IMPLEMENTASI *NAÏVE BAYES* PADA *TEXT-MINING* UNTUK ANALISIS SENTIMEN PARIWISATA

SKRIPSI

Disusun Oleh:

NATHANIEL FELIX FRADERIC

20191000053

TEKNIK INFORMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS BUDDHI DHARMA

TANGERANG

2023

IMPLEMENTASI *NAÏVE BAYES* PADA *TEXT-MINING* UNTUK ANALISIS SENTIMEN PARIWISATA

SKRIPSI

Diajukan sebagai salah satu syarat untuk kelengkapan kesarjanaan pada

Program Studi Teknik Informatika

Jenjang Pendidikan Strata 1

Disusun Oleh:

NATHANIEL FELIX FRADERIC

20191000053

TEKNIK INFORMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS BUDDHI DHARMA

TANGERANG

2023

LEMBAR PERSEMBAHAN

"Do as much as you can. The right thing comes at the right time."

(Nathaniel Felix)

Dengan mengucap puji syukur kepada Tuhan Yang Maha Esa, Skripsi ini kupersembahkan untuk:

1. Bapak Tjhai Tjin Fat dan Ibu Susie tercinta yang telah membesarkan aku dan selalu membimbing, mendukung, memotivasi, memberi apa yang terbaik bagiku serta selalu mendoakan aku untuk meraih kesuksesanku.
2. Kakak dan adik-adikku yang telah memberikan dukungan semangat serta dorongan yang senantiasa diberikan.
3. Bapak Aditiya Hermawan yang telah membimbing dan memberikan arahan dalam pembuatan skripsi ini.
4. Teman-teman kelompok belajar yang selalu berjuang bersama.

UNIVERSITAS BUDDHI DHARMA

LEMBAR PERNYATAAN KEASLIAN SKRIPSI

Yang bertanda tangan di bawah ini,

NIM	: 20191000053
Nama	: Nathaniel Felix Fraderic
Jenjang Studi	: Strata 1
Program Studi	: Teknik Informatika
Peminatan	: *Database*

Dengan ini saya menyatakan bahwa:

1. Skripsi ini adalah asli dan belum pernah diajukan untuk mendapat gelar akademik (Diploma/Sarjana) atau kelengkapan studi, baik di Universitas Buddhi Dharma maupun di Pereguruan Tinggi lainnya.
2. Skripsi ini saya buat sendiri tanpa bantuan dari pihak lain, kecuali arahan dosen pembimbing.
3. Dalam Skripsi ini tidak terdapat karya atau pendapat yang telah ditulis atau dipublikasikan orang lain, kecuali secara tertulis dengan jelas dan dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan dicantumkan daftar pustaka.
4. Dalam Skripsi ini tidak terdapat pemalsuan (kebohongan), seperti buku, artikel, jurnal, data sekunder, pengolahan data, dan pemalsuan tanda tangan dosen atau Ketua Progtram Studi Universitas Buddhi Dharma yang dibuktikan dengan keasliannya.
5. Lembar pernyataan ini saya buat dengan sesungguhnya, tanpa paksaan dan apabila dikemudian hari atau pada waktu lainnya terdapat penyimpangan dan ketidak benaran dalam pernyataan ini, saya bersedia menerima sanksi akademik berupa pencabutan gelar akademik yang telah saya peroleh karena Skripsi ini serta sanksi lainnya sesuai dengan peraturan dan norma yang berlaku.

Tangerang, 7 Agustus 2023

Penulis,

Nathaniel Felix Fraderic

UNIVERSITAS BUDDHI DHARMA

LEMBAR PERSETUJUAN PUBLIKASI ILIMIAH

Yang bertanda tangan di bawah ini,

NIM	: 20191000053
Nama	: Nathaniel Felix Fraderic
Jenjang Studi	: Strata 1
Program Studi	: Teknik Informatika
Peminatan	: *Database*

Dengan ini menyetujui untuk memberikan ijin kepada pihak Universitas Buddhi Dharma, Hak Bebas Royalti Non – Eksklusif (Non-exclusive Royalty-Free Right) atas karya ilmiah kami yang berjudul: "IMPLEMENTASI *NAÏVE BAYES* PADA *TEXT-MINING* UNTUK ANALISIS SENTIMEN PARIWISATA", beserta alat yang diperlukan (apabila ada).
Dengan Hak Bebas Royalti Non – Eksklusif ini pihak Universitas Buddhi Dharma berhak menyimpan, mengalih-media atau format-kan, mengelolanya dalam pangkalan data (database), mendistribusikannya, dan menampilkan atau mempublikasikannya di internet atau media lain untuk kepentingan akademis tanpa perlu meminta ijin dari saya selama tetap mencantumkan nama saya sebagai penulis atau pencipta karya ilmiah tersebut.
Saya bersedia untuk menanggung secara pribadi, tanpa melibatkan pihak Universitas Buddhi Dharma, segala bentuk tuntutan hukum yang timbul atas pelanggaran Hak Cipta dalam karya ilmiah saya ini.
Demikian pernyataan ini saya buat dengan sebenarnya.

Tangerang, 7 Agustus 2023

Penulis,

Nathaniel Felix Fraderic

20191000053

UNIVERSITAS BUDDHI DHARMA

LEMBAR PENGESAHAN PEMBIMBING

IMPLEMENTASI *NAÏVE BAYES* PADA *TEXT-MINING* UNTUK ANALISIS SENTIMEN PARIWISATA

Dibuat Oleh :

NIM : 20191000053

Nama : Nathaniel Felix Fraderic

Telah disetujui untuk dipertahankan di hadapan Tim Penguji Ujian Komprehensif

Program Studi Teknik Informatika

Peminatan Basis Data

Tahun Akademik 2022/2023

Disahkan oleh,

Tangerang, 21 Juli 2023

Pembimbing,

Aditiya Hermawan, S.Kom., M.Kom

NIDN. 0406128801

UNIVERSITAS BUDDHI DHARMA

LEMBAR PENGESAHAN SKRIPSI

IMPLEMENTASI *NAÏVE BAYES* PADA *TEXT-MINING* UNTUK ANALISIS SENTIMEN PARIWISATA

Dibuat Oleh :

NIM : 20191000053

Nama : Nathaniel Felix Fraderic

Telah disetujui untuk dipertahankan di hadapan Tim Penguji Ujian Komprehensif

Program Studi Teknik Informatika

Peminatan Basis Data

Tahun Akademik 2022/2023

Disahkan oleh,

Tangerang, 07 Agustus 2023

Dekan,	Ketua Program Studi,
Dr. Eng. Ir. Amin Suyitno, M.Eng	**Hartana Wijaya, S.Kom., M.Kom.**
NIDK : 8826333420	**NIDN: 0412058102**

LEMBAR PENGESAHAN TIM PENGUJI

Nama	:	Nathaniel Felix Fraderic
NIM	:	20191000053
Fakultas	:	Sains dan Teknologi
Judul Skripsi	:	IMPLEMENTASI *NAÏVE BAYES* PADA *TEXT-MINING* UNTUK ANALISIS SENTIMEN PARIWISATA

Dinyatakan LULUS setelah mempertahankan di depan Tim Penguji pada hari Senin, 7 Agustus 2023.

	Nama penguji :	Tanda Tangan :
Ketua Sidang :	**Desiyanna Lasut, S.Kom., M.Kom** NIDN. 0402128601	.....
Penguji I :	**Susanto Hariyanto, S.Kom., M.Kom** NIDN. 0428128601	.....
Penguji II :	**Aditiya Hermawan, S.Kom., M.Kom** NIDN.0406128801	.....

Mengetahui,
Dekan Fakultas Sains dan Teknologi

Dr. Eng. Ir. Amin Suyitno, M.Eng
NIDK : 8826333420

KATA PENGANTAR

Dengan mengucapkan Puji Syukur kepada Tuhan Yang Maha Esa, yang telah memberikan Rahmat dan karunia-Nya kepada penulis sehingga dapat menyusun dan menyelesaikan Skripsi ini dengan judul "**IMPLEMENTASI *NAÏVE BAYES* PADA *TEXT-MINING* UNTUK ANALISIS SENTIMEN PARIWISATA"**. Tujuan utama dari pembuatan Skripsi ini adalah sebagi salah satu syarat kelengkapan dalam menyelesaikan program pendidikan Strata 1 Program Studi Teknik Informatika di Universitas Buddhi Dharma. Dalam penyusunan Skripsi ini penulis banyak menerima bantuan dan dorongan baik moril maupun materiil dari berbagai pihak, maka pada kesempatan ini penulis menyampaikan rasa terima kasih yang sebesar-besarnya kepada:

1. Ibu Dr. Limajatini, S.E., M.M., B.K.P, sebagai Rektor Universitas Buddhi Dharma
2. Bapak Dr. Eng, Ir. Amin Suyitno, M.Eng, Dekan Fakultas Sains dan Teknologi
3. Bapak Hartana Wijaya, S.Kom., M.Kom, sebagai Ketua Program Studi Teknik Informatika
4. Bapak Aditiya Hermawan, S.Kom., M.Kom., sebagai pembimbing yang telah membantu dan memberikan dukungan serta harapan untuk menyelesaikan penulisan Skripsi ini.
5. Orang tua dan keluarga yang selalu memberikan dukungan baik moril dan materiil.
6. Teman-teman yang selalu membantu dan memberikan semangat

Serta semua pihak yang terlalu banyak untuk disebutkan satu-persatu sehingga terwujudnya penulisan ini. Penulis menyadari bahwa penulisan Skripsi ini masih belum sempurna, untuk itu penulis mohon kritik dan saran yang bersifat membangun demi kesempurnaan penulisan di masa yang akan datang.

Akhir kata semoga Skripsi ini dapat berguna bagi penulis khususnya dan bagi para pembaca yang berminat pada umumnya.

Tangerang, 7 Agustus 2023

Penulis

Implementasi *Naïve Bayes* Pada *Text-Mining* Untuk Analisis Sentimen Pariwisata
107 Halaman + xi / 35 Tabel / 26 Gambar / 3 Lampiran

ABSTRAK

Sektor Pariwisata sampai saat ini masih menjadi sektor alternatif yang diunggulkan untuk mendorong perekonomian Indonesia. Namun dalam pengembangan sektor pariwisata di Indonesia masih terdapat masalah yaitu komunikasi dan publikasi yang masih kurang. Untuk membantu meningkatkan kualitas tempat wisata dibutuhkan analisa tanggapan dari pengunjung untuk menilai respon masyarakat. Maka dari itu dibutuhkan teknik analisis data yang secara otomatis memproses tanggapan dari pengunjung tersebut. T*ext mining* dibutuhkan untuk melakukan analisa sentimen dengan cepat. Data yang diambil untuk proses ini dilakukan dengan bantuan API pada platform *twitter*. *Twitter* dijadikan objek penelitian karena *twitter* memberikan akses pengambilan data secara cepat dengan menggunakan kata kunci atau hashtag. Bentuk Implementasi yang digunakan dalam penelitian ini adalah website. Website ini dibuat dengan bahasa pemrograman *Python* dan PHP. Proses dari *text mining* ini menggunakan algoritma *Naïve Bayes* yang berlangsung pada *backend* website. Hasil dari website ini menampilkan hasil analisis dari tanggapan masyarakat terhadap suatu objek wisata dalam bentuk grafik dengan presentase dari kelas tersebut. Nilai *accuracy* yang dihasilkan oleh proses text mining adalah 87%. Berdasakan hasil evaluasi yang telah dilakukan terhadap 36 responden yang disebarkan secara publik mendapatkan tingkat kepuasan sebesar 81,48%. Dengan adanya aplikasi analisis sentimen pariwisata ini, diharapkan dapat membantu pengguna dalam mencari rekomendasi tempat wisata berdasarkan sentimen. Dalam penelitian selanjutnya, dapat menambahkan list dari kata-kata positif dan negatif dalam bahasa Indonesia.

Kata kunci: *Naïve bayes, Text-Mining, Twitter, Pariwisata, Sentimen*

Implementation of Naïve Bayes in Text-Mining for Tourism Sentiment Analysis
107 Pages + xi / 35 Table / 26 Images / 3 Libraries

ABSTRACT

The tourism sector is still the leading alternative sector to boost the Indonesian economy. However, in the development of the tourism sector in Indonesia there are still problems, namely communication and publications that are still lacking. To help improve the quality of tourist attractions, an analysis of responses from visitors is needed to assess the community's response. Therefore a data analysis technique is needed that automatically processes the responses from these visitors. Text mining is needed to do sentiment analysis quickly. The data taken for this process is carried out with the help of the API on the Twitter platform. Twitter is used as an object of research because Twitter provides access to data retrieval quickly using keywords or hashtags. The form of implementation used in this study is a website. This website is made with Python and PHP programming languages. The process of this text mining uses the Naïve Bayes algorithm which takes place on the website's backend. The results of this website display the results of an analysis of people's responses to a tourist attraction in graphical form with the percentage of that class. The accuracy value generated by the text mining process is 87%. Based on the evaluation results that have been carried out on 30 respondents who distributed publicly, they get a satisfaction level of 81,48%. With this tourism sentiment analysis application, it is hoped that it can assist users in finding recommendations for tourist attractions based on sentiment. In future research, you can add a list of positive and negative words in Indonesian.

Keywords: *Naïve bayes, Text-Mining, Twitter, Tourism, Sentiment*

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

BAB I

PENDAHULUAN

1.1 Latar Belakang

Sektor Pariwisata sampai saat ini masih menjadi sektor alternatif yang diunggulkan untuk mendorong perekonomian Indonesia. Kontribusi pariwisata Indonesia terhadap PDB pada tahun 2020 adalah sebesar (4.05% dari PDB) mengalami penurunan pada tahun sebelumnya yang mencapai 4.7%. Hal ini dikarenakan pandemi Covid-19 yang membuat kontribusi sektor pariwisata terhadap produk domestik boruto (PDB) mengalami penurunan. (Azzahra, 2022)

Namun dalam pengembangan sektor pariwisata di Indonesia ini masih terdapat beberapa masalah yang perlu diselesaikan sehingga sektor pariwisata ini dapat meningkatkan pertumbuhan ekonomi di Indonesia. Salah satu faktor masalah tersebut adalah komunikasi dan publikasi yang masih kurang (SBM, 2020). Perkembangan sektor pariwisata ini membutuhkan komunikasi dan publikasi yang baik agar dapat membantu promosi tempat wisata tersebut agar dapat meningkatkan jumlah masyarakat yang ingin berkunjung. Salah satunya adalah dengan mendapatkan tanggapan yang baik dan respon positif dari pengunjung yang pariwisata tempat tersebut. Sejumlah besar informasi dari tanggapan pengunjung sangat banyak dan beragam, selain itu analisis secara manual tanpa bantuan teknik analisis data yang dibantu komputer sangat menyita waktu dan kurang efisien untuk analisis dalam jumlah banyak. Oleh karena itu dibutuhkan sistem yang secara otomatis untuk pemrosesan tanggapan dari pengunjung tersebut.

Untuk penyelesaian permasalahan diatas maka, perlu dilakukan pemrosesan untuk menilai respon dari pengunjung berdasarkan sentimennya yaitu dengan Analisis sentimen atau *Opinion Mining*. *Opinion Mining* atau analisis sentimen dapat dianggap

sebagai perpaduan antara *text mining* dan *natural language processing.* Analisis sentimen merupakan pendekatan yang efektif untuk mengidentifikasi dan menganalisis persepsi pengguna terhadap aplikasi digital seperti sistem informasi berbasis website dan media sosial untuk mengoptimalkan pemasaran produk dan layanan divisi sesuai dengan permintaan pasar. Dalam konteks pemasaran destinasi pariwisata, persepsi konsumen terhadap tempat wisata perlu diidentifikasi untuk menganalisis keputusan untuk berkunjung ke destinasi wisata (Surgawi, 2016). Keberadaan wisatawan digital secara nyata menunjukkan bahwa terdapat peran dari media sebagai pembentuk minat bagi wisatawan digital untuk meningkatkan preferensi desninasi wisata. Kemudian intensitas interaksi antar wisatawan digital membentuk habitus baru di Internet atau media sosial yaitu pola penyebaran secara tematik berdasarkan dari popularitas topik (Singgalen, 2021). Dengan adanya tanggapan dari pengunjung wisata akan sangat membantu terhadap perkembangan objek wisata tersebut dengan meningkatkan popularitas tempat tersebut dan memberikan penilaian terhadap objek wisata tersebut. Hal ini menunjukan bahwa sentimen positif pengguna dalam bentuk informasi yang dipublikasikan dapat mempengaruhi keputusan pengguna untuk berkunjung ke destinasi wisata tertentu dan demikian juga sebaliknya. Hal ini dikarenakan banyaknya ulasan dari pengunjung wisata yang diberikan pada platform Web yang menghasilkan jumlah data yang besar terkait dengan opini pengunjung wisata terhadap suatu objek Pariwisata.

Menurut data yang dirilis We Are Social, jumlah pengguna *twitter* di Indonesia tercatat sebanyak 18,45 juta, dengan jumlah tersebut membuat Indonesia menempati peringkat ke-lima dari pengguna *Twitter* terbanyak di seluruh dunia (Ayu Rizaty, 2022). Dengan memanfaatkan kemajuan perkembangan teknologi informasi saat ini, kita dapat mendapatkan data menggunakan berbagai platform diantaranya merupakan

sosial media. Salah satu sosial media yang sering digunakan untuk melihat dan memberikan komentar terhadap objek wisata adalah *Twitter*. *Twitter* dipilih karena teks yang dapat dalam *Twitter* dapat ditambang dengan mudah menggunakan *API* yang tersedia secara public yang dapat diakses oleh semua pengguna *Twitter*. Dalam *Twitter* dapat menampilkan dengan informasi wisata diantaranya terdapat nama, foto, deskripsi, lokasi dan ulasan tentang destinasi wisata dari pengunjung yang akan sangat membantu pengguna lain untuk menentukan preferensi destinasi wisata yang ingin dikunjungi. Sehingga diharapkan dapat menarik perhatian pengunjung wisata baik dari wisatawan lokal maupun mancanegara.

Dari masalah yang dijelaskan pada bidang sektor pariwisata maka diperlukan untuk mendapatkan data komentar dari pengunjung destinasi wisata berbahasa indonesia. Kemudian dari data tersebut dilakukan Analisis Sentimen terhadap data yang berisi komentar pengunjung tersebut agar dapat mengetahui perspektif pengunjung terhadap objek destinasi wisata. Setelah dilakukan visualisasi hasil pengolahan dan analisis dapat di visualisasikan agar dapat membantu dalam proses pengambilan keputusan pengunjung terhadap objek destinasi wisata.

Analisis Sentimen yang sudah diolah dan divisualisasikan dapat membantu pengguna untuk melihat jumlah review dan opini dari pengunjung pariwisata sangat banyak dan beragam oleh karena itu akan menyulitkan dan memakan waktu untuk membaca dan mengulas secara keseluruhan untuk menarik kesimpulan dari data yang beragam. Oleh karena itu dapat dilakukan perancangan sistem yang secara otomatis akan mengelompokan opini dari pengunjung wisata yang ada sesuai dengan kelompok kelasnya. Kelas Sentimen dibagi menjadi 3 kelas karena mendapatkan hasil yang lebih tinggi dibanding menggunakan pengujian dari 5 kelas (Gunawan et al., 2018). Kelas

sentimen dibagi menjadi 3 kelas positif, netral, negatif sehingga pengguna dapat mendapatkan informasi opini sesuai yang diharapkan.

Dengan adanya sistem analisis sentimen ini diharapkan dapat membantu pengunjung destinasi wisata untuk melihat rating dan review dari suatu desninasi wisata sebagai acuan dalam pembuatan keputusan yang lebih baik. Sistem ini juga diharapkan dapat berperan dalam perkembangan pariwisata di Indonesia dengan memberikan kemudahan untuk wisatawan dalam memilih destinasi wisata dan pengelola usaha dalam menilai respon pengunjung sehingga dapat meningkatkan kualitas layanan pariwisata di Indonesia.

Berdasarkan uraian diatas, pemanfaatan media sosial dalam mengetahui respon pengunjung terhadap objek pariwisata dapat dilakukan melalui analisis sentimen. Analisis yang dilakukan menggunakan metode pengklasifikasian yaitu *Naïve bayes*. Teknik analisa ini sudah diterapkan pada penelitian sebelumnya untuk menemukan teks tambang dengan ketelitian tinggi seperti:

Penerapan metode *Naïve Bayes* ini telah digunakan pada penelitian (Fanissa et al., 2018), Berdasarkan pengujian, dengan menggunakan metode Query Expansion Ranking dengan algoritma *Naïve Bayes* menghasilkan *accuracy* tertinggi sebesar 86.6 pada seleksi fitur 75% terhadap sistem yang dibuat pada penelitian analisis sentimen Pariwisata di Kota Malang dan juga metode ini telah diterapkan pada penelitian (Somantri & Dairoh, 2019), dari pengujian yang dilakukan dengan menggunakan metode *Naïve Bayes* dan Decision Tree, tingkat *accuracy Naïve Bayes* menghasilkan 77,50% lebih baik dibandingkan dengan menggunakan Decision Tree yang menghasilkan tingkat *accuracy* 60,83% .

Berdasarkan latar belakang yang ada maka, penulis menggunakan algoritma pengklasifikasian *Naïve Bayes* untuk analisa sentimen pada *Twitter* dengan judul

penelitian "**IMPLEMENTASI NAIVE BAYES PADA *TEXT-MINING* UNTUK ANALISIS SENTIMEN PARIWISATA"**

1.2 Identifikasi Masalah

Berdasarkan latar belakang tertulis, kami memberikan informasi berikut tentang masalah yang akan digunakan sebagai bahan penelitian:

1. Penelitian ini dilakukan untuk membantu masyarakat sehingga dapat mengakses informasi tentang tempat wisata dengan lebih cepat, dan dapat mengklasifikasikan sentimen berdasarkan opini wisatawan.

1.3 Rumusan Masalah

Dengan adanya latar belakang tersebut, maka penulis mengambil perumusan masalah sebagai berikut :

a. Bagaimana mengimplementasi *opinion mining* dengan algoritma *Naïve Bayes* dapat membantu pengguna mendapatkan rekomendasi objek wisata dengan menganalisis sentimen opini publik secara cepat dan akurat?

1.4 Ruang Lingkup

Untuk membatasi ruang lingkup penelitian maka terdapat beberapa aturan dan pembahasan sebagai berikut :

a. Data yang dapat diproses dan diambil dari *Twitter* objek wisata di sekitar Yogyakarta.
b. Data Testing yang digunakan didapat dari *Twitter* yaitu dari API *Twitter* berdasarkan 5 hashtag (Jalan Malioboro, Kraton Yogyakarta, Candi Borobudur, Candi Prambanan, Tugu Yogyakarta) dari objek wisata di sekitar Yogyakarta.
c. Sentimen yang akan dianalisa dalam bahasa Indonesia.
d. Hasil dari *text mining* berupa sentimen positif, negatif dan netral dari *Tweet* tersebut.

e. Perancangan sistem informasi dibuat menggunakan bahasa pemrograman PHP dan database MySQL.

1.5 Tujuan dan Manfaat

1.5.1 Tujuan

Adapun tujuan dari pembuatan aplikasi ini adalah sebagai berikut :

a. Menerapkan algoritma *Naïve Bayes* pada *text mining* untuk mencari rekomendasi destinasi wisata dengan cepat dan akurat.
b. Mempermudah dan mempercepat pengguna dalam mencari destinasi wisata berdasarkan ulasan dan penilaian dari publik dalam bahasa Indonesia sehingga tidak menghabiskan waktu.
c. Mengklasifikasikan *tweets* berdasarkan sentimen positif, negatif dan netral.

1.5.2 Manfaat

Selain itu juga terdapat beberapa manfaat dalam penggunaan aplikasi ini, antara lain :

a. Algoritma *Naïve Bayes* ini dapat menentukan rekomendasi destinasi wisata bahasa Indonesia dengan cepat dan mempunyai *accuracy* tinggi.
b. Memberikan efisiensi waktu, terutama bagi pengunjung yang tidak memiliki waktu untuk melakukan research tersendiri.
c. Hasil dari *tweets* yang ditampilkan akan membantu pengguna dan pemilik usaha dalam melakukan penilaian terhadap objek wisata tersebut berdasarkan hasil analisa sentimen.

1.6 Metode Penelitian

Penulis menggunakan beberapa metode penilitian, antara lain :

1.6.1 Metode Pengumpulan Data

Data digunakan dalam pengujian ini diperoleh melalui media sosial *Twitter*. Data training yang akan digunakan pada penelitian ini adalah data yang telah telah dilabelkan dengan menggunakan list kata-kata negatif dan positif dari https://github.com/masdevid/ID-OpinionWords. Setelah itu, data dipisahkan menjadi data training sebesar 75% dan data testing sebesar 25%. Lalu data testing diambil yang dikumpulkan secara langsung dengan memanfaatkan *API Twitter* untuk mendapatkan informasi berupa tweet dari user.

1.6.2 Metode *Text mining*

Dalam tahap proses *text mining* dimulai dari tahapan yaitu :

a. Text Pre-processing

Dalam tahapan ini teks yang sudah dikumpulkan akan dilakukan beberapa tahapan pemrosesan yaitu *case folding, noise removal tokenization, filtering (stopword removal)* serta *stemming.*

b. Pelabelan

Dalam tahap ini dilakukan pelabelan dengan menghitung polaritas dari suatu kata menggunakan list kata-kata positif dan negatif dari *https://github.com/masdevid/ID-OpinionWords* . Kemudian dari hasil skor polaritas dari text memberikan informasi terkait kelas yang dapat dimasuki oleh setiap kalimat, yaitu (1,0,-1) untuk skor polaritas > 0 dikategorikan ke dalam kelompok positif, skor polaritas = 0 dikategorikan ke dalam kelompok netral, dan < 0 dikategorikan ke dalam kelompok negatif.

c. Klasifikasi Model menggunakan *Naïve Bayes Classifier*

Dalam tahapan ini dibuat model untuk menghitung skor dari probabilitas masing-masing kata untuk menentukan kelas masing-masing dokumen.

d. Evaluasi

Evaluasi dilakukan dengan menghitung nilai confusion matrix yang menghasilkan nilai *accuracy, precision, recall,* dan *F1-Score*. Hasil evaluasi akan divalidasi menggunakan *10 K-Fold Cross Validation.*

1.7 Sistematika Penulisan

BAB I PENDAHULUAN

Bab ini berisikan tentang latar belakang, identifikasi masalah, rumusan masalah, tujuan dan manfaat penelitian, ruang lingkup, metodologi penelitian, dan sistematika penulisan.

BAB II LANDASAN TEORI

Bab ini berisikan tentang teori-teori yang diambil dari jurnal-jurnal yang berkaitan dengan penyusunan laporan skripsi serta beberapa literatur yang berhubungan dengan penelitian ini.

BAB III ANALISA MASALAH & PERANCANGAN APLIKASI

Bab ini berisikan analisa kebutuhan, konstruksi algoritma atau metode, perancangan database, dan perancangan layar dan menu.

BAB IV PENGUJIAN DAN IMPELEMENTASI

Bab ini berisikan tampilan program, pengujian sistem dengan blackbox testing, pengolahan data kuesioner, dan evaluasi.

BAB V SIMPULAN DAN SARAN

Berisikan kesimpulan dan saran yang berkaitan dengan hasil penelitian yang sudah dilakukan berdasarkan yang telah diuraikan pada bab-bab sebelumnya.

BAB II

LANDASAN TEORI

2.1 Teori Umum

2.1.1 Data

Menurut Kamus Besar Bahasa Indonesia (KBBI), data adalah keterangan yang benar dan nyata sehingga dapat dijadikan bahan dasar kajian (analisis atau kesimpulan) dalam suatu penelitian (Abdillah et al., 2021).

Menurut Maria Teresa Biagetti, University of Rome,Italy (Rusdiana et al., 2014) menyimpulkan bahwa data adalah sebagai berikut:

" Datum/data merupakan sesuatu yang dapat memperluas pengetahuan manusia atau memperluas bidang pengetahuan ilmiah, teoretis, atau praktikal atau yang dapat direkam. Dan lebih lanjut data dapat membangkitkan informasi dan pengetahuan didalam pikiran kita."

Menurut (Abdillah et al., 2021), data dapat diklasifikasikan berdasarkan teknik pengumpulan data dan sumbernya, dapat dibedakan yaitu :

a. Data Kualitatif

Data Kualitatif adalah data yang berupa kata-kata atau kalimat, dan bukan berbentuk angka. Data kualitatif ini dapat diperoleh melalui berbagai macam teknik pengumpulan data misalnya wawancara, analisis dokumen, diskusi atau observasi.

b. Data Kuantitatif

Data kuantitatif adalah data yang berupa angka atau bilangan. Data kuantitatif dapat dianalisis menggunakan teknik perhitungan matematika atau statistika.

c. Data Primer

Data primer merupakan data yang secara langsung didapat dari sumber data yang dikumpulkan oleh peneliti secara langsung dari asal datanya.

d. Data Sekunder

Data ini dikumpulkan tidak langsung diambil dari sumber datanya. Maksudnya data yang diperoleh dikumpulkan dari berbagai sumber yang telah ada.

Dalam penelitian ini, sumber data yang digunakan berasal dari *tweet* pada *twitter* yang merupakan jenis data primer, pengambilan data *tweet* tersebut menggunakan bantuan *Public API* yang sudah disediakan olah *twitter*. Sifat data juga bersifat kualitatif, karena data yang diambil berupa tweet yang menggunakan kata-kata.

2.1.2 *Data Mining*

Menurut sumber (Larose, 2014), beberapa arti *data mining* adalah sebagai berikut:

- " *Data mining* merupakan kumpulan data observasional untuk menemukan relasi yang tidak terduga dan meringkas data dengan metode baru yang dapat dipahami dan bermanfaat bagi pemilik data".
- "*Data mining* adalah salah satu bidang indisipliner yang menggabungkan teknik dari machine learning, pengenalan pola,statistik, database, dan visualisasi untuk mengatasi masalah dalam mengkestrak informasi dari basis data yang besar".

Menurut (Larose, 2014), terdapat beberapa proses yang dilakukan dalam *data mining*, yaitu:

a. Deskripsi

   Untuk menemukan pola tersembunyi dan menntukan pola kedalam suatu aturan yang dapat dipahami oleh para ahli.

b. Estimasi

   Proses ini mirip seperti prediksi kecuali dalam variabel estimasi lebih mengarah ke numerik.

c. Prediksi

Untuk mengklasifikasikan berdasarkan pola atau perilaku yang diperkirakan akan datang.

d. Klasifikasi

Untuk menemukan model fungsional dan mendeskripsikan data berdasarkan kelas-kelasnya.

e. Klustering

Proses dalam menemukan kelompok data tanpa bergantung pada kelas tertentu untuk menemukan sebuah objek.

f. Asosiasi

Untuk menemukan kedekatan atribut yang muncul dalam suatu waktu tertentu.

Dalam *data mining* terdapat tahapan yang dilakukan pada proses *data mining*, yang diawali dengan seleksi data sumber ke data target, tahap preprocessing untuk meningkatkan kualitas data, transformasi, serta tahap interpretasi , evaluasi dan pengembangan pengetahuan baru. Yang dijelaskan secara rinci (Džeroski, 2009) sebagai berikut :

a. Seleksi data

Merupakan proses penyeleksian data dari sekumpulan data operasional yang dilakukan sebelum memasuki tahap penggalian informasi dalam KDD (Knowledge Discovery in Databases) dimulai.

b. Pre-processing/pembersihan

Merupakan proses sebelum *data mining* dilakukan, dalam proses ini dilakukan proses pembersihan data yang menjadi fokus KDD. Proses pembersihan ini sebagai berikut membuang data yang mempunyai duplikat, memeriksa data yang tidak konsisten, dan memperbaiki error pada data.

c. Coding

Merupakan suatu cara untuk mentransformasi data yang telah dipilih, sehingga data tersebut dapat digunakan untuk proses *data mining*.

d. *Data mining*

Merupakan proses untuk mencari pola dan informasi pada data terpilih dengan menggunakan metode tertentu.

e. Interpretasi/evaluasi

Pada tahap ini dilakukan pemeriksaan pola atau informasi yang ditemukan serta proses visualisasi pola/model yang diekstraksi.

2.1.3 Analisis Sentimen

Analisis Sentimen atau yang juga dikenal sebagai opinion mining adalah sebuah bidang studi yang menganalisis opini, evaluasi, sikap, penilaian serta perasaan seseorang tentang hal-hal seperti produk, layanan, organisasi, individu, masalah, peristiwa, topik, dan atributnya. Dalam hal ini, mewakili ruang masalah yang besar. Misalnya *sentiment analysis, opinion mining, opinion analysis, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis dan review mining* (Liu, 2015).

Analisis sentimen ini merupakan sebuah proses menganalisis teks untuk menentukan pesan tersebut berdasarkan emosi yang terlihat seperti positif, negatif atau netral. Analisis sentimen ini dapat menentukan sikap penulis terhadap suatu topik dalam suatu teks secara otomatis. Sehingga data sentimen yang telah di analisis dapat berguna untuk perusahaan yang dapat berguna untuk meningkatkan mutu dan reputasi mereka.

2.1.4 *Twitter*

Twitter merupakan platform sosial media untuk membantu penggunanya mengirimkan dan membaca pesan berbasis teks. *Twitter* didirikan oleh Jack Dorsey

pada bulan Maret 2006 (Sulandari et al., 2018). Saat pertama dirilis *Twitter* membatasi pengguna untuk mengirim pesan hanya sebatas 140 karakter pada satu kali pengiriman, tetapi pada tanggal November 2017 bertambah hingga 280 karakter pada satu kali *tweet*.

Twitter memiliki beberapa fitur utama seperti kicauan (*tweet)* merupakan fitur yang dapat digunakan pengguna untuk membagikan teks,gambar, video maupun gif kepada sesama pengguna *twitter*. Kemudian ada fitur *hashtag '#'* yang dapat digunakan pengguna untuk mencari berdasarkan kategori atau topik yang saling berhubungan, hashtag juga dapat menjadi trending topik di *twitter* untuk melihat pencarian terpopuler saat tersebut.

Twitter juga memiliki API yang dapat diakses secara terbatas oleh publik. *Twitter API* merupakan cara *twitter* untuk mengirimkan atau berhubungan satu sama lain untuk bertukar informasi. *Twitter* mengizinkan akses ke bagian API untuk membangun perangkat lunak baru yang terintegrasi dengan *twitter* (Twitter, 2022).

2.1.5 Website

Website pertama kali dibuat oleh Berners-Lee dan timnya pada tahun 1980-an dalam projek *World Wide Web* (W3)(Kaban & Sembiring, 2021). Tujuan awal tim Beners-Lee membuat web adalah untuk memudahkan para peneliti untuk bertukar informasi pada saat itu. Kemudian pada 30 April 1993, website mulai diperkenalkan kepada publik dan dapat diakses oleh siapapun secara gratis.

Website merupakan halaman yang berisi kumpulan informasi tertentu, yang dapat diakses melalui internet. Website memiliki unsur-unsur yang memungkinkan pengguna dapat mengakses melalui internet seperti berikut, domain (alamat sebuah website), hosting (server dimana semua file website dapat disimpan serta diakses

melalui internet), konten (berupa gambar, video, dan teks), code (HTML,Javascript,dan CSS), dan tampilan website (UI/UX).

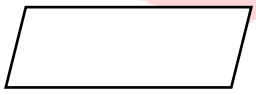
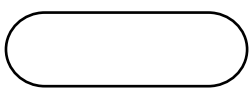
Pada penelitian ini, website digunakan untuk sebagai media untuk menampilkan hasil dari implementasi analisa sentimen pariwisata yang telah diklasifikasikan berdasarkan sentimennya, yang dapat diakses melalui perangkat komputer yang terhubung dengan internet.

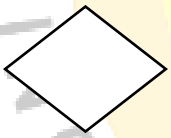
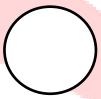
2.2 Teori Khusus

2.2.1 Flowchart

Flowchart adalah algoritma yang ditulis dengan bentuk-bentuk dan simbol. *Flowchart* merupakan suatu representasi penggambaran dari suatu algoritma yang biasa disebut diagram alur yang menggunakan bentuk-bentuk yang berbeda sesuai dengan jenis operasi yang digunakan, kemudian bentuk-bentuk ini dihubungkan dengan garis panah uang menunjukan aliran yang dilalui, aliran ini biasa disebut dengan garis aliran. Secara umum, *Flowchart* mempunyai dua jenis, yaitu *flowchart* program dan *flowchart sistem. Flowchart* mempunyai elemen atau bentuk standar yang sering digunakan (Irawan, 2022):

Tabel 2.1 Komponen Elemen Flowchart

Simbol	Nama	Keterangan
	*Input/Output* (Data)	Simbol ini digunakan untuk setiap proses *input* atau *output* data
	Terminator	Simbol untuk “mulai (*start)* ” atau “selesai (*end/stop)*” dari suatu proses algoritma

	Proses	Simbol ini digunakan pada saat melakukan suatu proses perhitungan dan pengolahan data dalam algoritma
	Garis Alir (*Flow Line*)	Simbol ini digunakan sebagai penanda arah aliran algoritma
	Persiapan (*Preparation*)	Simbol ini digunakan dalam proses pemberian nilai awal atau inisialisai dari algoritma
	*Predefined Process* (Sub Program)	Simbol ini digunakan sebagai proses dalam menjalankan sub program dan awal sub program
	Kondisional (*Decision*)	Simbol ini digunakan dalam kondisi, perbandingan pernyataan, penyeleksian data untuk langkah selanjutnya
	*On Page Connector*	Simbol ini digunakan untuk penghubung bagian-bagian dari *Flowchart* yang berada dalam satu halaman
	*Off Page Connector*	Simbol ini digunakan untuk menghubungkan *Flowchart* pada halaman yang berbeda
	Dokumen	Simbol ini digunakan untuk menjelaskan sebuah dokumen

Sumber : (Irawan, 2022)

2.2.2 *Text Mining*

Text Mining adalah suatu cara untuk menemukan informasi baru yang sebelumnya tidak diketahui oleh komputer dengan mengekstrak suatu informasi dari berbagai sumber, dengan saling menghubungkan informasi yang diekstraksi bersama-sama sehingga membentuk fakta baru atau hipotesis baru agar dapat dieksplorasi lebih lanjut (Hearst, 2017).

Text mining merupakan suatu proses untuk menambang data dalam bentuk text dan sumber data biasanya diambil dari dokumen. Tujuannya adalah untuk menemukan kata-kata yang dapat menggambarkan isi dokumen sehingga hubungan antar dokumen tersebut tersebut dapat dianalisis.

Dari pengertian diatas dapat disimpulkan bahwa *text mining* adalah teknik untuk menemukan pola didalam teks, dan menemukan informasi utama yang mewakili isi dari teks tersebut. Dalam penelitian ini, dilakukan *text mining* memakai data yang bersumber dari *tweet* yang ada pada media sosial *twitter* yang kemudian akan dilakukan klasifikasi dalam bentuk positif, negatif dan netral.

2.2.3 Machine Learning

Machine Learning (ML) merupakan salah satu cabang *Artificial Intelligence* (AI). Machine Learning merupakan teknologi yang dapat mempelajari data yang ada tanpa adanya perintah dari pengguna atau dapat belajar sendiri, ML ini dikembangkan dengan dasaran ilmu statistika, matematika, dan *datamining* sehingga dapat belajar dengan menganalisa data tanpa diperlukan perintah atau diprogram ulang.

Tujuan utama dari *machine learning* adalah mempelajari, merekayasa, dan meningkatkan dan mengembangkan model matematika yang dapat dilatih terus menerus untuk memprediksi dan membuat keputusan (Taherdoost, 2023). Dalam *Machine learning* terdapat dua macam teknik pembelajaran, sebagai berikut:

a. *Supervised Learning*

Teknik *Supervised learning* adalah teknik pembelajaran mesin yang diterapkan pada informasi yang telah ada pada data dengan memberikan label tertentu. Kemudian teknik ini dapat memberikan target terhadap *output* yang dilakukan dengan membandingkan dengan hasil dimasa lalu. Tujuannya adalah melatih sistem yang bekerja dengan sampel data yang belum terlihat sebelumnya

b. *Unsupervised Learning*

Teknik *Unsupervised learning* merupakan teknik yang diterapkan pada *machine learning* yang tidak memiliki informasi dan dapat diterapkan secara langsung, tujuan dari teknik pembelajaran ini untuk menemukan struktur atau pola tersembunyi pada data yang tidak memiliki label.

2.2.4 Text Preprocessing

Dalam dunia *data mining*, terdapat berbagai macam jenis data, diantaranya adalah data yang tidak tersruktur/ *unstructured data* seperti, audio,video,gambar,text, dan lain-lain. Semua data tersebut mempunyai struktur yang berbeda beda sehingga sangat rumit dibandingkan dengan data yang terstruktur yang memiliki bentuk dasar yaitu angka biner 0 dan 1. Oleh karena itu, proses Preprocessing text ini memegang peran penting dalam penerapan *text mining*. Untuk menemukan makna dibalik data-data tersebut. Pengolahan tersebut termasuk proses untuk menyeleksi data text agar menjadi lebih terstruktur dengan melalui tahapan-tahapan antara lain sebagai berikut :

a. *Case Folding*

Dalam proses ini dilakukan pengubahan data tidak terstruktur yaitu penggunaan huruf kapital yang tidak konsisten. Sehingga tahapan ini dilakukan proses untuk mengubah huruf kapital menjadi huruf kecil *(lowercase).*

b. *Tokenizing*

Pada tahapan ini merupakan proses untuk memecah kalimat-kalimat menjadi kata atau yang disebut dengan token. Sehingga kita dapat membedakan antara yang merupakan pemisah kata atau tidak.

c. *Filtering*

Merupakan tahapan untuk proses removing number atau menghilangkan angka, dan removing punctuation yaitu menghilangkan tanda baca, simbol-simbol, alamat website, white space, dan sebagainya.

d. *Stopword Removal*

Merupakan tahapan untuk menghapus kata-kata stop word yaitu memisahkan kata umum dan kata yang tidak memiliki makna dan menghilangkannya. Contohnya kata penghubung seperti “di” dan “seperti” . Dengan dilakukannya penghilangan stopword ini dapat mengurangi ukuran index dan mempercepat waktu pemrosesan.

2.2.5 Application Programming Interface (API)

a. Pengertian API

Application Programming Interface (API) adalah interface yang berfungsi sebagai perantara yang memungkinkan perusahaan untuk membuka data dan fungsionalitas aplikasi mereka kepada pihak pengembang, mitra bisnis, ataupun depertemen dalam perusahaan mereka. API ini berfungsi sebagai perantara untuk berkomunikasi satu sama lain. API juga merupakan salah satu cara dari program komputer berinteraksi satu sama lain dengan meminta dan mengirimkan suatu informasi yang dilakukan dengan mengizinkan aplikasi perangkat lunak untuk memanggil endpoint (alamat unik yang sesuai dengan

2.2.6 Klasifikasi

Menurut Kamus Besar Bahasa Indonesia (KBBI), klasifikasi adalah sistem penyusunan dalam bentuk kelompok atau golongan menurut jenis atau standar yang ditetapkan (Wanto et al., 2020). Sedangkan klasifikasi data merupakan proses mengasosiasi sifat-sifat/karakteristik metadata ke dalam setiap aset digital, untuk mengidentifikasi jenis data yang berhubungan dengan aset tersebut.

Klasifikasi berdasarkan jenis datanya dapat dibagi menjadi dua, yaitu:

a. Data Hitung

Pengklasifikasian ini merupakan suatu data yang didapat dari proses perhitungan dalam bentuk presentase maupun jumlah total nilai hitung data yang tercatat dan tersimpan. Misalnya, data pemilu, data karyawan kantor, data nasabah perbankan dan lain-lain.

b. Data Ukur

Pengklasifikasian ini merupakan suatu cara data menunjukan ukuran dan kapasitas dari suatu nilai. Misalnya, suhu pada thermometer, nilai IPK mahasiswa dan lain-lain.

2.2.7 *Naïve bayes*

Naïve Bayes adalah metode pengklasifikasian yang kuat dan mudah dilatih untuk menentukan probabilitas suatu hasil dari serangkaian kondisi menggunakan teorema Bayes (Taherdoost, 2023).

Klasifikasi *Naïve Bayes* merupakan klasifikasi yang bersifat *supervised learning* karena memiliki *supervisor* (klasifikasi dilakukan secara manual pada data yang digunakan dalam proses pelatihan) dalam proses pembelajaran atau learning (Gunawan et al., 2018). *Naïve Bayes* merupakan metode dalam pengklasifikasian yang sering digunakan dalam sentimen analisis karena penerapannya sederhana dan mudah dalam melakukan pengklasifikasian dokumen.

Penerapan metode klasifikasi *Naïve Bayes* dengan mempertimbangkan dua probabilitas A dan B, yang dapat dikorelasikan dengan P(A) dan P(B) dengan probabilitas bersyarat P(A|B) dan P(B|A) yang dirumuskan menggunakan persamaan:

$$\mathrm{P(A}|B)=\frac{p(B|A)\ P(A)}{p(B)}$$

Keterangan:

A : Hipotesis data merupakan suatu class spesifik.

B : Data dengan kelas yang masih belum diketahui.

P(A | B) : Probabilitas hipotesis berdasarkan kondisi.

P(A) : Probabilitas hipotesis.

P(B | A) : Probabilitas berdasarkan kondisi hipotesis.

P(B) : Probabilitas terjadinya B.

Jika Probabilitas perkiraan data latih yang digunakan untuk model Naïve Bayes menghasilkan nilai nol, maka hasil klasifikasi akan kurang baik. Untuk menghindari adanya probabilitas nol maka digunakan *add-one* atau *Laplacian Smoothing.* Ada penambahan +1 pada pembilang, dan |V| pada penyebut (Noto & Saputro, 2022). Berikut merupakan rumus dari *Naïve bayes* dengan menambahkan *Laplacian Smoothing* :

$$\mathrm{P(C}|F_i)=\frac{p(F_i|C)*P(C)+k}{P(F_i)+k(F_i)}$$

Keterangan:

$\mathrm{P(C}|F_i)$: Hipotesis kelas *C* terjadi jika bukti F_i terjadi (probabilitas posterior).

P(C) : Probabilitas hipotesis kelas C terjadi.

P($F_i|C$): Hipotesis kelas *C* terjadi jika bukti F_i terjadi (probabilitas posterior).

P(F_i) : Probabilitas terjadinya F_i

K : Nilai pelancaran *Laplace*.

$K(F_i)$: Jumlah kelas atribut F_i.

2.2.8 Confusion matrix

Confusion matrix digunakan untuk menggambarkan kinerja model klasifikasi pada kumpulan data yang nilai sebenarnya diketahui (Senthilselvi et al., 2021). *Confusion matrix* adalah tabel yang mengkategorikan prediksi dengan nilai yang cocok dengan nilai sebenarnya. *Confusion matrix* adalah tabel dengan 4 kombinasi berbeda dari nilai yang diprediksi dan nilai aktual. Ada 4 istilah yang merupakan hasil dari *confusion matrix* yaitu True Negative (TN), False Positive (FP), False rumu

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} * 100\%$$

$$Presisi = \frac{TP}{FP+TP} * 100\%$$

$$Recall = \frac{TP}{FN+TP} * 100\%$$

Sumber : (Senthilselvi et al., 2021)

Gambar 2.1 Rumus evaluasi *confusion matrix*

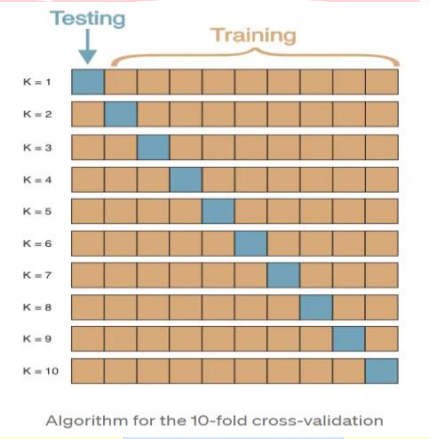
2.2.9 K-Fold Cross Validation

K-Fold cross validation adalah salah satu metode validasi model yang paling banyak digunakan dalam pemodelan statistik dan pembelajaran mesin. Tujuannya adalah untuk mengevaluasi keefektifan model dengan menguji model dengan beberapa himpunan bagian data yang berbeda.

Dalam validasi silang K-Fold, data dibagi secara acak menjadi k himpunan bagian berukuran sama. Setiap partisi disebut "fold". Model dilatih dengan lipatan misalnya, k-1 dan diuji dengan sisa lipatan. Proses ini diulang sebanyak k kali,

sehingga setiap lipatan digunakan 1 kali sebagai data uji (Belyadi & Haghighat, 2021).

Selama proses validasi silang, metrik evaluasi seperti skor presisi, *accuracy*, *recall* atau *F1-score* dihitung untuk setiap iterasi. Kemudian nilai-nilai metrik tersebut dirata-ratakan untuk mendapatkan perkiraan kinerja model secara keseluruhan.



Sumber: (Belyadi & Haghighat, 2021)

Gambar 2.2 *10 K-fold cross validation*

2.3 Teori Perancangan

2.3.1 HTML

HTML pada awalnya merupakan hasil dari pengembangan sebuah sistem untuk membagi dokumen yang ditujukan untuk ilmuan CERN (Organisasi Eropa untuk Riset Nuklir) yang dikembangkan oleh Berners-Lee dan tim pada tahun 1980 (Kaban & Sembiring, 2021). HTML merupakan sistem *markup* berbasis internet. Fungsi utama dari HTML adalah sebagai sistem untuk mengelola data dan informasi dari suatu dokumen shingga dapat diakses dan ditampilkan melalui internet melalui media website, tanpa adanya HTML browser internet tidak akan dapat menampilkan konten didalam website.

HTML dapat dikombinasikan dengan penggunaaan bahasa pemrograman lain antaralain, bahasa CSS (Cascade Style Sheet) serta JavaScript. Peran HTML untuk menyusun kerangka dan struktur pada halaman website. Kemudian, setelah itu CSS membantu untuk menyusun tampilan dan desain yang meliputi warna, format font, dan susunan layout dan lain-lain. Kemudian JavaScript berperan sebagai dasar logika pemrograman untuk memberikan website untuk bekerja secara interaktif dengan pengguna secara langsung.

2.3.2 PHP (Hypertext Preprocessor)

PHP (*Hypertext Preprocessor*) adalah bahasa pemrograman open source untuk web *Server-Side Scripting* (Adi, 2022). PHP adalah Bahasa pemrograman yang dapat diintegrasikan ke dalam HTML yang telah dikompilasi sehingga berada di server (server-side HTML *script*). PHP juga dapat dihubungkan ke database seperti MySQL. Bahasa pemrograman PHP dapat digunakan untuk membuat halaman web yang dinamis. Dinamis, artinya halaman yang akan ditampilkan dibuat saat klien meminta halaman tersebut. Dengan bantuan sistem mekanisme dinamis ini, dimungkinkan informasi yang diterima klien selalu diperbarui. Di PHP, semua *script* dieksekusi di server tempat skrip dieksekusi. PHP merupakan bahasa pemrograman yang banyak digunakan dalam pengembangan web karena memiliki beberapa keunggulan dibandingkan bahasa pemrograman sejenis lainnya. Berikut ini adalah kelebihan dari bahasa pemrograman PHP:

1. PHP adalah bahasa mutliplatform, artinya dapat bekerja diberbagai sistem operasi seperti,(Linux, Unix, Macintosh, Windows) dan dapat dieksekusi saat runtime melalui konsol dan juga dapat menjalankan perintah-perintah sistem lainnya.
2. PHP bersifat open source, artinya siapa saja bisa menggunakannya dengan bebas.

3. Banyak server yang telah support untuk bahasa pemrograman PHP seperti (Apache, IIS, Lighttpd, nginx, Xitam) dengan pengaturan relatif sederhana, dan banyak yang menyediakan dalam bentuk package (PHP, MySQL dan Web Server).
4. Di sisi pengembangan, karena banyak milis, lebih mudah komunitas dan pengembang siap membantu pengembangan.
5. Menurut pengertiannya, PHP adalah bahasa scripting yang paling mudah karena memiliki banyak referensi.
6. Banyak program dan program PHP gratis dan siap pakai seperti Wordpress, PrestaShop dan lainnya.
7. Dapat mendukung banyak database seperti MySQL, Oracle, MS-SQL, dll.

2.3.3 Bootstrap

Bootstrap adalah *framework* front-end yang intuitif dan kuat untuk membangun aplikasi web lebih cepat dan lebih mudah (Pranaya & Hendra, 2019). *Bootstrap* menggunakan HTML, CSS dan Javascript. *Bootstrap* juga merupakan proyek *open source*, sehingga pengembang bebas menggunakannya.

Bootstrap memungkinkan developer untuk mengembangkan website dengan mudah dan cepat. Pengembang hanya perlu memanggil kelas untuk membuat berbagai fungsi di situs web seperti tombol, panel, tabel, notifikasi, dll.

2.3.4 XAMPP

XAMPP adalah bahasa pemrograman *open source* yang dibuat oleh grup Apache. Penyebaran Apache untuk server Apache, MariaDB, PHP dan Perl adalah bagian dari *XAMPP*. Basis dibentuk oleh server lokal atau lokal. Baik laptop maupun PC bisa menggunakan server lokal ini. *XAMPP* digunakan untuk menguji klien atau situs web sebelum menyebarkannya ke server web jarak jauh.

Menurut (Khozaimi, 2020), *XAMPP* adalah sistem perangkat lunak yang skema namanya berasal dari singkatan Apache, MySQL atau MariaDB, PHP dan Perl. Setiap huruf dalam nama XAMPP memiliki arti terjemahan sebagai berikut:

a. **X**: Xampp adalah perangkat lunak yang bersifat *cross platform*, yang dapat digunakan untuk sistem operasi seperti, Windows, Linux dan Mac.
b. **A**: Xampp terinstall pada Web server Apache Web server.
c. **M**: Xampp terinstal pada DBMS Mysql.
d. **P**: Xampp terinstall pada PHP.
e. **P**: Xampp terinstall dalam bahasa Pearl.

2.3.5 MySQL

Database MySQL merupakan suatu software database yang berbentuk *Relational Database Management System* (RDBMS) (Indrawan, 2021). MySQL merupakan suatu program database server yang dapat mengirim dan menerima data dengan sangat cepat, dapat digunakan *multi-user* dan menggunakan perintah dasar dari SQL (*Structured Query Language*). MySQL dapat digunakan dengan gratis karena memiliki lisensi FreeSoftware dibawah lisensi GNU/GPL (*General Public License*). MySQL dapat juga digunakan sebagai server, dan program kita sebagai *client*. MySQL dapat digunakan sebagai server dan *client* dari sebuah database.

2.3.6 Scikit-learn

Scikit-learn merupakan pustaka *open source* untuk pembelajaran mesin. Library ini mensupport penggunaan dari berbagai algoritma seperti KNN, SVM, Random forest dan *Naïve bayes*. *Scikit-learn* dapat membantu hal-hal seperti preprocessing, classification, regression, clustering dan seleksi model (Nordeen, 2020). *Scikit-learn* digunakan untuk pemodelan statistik, pembelajaran mesin, dan

tugas ilmu data. *Scikit-learn* menyediakan berbagai algoritma dan fungsi untuk membangun, mengevaluasi, dan menerapkan model ke data.

2.3.7 *Twitter* API

Twitter merupakan platform media sosial yang berbagi informasi seluas mungkin, Oleh karena itu, *twitter* memberikan akses kepada perusahaan, pengembang, dan pengguna yang terprogram ke data *twitter* melalui API. *Twitter* mengizinkan akses informasi melalui API yang memungkinkan pengembang membangun sebuah perangkat lunak yang terhubung dengan *twitter*. *Twitter* juga mendukung API yang memungkinkan pengguna untuk mengelola informasi *Twitter* non-publik mereka sendiri (Direct Message) dan memberi izin untuk pengembang untuk mengaksesnya (*Twitter*, 2022).

Twitter memberikan akses API kepada pengembang untuk membuat aplikasi yang terintegrasi kepada *twitter* dengan mendaftarkan aplikasi tersebut. API merupakan cara komputer untuk berkomunikasi satu sama lain sehingga dapat meminta dan menyampaikan informasi. Proses ini dapat dilakukan dengan memberi izin perangkat lunak untuk memanggil endpoint. Secara default, aplikasi hanya diizinkan hanya untuk mengakses informasi publik di *twitter*.

Terdapat lima grup utama endpoint dalam *Twiter API* yaitu:

a. Account dan users

   Twitter mengizinkan para pengembang untuk melakukan pengelolaan profil dan pengaturan akun untuk membisukan, memblokir *users* serta mengelola pengikut dan pengguna.

b. Tweet dan replies

   Twitter mengizinkan pengembang untuk mengakses *tweet* dengan mencari kata kunci tertentu dan meminta sampel *tweet* dari akun tertentu.

c. Direct Message

Mengizinkan untuk mengakses percakapan pribadi yang secara eksplisit memberikan izin kepada aplikasi tertentu.

d. Ads

Pengembang dapat menggunakan *tweet* publik untuk mengidentifikasi minat dan topik pengguna sebagai alat bisnis untuk menjalankan iklan.

e. Publisher tools dan SDKs

Twitter menyediakan alat bagi pengembang untuk menyematkan timeline *twitter*, tombol berbagi dan konten pada halaman web *twitter*.

Pengambilan data melalui *twitter* akan dilakukan setelah mendapatkan API dari *twitter*, selanjutnya kita akan memperoleh hasil dari API *twitter* berupa :

a. consumer_key,
b. consumer_secret_key,
c. access_token,
d. acces_secret_key

2.4 Tinjauan Studi

***a. Tourism Companies Assessment via Social Media Using Sentiment Analysis* oleh Nadia F. AL-Bakri1, Janan Farag Yonan, Ahmed T. Sadiq, Ali Sami Abid**

Penelitian pada tahun 2021 yang dilakukan oleh Nadia F. AL-Bakri1, Janan Farag Yonan, Ahmed T. Sadiq, Ali Sami Abid yang berjudul "*Tourism Companies Assessment via Social Media Using Sentiment Analysis"* bertujuan untuk menilai keberhasilan perusahan pariwisata di Irak.

Untuk menganalisa sentimen pengguna Facebook, dilakukan model ekstraksi yang ditujukan pada satu set ulasan dari halaman publik `Tampilan Menentang`

di Facebook menggunakan software yaitu QSR Nvivo 11 untuk menganalisis data yang tidak terstruktur.

Dikumpulkan sejumlah 14200 komentar facebook dalam bahasa Arab dari 71 halaman perusahaan di Facebook, data yang dikumpulkan akan dipisahkan sesuai masing-masing perusahaan untuk diproses lebih lanjut.

Analisis yang dilakukan menggunakan tiga jenis algoritma machine learning yang berbeda yaitu Rough Set Theory (RST), *Naïve Bayes* (NB) and K-Nearest Neighbors (KNN).

Hasil penelitian ini adalah dengan membandingkan tiga algoritma pada machine learning yang berbeda didapatkan hasil sebagai berikut, Rought Set Theory memberikan *accuracy* kalsifikasi rasio lebih baik dibanding KNN dan *Naïve Bayes* yang menghasilkan *accuracy* sebesar 89.3% dan F-score sebesar 93.5%. Berdasarkan hasil pengujian diketahui bahwa dari 71 perusahaan pariwisata di Irak yang dievaluasi, 28% rating sangat baik, 26% rating baik, 31% rating sedang, 4 dari perusahaan ini memiliki peringkat penerimaan sedang dan 11 dari perusahaan tersebut memiliki peringkat buruk.

Kekuatan pada penelitian ini adalah komentar Facebook dapat digunakan untuk melihat penilaian terhadap suatu perusahaan di Iraq dengan menggunakan data yang banyak yang telah dikumpulkan untuk melakukan analisis sentimen.

Hal yang menjadi kelemahan pada penelitian ini, diidentifikasi bahwa bahasa Arab dan dialek Irak berdasarkan analisis sentimen untuk perusahaan pariwisata tidak diusulkan karena dari penelitian dari beberapa jurnal sebelumnya, belum ada studi khusus untuk pengambilan dari kata bahasa Arab untuk mengembangkan solusi untuk masalah penelitian dan tidak adanya beberapa fitur semantik dan tidak mencoba untuk mengklasifikasikan seluruh komentar karena hanya

mengklasifikasikan kata-kata Feature Selection menggunakan PSO (Particle Swarm Optimization).

b. Analisis Sentimen Pariwisata di Kota Malang Menggunakan Metode Naive Bayes dan Seleksi Fitur Query Expansion Ranking oleh Shima Fanissa1 , M. Ali Fauzi , Sigit Adinugroho

Penelitian pada tahun 2018 yang dilakukan oleh Shima Fanissa1 , M. Ali Fauzi , Sigit Adinugroho yang berjudul "Analisis Sentimen Pariwisata di Kota Malang Menggunakan Metode Naive Bayes dan Seleksi Fitur Query Expansion Ranking" bertujuan untuk menganalisis komentar dari masyarakat tentang objek wisata di Kota Malang untuk mencari sentimen positif dan negatif dari website TripAdvisor.

Dalam menganalisis sentiment pada website TripAdvisor menggunakan proses yaitu pengambilan data, preprocessing, seleksi fitur menggunakan metode *Query Expansion Ranking* ,dan pengklasifikasian menggunakan *Naïve bayes*.

Pengumpulan data dilakukan dengan cara mengambil data dari website TripAdvisor. Selanjutnya hasil data akan melalui proses preprocessing agar data tidak terstruktur dapat siap digunakan untuk proses klasifikasi.

Pada proses preprocessing ini terdiri dari tokenisasi dengan memisahkan dokumen menurut tokennya , filtering untuk menyeleksi kata-kata yang akan dibuang, dan stemming menggunakan library Sastrawi.

Setelah itu dilakukan pemberian seleksi fitur yang bertujuan untuk mengurangi dimensi fitur dengan metode Query Expansion Ranking dipadukan dengan metode Multi nomial *Naïve bayes*. Hasil penelitian ini dengan

menggunakan algoritma Query Expansion Ranking menghasilkan *accuracy* tertinggi sebesar 86.6 pada seleksi fitur 75%.

c. The influence of fake accounts on sentiment analysis related to COVID-19 in Indonesia oleh Rivanda Putra Pratama, Aris Tjahyanto

Penelitian pada tahun 2021 yang dilakukan oleh Rivanda Putra Pratama, Aris Tjahyanto yang berjudul "The influence of fake accounts on sentiment analysis related to COVID-19 in Indonesia" berfokus untuk melakukan analisis pada komentar berdasarkan subjek dari akun yang mempostingnya.

Dengan maraknya penggunaan akun palsu di media sosial menyebabkan kredibilitas dari opini berkurang. Berdasarkan permasalahan tersebut, penelitian ini melakukan eksperimen terhadap analisis sentimen menggunakan pendekatan machine learning untuk mengkategorikan akun palsu untuk melihat dampak akun palsu terhadap analisis sentimen.

Data yang didapat diambil menggunakan teknik crawling dari data tweet pada akun @KemenkesRI terkait dengan COVID-19 pada Januari 2021 hingga Juni 2021 sebanyak 4.170 tweet, sumber data ini dipilih karena terdapat banyak informasi,berita,masukan dan pengaduan masyarakat terkait COVID-19 pada akun ini.

Kemudian menggunakan tools Tweetbotornot berbasis bahasa R untuk mendeteksi data akun palsu atau asli. Selanjutnya didapatkan 1952 tweet dari akun palsu yang akan dihapus.

Setelah itu data dari akun asli akan dilakukan proses preprocessing, dengan menghilangkan stopwords, mengubah menjadi bentuk huruf kecil untuk

memudahkan proses pengklasifikasian. Lalu dilakukan proses labeling sesuai jenis tweet.

Dari hasil pengujian yang dilakukan diketahui bahwa klasifikasi sentimen pada kedua algoritma menunjukkan bahwa proses klasifikasi sentimen tanpa menggunakan akun palsu lebih baik daripada menggunakan semua data tweet. Dari hasil pengujian ini juga diketahui bahwa metode SVM memiliki nilai *accuracy* tertinggi sebesar 80,6% sedangkan *Naïve Bayes* hanya mendapat *accuracy* sebesar 59%.

d. Analisis Sentimen Penilaian Tempat Tujuan Wisata Kota Tegal Berbasis *Text mining* oleh Oman Somantri , Dairoh

Penelitian pada tahun 2019 yang dilakukan oleh Oman Somantri, Dairoh yang berjudul "Analisis Sentimen Penilaian Tempat Tujuan Wisata Kota Tegal Berbasis *Text mining* " bertujuan untuk mencari model sistem untuk memberikan informasi pendukung keputusan untuk para pengolah tempat wisata dan wisatawan yang ingin berkunjung berdasarkan review pengunjung sebelumnya.

Data penelitian yang didapat dari komentar-komentar dari opini pengunjung yang diperoleh melalui situs www.google.com/maps pada tahun 2017 sampai 2018. Data yang diperoleh berjumlah sebanyak 120 data file teks dengan isi jumlah kata yang berbeda tiap filenya.

Kemudian setelah data dalam jangka waktu tersebut diambil, dilakukan proses preprocessing. Tokenisasi yang merupakan proses penghilangan karakter simbol dan angka. Lalu dilakukan filterisasi token, dalam proses ini dilakukan pembatasan panjang karakter minimum dan maksimum. Setelah itu dilakukan proses untuk menghilangkan Stopword. Kemubdian dilakukan pembobotan TF-

IDF, untuk mendapatkan nilai bobot dari kata inputan teks yang akan dimasukan kedalam model yang akan digunakan.

Lalu hasil dari pembobotan, dilakukan proses validasi menggunakan K-fold Cross Validation. Proses ini digunakan untuk melihat tingkat *accuracy* dari model yang diusulkan.

Dari pengujian yang dilakukan dengan menggunakan metode *Naïve Bayes* dan Decision Tree, tingkat *accuracy Naïve Bayes* menghasilkan 77,50% lebih baik dibandingkan dengan menggunakan Decision Tree yang menghasilkan tingkat *accuracy* 60,83%.

e. Sentiment Analysis Objek Wisata Kalimantan Barat Pada Google Maps Menggunakan Metode Naive Bayes oleh Rifa'i A, Sujaini H, Prawira D

Penelitian berjudul ``Analisis Sentimen Objek Wisata Kalimantan Barat di Google Maps Menggunakan Metode Naive Bayes" yang dilakukan oleh Rifa`i A, Sujaini H, dan Prawira D pada tahun 2021 menyajikan analisis sentimen pada tempat wisata di Kalimantan Barat yang bertujuan untuk membangun sistem yang menyediakan eksekusi. Berdasarkan data peringkat Google Maps.

Untuk menganalisis sentimen pengguna Google Maps, data Google Maps di-crawl menggunakan Google Maps API.Kemudian dilakukan proses preprocessing teks menggunakan case, folding, tokenization, filtering, dan stemming. Kami kemudian melakukan pembobotan kata menggunakan TF-IDF.

Analisis dilakukan dengan menggunakan metode klasifikasi *Naïve bayes*. Sistem yang dibuat dalam penelitian ini adalah sistem berbasis web yang menggunakan bahasa pemrograman PHP untuk membuat website, bahasa

pemrograman *Python* untuk melakukan pengolahan data, dan MySQL sebagai databasenya.

Hasil Penelitian ini adalah Sistem ini dapat melakukan analisis sentimen dan mampu mengklasifikasikan sentimen, dan berdasarkan hasil pengujian terhadap 40 responden mendapatkan nilai 84,85% kedalam kategori baik sekali dan mendapatkan *accuracy* paling tinggi sebesar 76% berdasarkan pengujian confusion matrix.

Kekurangan penelitian ini, masih terdapat limitasi dalam sistem karena sistem masih belum mampu untuk mendapatkan klasifikasi ulasan dengan *accuracy* yang lebih tinggi, karena pengaruh karakteristik data testing dan data uji yang digunakan.

f. Sentiment analysis of *twitter* data related to Rinca Island Sentiment analysis of *twitter* data related to Rinca Island development using Doc2Vec and SVM and logistic regression as classifier oleh oleh Hidayat T, Ruldeviyani Y, Aditama A, Madya G, Nugraha A, Adisaputra M

Penelitian berjudul ``Sentiment analysis of *twitter* data related to Rinca Island development using Doc2Vec and SVM and logistic regression as classifier '' yang dilakukan oleh Hidayat T, Ruldeviyani Y, Aditama A, Madya G, Nugraha A, Adisaputra M pada tahun 2021 bertujuan untuk melakukan analisis terhadap sentimen masyarakat tentang pengembangan pada pulau Rinca.

Penelitian ini akan berfokus kepada klasifikasi sentimen yang diklasifikasikan sebagai positif,negatif, dan netral. Menggunakan dua model Doc2Vec yaitu model terdistribusi dan bag of word terdistribusi dan menggunakan svm serta logistic regression classifier. Penggunaan Doc2Vec ini bertujuan untuk menemukan kesamaan antar kalimat/paragraf.

Dilakukan pengumpulan data tweet sejumlah 8000 data tweet yang berisi kata kunci menggunakan *Twitter* API menggunakan proses crawling,kemudian dilakukan penghapusan data yang terdapat duplikat dan tidak relevan, setelah dilakukan pembersihan didapatkan 1038 data tweet. Setelah itu data diberikan tag dengan nilai -1,0,1 yaitu yang menolak,netral dan mendukung pembangunan. Kemudian dilakukan proses pre-processing text.

Analisis dilakukan dengan tiga algoritma pengklasifikasian yaitu SVM dan Logistic Regression. Dan dilakukan penerapan SVM dan Logistic Regression dengan menggunakan library Gensim dan modul Sklearn untuk model bag of word terdistribusi dan proses analisis sentimen menggunakan bahasa pemrograman *Python* dengan memori terdistribusi.

Hasil yang didapat dari analisis dengan model algoritma PV-DBOW dengan SVM, PV-DM dengan SVM, dan regresi logistik mendapat tingkat *accuracy* dan F1-score yang lebih tinggi dibandingkan model lainnya dan mendapatkan 87% sebagai hasil *accuracy* terbaik.

Kekurangan dari penelitian ini adalah data yang digunakan untuk pelatihan dan pengujian tidak seimbang sehingga data terpolarisasi sehingga saran untuk penelitian dimasa depan harus memastikan dataset yang dimiliki seimbang antara label serta penelitian harus mencoba untuk mengklasifikasikan berdasarkan topik tidak hanya berdasarkan sentimen.

g. Implementasi Metode K-Means dan *Naïve Bayes* Classifier untuk Analisis Sentimen Pemilihan Presiden (Pilpres) 2019 oleh Kurniawan I, dan Susanto A

Penelitian berjudul "Implementasi Metode K-Means dan *Naïve Bayes* Classifier untuk Analisis Sentimen Pemilihan Presiden (Pilpres) 2019 " yang dilakukan oleh Kurniawan I, dan Susanto A pada tahun 2019 bertujuan untuk

mengetahui opini masyarakat terhadap pemilu terhadap analisis dokumen text untuk sentimen positif dan negatif.

Untuk menganalisa sentimen di *Twitter*, dilakukan pengambilan data sebanyak 500 tweet dari dua hastag #2019GantiPresiden dan #2019TetapJokowi data diperoleh dengan menggunakan tools Rstudio dan *twitter* API dengan cara Scrapping.

Dalam menganalisasi sentiment pada *twitter* mereka melakukan proses sebagai berikut pengumpulan data, preprocessing, pelabelan, klasifikasi sentiment, hasil *accuracy* sentimen.

Setelah itu dilakukan analisa sentimen pada *twitter* menggunakan proses Pre-Processing data yaitu normalisasi untuk menghilangkan fitur yang tidak dibutuhkan ,case folding untuk mengubah huruf besa menjadi kecil agar mempermudah dalam proses membaca dokumen, tokenizing,pada tahapan ini kalimat tweet akan dilakukan pemisahan kata , stopword removal untuk menghapus kata penghubung dan stemming dilakukan untuk menghapus imbuhan dengan algoritma dari Nazief dan Adrian.

Hasil dari data yang telah dikumpulkan kemudian diklastering menjadi dua kelas yaitu positif dan negatif. Pengelompokan data training dilakukan penglabelan dokumen dengan memanfaatkan algoritma K-means dan data diambil sebanyak 500 data yang akan diuji yang memiliki 175 kelas positif dan 325 kelas negatif, setelah itu akan dilakukan pengujian menggunakan algoritma naive bayes.

Tahap akhir dilakukan pengujian pengklasifikasian tweet yang meliputi *accuracy,* dan error rate. Hasil dari penelitian ini yang menggunakan metode K-means untuk melakukan pembobotan dan Naive Bayes untuk klasifikasi

menghasilkan *accuracy* rata-rata sebesar 93.35% dan error rate rata-rata sebesar 6.66% dari 100 dan 150 data uji.

h. Sentiment Analysis of Social Media *Twitter* with Case of Anti-LGBT Campaign in Indonesia using *Naïve bayes*, Decision Tree and Random Forest Algorithm oleh Fitri V, Andreswari R, Hasibuan M

Penelitian pada tahun 2019 yang dilakukan oleh Fitri V, Andreswari R, Hasibuan M yang berjudul "Sentiment Analysis of Social Media *Twitter* with Case of Anti-LGBT Campaign in Indonesia using *Naïve bayes*, Decision Tree, and Random Forest Algorithm" bertujuan untuk melihat opini masyarakat terhadap kampanye Anti-LGBT di Indonesia.

Untuk melakukan analisa sentimen pengguna *Twitter* dilakukan pengambilan data menggunakan *Python*, data yang diambil berdasarkan kata kunci dari hashtag yang berkaitan.

Data yang diambil dilakukan pemrosesan menggunakan proses pre-processing yaitu case folding untuk mengubah huruf besa menjadi kecil agar mempermudah dalam proses membaca dokumen, tokenizing pada tahapan ini kalimat tweet akan dilakukan pemisahan kata , stopword removal untuk menghapus kata penghubung dan stemming dilakukan untuk menghapus imbuhan.

Setelah itu, data dipisahkan menjadi data training sebesar 75% dan data testing sebesar 25%. Kemudian dilakukan klasifikasi menggunakan algoritma Naive Bayes pada data training, setelah itu hasil klasifikasi diterapkan pada data testing.

Dilakukan analisa dengan data yang telah dibersihkan sebanyak 3.744 komentar. Analisa dilakukan dengan mencoba tiga metode berbeda yaitu *Naïve bayes*, decision tree, and random forest algorithm.

Hasil dari pengujian yang dilakukan dengan membandingkan tiga model algoritma yang berbeda didapatkan hasil sebagai berikut, Naive Bayes menghasilkan *accuracy* sebesar 83,43% dengan recall,precision dan F1-Measure antara 56% dan 65% memberikan *accuracy* yang lebih baik dibanding Decision Tree dan Random Forest yang menghasilkan *accuracy* sebesar 82,91% dengan recall,precision dan F1-Measure antara 27,66% dan 33,33%.

i. **Sentiment Analysis of Positive and Negative of YouTube Comments Using *Naïve Bayes* – Support Vector Machine (NBSVM) Classifier oleh Muhammad N, Bukhori S, Pandunata P**

Penelitian pada tahun 2019 yang dilakukan oleh Muhammad N, Bukhori S, Pandunata P yang berjudul "Sentiment Analysis of Positive and Negative of YouTube Comments Using *Naïve Bayes* –Support Vector Machine (NBSVM) Classifier" bertujuan untuk mendapatkan informasi sentimen terhadap komentar di Youtube.

Untuk melakukan analisis sentimen dilakukan pengumpulan data eksperimen yang didapat melalui komentar youtube menggunakan API pada youtube dengan teknik crawling dan menggunakan pendekatan NLP untuk mendapatkan hasil yang lebih akurat.

Data yang diperoleh didapat dari youtube dengan kategori pendidikan yaitu channel "Kok Bisa?" dengan jumlah subscriber 1.416.935 subscriber dan view 142.957.953.

Pengujian model dilakukan dengan menguji model klasifikasi Naive Bayes dan SVM. Penelitian ini menggunakan 3 skala data yang berbeda yaitu skenario data 60:40, skenario data 70:30 dan skenario data 80:20. Kemudian akan dilakukan

pengujian menggunakan confusion matrix untuk membandingkan *accuracy* tiap-tiap model.

Data yang akan digunakan menggunakan rasio yang dibagi menjadi 70% data latih dan 30% data uji. Data pelatihan yang digunakan berjumlah 233 komentar dengan dua label yaitu positif dan negatif. Sedangkan data uji berjumlah 100 komentar. Setelah itu dilakukan proses preprocessing dengan 4 tahapan yaitu casefolding, tokenization, filtering and stemming.

Selanjutnya dilakukan analisa dilakukan dengan metode klasifikasi Naive Bayes untuk menentukan probabilitas terjadinya data positif dan negatif pada data training. Kemudian *Naïve Bayes* dengan rasio jumlah log berfungsi untuk mengubah teks data menjadi vektor fitur yang akan diproses dalam Support Vector Machine.

Hasil klasifikasi *Naïve Bayes* dan SVM menjadi sangat optimal bila data training yang digunakan memiliki jumlah data yang bervariasi. Hasil dari pengujian yang diperoleh dari komentar video YouTube, yaitu kombinasi *Naïve Bayes* dan Support Vector Machine menghasilkan tingkat *accuracy* yang lebih baik dengan menggunakan data skala 7:3 yaitu 70% data pelatihan dan 30% data pengujian. Dengan menghasilkan yang tertinggi nilai performance test yaitu presisi 91%, recall of 83% dan skor f1 87%.

j. *Twitter* Sentiment Analysis towards COVID-19 Vaccines in the Philippines Using *Naïve Bayes* oleh Villavicencio C, Macrohon J, Inbaraj X,Jeng J, Hsieh J

Penelitian pada tahun 2021 yang dilakukan oleh Villavicencio C, Macrohon J, Inbaraj X,Jeng J, Hsieh J yang berjudul "*Twitter* Sentiment Analysis towards COVID-19 Vaccines in the Philippines Using *Naïve bayes*" bertujuan untuk menilai reaksi masyarakat terhadap Vaksin COVID-19 di Filipina.

Untuk melakukan analisis, dilakukan pengambilan data tweet dengan API pengambilan data diambil dari 1 hingga 31 maret 2021 yang merupakan bulan pertama dari program vaksinasi yang dilakukan pemerintah, dikumpulkan sebanyak 11.974 tweet. Setelah data dibersihkan terdapat sebanyak 993 tweet.

Kemudian dilakukan anotasi data secara manual untuk mengklasifikasikan tweet menjadi tiga yaitu: positif,netral, dan negatif. Setelah itu dilakukan pemrosesan data menggunakan operator yang terdapat pada rapid miner yaitu nominal to text, replace tag,transform case,filtering, tokenisasi , dan penghapusan stopwords. Pengklasifikasian ini menggunakan bahasa inggris dan tagalog.

Setelah itu dilakukan analisa menggunakan algoritma kalsifikasi naive bayes untuk mengklasifikasikan tweet berdasarkan polaritasnya. Hasil dari pengujian dari 993 tweet, sebanyak 828 atau 83,38%mendapatkan respon positif dan 82 atau 8,26% mendapatkan respon negatif. Hasil dari pengujian dan pelatihan data didapatkan bahwa model yang menerapkan algoritma Naive Bayes mendapatkan *accuracy* sebesar 81,77%.

k. Rangkuman Model Penelitian

Tabel 2.2 Rangkuman Model Penelitian

Peneliti	**Nama Jurnal**	**Tahun**	**Institusi**	**Judul dan Metode yang digunakan**	**Kesimpulan**
Nadia F. AL-Bakri1, Janan Farag	Baghdad Science Journal : Vol. 19 No. 2	2021	Ministry of Higher Educatio n &	***Tourism Companies Assessment via Social Media Using***	Berdasarkan hasil pengujian diketahui bahwa dari 71 perusahaan

Yonan, Ahmed T. Sadiq, Ali Sami Abid	(2022): Issue 2, Hal. 422-429		Scientific Research of Iraq	***Sentiment Analysis***	pariwisata di Irak yang dievaluasi, 28% rating sangat baik, 26% rating baik, 31% rating sedang, 4 dari perusahaan ini memiliki peringkat penerimaan sedang dan 11 dari perusahaan tersebut memiliki peringkat buruk.
Shima Fanissa1, M. Ali Fauzi , Sigit Adinugroho	Jurnal Teknologi Informasi Dan Ilmu Komputer : e-ISSN: 2548964X, Vol.	2018	Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas	Analisis Sentimen Pariwisata di Kota Malang Menggunakan Metode Naive Bayes dan Seleksi Fitur Query Expansion Ranking	Dalam menganalisis sentiment pada website TripAdvisor menggunakan pengklasifikasian menggunakan *Naïve bayes*. Setelah itu dilakukan

	2, No. 8, Hal. 2766 - 2770		Brawijaya		pemberian seleksi fitur yang bertujuan untuk mengurangi dimensi fitur dengan metode Query Expansion Ranking dipadukan dengan metode Multi nomial *Naïve bayes*. Hasil penelitian ini dengan menggunakanalgoritma Query Expansion Ranking menghasilkan *accuracy* tertinggi sebesar 86.6 pada seleksi fitur 75%.
Rivanda Putra Pratama,	Procedia Computer Science	2021	Department of Informati	The influence of fake accounts on	Penelitian ini melakukan eksperimen

Aris Tjahyanto	197 (2022) 143–150		on Systems, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia	sentiment analysis related to COVID-19 in Indonesia	terhadap analisis sentimen menggunakan pendekatan machine learning untuk mengkategorikan akun palsu untuk melihat dampak akun palsu terhadap analisis sentimen. Dari hasil pengujian yang dilakukan diketahui bahwa klasifikasi sentimen pada kedua algoritma menunjukkan bahwa proses klasifikasi sentimen tanpa menggunakan akun palsu lebih baik daripada

					menggunakan semua data tweet.
Oman Somantri, Dairoh	(Jurnal Edukasi dan Penelitian Informatika) ISSN: 2460-0741 Vol. 5 No.2	2019	Program Studi Teknik Informatika, Politeknik Harapan Bersama Tegal	Analisis Sentimen Penilaian Tempat Tujuan Wisata Kota Tegal Berbasis Text Mining	Mencari model sistem untuk memberikan informasi pendukung keputusan untuk para pengolah tempat wisata dan wisatawan yang ingin berkunjung berdasarkan review pengunjung sebelumnya yang didapat dari komentar google maps. Dari pengujian yang dilakukan dengan menggunakan metode *Naïve Bayes* dan Decision Tree,

					tingkat *accuracy* *Naïve Bayes* menghasilkan 77,50% lebih baik dibandingkan dengan menggunakan Decision Tree yang menghasilkan tingkat *accuracy* 60,83%.
Rifa'i A, Sujaini H, Prawira D	(Jurnal Edukasi dan Penelitian Informatika) ISSN: 2460-0741 Vol. 7 No.3	2021	Program Studi Informatika, Fakultas Teknik, Universitas Tanjungpura	Sentiment Analysis Objek Wisata Kalimantan Barat Pada Google Maps Menggunakan Metode Naive Bayes	Analisis sentimen pada tempat wisata di Kalimantan Barat yang bertujuan untuk membangun sistem yang menyediakan eksekusi. Berdasarkan data peringkat Google

					Maps. Sistem ini dapat melakukan analisis sentimen dan mampu mengklasifikasikan sentimen, dan berdasarkan hasil pengujian terhadap 40 responden mendapatkan nilai 84,85% kedalam kategori baik sekali dan mendapatkan *accuracy* paling tinggi sebesar 76% berdasarkan pengujian confusion matrix.
Hidayat T, Ruldeviyani Y,	Procedia Computer Science,	2021	Faculty of Computer	Sentiment analysis of *twitter* data related to	Analisis terhadap sentimen masyarakat tentang

Aditama A, Madya G, Nugraha A, Adisaputra M	Vol.197 Hal. 660-667		Science, Universitas Indonesia, Jakarta, Indonesia	Rinca Island Sentiment analysis of *twitter* data related to Rinca Island development using Doc2Vec and SVM and logistic regression as classifier	pengembangan pada pulau Rinca. Analisis dilakukan dengan tiga algoritma pengklasifikasian yaitu SVM dan Logistic Regression. Dan dilakukan penerapan SVM dan Logistic Regression dengan menggunakan library Gensim dan modul Sklearn untuk model bag of word terdistribusi dan proses analisis sentimen menggunakan bahasa pemrograman

					Python dengan memori terdistribusi. Hasil yang didapat dari analisis dengan model algoritma PV-DBOW dengan SVM, PV-DM dengan SVM,dan regresi logistik mendapat tingkat *accuracy* dan F1-score yang lebih tinggi dibandingkan model lainnya dan mendapatkan 87% sebagai hasil *accuracy* terbaik.
Kurniawan I, dan Susanto A	Jurnal Eksplorasi Informatika,	2019	Jurusan Teknik Informatika,	Implementasi Metode K-Means dan *Naïve Bayes*	Untuk menganalisa sentimen di *Twitter*, dilakukan

	e-ISSN: 2460-3694 Vol.9 Hal. 1-10		Fakultas Ilmu Komputer Universitas Dian Nuswantoro Semarang	Classifier untuk Analisis Sentimen Pemilihan Presiden (Pilpres) 2019	pengambilan data sebanyak 500 tweet dari dua hastag #2019GantiPresiden dan #2019TetapJokowi.Tahap akhir dilakukan pengujian pengklasifikasian tweet yang meliputi *accuracy*, dan error rate. Hasil dari penelitian ini yang menggunakan metode K-means untuk melakukan pembobotan dan Naive Bayes untuk klasifikasi menghasilkan *accuracy* rata-rata

					sebesar 93.35% dan error rate rata-rata sebesar 6.66% dari 100 dan 150 data uji.
Fitri V, Andreswari R, Hasibuan M	Procedia Computer Science, Vol.161 Hal.765-772	2019	Information Systems Department School of Industrial Engineering, Telkom University, Bandung, Indonesia	Sentiment analysis of social media *Twitter* with case of Anti-LGBT campaign in Indonesia using *Naïve bayes*, decision tree, and random forest algorithm	Melihat opini masyarakat terhadap kampanye Anti-LGBT di Indonesia. Untuk melakukan analisa sentimen pengguna *Twitter*. Hasil dari pengujian yang dilakukan dengan membandingkan tiga model algoritma yang berbeda didapatkan hasil sebagai berikut, Naive Bayes

					menghasilkan *accuracy* sebesar 83,43% dengan recall,precision dan F1-Measure antara 56% dan 65% memberikan *accuracy* yang lebih baik dibanding Decision Tree dan Random Forest yang menghasilkan *accuracy* sebesar 82,91% dengan recall,precision dan F1-Measure antara 27,66% dan 33,33%.
Muham mad N, Bukhori	Institute of Electrical	2019	Faculty of Compute	Sentiment Analysis of Positive and	Hasil klasifikasi *Naïve Bayes* dan SVM menjadi

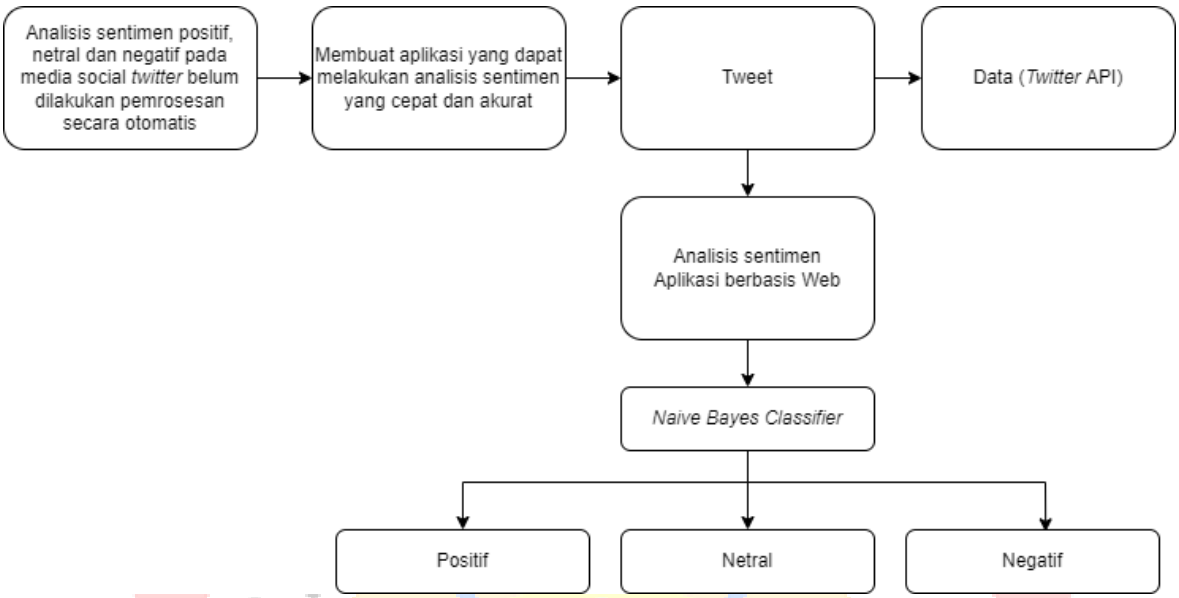
S, Pandunata P	and Electronics Engineers,		r Science University of Jember	Negative of YouTube Comments Using *Naïve Bayes* – Support Vector Machine (NBSVM) Classifier	sangat optimal bila data training yang digunakan memiliki jumlah data yang bervariasi. Hasil dari pengujian yang diperoleh dari komentar video YouTube,yaitu kombinasi *Naïve Bayes* dan Support Vector Machine menghasilkan tingkat *accuracy* yang lebih baik dengan menggunakan data skala 7:3 yaitu 70% data pelatihan dan 30% data pengujian. Dengan menghasilkan

					yang tertinggi nilai performance test yaitu presisi 91%, recall of 83% dan skor f1 87%.
Villavicencio C, Macrohon J, Inbaraj X,Jeng J, Hsieh J	Information (Switzerland), Vol.12 ,204	2021	Department of Information Engineering, I-Shou University, Kaohsiung City	*Twitter* sentiment analysis towards covid-19 vaccines in the Philippines using *Naïve bayes*	Hasil dari pengujian dari 993 tweet, sebanyak 828 atau 83,38% mendapatkan respon positif dan 82 atau 8,26% mendapatkan respon negatif. Hasil dari pengujian dan pelatihan data didapatkan bahwa model yang menerapkan algoritma Naive Bayes mendapatkan

					accuracy sebesar 81,77%.

Berdasarkan penelitian dari jurnal diatas secara keseluruhan, dapat disimpulkan bahwa berdasarkan perbandingan jurnal diatas, peneliti akan menggunakan metode *Naïve Bayes Classifier* karena metode *Naïve Bayes Classifier* ini sudah teruji memiliki presentase *accuracy* yang tinggi dan cocok dalam penerapan *Text-Mining* dalam berbagai topik.

2.5 Kerangka Pemikiran Penelitian



Gambar 2.3 Kerangka Pemikiran

BAB III

ANALISA DAN PERANCANGAN APLIKASI

3.1 Pengambilan Data

Dalam tahap ini, data dikumpulkan dari dua jenis data. *Data Training* yang berupa tweets yang dilabelkan menggunakan frasa-frasa yang mengandung sentimen positif dan negatif. Sedangkan *Data Testing* dikumpulkan dari API *Twitter*. Tahap pengumpulan data training ini dilakukan agar mesin pembelajaran ini dapat mengolah dan mempelajari frasa-frasa yang memiliki sentimen sehingga mesin dapat menentukan pengklasifikasian data *twitter* berdasarkan analisis dan model pembelajaran sehingga dapat menentukan frasa-frasa yang mengandung sentimen positif atau negatif.

Dalam tahap ini, *Data Training* yang digunakan sebanyak 2037 data akan diberi label sentimen akan diberikan berdasarkan list kata-kata yang sudah didefinisikan oleh Bing Liu dan telah dimodifikasi oleh Wahid, D.H dalam bahasa Indonesia digunakan berasal dari *Data Training* yang digunakan bersifat sekunder. Jumlah total list kata negatif yang akan dilatih dengan 2402 kata, dan jumlah total list kata positif adalah 1182 kata, sehingga jumlah total data list kata-kata positif dan negatif yang digunakan adalah 3633 kata. Baik kalimat positif maupun kalimat negatif tidak mengandung huruf kapital dan simbol. Jadi jika ingin membuat klasifikasi, informasi yang akan diverifikasi terlebih dahulu harus diubah menjadi huruf kecil. Setiap data yang akan diberikan label sentimen diberikan bobot berdasarkan dictionary atau kamus yang sudah didefinisikan sebelumnya bersumber dari https://github.com/masdevid/ID-OpinionWords. Contoh frasa ini dapat dilihat pada Tabel 3.1 dan Tabel 3.2.