyelamatkan uang Rakya…
4	2021-08-24 15:33:37	apriyantoHadiN1	@sudjiwotedjo Masih ruwet, ra mikir KPK
5	2021-08-24 15:33:15	AnjaniAneira	Ombudsman ga berhak ikut campur soal TWK KPK. Vaksinasi Tuntaskan Pandemi https://t.co/UF3pjC3lon
6	2021-08-24 15:33:14	Harlan34792342	@semvakmerah @KPK_RI @Arifuddin1112 @Resty__Cayah @UyokBack @msaid_didu @Ronnie_Rusli @RamliRizal…
7	2021-08-24 15:33:08	AnjaniAneira	Ombudsman ga berhak ikut campur soal TWK KPK. Vaksinasi Tuntaskan Pandemi https://t.co/DS9ZXOjbF1
8	2021-08-24 15:31:23	emefte	@Candraasmara85 Janganlah, nanti malu ketahuan jadi pegawai KPK kaya kawan @febridiansyah
9	2021-08-24 15:31:16	fahayabi	@hansssolo @KPK_RI Entarlahh.....kalau publik udah lupa 🤣
10	2021-08-24 15:30:35	Balkarshape221	Gubernur Sumbar mulai panik, soalnya bentar lagi jadi tahanan KPK terkait kasus sumbangan nya #KadrunPenipuHan…
11	2021-08-24 15:30:27	rubinasicilia	@detikcom Bagus emang tpi Aib dan Naif wat kita bangsa Demokrasi ini.. Krn pengalaman dan ide yg mrk lakukan itu s…
12	2021-08-24 15:29:53	ahduy_rg	Bukan karena orang pinter di KPK sudah dihabisin yah? KPK : OTT Berkurang Karena Info Dari Warga Makin Sedikit d…
13	2021-08-24 15:29:43	Sanusi94307223	@Muhamadalibaim @msaid_didu @KPK_RI @DivHumas_Polri Lah untuk apa pakai skck, jadi komisaris bumn saja bo…
14	2021-08-24 15:29:20	girisuprapdiono	Masak ? https://t.co/GunDwOhfRA
15	2021-08-24 15:29:12	BBerizik	@DiantyYasmin3 Sepertinya ada agenda benturkan kpk dengan presiden..
16	2021-08-24 15:28:47	OposisiCerdas	Harun Masiku Masih Gagal Ditangkap, KPK Berdalih Pandemi COVID-19 https://t.co/s4C35nDtMS
17	2021-08-24 15:28:26	gagah76	@republikaonline Kalo KPK gak Nemu si Harun Yo wis KPK ga ada harga diri nya lagi dah titik....
18	2021-08-24 15:28:21	positive_values	@BiLLRaY2019 Kualitas mentalitas kupluk ya kaya gini boro² jadi panutan/pemimpin .. moga² kasus korupsinya cepat t…
19	2021-08-24 15:28:09	rmol_id	@KPK_RI Periksa 3 Tersangka Kasus Pengadaan Tanah Munjul https://t.co/B3cwSDbroC

Gambar 3.2 Dataset yang didapat dari Twitter melalui proses *crawling*

3.3.3 *Data Preparation*

Tahap persiapan data dilakukan untuk membangun dataset akhir yang akan diproses pada tahap pemodelan berikutnya. Pada tahap inilah dilakukan *preprocessing* yaitu proses memetakan data dari yang sebelumnya berupa data mentah menjadi data yang lebih sesuai untuk data akhir pemodelan, seperti menghilangkan tanda baca, menghapus kata yang berulang dan kata yang sering muncul tetapi tidak terlalu memiliki makna dalam kalimat, serta mengubah kata menjadi huruf kecil. Setelah *preprocessing* data selesai dilakukan, tahap selanjutnya adalah *Labeling Data* yaitu proses memberikan label atau *class* pada *tweet* pengguna secara manual yang disimpan pada dataset. Pada penelitian ini terdapat 3 *class* atau label, yaitu positif, netral, dan negatif. Berikut adalah diagram alur tahap persiapan data :



Gambar 3.3 Diagram alur persiapan data

a. Cleaning Text

Cleaning text adalah proses menghilangkan tanda baca atau karakter serta menghapus kata atau kalimat yang berulang. Contoh dari proses *cleaning text* dapat dilihat pada tabel 3.3.

Tabel 3.3 *Cleaning Text*

No	***Input***	***Output***
1	Indonesia Corruption Watch (ICW) menilai pimpinan KPK di bawah Firli Bahuri merupakan kepemimpinan paling buruk sejak lembaga tersebut berdiri. "Betapa tidak, pimpinan yang seharusnya menjadi pelindung pegawai malah justru menjadi sutradara di balik pemberhentian paksa 51 pegawai KPK," ujar Kurnia Ramadhana, peneliti ICW	Indonesia Corruption Watch ICW menilai pimpinan KPK di bawah Firli Bahuri merupakan kepemimpinan paling buruk sejak lembaga tersebut berdiri Betapa tidak pimpinan yang seharusnya menjadi pelindung pegawai malah justru menjadi sutradara di balik pemberhentian paksa 51 pegawai KPK ujar Kurnia Ramadhana peneliti

b. Tokenization

Tokenization adalah proses untuk memisahkan kalimat yang didapat menjadi beberapa kata. Contoh dari proses *tokenization* dapat dilihat pada tabel 3.4.

Tabel 3.4 *Tokenization*

No	***Input***	***Output***
1	indonesia corruption watch menilai pimpinan kpk di bawah firli bahuri merupakan kepemimpinan paling buruk sejak lembaga tersebut berdiri pimpinan yang seharusnya menjadi pelindung pegawai justru menjadi	[‘indonesia’, ‘corruption’, ‘watch’, ‘menilai’, ‘pimpinan’, ‘kpk’, ‘di’, ‘bawah’, ‘firli’, ‘bahuri’, ‘merupakan’, ‘kepemimpinan’, ‘paling’, ‘buruk’, ‘sejak’, ‘lembaga’, ‘tersebut’, ‘berdiri’, ‘pimpinan’, ‘yang’, ‘seharusnya’,

No	*Input*	*Output*
	sutradara di balik pemberhentian paksa 51 pegawai kpk ujar kurnia ramadhana peneliti	‘menjadi’, ‘pelindung’, ‘pegawai’, ‘justru’, ‘menjadi’, ‘sutradara’, ‘di’, ‘balik’, ‘pemberhentian’, ‘paksa’, ‘51’, ‘pegawai’, ‘kpk’, ‘ujar’, ‘kurnia’, ‘ramadhana’, ‘peneliti’]

c. Stopwords Removal

Stopwords removal adalah proses untuk menghilangkan kata-kata yang sering muncul tetapi tidak terlalu memiliki makna dalam kalimat. Contoh dari proses *stopwords removal* dapat dilihat pada tabel 3.5.

Tabel 3.5 *Stopwords Removal*

No	*Input*	*Ouput*
1	indonesia corruption watch icw menilai pimpinan kpk di bawah firli bahuri merupakan kepemimpinan paling buruk sejak lembaga tersebut berdiri betapa tidak pimpinan yang seharusnya menjadi pelindung pegawai malah justru menjadi sutradara di balik pemberhentian paksa 51 pegawai kpk ujar kurnia ramadhana peneliti	indonesia corruption watch menilai pimpinan kpk di bawah firli bahuri merupakan kepemimpinan paling buruk sejak lembaga tersebut berdiri pimpinan yang seharusnya menjadi pelindung pegawai justru menjadi sutradara di balik pemberhentian paksa 51 pegawai kpk ujar kurnia ramadhana peneliti

d. Case Folding

Case folding adalah proses untuk mengubah setiap kata menjadi sama, contohnya menjadi huruf kecil menggunakan fungsi *lower case*. Contoh dari proses *case folding* dapat dilihat pada tabel 3.6.

Tabel 3.6 *Case Folding*

No	***Input***	***Output***
1	Indonesia Corruption Watch ICW menilai pimpinan KPK di bawah Firli Bahuri merupakan kepemimpinan paling buruk sejak lembaga tersebut berdiri Betapa tidak pimpinan yang seharusnya menjadi pelindung pegawai malah justru menjadi sutradara di balik pemberhentian paksa 51 pegawai KPK ujar Kurnia Ramadhana peneliti	indonesia corruption watch icw menilai pimpinan kpk di bawah firli bahuri merupakan kepemimpinan paling buruk sejak lembaga tersebut berdiri betapa tidak pimpinan yang seharusnya menjadi pelindung pegawai malah justru menjadi sutradara di balik pemberhentian paksa 51 pegawai kpk ujar kurnia ramadhana peneliti

3.3.4 *Modeling*

Dilakukannya tahap pemodelan adalah untuk membangun sebuah model yang tepat untuk memprediksi probabilitas data yang akan masuk serta mengelompokkannya sesuai dengan perhitungan probabilitas sebelumnya.

Pada proses ini, hal pertama yang dilakukan adalah inisialisasi nilai label atau *class* ke suatu polaritas. Setelah itu dilakukan *splitting* data. *Splitting* data merupakan sebuah proses yang dilakukan untuk membagi dataset menjadi dua bagian, yaitu data *training* dan data *testing*. Kemudian melakukan klasifikasi menggunakan algoritma *naïve bayes classifier* untuk memprediksi probabilitas suatu data yang akan masuk serta mengelompokkannya, sesuai dengan perhitungan probabilitas sebelumnya. Metode algoritma *naïve bayes* digolongkan menjadi tiga tipe berdasarkan fungsinya yaitu *Multinomial NB, Bernoulli NB, dan Gaussian NB.* Pada penelitian ini dilakukan sebuah pengujian untuk menentukan algoritma *naïve bayes* tipe mana yang lebih baik untuk proses klasifikasi. Contoh dari klasifikasi dapat dilihat pada tabel 3.7.

Tabel 3.7 Klasifikasi

No	*Tweet*	Label/Kelas
1.	"Mereka menganggap kinerja KPK baik karena banyak koruptor tertangkap. Operasi tangkap tangan jadi indikator yang paling kelihatan bagi publik tentang kinerja KPK," ujar Kunto dalam sebuah diskusi virtual	Positif
2.	Makanya KPK dinina bobokkan Negeri ini. Bersikap negarawan dianggap aneh, sedangkan berperilaku pengkhianat dianggap hebat	Negatif
3.	Dilansir rakyatdotnews, 700 Pegawai KPK Tes Urine Bersama BNN.	Netral

3.3.5 *Evaluation*

Setelah klasifikasi pada data dengan menggunakan algoritma *naïve bayes classifier* selesai dilakukan, selanjutnya melakukan evaluasi. Tahap evaluasi ini adalah untuk memberikan penilaian pada model yang telah dibangun sebelumnya. Evaluasi sangat diperlukan untuk mengetahui apakah pemodelan yang dibangun tersebut sudah cukup baik atau belum. Jika model yang diperoleh sudah cukup baik, maka dapat lanjut ke tahap *deployment* dan apabila model yang dibangun belum cukup baik, dapat dilakukan pengulangan kembali dalam pembuatan modelnya.

3.3.6 *Deployment*

Pada tahap penyebaran, model yang telah selesai dibangun beserta dengan tampilan *website* untuk visualisasi datanya akan disampaikan kepada *customer* agar *customer* juga dapat menilai hasilnya. Pada tahap ini juga laporan hasil penelitian dibuat, yang isinya adalah mencakup hasil akhir dari proses *data mining* tersebut.

a. Visualisasi data

Setelah seluruh data telah dianalisis, data tersebut akan divisualisasikan ke dalam bentuk grafik dan juga tabel yang akan ditampilkan pada sebuah *website*. Selain menampilkan dataset yang telah dilatih sebagai model, *website* yang dikembangkan juga dapat mengambil data-data baru pada Twitter serta menampilkannya pada *website* lengkap beserta dengan labelnya.

b. Penerapan *testing* data

Pada penelitian ini ketika data baru berhasil diambil, model akan terus mengalami perubahan dikarenakan kosa kata yang didapat selalu bertambah. Maka dari itu, *testing* data dilakukan untuk melihat performa model yang telah dibangun.

c. Pengujian *usability website*

Setelah *website* visualisasi data dibuat, tahap yang terakhir adalah pengujian *usability website.* Pengujian pada penelitian ini menggunakan kuesioner SUS (*System Usability Scale)* yang terdiri dari 10 item pertanyaan. Tampilan *website* visualisasi data ini, akan disampaikan kepada customer agar customer dapat menilai hasilnya apakah tampilan *website* tersebut sudah cukup baik ataukah belum.

V. SIMPULAN DAN SARAN

5.1 Simpulan

Kesimpulan yang diperoleh berdasarkan hasil penelitian yang telah dilakukan adalah sebagai berikut:

1. Sebuah sistem analisis untuk mengetahui sentimen mengenai kinerja dari Komisi Pemberantasan Korupsi berdasarkan *tweet* masyarakat di media sosial Twitter telah berhasil dibuat dengan menggunakan bahasa Python.
2. Proses klasifikasi menggunakan algoritma *multinomial naïve bayes* dengan perbandingan dataset 80:20, didapatkan nilai akurasi sebesar 0,64 atau sekitar 64%. Nilai *precision* terbesar didapat pada dataset yang berlabel positif yaitu sebesar 0.69. Nilai *recall* terbesar didapat pada dataset yang berlabel negatif yaitu sebesar 0.89 dari keseluruhan data berlabel negatif. Nilai F1-score terbesar didapat pada dataset yang berlabel negatif yaitu sebesar 0.74. Jadi, dapat diartikan sistem yang digunakan pada penelitian ini cukup baik untuk memprediksi sentimen *tweet* mengenai KPK pada media sosial Twitter.
3. *Website* untuk visualisasi data yang telah dikembangkan pada penelitian ini, dapat langsung mengambil data baru yang kemudian data tersebut akan diklasifikasikan dan divisualisasikan ke dalam bentuk tabel dan juga grafik.

5.2 Saran

Saran yang dapat diberikan untuk penelitian selanjutnya berdasarkan penelitian yang telah dilakukan adalah sebagai berikut:

1. Melakukan pengujian sistem lebih lanjut dengan menggunakan metode klasifikasi yang lain, seperti *Support Vector Machine* (SVM), *K-Nearest Neighbor* (KNN), ataupun *Decision Tree* agar dapat diketahui metode mana yang lebih baik untuk klasifikasi.
2. Menerapkan sistem lebih lanjut agar dapat mencari kata kunci yang lain, tidak hanya berfokus pada KPK.
3. Melakukan penambahan beberapa fitur pada *website* seperti pengoreksian label dan juga rentang waktu pada grafik agar *website* dapat lebih baik lagi.

DAFTAR PUSTAKA

DAFTAR PUSTAKA

[1] U. M. Sosiawan, "Peran Komisi Pemberantasan Korupsi (KPK) Dalam Pencegahan dan Pemberantasan Korupsi," *Jurnal Penelitian Hukum De Jure*, vol. 19, no. 4, hlm. 517, 2019, doi: 10.30641/dejure.2019.v19.517-538.

[2] A. Go, R. Bhayani, dan L. Huang, "Twitter Sentiment Classification using Distant Supervision," *Processing*, vol., hlm. 1–6, 2009.

[3] S. S. Arote dan R. L. Paikrao, "A Modified Approach Towards Personalized Travel Recommendation System Using Sentiment Analysis," *2018 International Conference On Advances in Communication and Computing Technology (ICACCT)*, hlm. 203–207, 2018, doi: 10.1109/ICACCT.2018.8529327.

[4] D. Xhemali, C. J. Hinde, dan R. G. Stone, "Naïve Bayes vs . Decision Trees vs . Neural Networks in the Classification of Training Web Pages," *IJCSI International Journal of Computer Science Issues*, vol. 4, no. 1, hlm. 16–23, 2009.

[5] B. D. Prasetya, F. S. Pamungkas, dan I. Kharisudin, "Pemodelan dan Peramalan Data Saham dengan Analisis Time Series menggunakan Python," *PRISMA, Prosiding Seminar Nasional Matematika*, vol. 3, hlm. 714–718, 2020.

[6] D. D. Putri, G. F. Nama, dan W. E. Sulistiono, "Analisis Sentimen Kinerja Dewan Perwakilan Rakyat (DPR) Pada Twitter Menggunakan Metode Naive Bayes Classifier," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 10, no. 1, hlm. 34–40, Jan 2022, doi: 10.23960/jitet.v10i1.2262.

[7] W. Bimananda, I. Riski, K. Dwi, R. Nooraeni, T. Siahaan, dan Y. Dhea, "Analisis Text Mining dari Cuitan Twitter Mengenai Infrastruktur di Indonesia dengan Metode Klasifikasi Naïve Bayes," *Eigen Mathematics Journal*, vol. 2, no. 2, hlm. 92–101, 2019, doi: 10.29303/emj.v1i2.36.

[8] F. Ratnawati, "Implementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter," *Jurnal Inovtek Polbeng - Seri Informatika*, vol. 3, no. 1, hlm. 50–59, 2018.

[9] A. Deviyanto dan M. D. R. Wahyudi, "Penerapan Analisis Sentimen Pada Pengguna Twitter Menggunakan Metode K-Nearest Neighbor," *JISKA (Jurnal Informatika Sunan Kalijaga)*, vol. 3, no. 1, hlm. 1–13, 2018, doi: 10.14421/jiska.2018.31-01.

[10] B. M. Pintoko dan K. Muslim, "Analisis Sentimen Jasa Transportasi Online pada Twitter Menggunakan Metode Naive Bayes Classifier," *e-Proceeding of Engineering*, vol. 5, no. 3, hlm. 8121–8130, 2018.

[11] twitter.com, "Tentang API Twitter," *https://help.twitter.com/id/rules-and-policies/twitter-api*. https://help.twitter.com/id/rules-and-policies/twitter-api (diakses Jun 13, 2021).

[12] P. Sai dan B. Balachander, "Sentimental analysis of twitter data using tweepy and textblob," *International Journal of Advanced Science and Technology*, vol. 29, no. 3, hlm. 6537–6544, 2020.

[13] A. T. Jaka, “Preprocessing Text untuk Meminimalisir Kata yang Tidak Berarti dalam Proses Text Mining,” *Jurnal Informatika UPGRIS*, vol. 1, hlm. 1–9, 2015.
[14] M. C. Kirana, N. P. Perkasa, M. Z. Lubis, dan M. Fani, “Visualisasi Kualitas Penyebaran Informasi Gempa Bumi di Indonesia Menggunakan Twitter,” *Journal of Applied Informatics and Computing (JAIC)*, vol. 3, no. 1, hlm. 23–32, 2019, doi: 10.30871/jaic.v0i0.1246.
[15] D. Heksaputra, Y. Azani, Z. Naimah, dan L. Iswari, “Penentuan Pengaruh Iklim Terhadap Pertumbuhan Tanaman dengan Naïve Bayes,” *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*, hlm. 34–39, 2013, [Daring]. Available: https://media.neliti.com/media/publications/88595-ID-penentuan-pengaruh-iklim-terhadap-pertum.pdf
[16] F. Handayani dan F. S. Pribadi, “Implementasi Algoritma Naive Bayes Classifier dalam Pengklasifikasian Teks Otomatis Pengaduan dan Pelaporan Masyarakat melalui Layanan Call Center 110,” *Jurnal Teknik Elektro*, vol. 7, no. 1, hlm. 19–24, 2015, doi: 10.15294/jte.v7i1.8585.
[17] H. K. C. A. Pratama, W. Suharso, dan Q. A’yun, “Pengklasifikasian Kanker Payudara Dan Kanker Paru-Paru Dengan Metode Gaussian Naïve Bayes , Multinomial Naïve Bayes , Dan Bernoulli Naïve Bayes,” *Jurnal Smart Teknologi*, vol. 3, no. 4, hlm. 350–355, 2022, Diakses: Sep 20, 2022. [Daring]. Available: http://jurnal.unmuhjember.ac.id/index.php/JST/article/view/7592/3886
[18] I. F. Rozi, A. T. Firdausi, dan K. Islamiyah, “Analisis Sentimen Pada Twitter Mengenai Pasca Bencana Menggunakan Metode Naïve Bayes Dengan Fitur N-Gram,” *JIP (Jurnal Informatika Polinema)*, vol. 6, no. 2, hlm. 33–39, 2020, doi: 10.33795/jip.v6i2.316.
[19] A. Ingargiola, “What is the Jupyter Notebook?,” *https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/what_is_jupyter.html*, 2015. https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/what_is_jupyter.html (diakses Sep 20, 2022).
[20] golfsql, “Apa Itu PostgreSQL? Bagaimana Cara Kerja PostgreSQL?,” *https://www.pgsql.com/apa-itu-postgresql-bagaimana-cara-kerja-postgresql/*, Nov 22, 2021. https://www.pgsql.com/apa-itu-postgresql-bagaimana-cara-kerja-postgresql/ (diakses Apr 04, 2022).
[21] P. Indonesia, “Jenis Website,” *https://www.proweb.co.id/articles/web_design/jenis_website.html*, Jan 14, 2011. https://www.proweb.co.id/articles/web_design/jenis_website.html (diakses Jun 23, 2021).
[22] B. Kumar, “What is Python Django and used for,” *https://pythonguides.com/what-is-python-django/*, Agu 04, 2021. https://pythonguides.com/what-is-python-django/ (diakses Sep 20, 2022).
[23] A. Bangor, P. Kortum, dan J. Miller, “Determining what individual SUS scores mean; adding an adjective rating,” *Journal of usability studies (JUS)*, vol. 4, no. 3, hlm. 114–123, 2009.
[24] N. Hotz, “What is CRISP DM?,” *https://www.datascience-pm.com/crisp-dm-2/*, Agu 08, 2022. https://www.datascience-pm.com/crisp-dm-2/