OPTIMASI KINERJA *NAIVE BAYES* MENGGUNAKAN FITUR SELEKSI CHI SQUARE DAN INFORMATION GAIN DALAM MENINGKATKAN HASIL AKURASI ANALISIS SENTIMEN *CYBERBULLYING*

TESIS

Oleh:

Muhammad Yasir
1911600961

PROGRAM STUDI MAGISTER ILMU KOMPUTER
FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS BUDI LUHUR

JAKARTA
GASAL 2021/2022

OPTIMASI KINERJA *NAIVE BAYES* MENGGUNAKAN FITUR SELEKSI CHI SQUARE DAN INFORMATION GAIN DALAM MENINGKATKAN HASIL AKURASI ANALISIS SENTIMEN *CYBERBULLYING*

TESIS

Diajukan untuk memenuhi salah satu persyaratan memperoleh gelar Magister Ilmu Komputer (MKOM)

Oleh:

Muhammad Yasir
1911600961

PROGRAM STUDI MAGISTER ILMU KOMPUTER
FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS BUDI LUHUR

JAKARTA
GASAL 2021/2022

LEMBAR PENGESAHAN

PROGRAM STUDI MAGISTER ILMU KOMPUTER
FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS BUDI LUHUR

LEMBAR PENGESAHAN

Nama	: Muhammad Yasir
Nomor Induk Mahasiswa	: 1911600961
Program Studi	: Magister Ilmu Komputer
Bidang Peminatan	: Teknologi Sistem Informasi
Jenjang Studi	: Strata 2
Judul	: OPTIMASI KINERJA NAIVE BAYES MENGGUNAKAN FITUR SELEKSI CHI SQUARE DAN INFORMATION GAIN DALAM MENINGKATKAN HASIL AKURASI ANALISIS SENTIMEN CYBERBULLYING

Laporan Tugas Akhir ini telah disetujui, disahkan dan direkam secara elektronik sehingga tidak memerlukan tanda tangan tim penguji.

Jakarta, Selasa 18 Januari 2022

Tim Penguji:

Ketua	: Dr. Hari Soetanto, S.Kom, M.Sc
Anggota	: Dr. Imelda, S.Kom., M.Kom.
Pembimbing	: Denni Kurniawan, S.T, M.T.I, Ph.D
Ketua Program Studi	: Dr. Rusdah, S.Kom., M.Kom.

LEMBAR PERNYATAAN

PROGRAM STUDI MAGISTER ILMU KOMPUTER
FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS BUDI LUHUR

SURAT PERNYATAAN TIDAK PLAGIAT DAN PERSETUJUAN PUBLIKASI

Saya yang bertanda tangan dibawah ini:

Nama	:	Muhammad Yasir
Nomor Induk Mahasiswa	:	1911600961
Program Studi	:	Magister Ilmu Komputer
Bidang Peminatan	:	Teknologi Sistem Informasi
Jenjang Studi	:	Strata 2

Menyatakan bahwa TESIS yang berjudul:

OPTIMASI KINERJA NAIVE BAYES MENGGUNAKAN FITUR SELEKSI CHI SQUARE DAN INFORMATION GAIN DALAM MENINGKATKAN HASIL AKURASI ANALISIS SENTIMEN CYBERBULLYING.

Merupakan:

1. Karya tulis saya sebagai laporan tesis yang asli dan belum pernah diajukan untuk mendapatkan gelar sarjana, baik di Universitas Budi Luhur maupun di perguruan tinggi lainnya.
2. Karya tulis ini bukan saduran / terjemahan, murni gagasan, rumusan dan pelaksanan penelitian / implementasi saya sendiri, tanpa bantuan pihak lain, kecuali arahan pembimbing akademik dan nara sumber di organisasi tempat riset.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat yang telah ditulis atau dipublikasikan orang lain, kecuali secara tertulis dengan dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan dicantumkan dalam daftar pustaka.
4. Saya menyerahkan hak milik atas karya tulis ini kepada Universitas Budi Luhur, dan oleh karenanya Universitas Budi Luhur berhak melakukan pengelolaan atas karya tulis ini sesuai dengan norma hukum dan etika yang berlaku.

Pernyataan ini saya buat dengan sesungguhnyanya dan apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka saya bersedia menerima sanksi akademik berupa pencabutan gelar yang telah diperoleh karena karya tulis ini, serta sanksi lainnya sesuai dengan norma yang berlaku di Universitas Budi Luhur.

Jakarta, 18 Januari 2022

Muhammad Yasir

ABSTRAK

OPTIMASI KINERJA *NAIVE BAYES* MENGGUNAKAN FITUR SELEKSI CHI SQUARE DAN INFORMATION GAIN DALAM MENINGKATKAN HASIL AKURASI ANALISIS SENTIMEN *CYBERBULLYING*

Oleh: Muhammad Yasir (1911600961)

Dalam mengekpresikan pendapat di media sosial terdapat tantangan yang nyata yang harus dihadapi, seperti pengguna *twitter* yang melakukan penyebaran *tweet* atau komentar berita bohong (*hoax*), penyebaran pencemaran nama baik, penyebaran ujaran kebencian, *body shaming* dan rasisme yang dimana hal tersebut merupakan tindakan *cyberbullying*. KPAI mencatat dalam kurun waktu 9 tahun, dari 2011 sampai 2019, ada 37.381 pengaduan kekerasan terhadap anak. Untuk *bullying* baik di pendidikan maupun di sosial media, angkanya mencapai 2.473 laporan dan trennya terus meningkat. Topik *cyberbullying* tersebut menjadi sumber data dan dasar dalam penelitian ini. Tujuan penelitian ini adalah melakukan analisis sentimen yang dapat mengklasifikasi sentiment *tweet* yang memiliki unsur *cyberbullying*. Hal tersebut penting dilakukan karena *cyberbullying* merupakan salah satu tindakan kejahatan yang melanggar UU ITE nomor 11 tahun 2008. Ada beberapa tantangan dalam proses analisis sentimen seperti *polarity shift, binary clasification, data sparsity* dan *accuracy*. Fokus pada penelitian ini adalah melakukan *experiment research* terhadapa kinerja *naïve bayes* dalam meningkatkan hasil akurasi analisis sentimen. Metode dalam penelitian ini menggunakan naïve bayes yang dikombinasikan dengan fitur seleksi *chi square* dan fitur seleksi *information gain* dalam proses analisis sentimen. Hasil penelitian yang didapatkan adalah nilai hasil akurasi metode *naive bayes* tanpa fitur seleksi (86%), *naive bayes* + fitur seleksi *chi square* (90%), *naive bayes* + fitur seleksi *information gain* (89%). Dapat disimpulkan, metode *naive bayes* yang dikombinasikan dengan fitur seleksi membuktikan terjadi peningkatan akurasi nilai hasil akurasi analisis sentimen sebesar 4%.

Kata Kunci: Analisis Sentimen, *Cyberbullying, Naïve Bayes, Chi Square, Information Gain*

ABSTRACT

NAIVE BAYES PERFORMANCE OPTIMIZATION USING CHI SQUARE SELECTION AND INFORMATION GAIN FEATURES TO INCREASE THE ACCURACY RESULTS OF CYBERBULLYING SENTIMENT ANALYSIS

By: Muhammad Yasir (1911600961)

In expressing opinions on social media, there are real challenges that must be faced, such as Twitter users who spread *tweet*s or comments on fake news (hoaxes), spread defamation, spread hate speech, body shaming and racism which are acts of cyber bullying. . KPAI noted that in a period of 9 years, from 2011 to 2019, there were 37,381 complaints of violence against children. For bullying both in education and on social media, the number reached 2,473 reports and the trend continues to increase. The topic of Cyberbullying is the source of data and the basis for this research. The purpose of this study is to conduct a sentiment analysis that can classify *tweet* sentiments that have elements of *cyberbullying*. This is important because cyberbullying is a crime that violates UU ITE number 11 of 2008. There are several challenges in the sentiment analysis process, such as polarity shift, binary classification, data sparsity and accuracy. The focus of this research is to conduct experimental research on the performance of nave Bayes in improving the accuracy of sentiment analysis. The method in this study uses nave Bayes combined with feature selection Chi Square Statistics and feature selection information gain in the sentiment analysis process. The results obtained are the accuracy of the *naive bayes* method without feature selection (86%), *naive bayes* + chi square feature selection (90%), *naive bayes* + feature selection information gain (89%). It can be concluded, the *naive bayes* method combined with feature selection proves that there is an increase in the accuracy of the sentiment analysis accuracy of 4%.

Keywords: Sentiment Analysis, *Cyberbullying*, Nave Bayes, Chi Square, Information Acquisition

KATA PENGANTAR

Puji syukur kehadirat Allah SWT yang telah melimpahkan rahmat dan karunia-Nya, sehingga penulis dapat menyelesaikan tesis ini dengan baik. Adapun penelitian ini disusun untuk memenuhi Tesis Magister Ilmu Komputer (S2) Universitas Budi Luhur Jakarta.

Tesis ini dapat terselesaikan atas dukungan dan do'a dari berbagai pihak. Oleh sebab itu perkenankanlah penulis menyampaikan rasa syukur dan terima kasih yang tulus kepada:

1. Allah SWT, atas segala petunjuk dan kemudahan-Nya sehingga akhirnya penulis dapat menyelesaikan tesis ini.
2. Keluarga khususnya kedua orang tua, isteri, dan anak penulis yang selalu mendukung memberi dukungan dan semangat serta keindahan do'a, nasehat, motivasi, dan semua yang telah diberikan.
3. Bapak Dr. Ir. Wendi Usino, M.M, M.Sc., selaku Rektor Universitas Budi Luhur
4. Bapak Dr. Deni Mahdiana, S.Kom., M.M., M.Kom., selaku Dekan Fakultas Teknologi Informasi Universitas Budi Luhur.
5. Ibu Dr. Rusdah, S.Kom., M.Kom., selaku Ketua Program Studi Magister Ilmu Komputer Universitas Budi Luhur.
6. Bapak Prof. Ir. Dana Indra Sensuse, Ph.D., selaku dosen pembimbing proposal tesis.
7. Bapak Denni Kurniawan, S.T. M.T.I, Ph.D., selaku dosen pembimbing tesis
8. Segenap Dosen Pengajar, Sekretariat Magister Ilmu Komputer Universitas Budi Luhur.
9. Dr. Istianingsih Sastrodihardjo, M.S.Ak,CA,CSRA.,CACP & Dr. Robertus Suraji,S.S.,M.A, sebagai Bu Lik dan Pak lik terhebatku yang selalu memberi semangat, motivasi, kritik dan saran membangun pada penulis dalam menyelesaikan tesis ini
10. Semua Rekan Mahasiswa yang selalu memberikan motivasi dalam kegiatan perkuliahan dan penyusunan proposal tesis.

Penulis menyadari bahwa dalam penulisan proposal tesis ini masih terdapat kelemahan yang perlu diperkuat dan kekurangan yang perlu dilengkapi. Oleh karena itu dengan kerendahan hati penulis mengharapkan masukan, koreksi, dan saran untuk memperkuat kelemahan dan melengkapi kekurangan tersebut.

Jakarta, 18 Januari 2022

Muhammad Yasir

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR ISI

BAB I
PENDAHULUAN

1.1 Latar Belakang

Media sosial merupakan primadona baru dalam perkembangan media informasi terkini. Saat ini media sosial merupakan media komunikasi paling efektif, transparan dan efisien serta memiliki peran penting dalam perubahan dan pembaharuan (Rahadi, 2017). Dari media sosial kita dapat menemukan teman lama, berkenalan dan berinteraksi dengan mudah dengan orang lain. Media sosial seperti *facebook, instagram, twitter, youtube* juga dinilai menjadi suatu wadah untuk mengekpresikan pendapat dan mengekspresikan keadaan yang terjadi baik berupa karya, ide, gagasan, pendapat, kritik bahkan penjualan suatu produk.

Dalam mengekpresikan pendapat di media sosial terdapat tantangan yang nyata yang harus dihadapi, seperti oknum-oknum pengguna media sosial *twitter* yang melakukan penyebaran *tweet* atau komentar berita bohong (*hoax*), penyebaran pencemaran nama baik, penyebaran ujaran kebencian, *body shaming* dan rasisme yang dimana hal tersebut merupakan unsur-unsur tindakan *cyberbullying*. *Cyberbullying* merupakan bentuk ancaman yang dilakukan pelaku untuk melecehkan korbannya dengan menggunakan perangkat teknologi atau media sosial. Intimidasi atau pelecehan secara verbal secara terus menerus yang dilakukan di dunia maya menyebabkan korban yang terindimidasi mengalami gangguan emosional (Maulana and Ernawati, 2020) dan efek dari *cyberbullying* sebagian besar telah dieksplorasi di masalah kesehatan mental remaja (Nixon, 2014). KPAI mencatat dalam kurun waktu 9 tahun dari 2011 sampai 2019, ada 37.381 pengaduan kekerasan terhadap anak. Untuk *bullying* baik di pendidikan maupun di sosial media, angkanya mencapai 2.473 laporan dan trennya terus meningkat. (KPAI, 2020). Topik *Cyberbullying* tersebut menjadi sumber data dan dasar dalam penelitian ini. Fokus penelitian ini adalah melakukan analisis sentimen yang dapat mengklasifikasi sentimen *tweet* yang memiliki unsur *cyberbullying*. Hal tersebut penting dilakukan karena *cyberbullying* merupakan salah satu tindakan kejahatan yang melanggar UU ITE nomor 11 tahun 2008.

Penelitian mengenai sentimen *cyberbullying* sebelumnya, Fajar Agus Maulana dan Ernawati mereka melakukan penelitian untuk mengklasifikasi *cyberbullying* dari mention akun publik figur politik di Indonesia di *twitter* menggunakan *naïve bayes* dengan hasil akurasi sebesar 76% (Maulana and Ernawati, 2020) . Kemudian penelitian yang dilakukan Khorul zuhri dkk mereka melakukan penelitian untuk mengklasifikasi ujaran kebencian terhadap pilpres 2019 berdasarkan opini dari *twitter* menggunakan metode *naive bayes* dengan hasil akurasi 71% (Zuhri *et al.*, 2020). Dalam hal ini penulis menilai hasil akurasi penelitian sebelumnya masih bisa di tingkatkan.

Ada beberapa teknik dalam melakukan klasifikasi analisis sentimen dan beberapa pengklasifikasi yang paling banyak diterapkan adalah *Naive bayes* (NB),

Maximum Entropy (MaxEnt), *Support Vector Machines* (SVM), *Logistic Regression* (LR), *Random Forest* (RF), dan *Conditional Random Field* (CRF)(Giachanou and Crestani, 2016). Dan pada penelitian ini penulis menggunakan *naïve bayes* disamping metodenya *simple* dan sederhana, pengklasifikasian *naive bayes* untuk tujuan analisis sentimen memiliki akurasinya yang tinggi. Meskipun teorema sederhana, ia melakukan hampir sama baiknya dengan banyak pendekatan yang kompleks lainnya (Sintaha and Mostakim, 2019).

Ada beberapa tantangan dalam analisis sentimen seperti *polarity shift, binary clasification, data sparsity* dan *accuracy* (Abirami and Gayathri, 2017). Dan *twitter* adalah domain baru untuk sentimen analisis yang sangat menantang. Salah satu tantangan utama adalah batas panjang 140 karakter. Karena keterbatasan ini orang tidak akan mengungkapkan pendapat mereka dengan jelas. masalah ini terkait erat dengan keakuratan analisis sentimen (Abirami and Gayathri, 2017). Selain itu, panjang pendeknya dan jenis media informal menyebabkan munculnya informalitas tekstual yang sering ditemukan di *twitter* (Giachanou and Crestani, 2016). Pada penelitian ini, penulis fokus pada bagaimana meningkatkan akurasi dalam proses analisis sentimen dengan menggunakan metode *naïve bayes* dengan tantangan keterbatasan penulisan huruf atau kalimat yang diberlakukan di media sosial khususnya *twitter*, lalu mengkombinasikan dengan fitur seleksi kemudian membandingkan hasil akurasinya.

Fitur seleksi penting dalam analisis sentimen untuk memilih fitur informatif dan relevan yang membantu pengklasifikasian untuk menghasilkan hasil klasifikasi yang akurat, dan dapat mengurangi beban pemrosesan tanpa mengurangi kinerjanya (Jiang *et al.*, 2007). Dalam proses analisis sentimen metode fitur seleksi yang sering digunakan adalah *Document Frequency* (DF), *Mutual Information* (MI), *Odds Ratio* (OR), *Categorical Proportional Differences* (CPD), dan *Information Gain* (IG), serta *Chi Square Statistic*, yang dianggap sederhana, cepat dan efektif (Hung *et al.*, 2015).

Naïve Bayes (NB) adalah algoritma *machine learning* sederhana dan efisien yang memberikan keakuratan klasifikasi yang kompetitif. Ditambah dengan efisiensi komputasi dan banyak fitur lain yang di inginkan, hal ini menyebabkan banyak penerapan *naïve bayes* secara luas dalam praktiknya (Webb, 2016). dan kemudian dalam proses fitur seleksi, *Chi square* x^2 *statistic (Chi* x^2*)* dan *Information Gain* (IG) merupakan metode terbaik dari fitur seleksi (Hung *et al.*, 2015). Pada penelitian ini penulis menggunakan metode *naïve bayes* yang dikombinasikan dengan fitur seleksi *Chi square* x^2 dan *Information Gain* yang kemudian dilakukan komparasi hasil akurasi dalam proses analisis sentimen. Hasil penelitian vidhya dkk menjelaskan bahwa teknik *naïve bayes* berkinerja lebih baik dan menghasilkan akurasi klasifikasi yang lebih tinggi bila dikombinasikan dengan teknik lainnya (A and Aghila, 2010). Diharapkan dengan adanya penambahan seleksi fitur dapat membantu meningkatkan tingkat akurasi pada proses analisis sentimen *cyberbullying*.

Atas dasar urian tersebut, penulis mengangkat penelitian dengan judul ”Optimasi Kinerja *Naive bayes* Menggunakan Fitur Seleksi *Chi Square* Dan *Information Gain* Dalam Meningkatkan Hasil Akurasi Analisis Sentimen *Cyberbullying*

1.2 Masalah Penelitian

1.2.1 Identifikasi Masalah

Identifikasi masalah dari latar belakang diatas adalah batasan panjang penulisan yang berlaku di media sosial *twitter* menyebabkan keterbatasan orang tidak mengungkapkan pendapat mereka dengan jelas. masalah ini terkait erat dengan keakuratan akurasi klasifikasi analisis sentimen yang memiliki unsur *cyberbullying*, serta bagaimana mengoptimalkan kinerja metode *naïve bayes* dengan fitur seleksi *chi square vs naïve bayes* dengan fitur seleksi *Information Gain* dalam meningkatkan nilai akurasi analisis sentimen diatas 76% seperti yang sudah dilakukan penelitian-penelitian sebelumnya.

1.2.2 Rumusan Masalah

Dari uraian identifikasi masalah tersebut, peneliti melakukan rumusan masalah yaitu bagaimana mengevaluasi kinerja *naïve bayes* dengan mengkombinasikan *naïve bayes* dengan fitur seleksi *chi square* vs *naïve bayes* dengan fitur seleksi *Information Gain* dalam meningkatkan nilai akurasi pada proses analisis sentimen *cyberbullying* diatas 76%?

1.2.3 Batasan Masalah

Agar penelitian ini menjadi terarah, maka penulis perlu membatasi ruang lingkup penelitian sebagai berikut:

1. Objek yang digunakan dalam penelitian adalah data *twitter* dengan kata kunci “saipul jamil” yang diambil mulai tanggal 2 September s/d 09 September 2021 menggunakan *API twitter*.
2. Pemodelan analisis sentimen menggunakan metode *naïve bayes*
3. Fitur seleksi menggunakan fitur seleksi *Chi Square* dengan fitur seleksi *Information Gain*
4. Dalam penelitian ini menggunakan bahasa pemograman Python
5. Tahapan pengembangan menggunakan Crisp DM

1.3 Tujuan Dan Manfaat Penelitian

1.3.1 Tujuan Penelitian

Adapun tujuan penelitian ini adalah mengevaluasi kinerja *naïve bayes* dengan mengkombinasikan dengan fitur seleksi (*chi square vs Information Gain)* dalam meningkatkan nilai akurasi pada proses analisis sentimen diatas 76%.

1.3.2 Manfaat Penelitian

Bedasarkan tujuan penelitian tersebut manfaat penelitian ini diharapkan:

1. Dapat memberikan kontribusi pengetahuan bagi pengembangan ilmu teknologi informasi, khususnya dalam pengembangan ilmu *text mining* yaitu sentimen analisis.

2. Dapat membantu memberikan informasikan pada masyarakat atau KPAI mengenai *trend* tindakan *cyberbullying* di Indonesia di media sosial khususnya di *twitter*

1.4 Tata Urut Penulisan

Sistematika penulisan dalam penulisan tesis ini, adalah sebagai berikut:

BAB I PENDAHULUAN

Berisi penjelasan latar belakang permasalahan yang akan diteliti, masalah penelitian, tujuan penelitian, batasan ruang lingkup dan manfaat penelitian serta tata urut penulisan penelitian.

BAB II LANDASAN TEORI DAN KERANGKA PEMIKIRAN

Bab ini membahas mengenai teori-teori yang menjadi dasar penelitian meliputi tinjauan pustaka, tinjauan studi, tinjauan obyek penelitian, kerangka konsep / pola pikir pemecahan masalah.

BAB III METODOLOGI DAN RANCANGAN PENELITIAN

Bab ini menjelaskan mengenai metode penelitian, metode pengumpulan data / *sampling*, teknik analisis, langkah-langkah penelitian, dan jadwal penelitian

BAB IV PEMBAHASAN PENELITIAN

Bab ini menjelaskan mengenai pembahasan analisis dan evaluasi dari model yang dibangun

BAB IV PENUTUP

Bab ini menjelaskan mengenai kesimpulan dan saran mengenai penelitian agar penelitian selanjutnya yang dapat memperbaiki penelitian ini.

BAB 2
LANDASAN TEORI DAN KERANGKA PENELITIAN

2.1 Tinjauan Pustaka

2.1.1 Media Sosial

Media sosial adalah sebuah *platform* di internet yang memungkinkan pengguna untuk membuat dan berbagi konten dalam berbagai konteks seperti informasi, pendidikan, sindiran, kritik, dan sebagainya. kepada khalayak ramai (Widiastuti, 2018)

Media sosial merupakan bentuk media baru yang memungkinkan terjadinya digitalisasi, konvergensi, interaktif, dan pengembangan jaringan yang berkaitan dengan pembuatan pesan dan penyampaian pesan. Kemampuannya untuk menawarkan interaktivitas memungkinkan pengguna media baru untuk memilih informasi apa yang dikonsumsi, sambil mengontrol *output* informasi yang dihasilkan dan membuat pilihan yang mereka inginkan. (Watie, 2016).

Media sosial memiliki kekuatan untuk mempengaruhi opini publik yang berkembang di masyarakat. Menggalang dukungan atau gerakan massa bisa terbentuk karena kekuatan media *online* karena apa yang ada di media sosial terbukti mampu membentuk opini, sikap dan perilaku publik atau masyarakat.

Berikut di bawah ini klasifikasi macam-macam jejaring sosial berdasarkan fungsi dan kegunaannya (Putri *et al*., 2016):

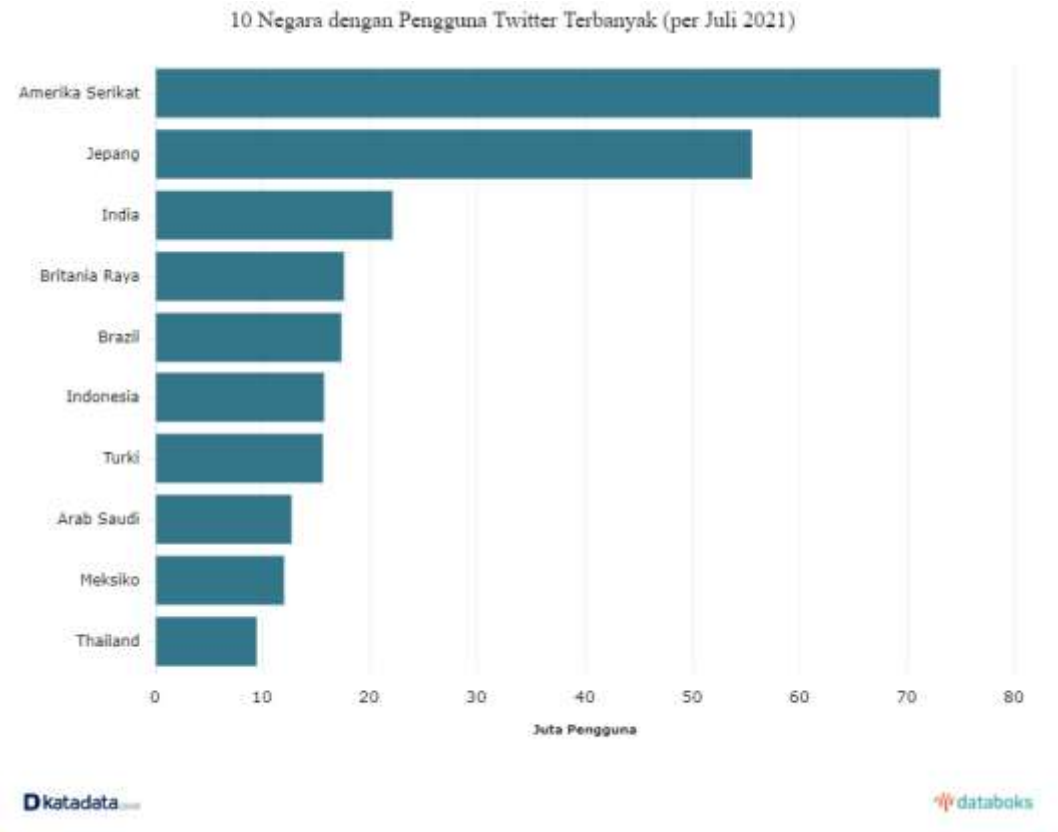
a. Konten kolaborasi (contohnya, *Wikipedia*)
b. Blog dan *microblog* (contohnya, *Twitter*)
c. Situs jejaring sosial berita (contohnya, *Digg*)
d. Konten Video (contohnya, *YouTube*)
e. Situs jejaringan sosial (contohnya, *Facebook*)
f. *Game* dunia maya (contohnya, *World of Warcraft*)
g. Situs dunia sosial virtual (contohnya, *Second Life*)

2.1.2 *Twitter*

Twitter didirikan oleh Jack Dorsey, *Twitter* merupakan layanan jejaring sosial dan mikrobloging yang memungkinkan penggunanya untuk mengirim dan membaca pesan berbasis teks hingga 140 karakter, yang dikenal dengan sebutan kicauan (*tweet*) (Zukhrufillah, 2018).

Analisis Sentimen telah dipelajari di banyak media, termasuk *review*, forum diskusi, dan blog. Baru-baru ini, para peneliti mulai menganalisis pendapat yang diungkapkan dalam mikroblog karena mengandung banyak teks opini. *Twitter* adalah salah satu situs *microblogging* paling populer yang memiliki banyak orang yang berbagi pemikiran, pendapat, dan jenis informasi lainnya. Kalimat ini mengatakan bahwa informasi yang diposting di *twitter* sering kali berisi opini

tentang produk, layanan, selebriti, acara, atau apa pun yang menarik minat pengguna. Karena popularitasnya yang semakin meningkat, t*witter* baru-baru ini menarik minat banyak peneliti yang menganalisis data t*witter* untuk berbagai tugas berbeda seperti membuat prediksi.



Gambar 2. 1 Negara Pengguna *Twitter* di Dunia (Databoks, 2021)

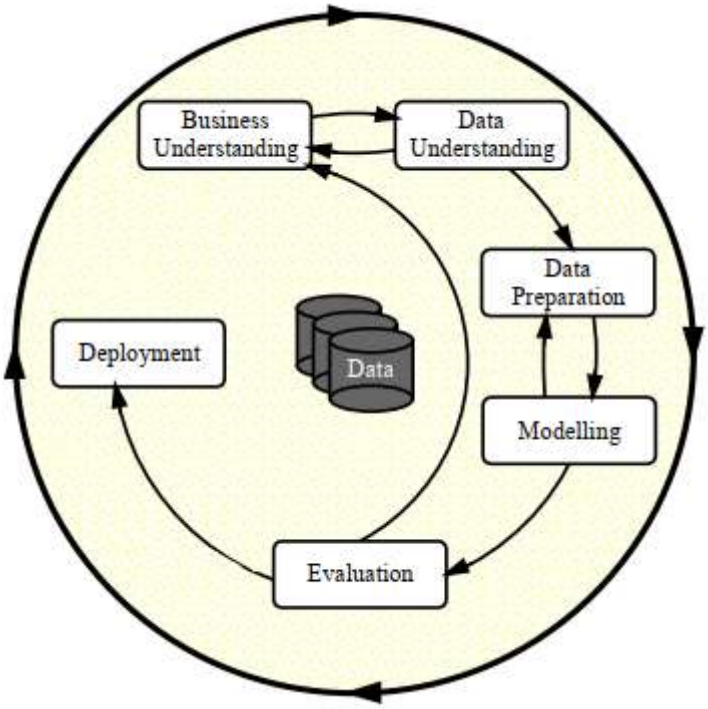
Tantangan Sentimen analisis *twitter* dijelaskan sebagai berikut (Giachanou and Crestani, 2016):

a. Panjang Teks: Salah satu karakteristik unik dari *tweet* adalah panjangnya yang pendek, yang bisa mencapai 140 karakter.
b. Relevansi Topik: Sebagian besar pekerjaan yang dilakukan analisis sentimen *twitter* bertujuan untuk mengklasifikasikan orientasi sentimen dari sebuah *tweet* tanpa mempertimbangkan relevansi topikal. Untuk menangkap relevansi topik sebuah *tweet*, banyak peneliti hanya menganggap keberadaan kata sebagai indikator relevansi topik
c. Bahasa Informal: Karena jenis komunikasi informal dan batasan panjangnya, bahasa yang digunakan di *twitter* sangat berbeda dengan bahasa yang digunakan dalam genre teks lain (web, blog, berita, dll.)
d. *Data Sparsity*: *Tweet* mengandung banyak *noise* karena penggunaan bahasa yang salah dan salah ejaan secara ekstensif. Fenomena ini, yang dikenal sebagai *data sparsity*.

2.1.3 *Text Mining* (CRISP-DM)

Text Mining adalah variasi dari bidang *data mining*, yang mencoba menemukan pola menarik dari kumpulan data. *Text mining* juga dikenal sebagai *Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text* (KDT). *Text Mining* adalah prosedur mensintesis informasi dengan menganalisis hubungan, pola, dan aturan dari data tekstual dari sumber yang terstuktur atau tidak terstruktur menjadi wawasan atau pengetahuan baru (Allahyari *et al.*, 2017). Dalam penelitian ini penulis menggunakan metodologi CRISP-DM, dalam siklus pengembangannya CRISP-DM dianggap sebagai metodologi *data mining* terlengkap dalam hal pemenuhan kebutuhan proyek industri, dan telah menjadi yang paling luas penggunaannya dalam proyek *data mining*. Singkatnya, CRISP-DM dianggap sebagai standar *de facto* untuk proyek analitik, *data mining*, dan *data science* (Schröer, Kruse and Gómez, 2021).

Tahapan penelitan menggunakan CRISP-DM sebagai berikut:



***Gambar 2. 2* Tahapan CRISP-DM(Schröer, Kruse and Gómez, 2021)**

A. Bussiner Understanding

Fase ini dimulai dengan memahami tujuan dan persyaratan proyek dari perspektif bisnis dan kemudian mengubah pengetahuan ini menjadi definisi masalah penambangan data dan rencana proyek awal yang dirancang untuk mencapai tujuan.

B. Data Understanding

Proses pemahaman data dimulai dengan pengumpulan data dan dilanjutkan dengan kegiatan mengidentifikasi data, mengidentifikasi masalah kualitas data,

menemukan wawasan pertama ke dalam data, atau mendeteksi subset yang menarik untuk membentuk hipotesis untuk informasi tersembunyi.

Di tahap ini penulis melakukan pelabelan dataset manual dengan bantuan anator, ada baiknya dalam proses pelabelan ini melibatkan ahli atau pakar bahasa sesuai bahasa yang digunakan. Tujuannya adalah ahli atau pakar bahasa tersebut lebih memahami pemaknaan dari setiap teks yang tertulis (Hadna, Santosa and Winarno, 2016)

C. Data Preparation

Tahap persiapan data mencakup semua kegiatan untuk membangun dataset akhir dari data mentah awal. Tugas persiapan data kemungkinan akan dilakukan beberapa kali, tetapi tidak dalam urutan yang ditentukan. Tugas seperti tabel, catatan, dan pemilihan atribut, *preprocessing* data, konstruksi atribut baru, dan transformasi data untuk pemodelan.

Gambar 2. 3 Tahapan Pembersihan data *Twitter*
(Giachanou and Crestani, 2016)

Tahap *preprocessing* atau pembersihan data *twitter* di atas adalah skenario standar *preprocessing data text*. Beberapa karakteristik unik *tweet* memerlukan langkah yang berbeda untuk mengatasinya (Giachanou and Crestani, 2016).
Tahapan *preprocessing twitter* sebagai berikut (Kane, Mishra and Dutta, 2016):

1. *Tokenize*: Tahap ini merupakan proses pembagian kata. Kalimat akan dibagi menjadi beberapa bagian yang disebut token. Token dapat berupa kata, frasa, atau elemen makna lainnya. Kita bisa menggunakan *library nltk.tokenize* Kita bisa menggunakan *library nltk.tokenize* untuk membagi kata menjadi kelompok huruf.
2. *Stopword:* Tahap ini merupakan proses untuk menghilangkan kata-kata umum dan sering yang tidak memiliki pengaruh signifikan dalam sebuah kalimat. sehingga data dapat diproses lebih efisien pada tahap selanjutnya. Mengimpor daftar *stopwords* bisa menggunakan dari *library nltk.corpus*.
3. *Case Folding :* Tahap ini merupakan proses mengubah kata menjadi bentuk yang sama. Pada langkah ini, mengonversi semua kata menjadi huruf kecil menggunakan metode "*lower case*" Python.
4. *Remove Punctuation:* Tahap ini merupakan proses menghapus karakter, angka, *string ASCII*, dan tanda baca. Pesan *twitter* biasanya berisi simbol, angka, dan tanda baca. Semua ini dihapus menggunakan sintaks ekspresi reguler
5. *Stemming:* Tahap ini adalah proses mendapatkan dasar atau akar kata dengan menghilangkan imbuhan dan sufiks. Penelitian ini menggunakan *library* python sastrawi untuk menghilangkan kata-kata imbuhan dalam bahasa Indonesia ke bentuk dasarnya.

D. Modeling

Pada fase ini, berbagai teknik pemodelan dipilih dan diterapkan, dan parameter dikalibrasi ke nilai optimal. Biasanya, ada beberapa teknik untuk jenis masalah *data mining* yang sama. Beberapa metode memerlukan format data tertentu. Pada tahap ini penulis memilih metode *naïve bayes* yang di kombinasikan dengan fitur seleksi.

E. Evaluation

Pada tahapan ini, telah membangun satu atau lebih model yang tampak berkualitas tinggi dari perspektif analisis data. Sebelum melanjutkan ke implementasi akhir model, penting untuk mengevaluasi model secara lebih menyeluruh, dan meninjau langkah-langkah yang diambil untuk membangun model, untuk memastikan bahwa model mencapai tujuan bisnis dengan benar. Tujuan utamanya adalah untuk menentukan apakah ada beberapa masalah yang belum dipertimbangkan secara memadai. Pada akhir fase ini, keputusan harus dibuat tentang bagaimana menggunakan hasil *data mining*

F. Deployment

Pemodelan bukanlah akhir dari proyek. Secara umum, pengetahuan yang diperoleh perlu diatur dan disajikan sedemikian rupa sehingga pengguna dapat menggunakannya. Bergantung pada apa yang butuhkan pengguna, fase penerapan bisa sesederhana membuat laporan atau serumit menerapkan proses penambangan data berulang.

2.1.4 Analisis Sentimen

Menggali opini dan sentimen dari media sosial sangat sulit karena banyaknya data yang dihasilkan oleh berbagai sumber. Informasi opini tentang suatu topik disembunyikan dalam data dan oleh karena itu hampir tidak mungkin bagi seseorang untuk melihat melalui berbagai sumber dan mengekstrak informasi yang berguna. Oleh karena itu, para peneliti mulai menyelidiki dan mengembangkan pendekatan yang dapat secara otomatis mendeteksi polaritas teks dan dapat mengekstrak informasi opini secara efektif bahkan dalam jumlah data yang besar dengan melakukan analisis sentimen (Giachanou and Crestani, 2016).

Analisis sentimen mengacu pada domain *natural language processing* (NLP) yang berhubungan dengan studi komputasi opini, sentimen, dan emosi yang diungkapkan dalam teks. Analisis Sentimen bertujuan untuk mempelajari opini, sikap, dan emosi orang terhadap suatu entitas. Entitas dapat mewakili individu, peristiwa, atau topik(Venugopalan and Gupta, 2015).

Analisis sentimen dapat didefinisikan sebagai proses yang mengotomatiskan penambangan sikap, opini, pandangan, dan emosi dari teks, ucapan, *tweet*, dan sumber basis data melalui *Natural Language Processing* (NLP). Analisis sentimen melibatkan pengelompokan opini dalam teks ke dalam kategori seperti "positif" atau "negatif" atau "netral"(A. and Sonawane, 2016).

Tugas analisis sentimen bertujuan untuk menentukan sikap seorang pengguna terhadap topik tertentu berdasarkan pemrosesan bahasa alami (NLP). Sikap disini

mungkin penilaian atau evaluasi mereka, keadaan afektif mereka atau komunikasi emosional yang dimaksudkan (Li and Liu, 2010).

Ada beberapa tantangan dalam analisis sentimen sebagai berikut (Abirami *et al.,* 2017):

A. *Polarity Shift*, adalah masalah yang paling penting untuk ditangani dalam Analisis Sentimen. Pergeseran polaritas berarti bahwa polaritas (Sentimen) kalimat dihitung dengan cara yang berbeda dari polaritas yang sebenarnya dinyatakan dalam Kalimat.
B. *Binary Clasification*, adalah masalah penting lain yang harus ditangani di mana polaritas ulasan yang diberikan diklasifikasikan hanya menggunakan "Positif", "Negatif" mengabaikan "Netral". Jenis masalah ini terutama muncul ketika klasifikasi sentimen murni didasarkan pada algoritma *machine learning*. Penambangan opini yang hanya mempertimbangkan positif dan negatif tidak akan memiliki akurasi yang baik saat ini
C. *Data Sparsity*, Masalah ketiga adalah masalah *data sparsity* yang disebabkan karena batasan karakter yang diberlakukan di situs *web microblogging* /media sosial. Misalnya batas maksimum karakter di *twitter* adalah 140. Karena batasan ini orang tidak akan mengungkapkan pendapat mereka secara jelas. Ketiga masalah ini terkait erat dengan keakuratan analisis sentimen.

2.1.5 *Chi Square* (Chi X^2)

Chi square digunakan untuk mengetahui apakah ada korelasi antara variabel nonnumerik yang sering digunakan dalam studi statistik. Pengujian kecocokan, pengujian independensi, dan pengujian homogenitas yang dikembangkan oleh Pearson adalah kontribusi paling signifikan yang dia buat untuk teori statistik modern. Dua tujuan khusus dari pengujian *Chi square* adalah untuk menguji hipotesis bahwa tidak ada korelasi antara dua atau lebih kelompok, populasi atau kriteria, dan untuk menguji sejauh mana distribusi data yang diamati sesuai dengan distribusi yang diharapkan (Nihan, 2020).

Chi square merupakan seleksi fitur yang melihat ketergantungan *term* dengan kategorinya beberapa syarat uji *chi square* dapat digunakan yaitu (Nisa *et al.,* 2019):

A. Tidak ada sel dengan nilai *actual count* atau frekuensi kenyataan ($F0$) yang 0 (nol);
B. Apabila tabel kontingensi berbentuk 2x2, maka tidak boleh ada sel dengan *expected count* atau frekuensi harapan (Fh) kurang dari 5;
C. Apabila bentuk tabel lebih dari 2 x 2, misal 2 x 3, maka jumlah sel dengan frekuensi harapan (Fh) yang kurang dari 5 tidak boleh lebih dari 20%

Rumus untuk menghitung statistik *Chi square* adalah (Singhal and Rana, 2015):

$$x^2 = \sum_{i=1}^{n} \frac{(O_1 - E_1)^2}{E_1} \quad (1)$$

di mana:
x^2= Distribusi *Chi square*
O_1= Nilai observasi (pengamatan) ke-i
E_1 = Nilai ekspektasi ke-i

Contoh perhitungan *chi square* (NURYADI *et al.*, 2017) :

Dalam suatu eksperimen genetika menurut Mendell telah diketemukan bahwa semacam karakteristik diturunkan meurut perbandingan 1:3:3:9 untuk kategorikategori A, B, C, dan D. Akhir-akhir ini dilakukan 160 kali pengamatan dan terdapat 5 kategori A, 23 kategori B, 32 kategori C dan 100 kategori D. Dengan menggunakan $\propto$= 0,05, apakah data di atas menguatkan teori genetika tersebut?

Penyelesaian :

Berdasarkan teori, diharapkan terdapat 1/16 X 160 = 10 kategori A, masing-masing 30 kategori B dan C, dan 90 kategori D. Data hasil pengamatan dan yang diharapkan adalah sebagai berikut :

Tabel 2. 1* Contoh Implementasi *Chi Square

Kategori	A	B	C	D
Pengamatan (O_1)	5	23	32	100
Diharapkan (E_1)	10	30	30	90

Dari rumus di dapatkan :

$$x^2 = \sum_{i=4}^{4} \frac{(O_1 - E_1)^2}{E_1} \quad (2)$$

$$= \frac{(5-10)^2}{10} + \frac{(23-30)^2}{30} + \frac{(32-30)^2}{30} + \frac{(100-90)^2}{90} = 5{,}38$$

Dari tabel distribusi *chi square* diperoleh $x^2{}_{1-0,05;(4-1)}$ = 7,81. Sehingga pengujian memperlihatkan H_0 diterima yang artinya teori menurut Mendell benar. Atau dengan cara :

Tabel 2. 2* Contoh Perhitungan *Chi Square

Kategori	O	E	$\frac{(O-E)^2}{E}$
A	5	10	2,5
B	23	30	1,63333
C	32	30	0,13333

D	100	90	1,11111
Total	160	160	5,37778

Dengan cara tersebut, maka diperoleh x^2 = 5,377 atau 5, 38. Derajad kebebasan (db) uji tersebut adalah jumlah kategori (k) dikurangi 1 = 4 – 1 = 3. Pada taraf signifikasi (α) = 5% harga x^2 tabel = 7,81. Karena x^2 hitung < x^2 tabel, maka hipotesis nol diterima.

2.1.6 *Information Gain* (IG)

Metode fitur seleksi *information Gain* adalah salah satu dari metode pemilihan fitur yang paling populer (Ahmad *et al.,* 2019). Rumus Formula *Information Gain* sebagai berikut:

$$IG(t) = \sum_{i=1}^{m} p(c_1) log\ p(c_1) + p(t) \sum_{i=1}^{m} p(c_1 \mid t)\ log\ p(c_1 \mid t) + p(t) \sum_{i=1}^{m} p(c_1 \mid t)\ log\ p(c_1 \mid t) \quad (3)$$

Dalam rumus $p(c_1)$ mewakili probabilitas bahwa *sample* arbitrer termasuk dalam kategori (c_1) , $p(t)$ mewakili probabilitas bahwa istilah tersebut ada dalam sebuah teks, $p(c_1 \mid t)$menunjukkan probabilitas bahwa sebuah teks termasuk dalam kategori (c_1) sedangkan $p(t)$ muncul dalam teks ini, menunjukkan probabilitas bahwa istilah t tidak ada dalam teks, $p(c_1 \mid t)$ mewakili probabilitas bahwa teks termasuk dalam kategori c_1 sedangkan t tidak muncul dalam teks ini., $p(c_1 \mid t)$ mewakili jumlah total kategori Ternyata, nilai IG (t) yang lebih besar menunjukkan lebih banyak informasi yang diberikan oleh kata fitur ke kategori tersebut, dan semakin penting kata fitur tersebut (Wu and Xu, 2016).

2.1.7 *Naïve Bayes* (NB)

Pengklasifikasi *naive bayes* secara mengejutkan efektif dalam praktik karena keputusan klasifikasinya mungkin sering benar bahkan jika perkiraan probabilitasnya tidak akurat. Meskipun beberapa kondisi optimalitas *naive bayes* telah diidentifikasi di masa lalu, pemahaman yang lebih dalam tentang karakteristik data yang mempengaruhi kinerja *naive bayes* masih diperlukan (Khairati *et al.*, 2019).

Naïve Bayes menyediakan mekanisme untuk menggunakan informasi dalam data *sample* untuk memperkirakan probabilitas posterior P(H | X) dari setiap kelas H yang diberikan objek X. Setelah memiliki perkiraan seperti itu, baru dapat menggunakannya untuk klasifikasi atau aplikasi pendukung keputusan lainnya (Webb, 2016).

Formula *naïve bayes* sebagai berikut (Bustami, 2010):

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

X : Data dengan *Class* yang belim diketahui
H : Hipotesis data X meruapakan suatu *class* spesifik
P (H|X) : Probabilitas hipotesis H berdasarkan kondisi X (*posteriori probability*)
P (H) : Probabilitas hipotesis H (*prior probability*)
P (X|H) : Probabilitas X berdasarkan kondisi pada hipotesis H
P (X) : Probabilitas X

2.1.8 *Confusion Matrix*

Confusion Matrix menjelaskan ukuran n x n yang terkait dengan pengklasifikasi mewakili klasifikasi yang diprediksi dan aktual, di mana n adalah jumlah kelas yang berbeda (Visa Sofia, 2011).

Tabel 2. 3 The confusion matrix for two-class classification

	Predictive Negative	*Predictive Positive*
Actual Negative	*A*	*B*
Actual Positive	*C*	*D*

Tabel 2.3 menunjukkan *Confusion Matrix* untuk n = 2, yang entrinya memiliki arti sebagai berikut:

a adalah jumlah prediksi negatif yang benar;
b adalah jumlah prediksi positif palsu;
c adalah jumlah prediksi negatif palsu;
d adalah jumlah prediksi positif yang benar
Keakuratan prediksi dan kesalahan klasifikasi dapat diperoleh dari matriks ini sebagai berikut:

$$Akurasi = \frac{a+d}{a+b+c+d} \quad (5)$$

$$error = \frac{b+c}{a+b+c+d} \quad (6)$$

2.1.9 *Cyberbullying*

Cyberbullying adalah tindakan ancaman yang dilakukan pelaku untuk melecehkan korbannya dengan menggunakan perangkat teknologi atau media sosial. Intimidasi atau pelecehan secara verbal secara terus menerus yang dilakukan di dunia maya menyebabkan korban yang terindimidasi mengalami gangguan emosional (Maulana and Ernawati, 2020). dan *Cyberbullying* memiliki berbagai

bentuk, seperti menyebarkan ujaran kebencian atas dasar rasisme, gender, disabilitas, agama, dan seksualitas; mempermalukan seseorang; Pengasingan sosial; menguntit; mengancam seseorang secara *online*; dan menampilkan informasi pribadi tentang individu yang dibagikan secara rahasia (Talpur and O'Sullivan, 2020).

Dari penjelasan diatas dapat disimpulkan bahwa *cyberbullying* merupakan tindakan intimidasi atau ancaman seperti melakukan ujaran kebencian, mempermalukan seseorang, rasisme, *body shaming*, yang dilakukan oknum pengguna media sosial menggunakan perangkat teknologi.

***Gambar 2. 4* Riset Microsoft *Cyberbullying* di Indonesia (Microsoft, 2021)**

Ada beberapa tipe jenis perilaku *cyberbullying* dapat dikenali (Alanazi and Alves-foss, 2020):

1. *Flooding* melibatkan pelaku intimidasi yang mengirimkan komentar / postingan yang tidak masuk akal berulang kali agar tidak memungkinkan korban yang ditargetkan untuk berpartisipasi dalam percakapan
2. Penyamaran melibatkan pelaku intimidasi yang berpura-pura menjadi atau menyamar sebagai korban target
3. *Flaming / Bashing* melibatkan perkelahian online di mana pelaku intimidasi mengirim dan/atau memposting konten yang menghina, menyakitkan, dan vulgar kepada korban yang ditargetkan secara pribadi atau publik dalam grup *online*
4. *Trolling* melibatkan dengan sengaja menerbitkan komentar yang tidak sesuai dengan komentar lain untuk memicu argumen atau emosi negatif meskipun komentar itu sendiri mungkin tidak vulgar atau menyakitkan.
5. Pelecehan adalah jenis percakapan di mana pelaku intimidasi sering mengirimkan pesan yang menghina dan kasar kepada korban secara pribadi.
6. Penghinaan terjadi ketika pelaku intimidasi mengirimkan atau mempublikasikan gosip atau pernyataan palsu tentang korban untuk merusak persahabatan/reputasi korban
7. *Outing* terjadi ketika seorang pengganggu memposting atau mempublikasikan informasi pribadi atau memalukan di ruang obrolan atau forum publik. Jenis *cyberbullying* ini mirip dengan fitnah. Namun, dalam sebuah tamasya, mungkin ada hubungan antara pelaku dan korban.
8. Pengecualian melibatkan dengan sengaja mengecualikan seseorang dari grup *online* (mengucilkan orang lain). Jenis *cyberbullying* ini terjadi di kalangan remaja dan remaja lebih menonjol.

2.1.10 *Python*

Python adalah bahasa pemrograman tingkat tinggi serta *powerfull* yang memiliki struktur data tingkat tinggi yang efisien dan pendekatan yang sederhana namun efektif untuk pemrograman berorientasi objek (Rossum and Drake, 2003). Sintaks *python* yang elegan dan pengetikan dinamis, bersama dengan sifatnya yang ditafsirkan, menjadikannya bahasa yang ideal untuk pembuatan skrip dan pengembangan aplikasi yang cepat di banyak area di sebagian besar *platform* (van Rossum, 2018).

Fitur penting dalam bahasa pemograman *python* (Kuhlman, 2013) sebagai berikut:

1. Tipe data tingkat tinggi bawaan: string, daftar, kamus, dll.
2. Struktur kontrol yang biasa: *if, if else, if elif else, while,* ditambah koleksi iterator yang kuat (*for*).
3. Beberapa tingkat struktur organisasi: fungsi, kelas, modul, dan paket. Ini membantu dalam mengatur kode. Contoh yang sangat bagus dan besar adalah pustaka standar *Python*

4. Kompilasi dengan cepat ke kode *byte*. Kode sumber dikompilasi ke kode byte tanpa langkah kompilasi terpisah. Modul kode sumber juga dapat "dikompilasi" ke file kode byte.
5. *Python* berorientasi objek menyediakan cara yang konsisten untuk menggunakan objek: semuanya adalah objek. Dan, dalam Python mudah untuk mengimplementasikan tipe objek baru (disebut kelas dalam pemrograman berorientasi objek)
6. Ekstensi dalam C dan C++ Modul ekstensi dan jenis ekstensi dapat ditulis dengan tangan. Ada juga alat yang membantu dalam hal ini, misalnya, SWIG, sip, Pyrex
7. Jython adalah versi Python yang "bermain baik dengan" Java

2.1.11 *Flask*

Flask adalah kerangka kerja web mikro yang ditulis dalam bahasa pemrograman Python. Flask dibuat pada tahun 2004 oleh Armin Ronacher. Flask dilisensikan di bawah lisensi BSD tiga bagian, Flask dirancang untuk membuat aplikasi *web* cepat dan mudah, dengan kemampuan untuk menskalakan aplikasi yang kompleks. Ini dimulai sebagai pembungkus sederhana di sekitar alat dan jinja dan telah berkembang menjadi salah satu kerangka kerja aplikasi *web python* paling populer (Mufid *et al.*, 2019).

Framework Flask adalah kerangka kerja *web* dari bahasa Python. Flask menyediakan *library* dan set kode yang dapat Anda gunakan untuk membangun situs *web*, tanpa harus melakukan semuanya dari awal. Karena fitur-fiturnya yang sederhana, Flask akan lebih ringan dan tidak terlalu bergantung pada banyak perpustakaan eksternal (Mufid *et al.*, 2019)

Dari penjelasan tersebut dapat disimpulkan *Framework* Flask adalah *framework* web dari bahasa Python yang menyediakan perpustakaan dan kumpulan kode yang dapat digunakan untuk membangun situs *web*, tanpa perlu melakukan semuanya dari awal. Karena fiturnya yang sederhana, flask akan lebih ringan dan tidak bergantung pada banyak library eksternal yang perlu diperhatikan.

Secara umum, flask menyediakan 'Wekzeug' yang berguna untuk menerima permintaan (url) dan merespons. Werkzeug adalah kumpulan perpustakaan yang dapat digunakan untuk membuat aplikasi web yang kompatibel dengan WSGI (*Web Server Gateway Interface*) dengan Python. Server WSGI (*Web Server Gateway Interface*) diperlukan untuk aplikasi *web* Python karena server *web* tidak dapat berkomunikasi langsung dengan Python. WSGI adalah antarmuka antara *server* web dan aplikasi *web* berbasis Python (Patrick Kennedy, 2021).

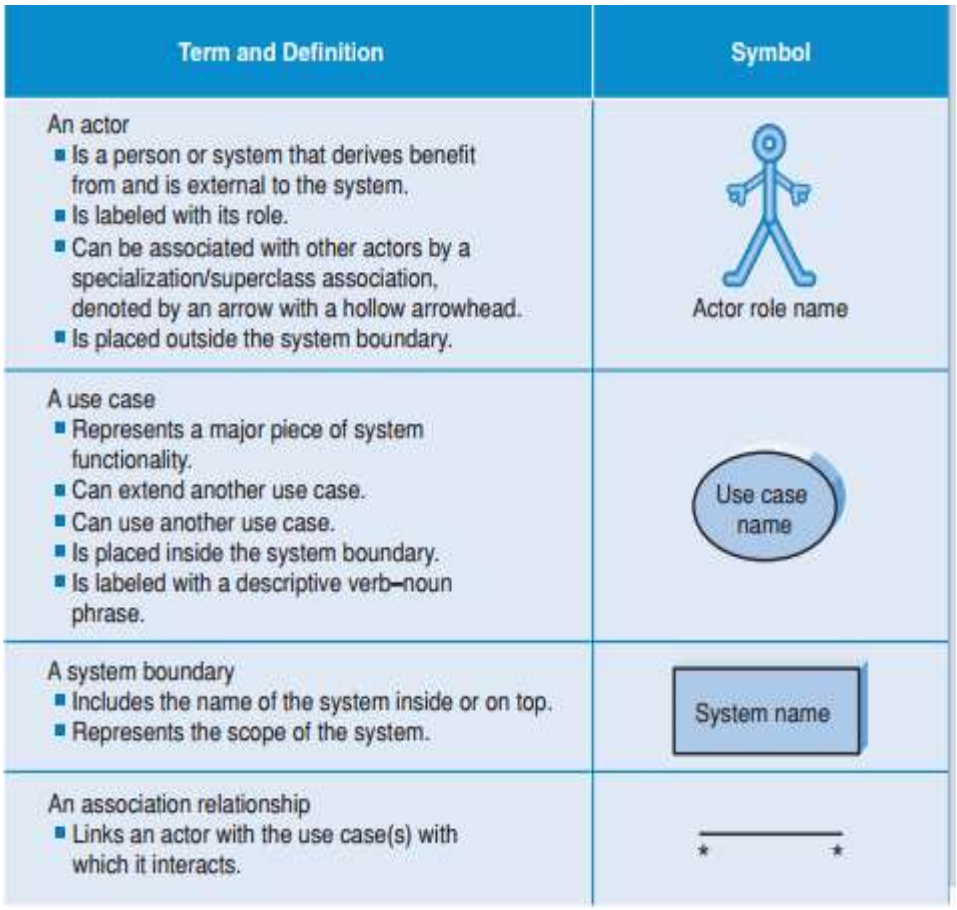
2.1.12 *Unified Modeling Language* (UML)

Unified Modeling Language (UML) adalah bahasa pemodelan visual tujuan umum yang digunakan untuk menentukan, memvisualisasikan, membangun dan mendokumentasikan artefak dari sistem perangkat lunak (Rumbaugh, 2013). UML menangkap informasi tentang struktur dan perilaku dinamis dari suatu sistem.

Sistem adalah model sebagai kumpulan objek diskrit yang berinteraksi untuk melakukan pekerjaan yang pada akhirnya menguntungkan pengguna. Tujuan dari *Unified Modeling Language* adalah untuk menyediakan kosakata istilah berbasis objek dan teknik diagram yang cukup kaya untuk memodelkan proyek pengembangan sistem dari analisis ke desain(Alan Dennis, Barbara Haley Wixom and David Tegarden, Barbara Haley Wixom, 2012)

2.1.12.1 *Use Case Diagram*

Use case Diagram adalah driver utama untuk semua teknik diagram UML. *Use case* mengkomunikasikan pada tingkat tinggi apa yang perlu dilakukan sistem, dan setiap teknik diagram UML dibangun di atas ini dengan menghadirkan fungsionalitas dengan cara yang berbeda, setiap tampilan melayani tujuan yang berbeda (Alan Dennis, Barbara Haley Wixom and David Tegarden, Barbara Haley Wixom, 2012)



Gambar 2. 5 Elemen Use case Diagram (Alan Dennis, Barbara Haley Wixom and David Tegarden, Barbara Haley Wixom, 2012)

Diagram *use case* menggambarkan dengan cara yang sangat sederhana fungsi utama sistem dan berbagai jenis pengguna yang akan berinteraksi dengannya. Berikut ini merupakan elemen *use case* diagram menurut (Alan Dennis, Barbara Haley Wixom and David Tegarden, Barbara Haley Wixom, 2012):

A. Aktor

Figur *stickman* berlabel pada diagram mewakili aktor. Sosok *stickman* berlabel pada diagram mewakili aktor. Aktor adalah orang atau sistem lain yang berinteraksi dengan dan memperoleh nilai dari sistem. Aktor bukanlah pengguna tertentu, tetapi peran yang dapat dimainkan pengguna saat berinteraksi dengan sistem. Aktor berada di luar sistem dan memulai use case.

B. Hubungan Asosiasi (*Association Relationship*)

Garis yang ditarik dari aktor ke *use case* mewakili sebuah asosiasi. Asosiasi biasanya mewakili komunikasi dua arah antara *use case* dan aktor, kemudian "*" ditampilkan di kedua ujung asosiasi mewakili multiplisitas

C. Use Case

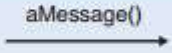
Use case, digambarkan dengan oval, adalah proses utama yang akan dilakukan sistem yang menguntungkan aktor dalam beberapa cara dan diberi label oleh frasa verba deskriptif. Ada kalanya satu *use case* akan menggunakan fungsionalitas atau memperluas fungsionalitas *use case* lain dalam diagram, dan ini ditunjukkan dengan *include* atau *extend relationship*.

D. Batasan Sistem (*Boundary System)*

Batasan sistem merupakan kotak yang mewakili sistem dan dengan jelas menggambarkan bagian mana dari diagram yang eksternal atau internal.

2.1.12.2 *Sequence* Diagram

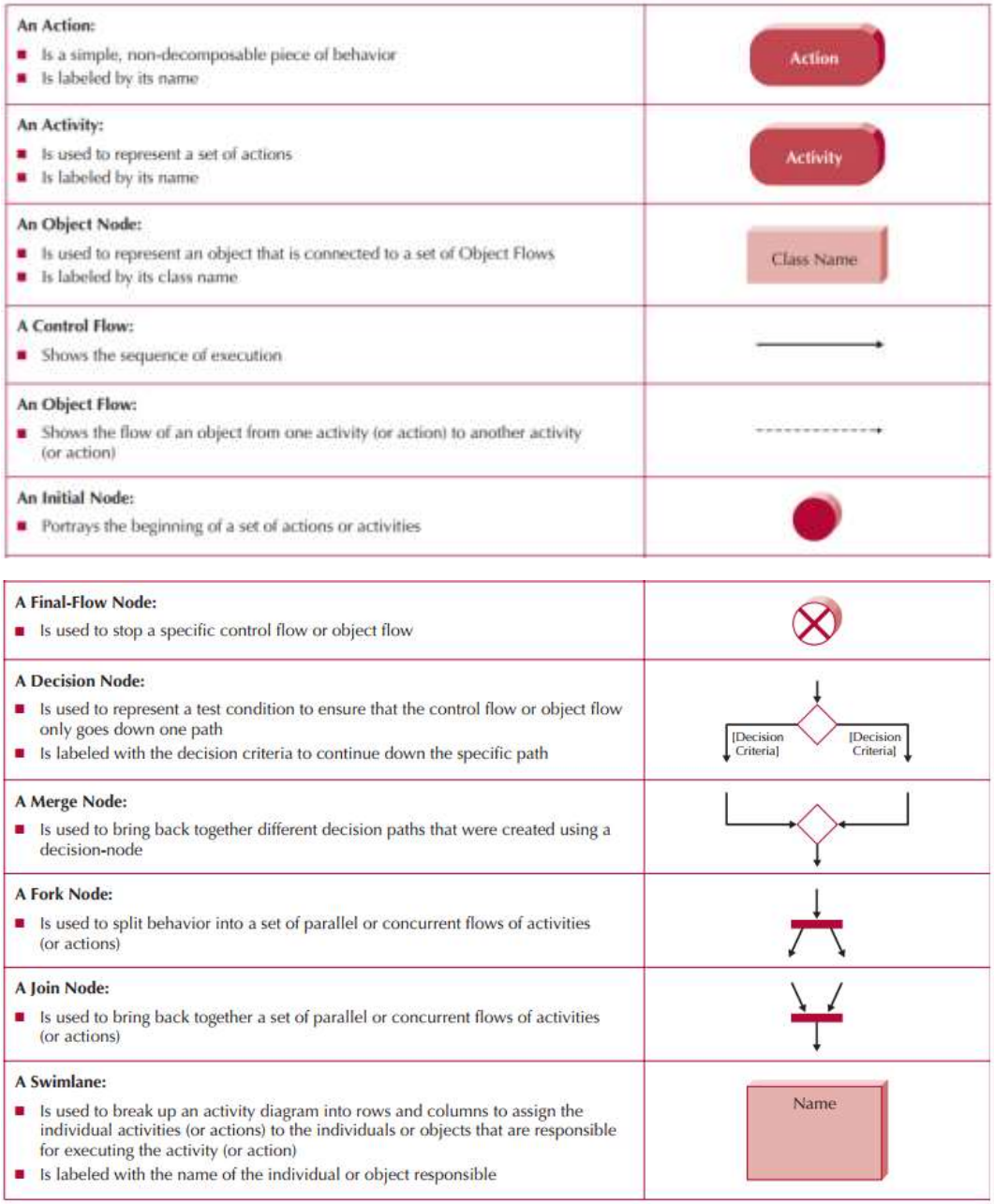
Diagram *Sequence* menggambarkan objek yang berpartisipasi dalam *use case* dan pesan yang lewat di antara mereka dari waktu ke waktu untuk satu *use case*. Diagram *sequence* adalah model dinamis yang mendukung pandangan dinamis dari sistem yang berkembang. Ini menunjukkan urutan eksplisit pesan yang dikirim antara objek dalam interaksi yang ditentukan. Diagram *sequence* dapat menjadi diagram urutan generik yang menunjukkan semua skenario yang mungkin untuk *use case*, tetapi biasanya setiap analis mengembangkan satu set diagram urutan instan, yang masing-masing menggambarkan skenario tunggal dalam *use case*.

Term and Definition	Symbol
An actor: ■ Is a person or system that derives benefit from and is external to the system. ■ Participates in a sequence by sending and/or receiving messages. ■ Is placed across the top of the diagram.	 anActor
An object: ■ Participates in a sequence by sending and/or receiving messages. ■ Is placed across the top of the diagram.	
A lifeline: ■ Denotes the life of an object during a sequence. ■ Contains an X at the point at which the class no longer interacts.	
A focus of control: ■ Is a long narrow rectangle placed atop a lifeline. ■ Denotes when an object is sending or receiving messages.	
A message: ■ Conveys information from one object to another one.	
Object destruction: ■ An *X* is placed at the end of an object's lifeline to show that it is going out of existence.	X

Gambar 2. 6 Elemen Diagram *Sequence* (Alan Dennis, Barbara Haley Wixom and David Tegarden, Barbara Haley Wixom, 2012)

2.1.12.3 *Activity Diagram*

Activity Diagram digunakan untuk memodelkan perilaku proses bisnis yang tidak bergantung pada objek. Diagram aktivitas dapat dilihat sebagai diagram aliran data yang canggih yang digunakan bersama dengan analisis terstruktur (Alan Dennis, Barbara Haley Wixom, 2012). Singkatnya *activity diagram* menggambarkan aktivitas utama dan hubungan di antara aktivitas dalam suatu proses.



Gambar 2. 7Elemen Diagram Activity
(Alan Dennis, Barbara Haley Wixom and David Tegarden, Barbara Haley Wixom, 2012)

2.1.13 HTML

HTML(*Hyper Text Markup Language*) adalah sekumpulan simbol-simbol atau tag-tag yang dituliskan dalam sebuah *file* yang dimaksudkan untuk menampilkan halaman pada *web browser*. Tag-tag HTML selalu diawali dengan <x> dan diakhiri dengan </x>. Sebuah halaman *website* akan diapit oleh tag <html>……</html>. File-file HTML selalu berakhiran dengan ekstensi *.htm atau

*.html. Jadi jika anda mengetik sebuah naskah dan menyimpannya dengan ekstensi *.html maka anda membuat *file* yang berformat HTML(Astamal, 2006)

2.1.14 CSS

CSS adalah kependekan dari *Cascading Style Sheet*, berfungsi untuk mempercantik penampilan HTML atau menentukan bagaimana elemen HTML ditampilkan, seperti menentukan posisi, merubah warna teks atau *background* dan lain sebagainya.

Di dalam CSS terdiri dari beberapa komponen yaitu (Ariona, 2016):

A. *Selector* adalah elemen/tag HTML yang ingin diberi *style*. Anda dapat menuliskan langsung nama tag yang ingin diberi *style* tanpa perlu menambahkan tanda <>. Pada contoh kode CSS di atas, kita akan memberi *style* pada seluruh tag h1 yang terdapat dalam *file* HTML.
B. *Property* adalah sifat-sifat yang ingin diterapkan pada *selector*, seperti warna *text*, warna *background*, jarak antar elemen, garis pinggir dan lain sebagainya.

2.1.15 *WEB Browser*

WEB Browser adalah suatu program atau *software* yang digunakan untuk menjelajahi internet atau untuk mencari informasi baik berupa gambar, teks, video, suara, dan animasi dari suatu *web* yang tersimpan didalam komputer (Ariona, 2016).

2.2 Tinjauan Studi

Dalam melaukan penelitian, penulis mencari, membaca dan mempelajari penelitian sejenis yang berkaitan dengan text mining khususnya membahas mengenai analisis sentiment *cyberbullying* untuk dijadikan bahan pembelajaran dan bahan perbandingan sebagai berikut:

1. Penelitian yang berjudul "Analisa Sentimen *Cyberbullying* Di Jejaring Sosial *Twitter* Dengan Algoritma *Naïve Bayes*" menjelaskan tentang bagaimana melakukan analisis sentimen untuk mendeteksi *tweet-tweet* yang mengandung unsur *cyberbullying* dan didapatkan hasil metode *naïve bayes* dapat digunakan untuk mengklasifikasikan *tweet* yang mengandung unsur *cyberbullying* dan evaluasi menggunakan *confusion matrix* didapatkan nilai akurasi sebesar 76%, *precission 76,09%, recall 97,22%* dan *specificity* sebesar 21,4% (Maulana and Ernawati, 2020).
2. Penelitian yang berjudul "*Analisis Sentimen Masyarakat Terhadap Pilpres 2019 Berdasarkan Opini Dari Twitter Menggunakan Metode Naive bayes Classifier*" menjelaskan bagaimana melakukan klasifikasi *cyberbullying* dalam bentuk ujaran kebencian terhadap pilpres 2019 menggunakan *naïve bayes* dan TF-IDF, evaluasi menggunakan *confusion matrix* didapatkan nilai akurasi sebesar 71% (Zuhri *et al.*, 2020)
3. Penelitian yang berjudul *"Cyberbullying Detection Using Machine Learning"* menjelaskan bagaimana melakukan pendeteksian *cyberbullying* menggunakan metode *machine learning* seperti *Random Forest, Naive bayes, SVM, Linear*

Regression. Hasil penelitian tersebut didaptkan nilai akurasi menggunakan metode *naïve bayes* 75,7% (Mehta *et al*., 2021).

4. Penelitian yang berjudul *"Filtering Impolite Words in Social Network Using Naïve Bayes Classifier"* penelitian ini menjelaskan bagaimana melakukan filter Kata-kata Sopan di Jejaring Sosial Menggunakan *Naïve Bayes Classifier*. Hasil dari penelitian tersebut didapatkan nilai presisi sebesar 66,67%, sistem recall sebesar 70%, sistem akurasi sebesar 74% dan *F-Measure* sebesar 68,29% (Lusiana, Gemini and Efendi, 2018)
5. Penelitian yang berjudul *"Detecting A Twitter Cyberbullying Using Machine Learning"* penelitian ini menjelaskan bagaimana komparasi metode *Naïve Bayes*, SVM + TF IDF dalam analisia sentimen *cyberbullying*. Hasil akurasi menggunakan metode *naïve bayes* 52,7%. dan menggunakan SVM 71,2% (Dalvi, Baliram Chavan and Halbe, 2020)
6. Penelitian yang berjudul "*Naive bayes Classifier on Twitter Sentiment Analysis BPJS of HEALTH*" penelitian ini menjelaskan bagaimana melakukan analisis sentimen pada BPJS kesehatan menggunakan metode *naïve bayes*. Hasil penelitian ini mendapatkan akurasi menggunakan metode *naïve bayes* 70% (Hidayati *et al*., 2019).
7. Penelitian yang berjudul "*Deteksi Cyberbullying pada Facebook Menggunakan Algoritma K-Nearest Neighbor"* penelitian ini menjelaskan bagaimana melakukan analisis sentimen pada media sosial Facebook menggunakan metode KNN. Hasil penelitian ini didapatkan nilai akurasi menggunakan metode KNN sebesar 71.43% (Hasan, 2021).
8. Penelitian yang berjudul "*Klasifikasi Naïve Bayes pada Analisis Sentimen atas Penolakan Dibukanya Larangan Ekspor Benih Lobster"* penelitian ini menjelaskan bagaimana melakukan analisis sentimen pada studi kasus Penolakan Dibukanya Larangan Ekspor Benih Lobster menggunakan metode *naïve bayes*. Hasil penelitian ini didapatkan nilai akurasi menggunakan metode metode *naïve bayes* 72.50% (Sulastri, 2020).
9. Penelitian yang berjudul *"Cyber-Bullying Detection Using Naive bayes And N-Gram"* penelitian ini menjelaskan bagaimana melakukan analisis sentimen cyberbulying menggunakan *naïve bayes* dikombinasikan dengan fitur N-GRAM. Hasil akurasi *Naïve Bayes + Uni-Gram 66.77%, Naïve Bayes + Bi-Gram 67.29%, Naïve Bayes + Tri-Gram 57.86%, Naïve Bayes + Ni-Gram 65.09%* (Sharma and Singh, 2020).
10. Penelitian yang berjudul *"Penggunaan Algoritma Klasifikasi Terhadap Analisa Sentimen Pemindahan Ibukota Dengan Pelabelan Otomatis"* Penelitian ini menjelaskan bagaimana melakukan analisis sentimen dengan pelabelan otomatis menggunakan Vader, metode yang digunakan SVM dan Naïve bayes, hasil penelitian ini didapatkan SVM menghasilkan nilai akurasi dan AUC yang paling baik yakni akurasi sebesar 76,40% dan AUC sebesar 0,771(Watori *et al.*, 2020).

***Tabel 2. 4* Tabel Tinjauan Studi**

No	Penulis	Judul	Masalah	Metode Pemecahan Masalah	Hasil Penelitian
1	Maulana and Ernawati, 2020	Analisa Sentimen *Cyberbullying* Di Jejaring Sosial *Twitter* Dengan Algoritma *Naïve Bayes*	Bagaimana melakukan analisis sentiment untuk mendeteksi *tweet-tweet* yang mengandung unsur *cyberbullying*	Metode *Naïve Bayes*	metode naïve bayes dapat digunakan untuk mengklasifikasikan *tweet* yang mengandung unsur *cyberbullying* dan evaluasi menggunakan *confusion matrix* didapatkan nilai akurasi sebesar *76%, precission 76,09%, recall 97,22%* dan *specificity* sebesar 21,4%
2	Zuhri *et al,*. 2020	Analisis Sentimen Masyarakat Terhadap Pilpres 2019 Berdasarkan Opini Dari *Twitter* Menggunakan Metode *Naive Bayes Classifier*	Bagaimana melakukan analisis sentimen untuk mendeteksi *tweet-tweet* yang mengandung unsur ujaran kebencian terhadap pilpres	*Naïve Bayes + TF IDF*	metode *naïve bayes + TF IDF* dapat digunakan untuk mengklasifikasikan *tweet* yang mengandung unsur *cyberbullying* dan evaluasi menggunakan *confusion matrix* didapatkan nilai akurasi sebesar 71%,
3	Mehta *et al.*, 2021	*Cyberbullying Detection Using Machine Learning*	Melakukan analisis sentimen *Cyberbullying* dengan beberapa metode	*Random Forest, Naive bayes, SVM, Linear Regression*	Hasil akurasi menggunakan metode *naïve bayes* 75,7%. Tertinggi menggunakan SVM 79%
4	Lusiana, Gemini and Efendi, 2018	*Filtering Impolite Words in Social Network*	Melakukan filter Kata-kata Sopan di Jejaring Sosial	*Naïve Bayes*	Hasil dari pengujian ini adalah sistem presisi sebesar 66,67%, sistem

		Using Naïve Bayes Classifier	Menggunakan *Naïve Bayes Classifier*		*recall* sebesar 70%, sistem akurasi sebesar 74% dan *F-Measure* sebesar 68,29%
5	Dalvi, Baliram Chavan and Halbe, 2020	*Detecting A Twitter Cyberbullying Using Machine Learning*	Komparasi metode *Naïve bayes* dengan SVM dalam analisia sentimen *cyberbullying*	*Naïve Bayes, SVM + TF IDF*	Hasil akurasi menggunakan metode *naïve bayes* 52,7%. dan menggunakan SVM 71,2%
6	Hidayati *et al.*, 2019	*Naive bayes Classifier on Twitter Sentiment Analysis BPJS of HEALTH*	Melakukan analisis sentimen pada bpjs kesehatan	*Naïve Bayes*	Hasil akurasi menggunakan metode *naïve bayes* 70%.
7	Hasan, 2021	Deteksi *Cyberbullying* pada *Facebook* Menggunakan Algoritma *K-Nearest Neighbor*	Melakukan analisis sentimen pada media sosial *Facebook*	KNN	Hasil akurasi menggunakan metode KNN 71.43%.
8	Sulastri, 2020	Klasifikasi *Naïve Bayes* pada Analisis Sentimen atas Penolakan Dibukanya Larangan Ekspor Benih Lobster	Melakukan analisis sentimen pada Penolakan Dibukanya Larangan Ekspor Benih Lobster	*Naïve Bayes*	Hasil akurasi menggunakan metode *naïve bayes* 72.50%.
9	Sharma and Singh, 2020	*Cyber-Bullying Detection Using Naive bayes And N-Gram*	Melakukan analisis sentimen *cyberbulying* menggunakan *NB + N-GRAM*	*Naïve Byaes + N-GRAM*	*Hasil akurasi Naïve Bayes + Uni-Gram with 66.77% Naïve Bayes + Bi-Gram with 67.29% Naïve Bayes + Tri-Gram with 57.86% Naïve Bayes + Ni-Gram with 65.09%*

10	Watori *et al.*, 2020	Penggunaan Algoritma Klasifikasi Terhadap Analisa Sentimen Pemindahan Ibukota Dengan Pelabelan Otomatis	Melakukan analisis sentimen dengan pelabelan otomatis dengan Vader	*SVM, Naïve Bayes*	Support Vector Machine menghasilkan nilai akurasi dan AUC yang paling baik yakni akurasi sebesar 76,40% dan AUC sebesar 0,771

Penelitian ini mempelajari konsep, teknik dan algoritma dari studi literatur yang telah dibahas. Dari tinjauan studi beberapa penelitian diatas sebelumnya, penulis megetahui bagaimana melakukan klasifikasi yang mengandung unsur *cyberbullying* dengan tingkat akurasi rata-rata 70%-76%.

Penulis menilai tingkat akurasi tersebut masih bisa di tingkatkan diatas 76% dengan cara mengkombinasikan dengan fitur seleksi. Seleksi fitur merupakan solusi dari masalah kategorisasi teks yang sudah banyak dikenal. Dalam kategorisasi teks, penting untuk memiliki pilihan fitur yang baik. Seleksi fitur adalah strategi yang dapat digunakan untuk meningkatkan akurasi kategorisasi, efektivitas, dan efisiensi komputasi (S. and B., 2017)

Berdasarkan ringkasan tinjauan studi, maka yang menjadi perbedaan dalam penelitian ini yaitu:

1. Penelitian ini menggunakan sumber data *twitter* dengan keyword "saipul jamil" yang di dapatkan melalui *API Twitter*
2. Penelitian ini mengembangkan 3 algoritma yaitu *naive bayes* yang di kombinasikan kemudian melakukan komparasi fitur seleksi menggunakan *chi square* dan *information gain* dengan alasan sebagai berikut:
   a. Kelebihan metode *naïve bayes* sederhana namun memiliki akurasi yang tinggi. Keuntungan menggunakan metode *naïve bayes* adalah hanya membutuhkan sedikit data latih untuk mengestimasi yang diperlukan untuk klasifikasi dan sensitif terhadap seleksi fitur (Nurhayati *et al.*, 2019).
   b. Masalah utama dalam klasifikasi teks adalah dimensi tinggi fitur ruang, sering terjadi pada teks yang memiliki puluhan ribu fitur. Sebagian besar fitur ini tidak relevan dan tidak berguna untuk klasifikasi teks bahkan dapat mengurangi tingkat akurasi, oleh karena itu pemilihan fitur yang sesuai sangat diperlukan (Nurhayati *et al.*, 2019) dan dalam proses fitur seleksi, *Chi square* x^2 *statistic* dan *Information Gain* (IG) merupakan metode terbaik dari fitur seleksi (Hung *et al.*, 2015)
3. Penelitian ini menggunakan bahasa pemograman python
4. Metode pengembangan menggunakan Model CRISP-DM, karena CRISP-DM sangat lengkap dan didokumentasikan. Semua tahapannya diatur, terstruktur, dan ditentukan, memungkinkan proyek dapat dengan mudah dipahami atau

direvisi (Azevedo and Santos, 2008) dan CRISP-DM dianggap sebagai standar de facto untuk proyek analitik, *data mining*, dan *data science* (Martinez-Plumed *et al.*, 2019)

2.3 Tinjauan Objek Penelitian

Pada penelitian ini, penulis mengangkat isu *cyberbullying* yang dilakukan oleh pengguna media sosial twitter terhadap penyanyi dangdut yang bernama Saipul Jamil.

Saipul divonis hukuman penjara di dua kasus. Pada 14 Juni 2016, Pengadilan Negara Jakarta Utara menjatuhkan hukuman 3 tahun kepada Saipul Jamil. Kala itu, hakim menyatakan pedangdut itu terbukti melanggar pasal 292 KUHP tentang perbuatan cabul karena mencabuli korban yang tinggal di rumahnya, dan korban saat itu masih usia dini.

Selain kasus pencabulan, Saipul Jamil juga diadili di kasus suap. Pangkal masalahnya adalah Saipul lewat pengacaranya menyogok majelis hakim. Belakangan, duit suap itu hanya dinikmati panitera pengganti Rohadi. Pada 2017, Saipul Jamil divonis 3 tahun bui. Saipul Jamil terbukti bersalah menyuap majelis hakim di PN Jakarta Utara sebesar Rp 250 juta. Hakim menyatakan uang Rp 250 juta dari rekening Saipul untuk mempengaruhi hakim PN Jakarta Pusat dalam putusan hakim dalam perkara pencabulan.

Pada tanggal 2 September 2021, Saipul Jamil resmi bebas dari Lapas Cipinang. Saipul Jamil bebas setelah mendapatkan remisi sebanyak 30 bulan yang seharusnya tidak layak ia dapatkan. Selain itu, yang menjadi kontroversi adalah mantan narapidana pencabulan anak diusia dini ini masih disambut meriah ketika keluar dari penjara dengan berkalung bunga dan melambaikan tangan menyampaikan apa yang ingin dilakukannya ketika keluar dari penjara. hal ini menjadi trending di media sosial khusus nya *twitter* untuk memboikot saipul jamil dan sampai ada ajakan petisi untuk melarang saipul jamil untuk tampil di TV.

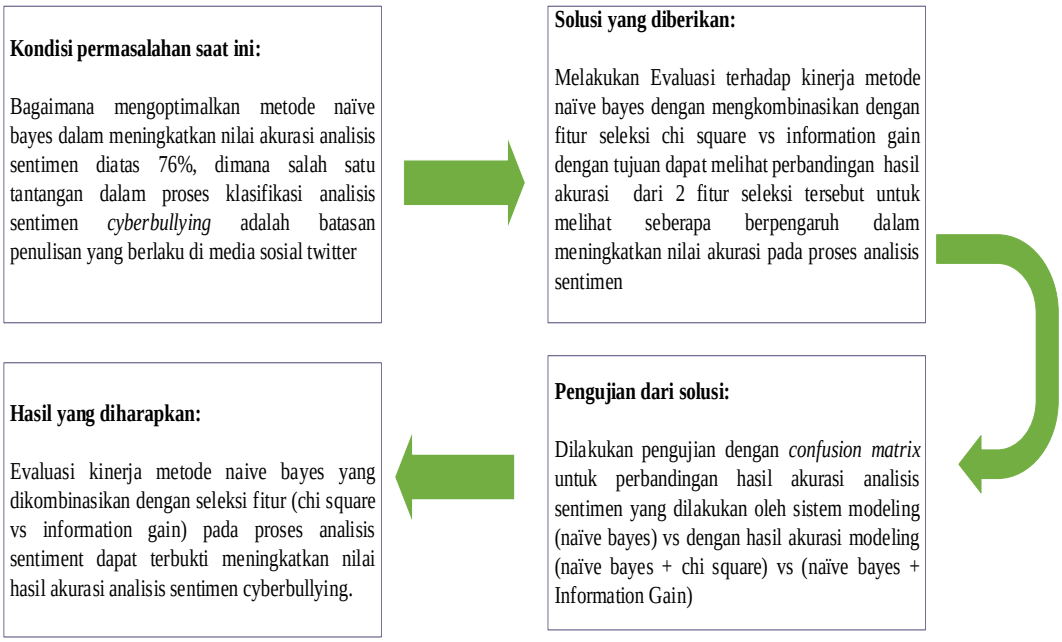
Gambar 2. 8 **Info Berita Saipul Jamil**

Penelitian ini menggunakan sumber data twitter dengan *keyword* "saipul jamil"yang di dapatkan melalui *API Twitter*. Data scraping di dapatkan 18067 *tweet* yang mengandung *keyword* tersebut

A	B	C
Created-At	From-User	Text
09/02/2021	Winner	welcome back bang @saipuljamil
09/02/2021	pray	@saipuljamil Bakalan nge-hap lagi ga, om?
09/02/2021	Jeni Nata Widianingrum	@saipuljamil Tuhan Maha Pemaaf dan menerima Taubat.Taubat artinya tdk melakukan perbuatan tercela yg sama lagi. Ini ha
09/02/2021	Kureta ID	Bebas dari Penjara, Saipul Jamil Ngaku Trauma
09/02/2021	Lee	@saipuljamil Untuk orangtua, tolong jgn hanya memperhatikan anak cewe. Dengan ini anak cowo juga harus di perhatikan da
09/02/2021	Bung FALDO DEMPUL.	@saipuljamil Semoga ga diulangi lagi perbuatan yg itu ya bro...
09/02/2021	Anonymous_2024	@saipuljamil Semoga bisa berubah lebih baik?
09/02/2021	Pejuang_RakyatJelata	@Anonymous_2024 @saipuljamil Aamiin YRA.
09/02/2021	igifagan	@saipuljamil Om jangan bool gw ya...
09/02/2021	Pejuang_RakyatJelata	@saipuljamil semoga jgn mengulangi lagi perbuatan yg tdk baik.
09/02/2021	mimin on the way	@saipuljamil Nyanyi dulu mase
09/02/2021	radio ga ga, radio goo go	@saipuljamil Alhamdulillah lulus gelarnya apa bg
09/02/2021	Mandailing_Sipit	@saipuljamil Siap2 dipanggil YT dan TV.. Dan jd motivator dan duta 'hap' seindonesia. Semangat...
09/02/2021	A?d?r?i?a?n?	Alhamdulillah tumpengan
09/02/2021	Wa	@saipuljamil Enak banget ya keluar penjara dapet sambutan kaya abis menang emas olimpiade
09/02/2021	cincau mas hiho // want	@stars4jwoo @saipuljamil maaf ya van aku cepuin
09/02/2021	EUNZIW..	@saipuljamil @RBTHIndonesia Kok dipenjara bisa maen hp? ?
09/02/2021	titit tembem	@saipuljamil @AlrasyidAad Selamat bang ipul
09/02/2021	Kamte Ku Ngena	@saipuljamil homo
09/02/2021	Becandain Aja	@eunziw @saipuljamil @RBTHIndonesia Nah iya mikir jg. boleh kali ya?
09/02/2021	mekomengo	@saipuljamil wilujeng tusbol

Gambar 2. 9 **Hasil Scrapping Dataset**

2.4 Kerangka Konsep / Pola Pemikiran Masalah



Gambar 2. 10 **Kerangka Konsep Penelitian**

2.5 Hipotesis

Berdasarkan kerangka konsep yang telah dijabarkan, dapat dirumuskan hipotesis sebagai berikut yaitu diduga metode *naive bayes* yang dikombinasikan dengan seleksi fitur (*NB + Chi square vs NB + Information Gain)* dapat meningkatkan nilai akurasi analisis sentimen *cyberbullying* di atas 76%.

BAB III
METODOLOGI DAN RANCANGAN PENELITIAN

3.1 Metode Penelitian

Metode yang digunakan dalam penelitian ini menggunakan Crips DM. dalam siklus pengembangannya CRISP-DM dianggap sebagai metodologi *data mining* terlengkap dalam hal pemenuhan kebutuhan proyek industri, dan telah menjadi yang paling luas penggunaannya dalam proyek *data mining*. Singkatnya, CRISP-DM dianggap sebagai standar *de facto* untuk proyek analitik, data *mining*, dan data *science* (Martinez-Plumed *et al.*, 2019).

Metode CRIPS-DM terdiri dari 6 tahapan yaitu:

1. *Bussines Understanding*
2. *Data Understanding*
3. *Data Preparation,*
4. *Modeling*
5. *Evaluation*
6. *Deployment.*

3.2 Metode Pemilihan Sampling

Metode pemilihan *sample* pada penelitian ini adalah dengan cara pengambilan data berdasarkan sumber yang menggunakan data *primer* yaitu data yang diperoleh dari sumber dengan teknik pengumpulan data. Pada tahap ini menggunakan metode *simple random sampling*. dikatakan *simple* atau sederhana sebab pengambilan *sample* anggota populasi dilakukan secara acak, tanpa memperhatikan strata yang terdapat dalam populasi tersebut dan untuk menentukan besar jumlah *sample*, peneilti menggunakan rumus slovin sebagai berikut :

$$n = \frac{N}{1 + Ne^2} \tag{7}$$

Keterangan :
N = Populasi.
n = *sample*.
e = Tingkat kesalahan penarikan *sample* 2% dan tingkat kepercayaan 98%

menurut survei databoks, indonesia masuk kedalam list daftar 10 negara pengguna twitter di dunia sebanyak 15,7 juta pengguna.

sehingga berdasarkan rumus slovin, *sample* yang diambil ada lah sebagai berikut :

$$n = \frac{15.700.000}{1 + 15.700.000,(0{,}02)^2}$$

$$n = \frac{15.700.000}{1 + 15.700.000.(0{,}0004)}$$

$$n = \frac{15.700.000}{6281}$$

$$n = 2.499{,}60197\ dibulatkan\ jadi\ 2500$$

Penulis melakukan *scrapping* data sebanyak 180067 twit, kemudian dilakukan proses *cleansing preprocessing* data dan mengambil data *sample* dari perhitungan diatas, maka *sample* yang diambil adalah 2500 *sample* untuk dilakukan *training*.

3.3 Metode Pengumpulan Data

3.3.1 Studi Pustaka

Mempelajari teori-teori yang digunakan dalam penelitian. Sumber teori utama berasal dari buku dan jurnal mengenai analisis sentimen, mengenai media sosial dan mengenai UU ITE nomor 11 tahun 2008 untuk memperkaya informasi dari media lainnya. Melalui studi pustaka, penulis menganalisa metode apa yang tepat untuk memecahkan permasalahan yang dikaji dalam penelitian ini dan tahapan apa saja yang akan dilakukan.

3.3.2 Studi Literatur

Dari proses studi literatur penulis mempelajari penelitian-penelitian yang berhubungan dengan analisis *cyberbullying* dan mencari metode yang tepat untuk menyelesaikan permasalahan yang telah di identifikasi sebelumnya.

3.3.3 Studi Observasi

Dalam penelitian ini menggunakan sumber data *primer*, yaitu data yang dikumpulkan langsung dari *twitter* dengan observasi. data dikumpulkan dengan cara *scraping twitter* menggunakan *API twitter* mulai tanggal 2 September s/d 09 September 2021. Pemilihan *sample* data menggunakan *keyword* "Saipul Jamil" dan diklasifikasn menjadi 2 label positif dan negatif.

Contoh Twit label Positif :

"Alhamdulillah semoga menjadi orang yang lebih baik lagi ya bang @saipuljamil... 👍👍👍"

Contoh Twit label Negatif :

"@saipuljamil Kok bisa ya orang bangga keluar dari penjara tanpa mikirin beban pikiran si korban yg bisa aja trauma ? Dasar Pedofil"

3.3 Teknik Analisis dan Pengujian Data

Pada Tahapan penelitian ini, penulis menggunakan *framework* CRIPS-DM terdiri dari 6 tahapan yaitu *bussines understanding, data understanding, data preparation, modeling, evaluation, deployment*.

1. *Bussines Understanding*

Tahap awal ini berfokus pada pemahaman tujuan dan persyaratan dan kemudian mengubah pengetahuan ini menjadi definisi masalah. Di penelitian ini, penulis melakukan *experiment research* bagaimana mengevaluasi kinerja metode *naïve bayes* dengan mengkombinasikan serta komparasi fitur seleksi *chi square* dengan *Information Gain* dalam meningkatkan nilai akurasi pada proses analisis sentimen diatas 76% yang telah dilakukan penelitian-penelitian sebelumnya dengan mempelajari studi pustaka, studi literatur sejenis dan observasi melakukan pencarian topik di *twitter.*

2. ***Data Understanding***

Tahap pemahaman data ini dimulai dengan pengumpulan data awal. Di penelitian ini penulis melakukan pengumpulan data awal terkait isu *cyberbullying* yang di tujukan kepada penyanyi dangdut saipul jamil dalam sebuah postingan *twitter* dari tanggal tanggal 2 September s/d 09 September 2021. Pengumpulan data menggunakan python dan didapatkan 18.0067 twit. Disini penulis berkoordinasi dengan ahli bahasa untuk membantu melabeli dataset (positif/negatif) yang telah dikumpulkan agar siap di olah

3. Data Preparation

Tahap persiapan data mencakup semua kegiatan untuk membangun *dataset* akhir. Pada tahap ini dilakukan *preprocessing text* sebagai berikut:

A. Tahap *Tokenize* : Proses memisahkan teks menjadi potongan-potongan kata
B. Tahap *Case Folding & Remove Punctuation* : Tahap *Case folding* dimulai dengan proses merubah teks menjadi huruf kecil sedangkan *remove punctuation* adalah proses menghilangkan tanda baca atau simbol yang ada dalam *dataset*
C. Tahap *Stopword* : Proses menghilangkan kata-kata yang tidak penting
D. Tahap *Stemming* : Proses mengubah kata berimbuhan menjadi kata dasar

4. *Modeling*

Pada tahapan ini, berbagai teknik pemodelan dipilih dan diterapkan, dan parameter dikalibrasi ke nilai optimal. Disini penulis sudah masuk dalam proses mining, *dataset* yang sudah siap olah dilakukan dan ditambahkan fitur seleksi *chi square* dan *information gain* kemudian membangun model *naïve bayes (nb), naïve bayes (nb) + chi square*, dan *naïve bayes (nb) + information gain.*

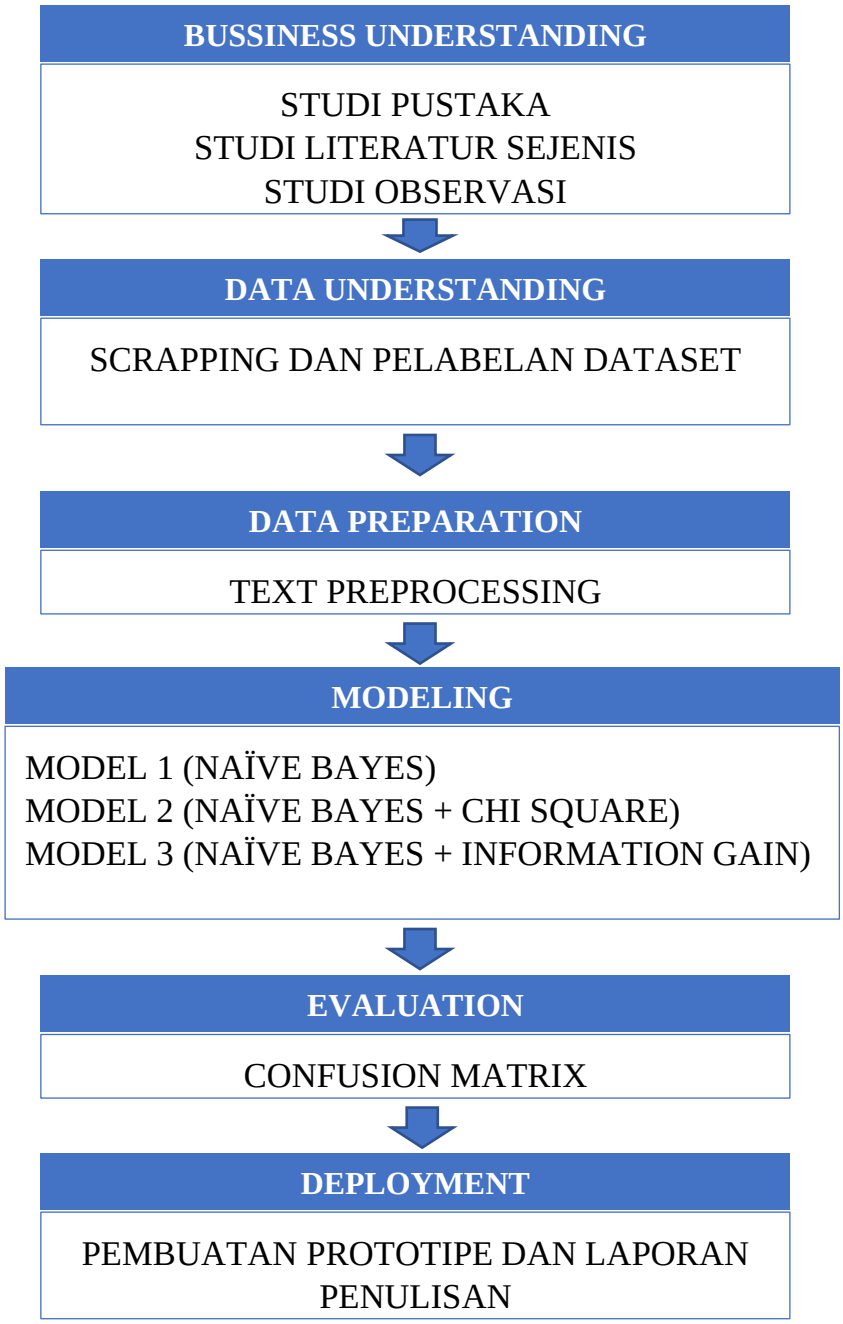
5. *Evaluation*

Pada Tahap ini penulis melakukan pembuatan laporan hasil penelitian, pembuatan laporan hasil dapat dilakukan setelah selesai evaluasi model dengan *confusion matrix*

6. *Deployment*

Pada tahapan ini penulis membuat rancangan penelitian berupa *usecase diagram, activity diagram dan sequence diagram*, kemudian membuat prototipe pengembangan model menggunakan FLASK.

3.4 Langkah-Langkah Penelitian



Gambar 3. 1 Langkah Penelitian

3.5 Jadwal Penelitian

Tabel 3. 1 Jadwal Penelitian

NO	Kegiatan	Bulan					
		Juli	Agustus	September	Oktober	November	Desember
1	Pemilihan Topik						
2	Pengumpulan data studi pustaka						
3	Pengumpulan data studi literatur						
4	Pengumpulan data observasi (dataset)						
5	Analis data & evaluasi						
6	Penyusunan Laporan penulisan						

BAB IV
PEMBAHASAN

4.1 Pembahasan Hasil Penelitian

Pada bab ini akan diuraikan mengenai hasil penelitian dan pengukuran kinerja model. Mengenai pengembangan aplikasi dibahas dalam *deployment* aplikasi untuk menunjukan bahwa hasil dari aplikasi yang dibuat sesuai dengan yang di harapkan. Untuk pengukuran kinerja model menggunakan metode *naive bayes* yang di kombinasikan dengan fitur seleksi *chi square* dan *Information gain* yang akan di jelaskan pada hasil penelitian ini.

1. Bussiness Understanding

Dalam penelitian ini, penulis melakukan penelitian *experiment research* untuk mengevaluasi kinerja *naïve bayes* yang dikombinasikan dengan fitur seleksi (*chi square vs Information Gain)* dalam meningkatkan nilai akurasi pada proses analisis sentimen.

2. Data Understanding

Tahap ini merupakan pendataan awal terkait *cyberbullying* penyanyi dangdut Saipul Jamil dari tanggal 2 September sampai dengan 9 September 2021. Pengumpulan data menggunakan kata kunci pencarian "saipul jamil" dan diperoleh 18.0067 *tweet*. Di sini penulis berkoordinasi dengan ahli bahasa untuk meminta bantuan melabelkan data yang dikumpulkan (positif / negatif) dan mempersiapkannya untuk diproses.

3. Data Preparation

Tahap persiapan data mencakup semua kegiatan untuk membangun *dataset* akhir. Pada tahap ini dilakukan *text preprocessing* data adalah sebagai berikut:

a. Tokenize

Tahap *Tokenize* merupakan proses memisahkan teks menjadi potongan-potongan kata yang menyusunnya. Dengan menggunakan `nltk.word_tokenize` pada python.

```
df['tokenize'] = df.apply(lambda row: nltk.word_tokenize(row[
'Text']), axis=1)
```

Gambar 4. 1 Implementasi *Tokenize*

Berikut merupakan hasil *text preprocessing* tahap *tokenize* yang terdapat pada tabel 4.1 berikut:

Tabel 4. 1* Hasil *Text Preprocessing Tokenization

Sebelum	Sesudah
Bebas dari Penjara, Saipul Jamil Ngaku Trauma\n@saipuljamil\n https://t.co/sgRJ4nXieH	['Bebas', 'dari', 'Penjara', ',', 'Saipul', 'Jamil', 'Ngaku', 'Trauma', '@', 'saipuljamil', 'https', ':', '//t.co/sgRJ4nXieH']
@saipuljamil Tuhan Maha Pemaaf dan menerima Taubat.Taubat artinya tdk melakukan perbuatan tercela yg sama lagi. Ini hanya Peringatan Bang Saipul Jamil ,Akibat perbuatan anda trauma seumur hidup diterima korban. Tdk bisa perbaiki kerusakan yg dibuat. Jadi Jgn sampai berbuat lagi ! https://t.co/iayMFyx0TT	['@', 'saipuljamil', 'Tuhan', 'Maha', 'Pemaaf', 'dan', 'menerima', 'Taubat.Taubat', 'artinya', 'tdk', 'melakukan', 'perbuatan', 'tercela', 'yg', 'sama', 'lagi', '.', 'Ini', 'hanya', 'Peringatan', 'Bang', 'Saipul', 'Jamil', ',', 'Akibat', 'perbuatan', 'anda', 'trauma', 'seumur', 'hidup', 'diterima', 'korban', '.', 'Tdk', 'bisa', 'perbaiki', 'kerusakan', 'yg', 'dibuat', '.', 'Jadi', 'Jgn', 'sampai', 'berbuat', 'lagi', '!', 'https', ':', '//t.co/iayMFyx0TT']

b. Case Folding & Remove Punctuation

Pada tahap ini merupakan proses mengubah sebuah kata menjadi bentuk yang sama atau huruf kecil. Pada langkah ini, penulis mengubah semua kata dalam kalimat menjadi huruf kecil menggunakan fungsi "*lower case*" python. Sedangkan Tahap *Remove Punctuation* merupakan Proses menghilangkan tanda baca atau simbol yang ada dalam data

```
import re
import string

def clean_text(text):
  text = text.lower() #lowercase atau case folding
  text = re.sub('@[^\s]+', '', text) #remove username
  text = re.sub('\[.*?\]', '', text) # remove square brackets
  text = re.sub('((www\.[^\s]+)|(https?://[^\s]+))', '', text
) # remove URLs
  text = re.sub('[%s]' % re.escape(string.punctuation), '', t
ext) # remove punctuation
  text = re.sub('\w*\d\w*', '', text)
  text = re.sub('[‘’“”…]', '', text)
  text = re.sub('\n', '', text)
  return text

clean1 = lambda x: clean_text(x)
```

Gambar 4. 2* Implementasi *Case Folding & Remove Punctuation

Berikut merupakan hasil *text preprocessing* tahap *case folding & remove punctuation* yang terdapat pada tabel 4.2 berikut:

Tabel 4. 2 Sebelum dan Sesudah *Case Folding & Remove Punctuation*

Sebelum	Sesudah
Bebas dari Penjara, Saipul Jamil Ngaku Trauma\n@saipuljamil\n https://t.co/sgRJ4nXieH	bebas dari penjara saipul jamil ngaku trauma
@saipuljamil Tuhan Maha Pemaaf dan menerima Taubat.Taubat artinya tdk melakukan perbuatan tercela yg sama lagi. Ini hanya Peringatan Bang Saipul Jamil ,Akibat perbuatan anda trauma seumur hidup diterima korban. Tdk bisa perbaiki kerusakan yg dibuat. Jadi Jgn sampai berbuat lagi ! https://t.co/iayMFyx0TT	tuhan maha pemaaf dan menerima taubattaubat artinya tdk melakukan perbuatan tercela yg sama lagi ini hanya peringatan bang saipul jamil akibat perbuatan anda trauma seumur hidup diterima korban tdk bisa perbaiki kerusakan yg dibuat jadi jgn sampai berbuat lagi

c. Stopword

Tahap ini merupakan proses untuk menghilangkan kata-kata umum dan sering yang tidak memiliki pengaruh signifikan dalam sebuah kalimat. sehingga data dapat diproses lebih efisien pada tahap selanjutnya. Mengimpor daftar *stopwords* menggunakan dari *library nltk.corpus*.

```
import nltk
nltk.download('punkt')
nltk.download('stopwords')

from nltk.corpus import stopwords
additional = ['saipul', 'jamil', 'rt', 'retweet']
sw = set().union(stopwords.words('indonesian'), additional
)
```

Gambar 4. 3 Implementasi Stopword

Berikut merupakan hasil *text preprocessing* tahap *stopword* yang terdapat pada tabel 4.3 berikut:

Tabel 4. 3 Sebelum & Sesudah *Stopword*

Sebelum	Sesudah
bebas dari penjara saipul jamil ngaku trauma	bebas penjara ngaku trauma

tuhan maha pemaaf dan menerima taubattaubat artinya tdk melakukan perbuatan tercela yg sama lagi ini hanya peringatan bang saipul jamil akibat perbuatan anda trauma seumur hidup diterima korban tdk bisa perbaiki kerusakan yg dibuat jadi jgn sampai berbuat lagi	tuhan maha pemaaf menerima taubattaubat tdk perbuatan tercela yg peringatan bang akibat perbuatan trauma seumur hidup diterima korban tdk perbaiki kerusakan yg jgn berbuat

d. Stemming

Tahap *Stemming* : Proses mengubah kata berimbuhan menjadi kata dasar. Pada tahap ini menggunakan *library* Python Sastrawi untuk menghilangkan kata-kata imbuhan dalam bahasa Indonesia ke bentuk dasarnya.

```
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory

factory = StemmerFactory()
stemming = factory.create_stemmer()

output = [(stemming.stem(token)) for token in text]
```

Gambar 4. 4 Implemetasi *Stemming*

Berikut merupakan hasil *text preprocessing* tahap *stemming* yang terdapat pada tabel 4.4 berikut:

Tabel 4. 4 Sebelum dan Sesudah *Stemming*

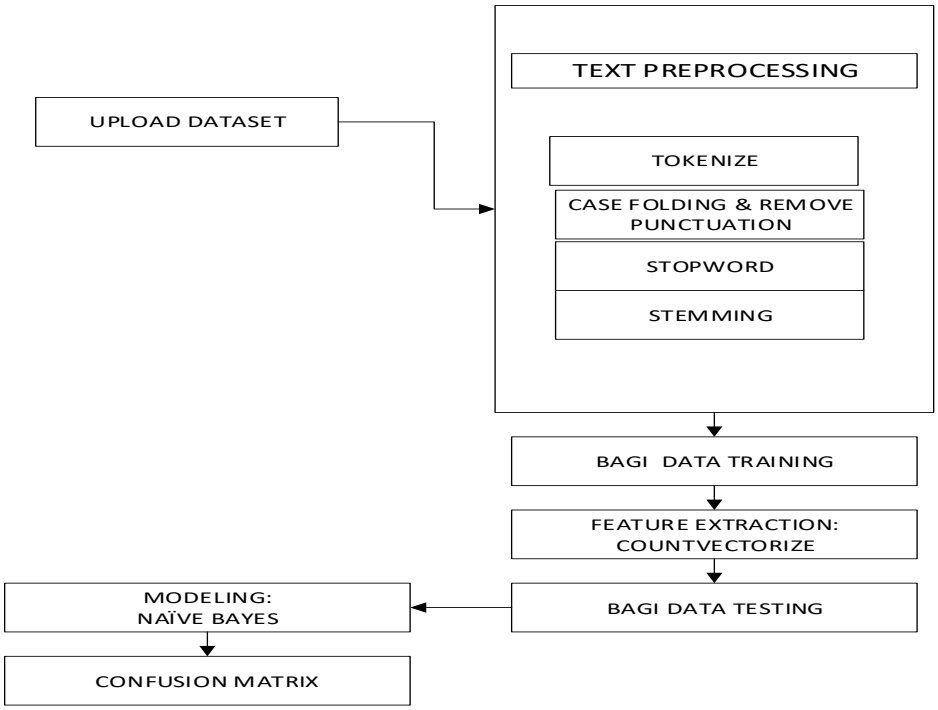
Sebelum	Sesudah
bebas penjara ngaku trauma	bebas penjara ngaku trauma
tuhan maha pemaaf menerima taubattaubat tdk perbuatan tercela yg peringatan bang akibat perbuatan trauma seumur hidup diterima korban tdk perbaiki kerusakan yg jgn berbuat	tuhan maha maaf terima taubattaubat tdk buat cela yg ingat bang akibat buat trauma umur hidup terima korban tdk baik rusa yg jgn buat

4. Modeling

Pada tahap ini dilakukan pemodelan *naïve bayes, naïve bayes + chi square, naïve bayes + information gain,* dan didapatkan hasil akurasi 3 model sebagai berikut :

A. Model *Naïve bayes*

Pada tahap ini merupakan proses mining menggunakan model *naïve bayes* tanpa fitur seleksi. Dengan alur sebagai berikut :



Gambar 4. 5 Alur Pemodelan *Naive bayes* Tanpa Fitur Seleksi

Tahap pertama adalah melakukan *upload dataset* mentah yang sudah diberi label (positif / negatif) untuk dilakukan proses *text preprocessing* menggunakan *google collab*. Selanjutnya, dilakukan *text preprocessing* data dengan melakukan *tokenize, case folding & remove punctuation, stopword, stemming*. Dari 180067 data mentah, Setelah melakukan *text preprocessing* tersisa 6787 data. Setelah dilakukan pembersihan atau *text preprocessing,* kemudian dilakukan konversi *label to number*. Fungsi ini untuk mengubah kedalam tipe data integer 1(positif),-1(negatif).

Kemudian 6787 data tersebut di bagi lagi menjadi 2500 untuk *sample* data yang didapatkan memalui perhitungan slovin. lalu di konversi ke *vector* menggunakan fitur *extraction counvectorize* untuk digunakan sebagai masukan dalam algoritma pembelajaran mesin. Setelah itu, dari 2500 data tersebut dilakukan split data untuk *testing*. Split yang digunakan adalah 80:20.

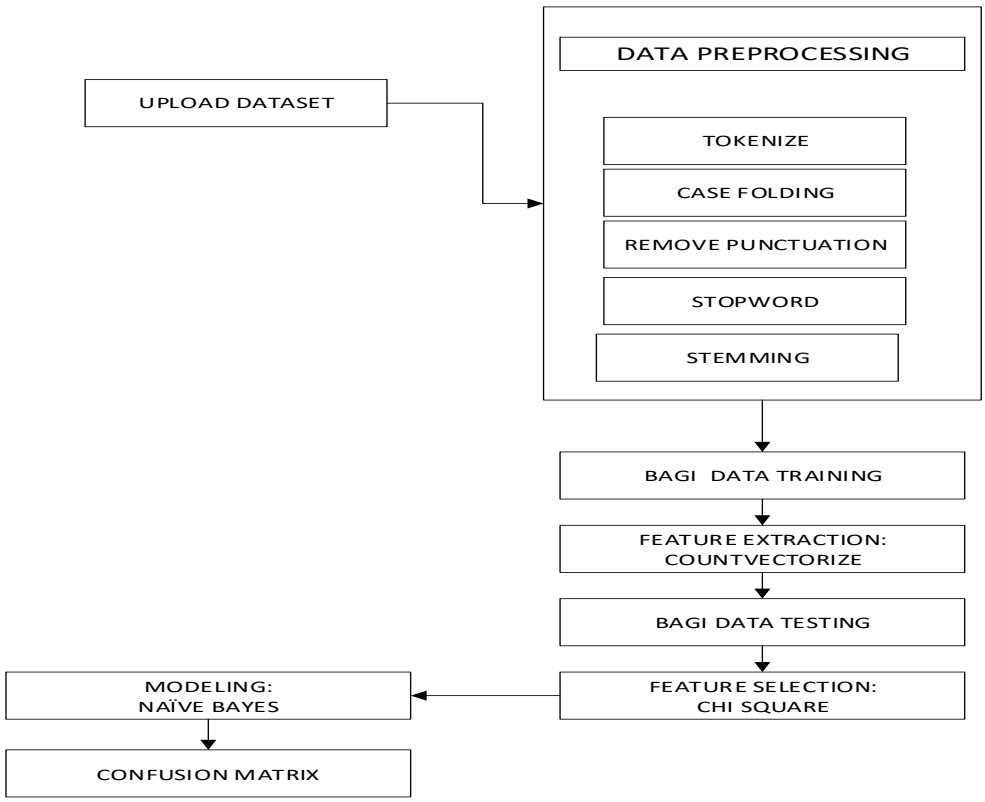
Tahap selanjutnya dilakukan *modeling* menggunakan metode *naïve bayes* tanpa fitur seleksi. Dan di dapatkan hasil akurasi sebesar 0.856 atau dibulatkan menjadi 0.86. Setelah proses *modeling,* tahap selanjutnya adalah melakukan *confusion matrix* untuk mengukur performa klasifikasi.

```
from sklearn.naive_bayes import MultinomialNB
MultinomialNB_model1 = MultinomialNB()

model1 = MultinomialNB_model1.fit(X_train, y_train)
prediction1 = model1.predict(X_test)
print('Akurasi model NB tanpa fitur seleksi: ', model1.score(
X_test, y_test))
```

Gambar 4. 6 Impelentasi Pemodelan *Naive bayes* Tanpa Fitur Seleksi

B. Model *Naïve Bayes + Chi Square*

Pada tahap ini merupakan proses *mining* menggunakan model *naïve bayes* dengan kombinasi fitur seleksi *chi square*. Dengan penjelasan alur sebagai berikut:



Gambar 4. 7 Alur Pemodelan *Naive bayes + Chi Square*

Tahap pertama adalah melakukan *upload dataset* mentah yang sudah diberi label (positif / negatif) untuk dilakukan proses *text preprocessing* menggunakan *google collab*. Selanjutnya, dilakukan *text preprocessing* data dengan melakukan *tokenize, case folding & remove punctuation, stopword, stemming*. Dari 180067 data mentah, Setelah melakukan *text preprocessing* tersisa 6787 data. Setelah dilakukan pembersihan atau *text preprocessing,* kemudian dilakukan konversi *label to number*. Fungsi ini untuk mengubah kedalam tipe data *integer* 1(positif), -1(negatif).

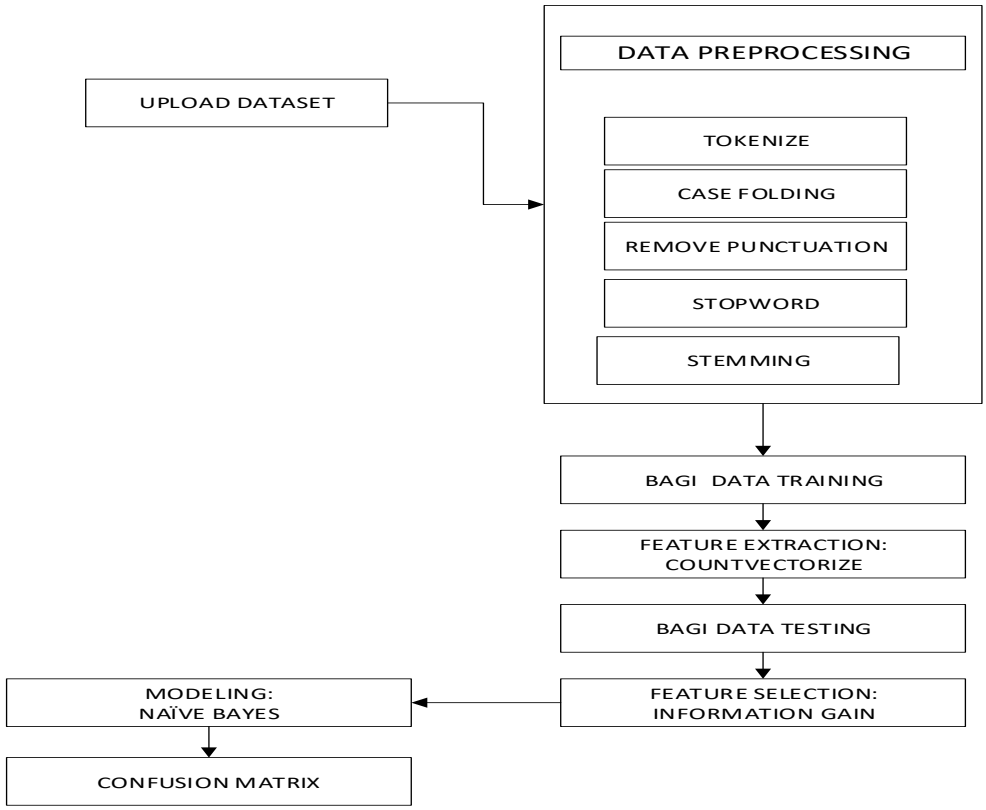
Kemudian 6787 data tersebut di bagi lagi menjadi 2500 untuk *sample* data yang didaptkan memalui perhitungan slovin. lalu di konversi ke *vector* menggunakan fitur *extraction counvectorize* untuk digunakan sebagai masukan dalam algoritme pembelajaran mesin. Setelah itu, dari 2500 data tersebut dilakukan split data untuk *testing*. Split yang digunakan adalah 80:20. Tahap selanjutnya dilakukan modeling menggunakan metode *naïve bayes* yang dikombinasikan dengan fitur seleksi *chi square*. Dan di dapatkan hasil akurasi sebesar 0.896 atau dibulatkan menjadi 0.90. Setelah proses *modeling*, tahap selanjutnya adalah melakukan *confusion matrix* untuk mengukur performa klasifikasi.

```
from sklearn.feature_selection import SelectKBest, chi2
chi2 = chi2(X_train, y_train)[0]
from sklearn.feature_selection import SelectKBest, chi2
ch2_result = []
ch2 = SelectKBest(chi2)

x_train_chi2_selected = ch2.fit_transform(X_train, y_train)
x_validation_chi2_selected = ch2.transform(X_test)

model2 = MultinomialNB_model2.fit(x_train_chi2_selected, y_tr
ain)
prediction2 = model2.predict(x_validation_chi2_selected)
print('Akurasi model NB+CHI: ', model2.score(x_validation_chi
2_selected, y_test))
```

Gambar 4. 8 Implementasi Pemodelan NB + Fitur Seleksi *Chi Square*

C. Model *Naïve Bayes + Information Gain*

Pada tahap ini merupakan proses *mining* menggunakan model *naïve bayes* dengan kombinasi fitur seleksi *Information Gain*. Dengan penjelasan alur sebagai berikut :



Gambar 4. 9 Alur Pemodelan *Naive bayes + Information Gain*

Tahap pertama adalah melakukan *upload dataset* mentah yang sudah diberi label (positif / negatif) untuk dilakukan proses *text preprocessing* menggunakan *google collab*. Selanjutnya, dilakukan *text preprocessing* data dengan melakukan *tokenize, case folding & remove punctuation, stopword, stemming*. Dari 180067 data mentah, Setelah melakukan *text preprocessing* tersisa 6787 data. Setelah dilakukan pembersihan atau *text preprocessing,* kemudian dilakukan konversi *label to number*. Fungsi ini untuk mengubah kedalam tipe data *integer* 1(positif), -1(negatif).

Kemudian 6787 data tersebut di bagi lagi menjadi 2500 untuk *sample* data yang didaptkan memalui perhitungan slovin. lalu di konversi ke *vector* menggunakan fitur *extraction counvectorize* untuk digunakan sebagai masukan dalam algoritme pembelajaran mesin. Setelah itu, dari 2500 data tersebut dilakukan split data untuk *testing*. Split yang digunakan adalah 80:20. Tahap selanjutnya dilakukan *modeling* menggunakan metode *naïve bayes* yang dikombinasikan dengan fitur seleksi *Information Gain*. Dan di dapatkan hasil akurasi sebesar 0.886 atau dibulatkan menjadi 0.88. Setelah proses *modeling*, tahap selanjutnya adalah melakukan *confusion matrix* untuk mengukur performa klasifikasi.

```
from sklearn.feature_selection import mutual_info_classif

mutual = SelectKBest(mutual_info_classif)
x_train_features_ig = mutual.fit_transform(X_train, y_train)
```

```
x_test_features_ig = mutual.transform(X_test)

model3 = MultinomialNB_model3.fit(x_train_features_ig, y_trai
n)
prediction3 = model3.predict(x_test_features_ig)
print('Akurasi model NB+IG: ', model3.score(x_test_features_i
g, y_test))
```

Gambar 4. 10 Implementasi Pemodelan NB + *Information Gain*

D. Hasil *Modeling*

```
print('Akurasi model NB tanpa fitur seleksi: ', model1.score(
X_test, y_test))
print('Akurasi model NB+CHI: ', model2.score(x_validation_chi
2_selected, y_test))
print('Akurasi model NB+IG: ', model3.score(x_test_features_i
g, y_test))


Akurasi model NB tanpa fitur seleksi:  0.856
Akurasi model NB+CHI:  0.896
Akurasi model NB+IG:  0.886
```

Gambar 4. 11 Hasil Akurasi Pemodelan

Dari pemodelan tersebut, ketiga metode klasifikasi yang dibandingkan dan diverifikasi dengan menguji nilai akurasi, dan didapatkan hasil metode *Naive bayes + Chi Square* mendapatkan nilai akurasi tertinggi.

Tabel 4. 5 Hasil Akurasi Pemodelan

Metode	Akurasi
Model 1 (NB tanpa fitur seleksi)	86%
Model 2 (NB + *Chi Square*)	90%
Model 3 (NB + *Information Gain)*	89%

5. Evaluation

Tahapan *evaluation* ini merupakan tahap laporan evaluasi model dengan menggunakan *confusion matrix* untuk mengukur performa klasifikasi dari ketiga metode sebagai berikut:

A. *Confusion Matrix Naïve Bayes* Tanpa Fitur Seleksi

Berikut adalah *clasification report* pada proses pemodelan *naïve bayes* tanpa fitur seleksi:

```
Classification report:
              precision    recall  f1-score   support
```

-1	0.92	0.82	0.87	285
1	0.79	0.90	0.84	215
accuracy			0.86	500
macro avg	0.85	0.86	0.86	500
weighted avg	0.86	0.86	0.86	500

Gambar 4. 12 *Clasification Report Naïve Bayes* Tanpa fitur Seleksi

Adapun perhitungan pada gambar 4.5 langkah awalnya adalah dengan menggunakan tabel *confusion matrix* berikut:

Tabel 4. 6 *Confution Matrix* Tanpa Fitur Seleksi

Confusion Matrix	*Predicted Class*		
Observed Class		*Positif*	*Negatif*
	Positif	TP = 234	FN = 21
	Negatif	FP = 51	TN = 194

Akurasi : $\frac{TP+TN}{TP+FP+TN+FN} = \frac{234+194}{234+51+21+194} = \frac{428}{500}$ = 0.856 x 100% = 85,6

= Dibulatkan 86%

Recall (neg) : $\frac{TP}{TP+FP} = \frac{234}{234+51} = \frac{234}{285}$ = 0.82105 x 100% = 82,1

= Dibulatkan 82%

Recall (pos) : $\frac{TN}{TN+FN} = \frac{194}{194+21} = \frac{194}{215}$ = 0.9023 x 100% = 90,23

= Dibulatkan 90%

Precision (neg) : $\frac{TP}{TP+FN} = \frac{234}{234+21} = \frac{234}{255}$ = 0.9176 x 100% = 91,76

= Dibulatkan 92%

Precision (pos) : $\frac{TN}{TN+FP} = \frac{194}{194+51} = \frac{194}{245}$ = 0.7918 x 100% = 79,18

= Dibulatkan 79%

B. Confution Matrix Naïve Bayes + Chi Square

Berikut adalah *Clasification Report* pada proses pemodelan naïve bayes yang dikombinasikan dengan fitur seleksi Chi Square:

Classification report:				
	precision	recall	f1-score	support
-1	0.88	0.95	0.91	285
1	0.93	0.82	0.87	215

```
              precision    recall  f1-score   support

    accuracy                           0.90       500
   macro avg       0.90      0.89      0.89       500
weighted avg       0.90      0.90      0.90        50
```

Gambar 4. 13 *Clasification Report Naive bayes + Chi Square*

Adapun perhitungan pada gambar 4.6 langkah awalnya adalah dengan menggunakan tabel *confusion matrix* berikut:

Tabel 4. 7 *Confution Matrix Naive bayes + Chi Square*

Confusion Matrix	*Predicted Class*		
Observed Class		*Positif*	*Negatif*
	Positif	TP = 271	FN = 38
	Negatif	FP = 14	TN = 177

Akurasi : $\frac{TP+TN}{TP+FP+TN+FN} = \frac{271+177}{271+14+177+38} = \frac{448}{500}$ = 0.896 x 100% = 89,6

= Dibulatkan 90%

Recall (neg) : $\frac{TP}{TP+FP} = \frac{271}{271+14} = \frac{271}{285}$ = 0.9508 x 100% = 95,08

= Dibulatkan 95%

Recall (pos) : $\frac{TN}{TN+FN} = \frac{177}{177+38} = \frac{177}{215}$ = 0.823 x 100% = 82,3

= Dibulatkan 82%

Precision (neg) : $\frac{TP}{TP+FN} = \frac{271}{271+38} = \frac{271}{309}$ = 0.877 x 100% = 87,7

= Dibulatkan 88%

Precision (pos) : $\frac{TN}{TN+FP} = \frac{177}{177+14} = \frac{177}{191}$ = 0.926 x 100% = 92,6

= Dibulatkan 93%

C. Confution Matrix Naïve Bayes + Information Gain

Berikut adalah *Clasification Report* pada proses pemodelan *naïve bayes* yang dikombinasikan dengan fitur seleksi *Information Gain*:

```
Classification report:
              precision    recall  f1-score   support

          -1       0.86      0.95      0.90       285
           1       0.92      0.80      0.86       215
```

accuracy			0.89	500
macro avg	0.89	0.88	0.88	500
weighted avg	0.89	0.89	0.88	500

Gambar 4. ***14 Clasification Report Naive bayes + Information Gain***

Adapun perhitungan pada gambar 4.7 langkah awalnya adalah dengan menggunakan tabel *confusion matrix* berikut:

Tabel 4. 8 ***Confution Matrix Naïve Bayes + Information Gain***

Confusion Matrix	*Predicted Class*		
Observed Class		*Positif*	*Negatif*
	Positif	TP = 271	FN = 43
	Negatif	FP = 14	TN = 172

Akurasi : $\frac{TP+TN}{TP+FP+TN+FN} = \frac{271+172}{271+14+172+43} = \frac{443}{500}$ = 0.886 x 100% = 88,6

= Dibulatkan 89%

Recall (neg) : $\frac{TP}{TP+FP} = \frac{271}{271+14} = \frac{271}{285}$ = 0.9508 x 100% = 95,08

= Dibulatkan 95%

Recall (pos) : $\frac{TN}{TN+FN} = \frac{172}{172+43} = \frac{172}{215}$ = 0.8 x 100% = 80

= 80%

Precision (neg) : $\frac{TP}{TP+FN} = \frac{271}{271+43} = \frac{271}{314}$ = 0.863 x 100% = 86,3

= Dibulatkan 86%

Precision (pos) : $\frac{TN}{TN+FP} = \frac{172}{172+14} = \frac{177}{186}$ = 0.924 x 100% = 92,4

= Dibulatkan 92%

Dari hasil perhitungan akurasi *confusion matrix* diatas dapat disimpulkan, yaitu:

a. *Naive bayes* tanpa fitur seleksi mendapatkan *accuracy (86.00%), recall positif (90%), recall negatif (82%) dan precision negatif (92%) precision positif (79%).*

b. *Naive bayes* yang dikombinasikan dengan fitur seleksi *chi square* mendapatkan *accuracy (90%), recall positif (82%), recall negatif (95%) dan precision negatif (88%) precision positif (93%).*

c. *Naive bayes* yang dikombinasikan dengan fitur seleksi *information gain* mendapatkan *accuarcy (89%), recall positif (80%), recall negatif (95%) dan precision negatif (86%) precision positif (92%).*

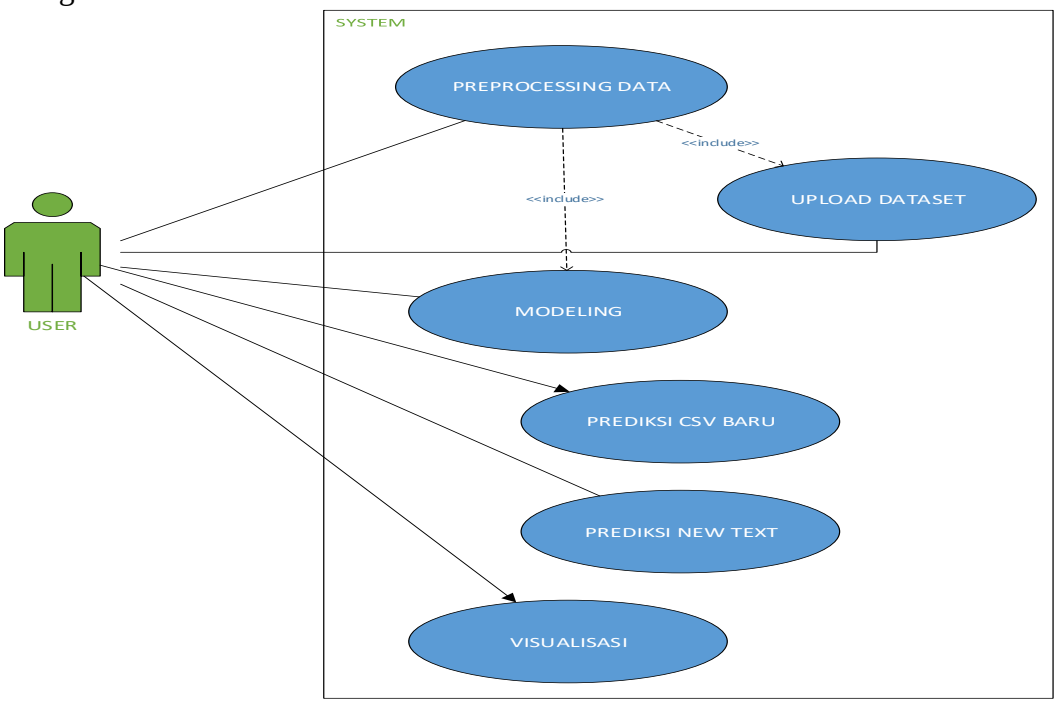
6. Deployment

Pada tahap ini menjelaskan rancangan *usecase diagram, activity diagram dan sequence diagram*, kemudian prototipe pengembangan model menggunakan FLASK.

1. *Design Model*

A. Use Case Diagram

Pada tahapan ini akan menjelaskan fungsionalitas suatu sistem yang dibangun sebagai berikut:

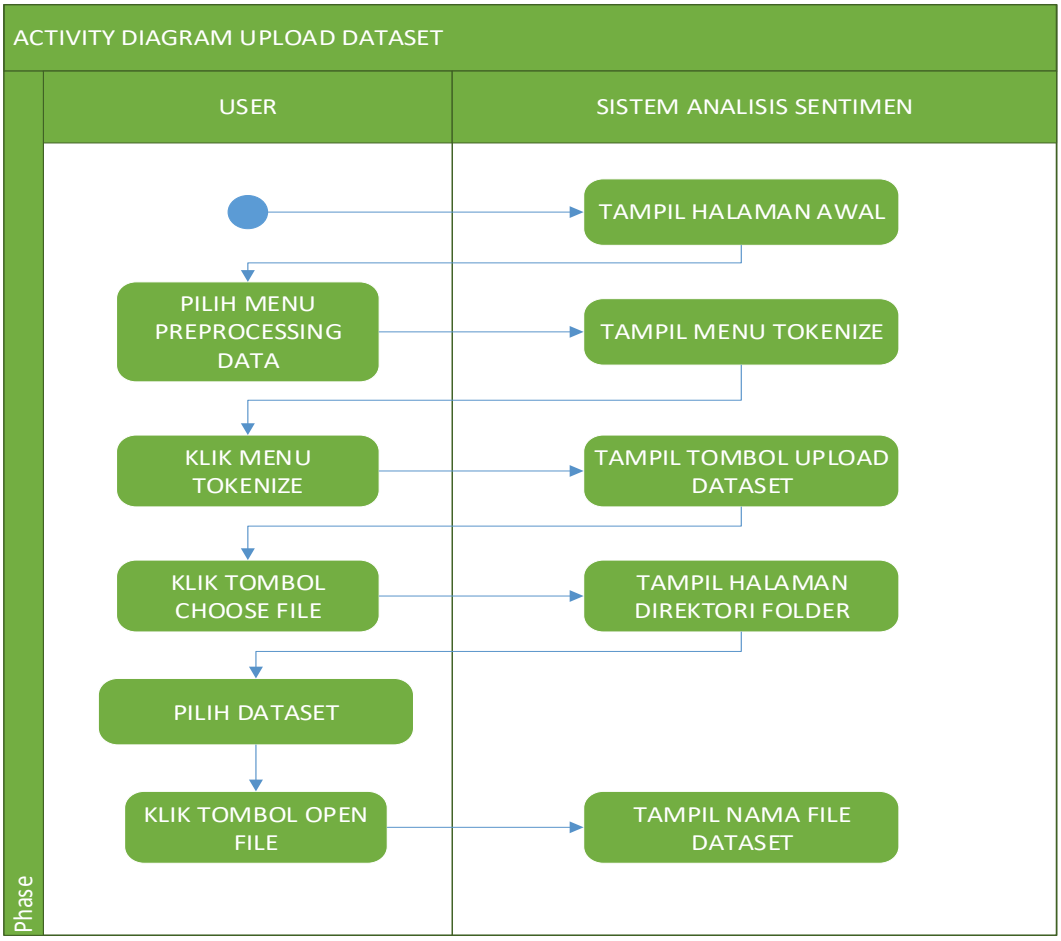


Gambar 4. 15 *Diagram Use Case*

B. Activity Diagram

Activity diagram pada tahap ini menggambarkan aktivitas utama dan hubungan di antara aktivitas dalam suatu proses. Berikut *activity diagram* dalam penelitian ini:

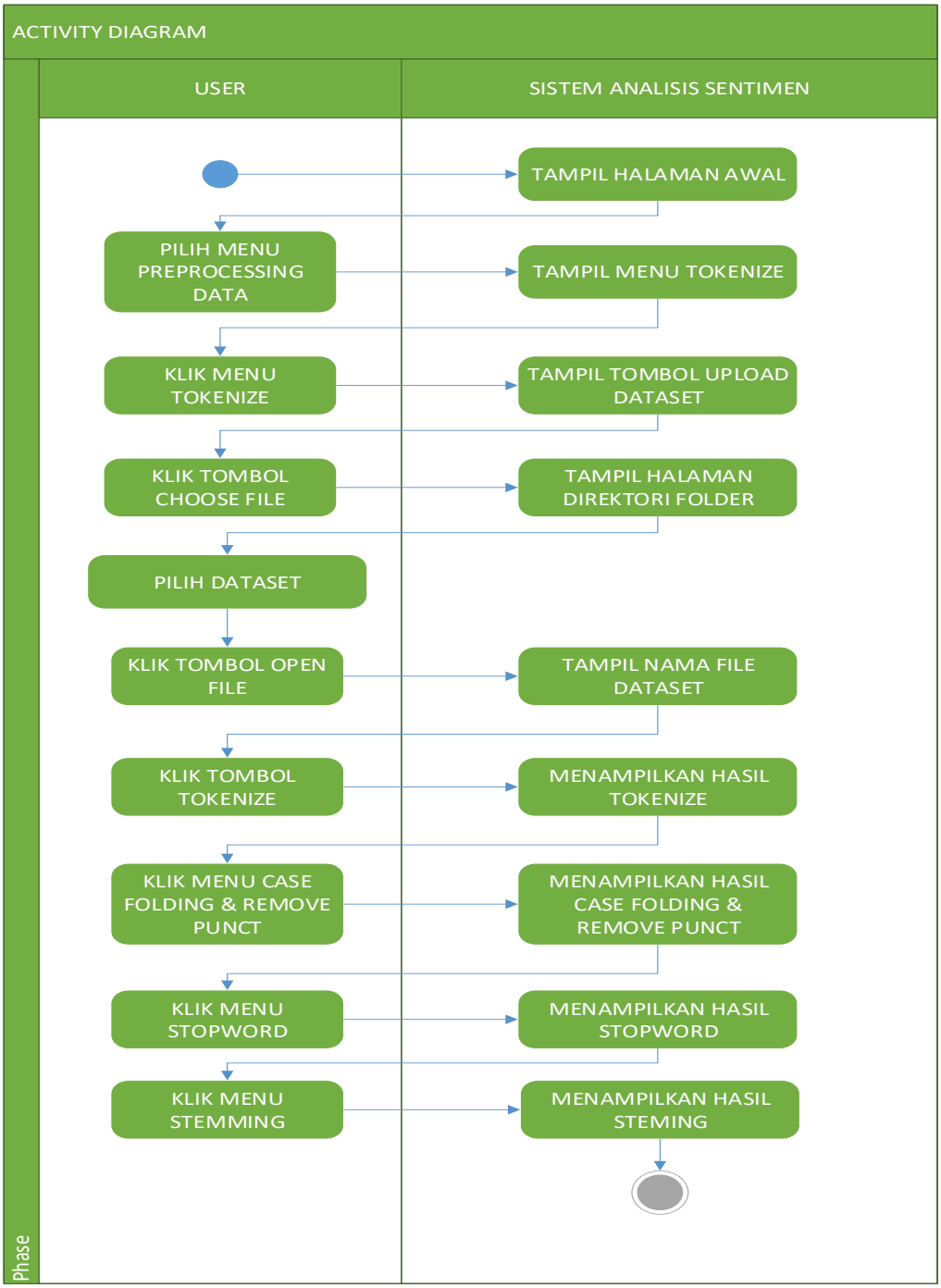
1. *Activity Diagram Upload Dataset*



Gambar 4. 16 ***Activity Diagram Upload Dataset***

Pada halaman awal, akan menampilkan beberapa menu. pertama pengguna memilih menu *preprocessing* data. kemudian sistem akan menampilkan *drop down* menu *tokenize* dan klik *tokenize*. sistem akan menampilkan tombol *upload dataset* yang akan di olah. pengguna klik tombol *choose file* sistem akan menampilkan halaman direktori *folder* lalu pilih *dataset* setelah itu klik tombol *open file* maka sistem akan menampilkan nama *file dataset* yang akan di olah

2. *Activity Diagram Preprocessing Data*

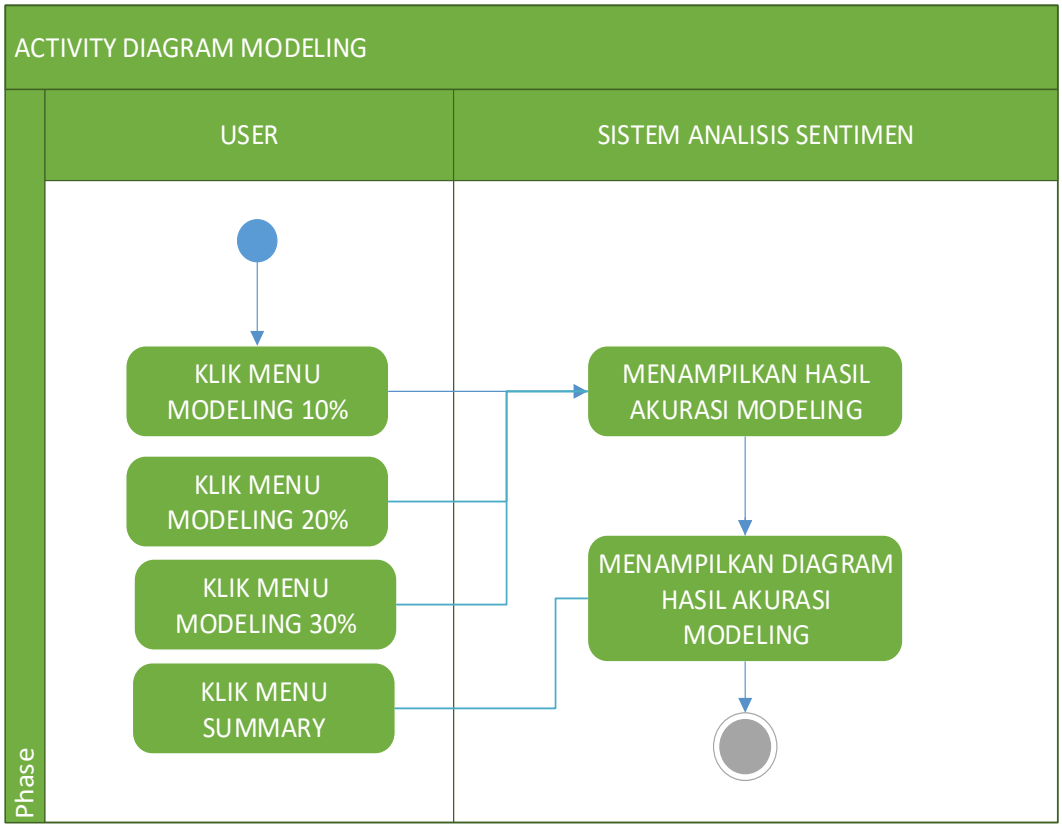


Gambar 4. 17 *Activity Diagram Preprocessing Data*

Diagram *activity* ini lanjutan dari diagram *activity* upload dataset. Langkahnya adalah Pada halaman awal, akan menampilkan beberapa menu. pertama pengguna memilih menu *preprocessing* data. kemudian sistem akan menampilkan *drop down* menu *tokenize* dan klik *tokenize*. sistem akan menampilkan tombol *upload dataset* yang akan di olah. pengguna klik tombol *choose file* sistem akan menampilkan halaman direktori *folder* lalu pilih *dataset* setelah itu klik tombol *open file* maka sistem akan menampilkan nama *file dataset* yang akan di olah.

Kemudian klik tombol *tokenize* sistem akan memproses dan menampilkan hasil *tokenize*, lanjut klik menu *case folding & remove punctuation*, sistem akan memproses dan menampilkan hasil *case folding & remove punctuation*, lanjut klik menu *stopword* sistem akan memproses dan menampilkan hasil *stopword*, setelah itu klik menu *stemming*, sistem akan menampilkan hasil *stemming*

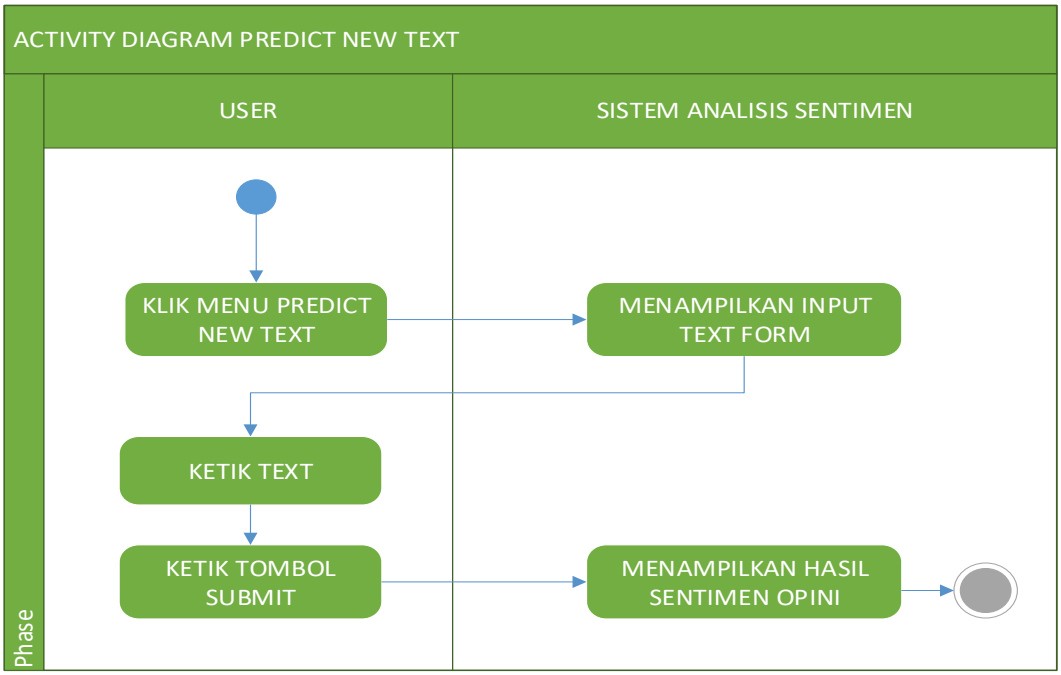
3. Activity Diagram Modeling



Gambar 4. 18 Activity Diagram Modeling

Pada tahapan ini,setelah pengguna menyelesaikan proses *preprocessing* data, lanjut klik menu *modeling*. Sistem akan memproses dan menampilkan hasil akurasi pemodelan dan menampilkan diagram hasil akurasi *modeling.*

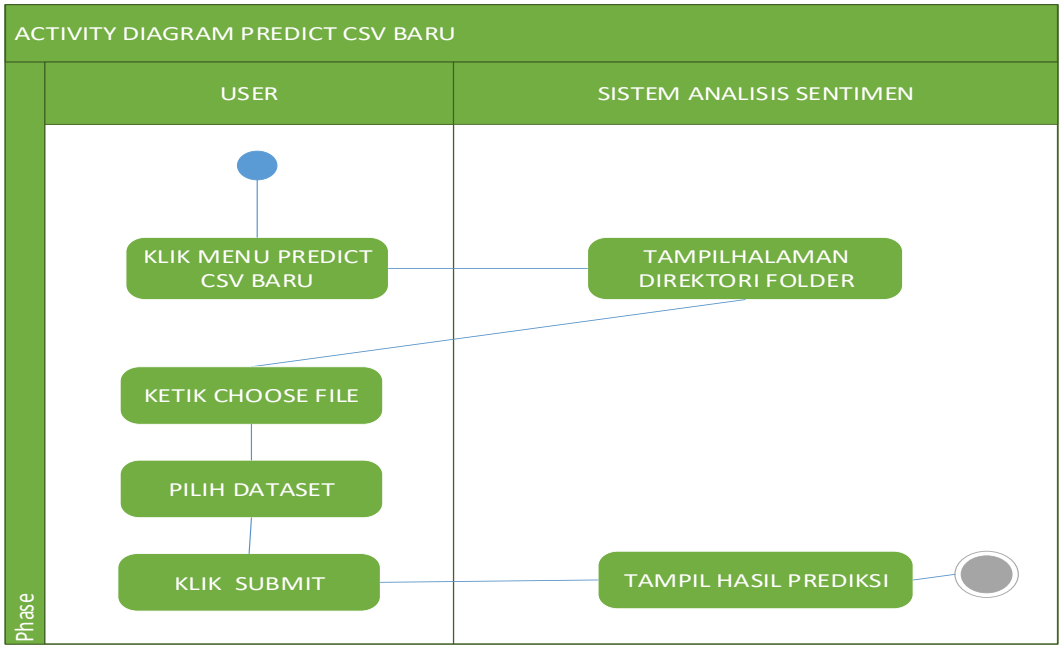
4. *Activity diagram Prediksi New Text*



Gambar 4. 19 Activity diagram Prediksi New Text

Pada Tahap ini, pengguna dapat melakukan prediksi text baru. Pertama pengguna memilih menu *predict new text*, sistem akan menampilkan *input text form*, lalu ketik teks atau opini yang di inginkan, kemudian klik tombol *submit*, lalu sistem akan menampilkan prediksi sentimen teks atau opini.

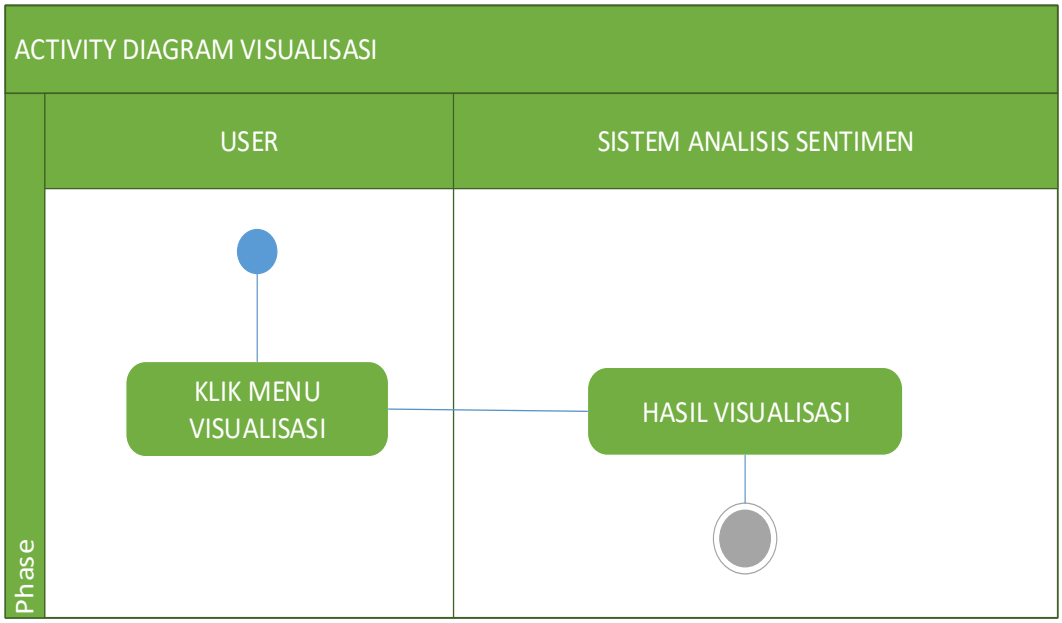
5. *Activity diagram Prediksi CSV Baru*



Gambar 4. 20 *Activity Diagram Predict* CSV Baru

Pada Tahap ini, pengguna dapat melakukan prediksi CSV baru. Pertama pengguna memilih menu *predict CSV Baru,* tampil halaman direktori *folder*, pilih *dataset*, klik *submit* kemudian sistem akan menampilkan hasil prediksi

6. *Activity Diagram Visualisasi*



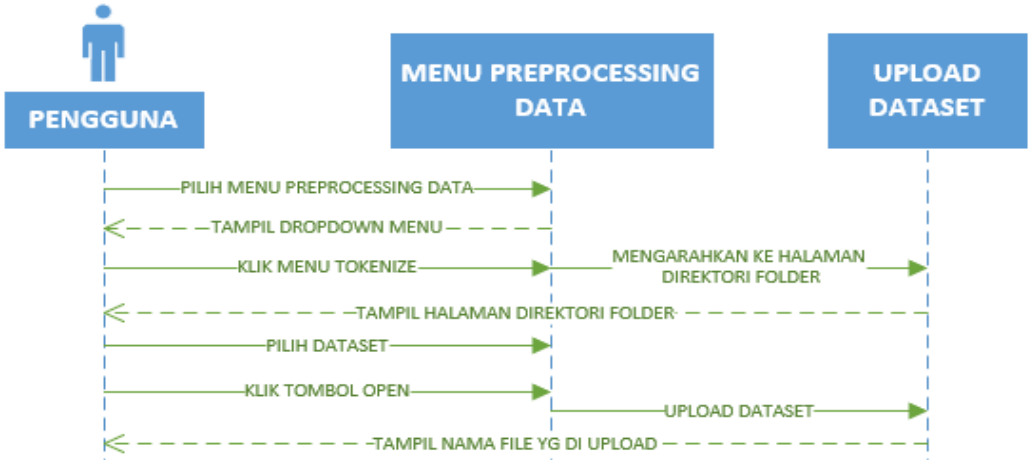
Gambar 4. 21 *Activity Diagram* Visualisasi

Pada Tahap ini, menjelaskan proses visualisasi, pertama pengguna memilih menu visualisasi kemudian sistem akan menampilkan hasil visualisasi data prediksi analisis sentimen cyberbullying saipul jamil

C. Sequence Diagram

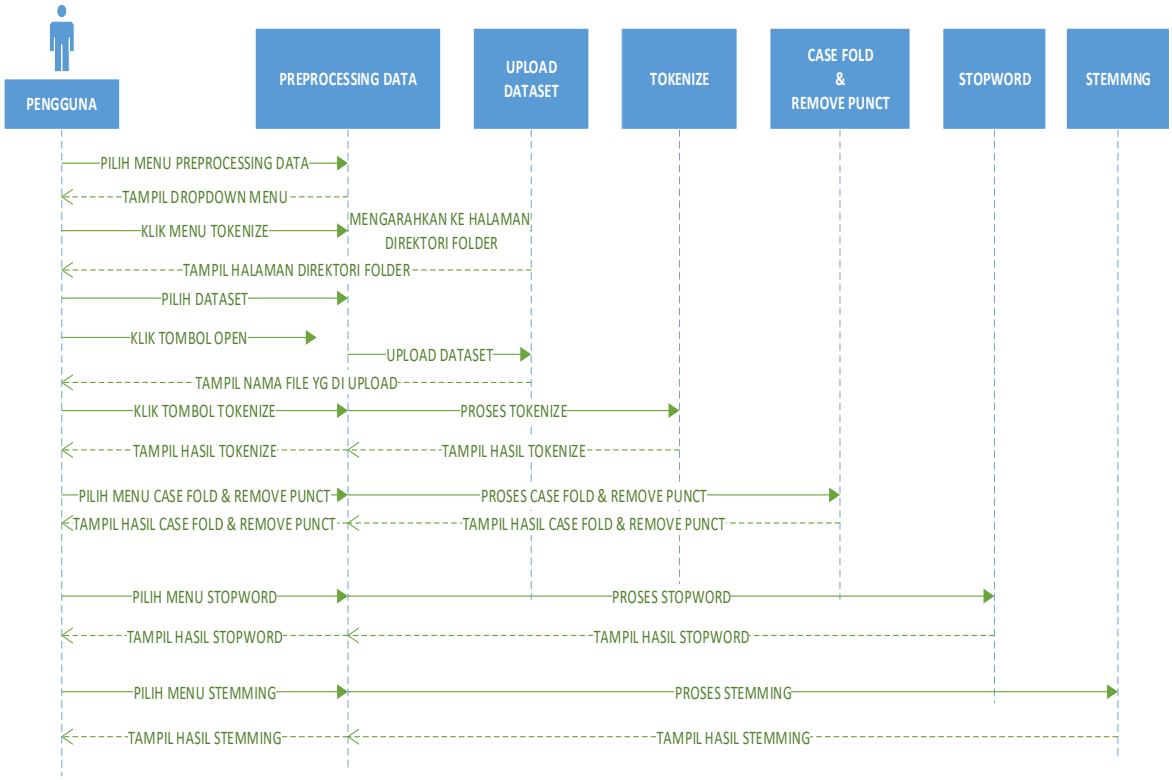
Diagram Sequence pada tahapan ini digunakan untuk menjelaskan dan menampilkan interaksi antar objek-objek sebagai berikut:

1. *Sequence Diagram Upload Dataset*



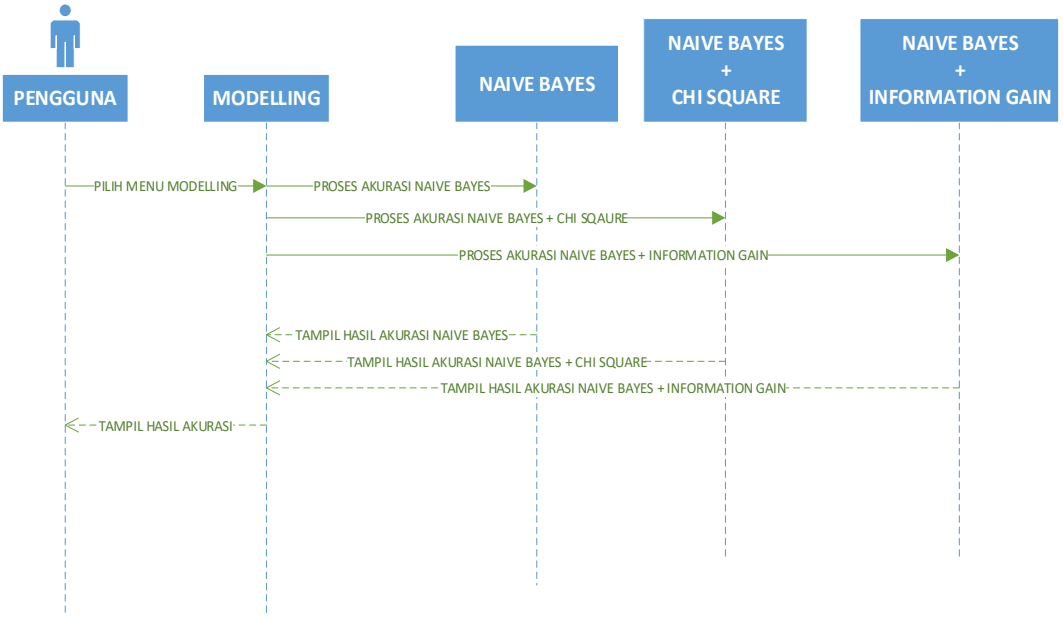
Gambar 4. 22 Sequence Diagram Upload Dataset

2. *Sequence Diagram Preprocessing Data*



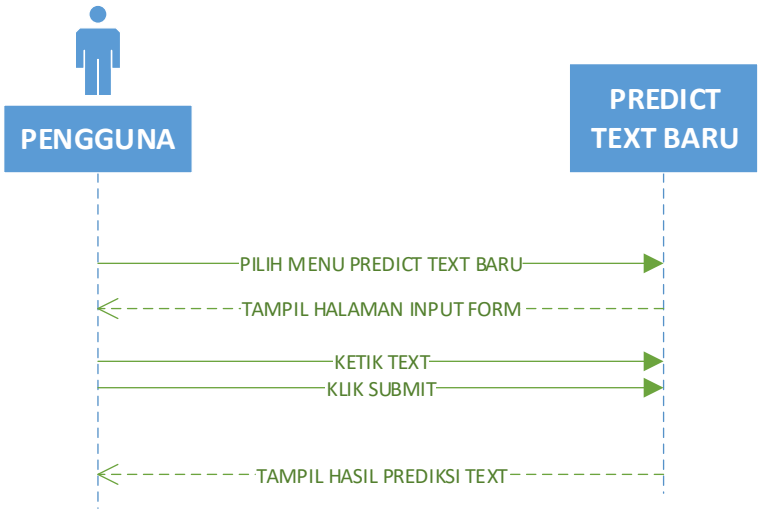
Gambar 4. 23 Sequence Diagram Preprocessing Data

3. *Sequence Diagram Modeling*



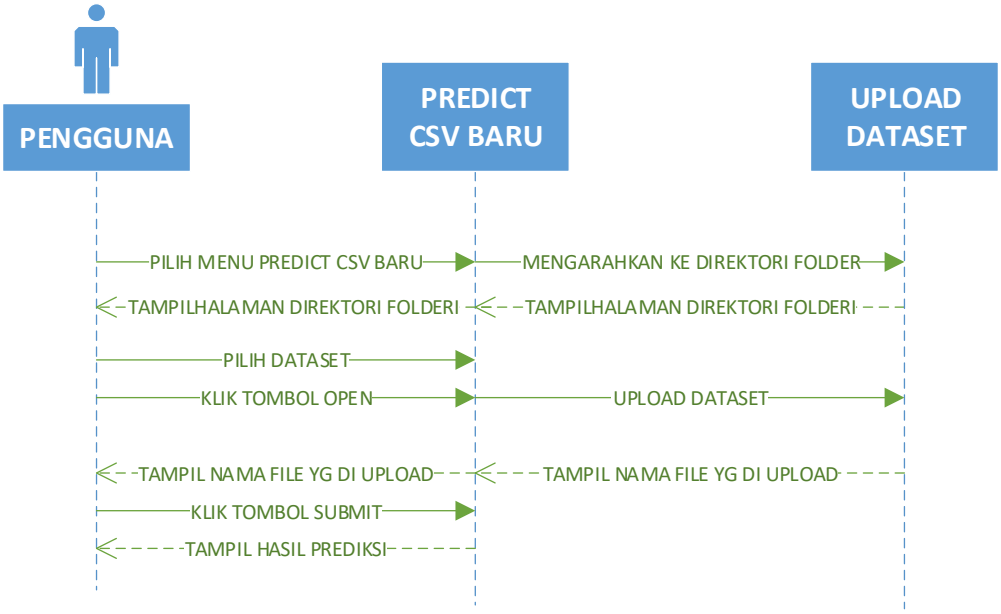
Gambar 4. 24 Sequence Diagram Modeling

4. *Sequence Diagram Predict Text Baru*



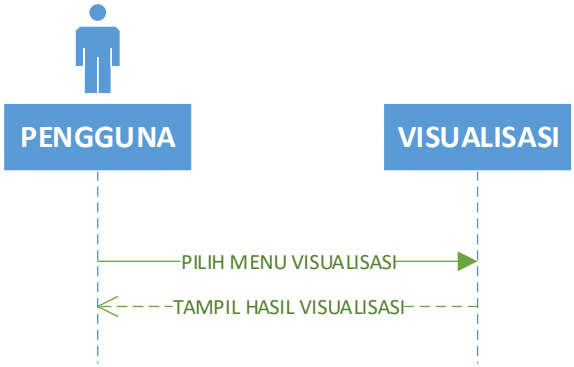
Gambar 4. 25 Sequence Diagram Predict Text Baru

5. *Sequence Diagram Predict CSV Baru*



***Gambar 4. 26 Sequence Diagram* Predict CSV Baru**

6. Sequence Diagram Visualisasi



Gambar 4. 27 *Sequence Diagram Visualisasi*

2. Prototipe

Pada tahap ini menjelaskan mengenai hasil pengembangan prototipe yang di buat. Aplikasi prototipe dibuat menggunakan bahasa python dengan *framework* FLASK dari model yang terpilih yaitu *Naïve Bayes + Chi Square*.

Gambar 4. 28 Halaman awal

Pada Gambar 4.23 merupakan halaman awal tampilan sistem, terdiri dari menu *Preprocessing Data, Modeling, Predict* Baru, *Predict Text* Baru dan Visualisasi Sentimen Analisis *Cyberbullying* Saipul Jamil

	Text	tokenize
0	welcome back bang @saipuljamil	['welcome', 'back', 'bang', '@', 'saipuljamil']
1	@saipuljamil Bakalan nge-hap lagi ga, sm?	['@', 'saipuljamil', 'Bakalan', 'nge-hap', 'lagi', 'ga', ',', 'sm', '?']
2	@saipuljamil Tuhan Maha Pemaaf dan menerima Taubat Taubat artinya tdk melakukan perbuatan tercela yg sama lagi. ini hanya Peringatan Bang Saipul Jamil. Akibat perbuatan anda trauma seumur hidup diterima korban. Tdk bisa perbaiki kerusakan yg dibuat. Jadi Jgn sampai berbuat lagi ! https://t.co/iayMFyxoTT	['@', 'saipuljamil', 'Tuhan', 'Maha', 'Pemaaf', 'dan', 'menerima', 'Taubat Taubat', 'artinya', 'tdk', 'melakukan', 'perbuatan', 'tercela', 'yg', 'sama', 'lagi', '.', 'ini', 'hanya', 'Peringatan', 'Bang', 'Saipul', 'Jamil', '.', 'Akibat', 'perbuatan', 'anda', 'trauma', 'seumur', 'hidup', 'diterima', 'korban', '.', 'Tdk', 'bisa', 'perbaiki', 'kerusakan', 'yg', 'dibuat', '.', 'Jadi', 'Jgn', 'sampai', 'berbuat', 'lagi', '!', 'https', ':', '//t.co/iayMFyxoTT']
3	Bebas dari Penjara, Saipul Jamil Ngaku 'Trauma'\n@saipuljamil\n https://t.co/sgRJqnXeeH	['Bebas', 'dari', 'Penjara', ',', 'Saipul', 'Jamil', 'Ngaku', 'Trauma', '@', 'saipuljamil', 'https', ':', '//t.co/sgRJqnXeeH']
4	@saipuljamil Untuk orangtua, tolong jgn hanya memperhatikan anak cewe. Dengan ini anak cowo juga harus di perhatikan dan dijaga. Dari PK.	['@', 'saipuljamil', 'Untuk', 'orangtua', ',', 'tolong', 'jgn', 'hanya', 'memperhatikan', 'anak', 'cewe', '.', 'Dengan', 'ini', 'anak', 'cowo', 'juga', 'harus', 'di', 'perhatikan', 'dan', 'dijaga', '.', 'Dari', 'PK', '.']
5	@saipuljamil Semoga ga diulangi lagi perbuatan yg itu ya bro	['@', 'saipuljamil', 'Semoga', 'ga', 'diulangi', 'lagi', 'perbuatan', 'yg', 'itu', 'ya', 'bro']
6	@saipuljamil Semoga bisa berubah lebih baik?	['@', 'saipuljamil', 'Semoga', 'bisa', 'berubah', 'lebih', 'baik', '?']
7	@Anonymous_2014 @saipuljamil Aamiin YRA	['@', 'Anonymous_2014', '@', 'saipuljamil', 'Aamiin', 'YRA']
8	@saipuljamil Om jangan bodi gw ya..	['@', 'saipuljamil', 'Om', 'jangan', 'bodi', 'gw', 'ya', '..']
9	@saipuljamil semoga jgn mengulangi lagi perbuatan yg tdk baik	['@', 'saipuljamil', 'semoga', 'jgn', 'mengulangi', 'lagi', 'perbuatan', 'yg', 'tdk', 'baik']

Gambar 4. 29 Tampilan Proses *Tokenize*

Pada Gambar 4.24 merupakan tampilan hasil *preprocessimg* data tahap *tokenize*, Langkahnya adalah Pada halaman awal, akan menampilkan beberapa menu. pertama pengguna memilih menu *preprocessing* data. kemudian sistem akan menampilkan *drop down* menu *tokenize* dan klik *tokenize*. sistem akan menampilkan tombol *upload dataset* yang akan di olah. pengguna klik tombol *choose file* sistem akan menampilkan halaman direktori *folder* lalu pilih *dataset* setelah itu klik tombol *open file* maka sistem akan menampilkan nama *file dataset* yang akan di olah. Kemudian klik tombol *tokenize* sistem akan memproses dan menampilkan hasil *tokenize*

10 Data pertama

	Text	tokenize	Case Folding & Remove Punc
0	welcome back bang @saipuljamil	['welcome', 'back', 'bang', '@', 'saipuljamil']	welcome back bang
1	@saipuljamil Bakalan nge-hap lagi ga, sm?	['@', 'saipuljamil', 'Bakalan', 'nge-hap', 'lagi', 'ga', ',', 'sm', '?']	bakalan ngehap lagi ga sm
2	@saipuljamil Tuhan Maha Pemaaf dan menerima Taubat Taubat artinya tdk melakukan perbuatan tercela yg sama lagi. ini hanya Peringatan Bang Saipul Jamil. Akibat perbuatan anda trauma seumur hidup diterima korban. Tdk bisa perbaiki kerusakan yg dibuat. Jadi Jgn sampai berbuat lagi ! https://t.co/iayMFyxoTT	['@', 'saipuljamil', 'Tuhan', 'Maha', 'Pemaaf', 'dan', 'menerima', 'Taubat Taubat', 'artinya', 'tdk', 'melakukan', 'perbuatan', 'tercela', 'yg', 'sama', 'lagi', '.', 'ini', 'hanya', 'Peringatan', 'Bang', 'Saipul', 'Jamil', '.', 'Akibat', 'perbuatan', 'anda', 'trauma', 'seumur', 'hidup', 'diterima', 'korban', '.', 'Tdk', 'bisa', 'perbaiki', 'kerusakan', 'yg', 'dibuat', '.', 'Jadi', 'Jgn', 'sampai', 'berbuat', 'lagi', '!', 'https', ':', '//t.co/iayMFyxoTT']	tuhan maha pemaaf dan menerima taubat taubat artinya tdk melakukan perbuatan tercela yg sama lagi ini hanya peringatan bang saipul jamil akibat perbuatan anda trauma seumur hidup diterima korban tdk bisa perbaiki kerusakan yg dibuat jadi jgn sampai berbuat lagi
3	Bebas dari Penjara, Saipul Jamil Ngaku 'Trauma'\n@saipuljamil\n https://t.co/sgRJqnXeeH	['Bebas', 'dari', 'Penjara', ',', 'Saipul', 'Jamil', 'Ngaku', 'Trauma', '@', 'saipuljamil', 'https', ':', '//t.co/sgRJqnXeeH']	bebas dari penjara saipuljamil ngaku trauma
4	@saipuljamil Untuk orangtua, tolong jgn hanya memperhatikan anak cewe. Dengan ini anak cowo juga harus di perhatikan dan dijaga. Dari PK.	['@', 'saipuljamil', 'Untuk', 'orangtua', ',', 'tolong', 'jgn', 'hanya', 'memperhatikan', 'anak', 'cewe', '.', 'Dengan', 'ini', 'anak', 'cowo', 'juga', 'harus', 'di', 'perhatikan', 'dan', 'dijaga', '.', 'Dari', 'PK', '.']	untuk orangtua tolong jgn hanya memperhatikan anak cewe dengan ini anak cowo juga harus di perhatikan dan dijaga dari pk
5	@saipuljamil Semoga ga diulangi lagi perbuatan yg itu ya bro	['@', 'saipuljamil', 'Semoga', 'ga', 'diulangi', 'lagi', 'perbuatan', 'yg', 'itu', 'ya', 'bro']	semoga ga diulangi lagi perbuatan yg itu ya bro
6	@saipuljamil Semoga bisa berubah lebih baik?	['@', 'saipuljamil', 'Semoga', 'bisa', 'berubah', 'lebih', 'baik', '?']	semoga bisa berubah lebih baik
7	@Anonymous_2014 @saipuljamil Aamiin YRA	['@', 'Anonymous_2014', '@', 'saipuljamil', 'Aamiin', 'YRA']	aamiin yra
8	@saipuljamil Om jangan bodi gw ya..	['@', 'saipuljamil', 'Om', 'jangan', 'bodi', 'gw', 'ya', '..']	om jangan bodi gw ya
9	@saipuljamil semoga jgn mengulangi lagi perbuatan yg tdk baik	['@', 'saipuljamil', 'semoga', 'jgn', 'mengulangi', 'lagi', 'perbuatan', 'yg', 'tdk', 'baik']	semoga jgn mengulangi lagi perbuatan yg tdk baik

Gambar 4. 30 Tampilan Proses *Case Folding & Remove Punctuation*

Gambar 4.25 merupakan lanjutan dari proses sebelumnya, Setelah menampilkan hasil *tokenize*. Untuk ke tahap ini *user* klik menu *Case Folding & Remove Punctuation*, kemudian sistem akan menampilkan tabel hasil *Case Folding & Remove Punctuation*

Gambar 4. 31 Tampilan Proses Stopword

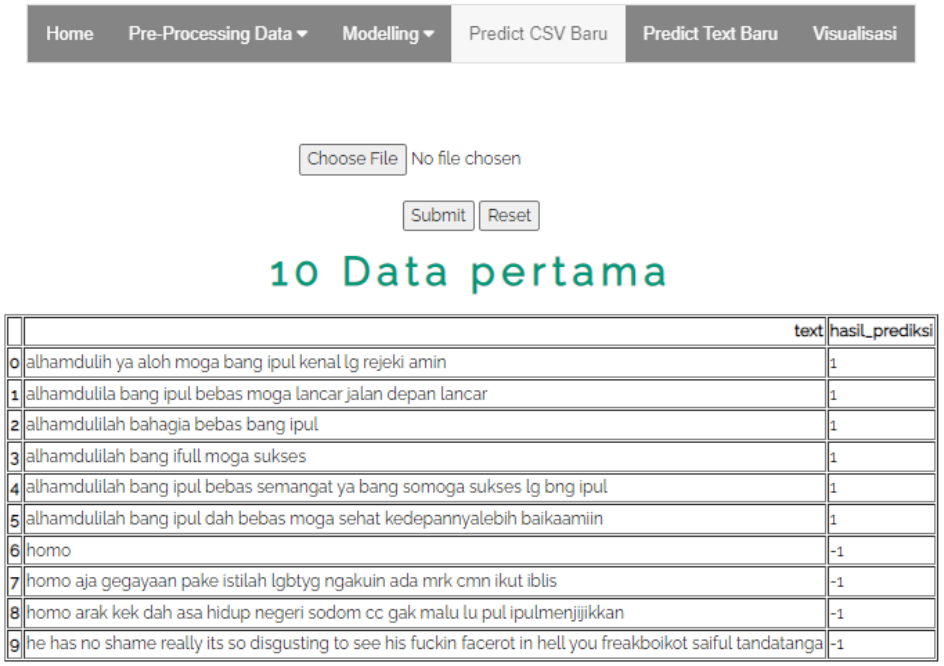
Gambar 4.26 merupakan lanjutan dari proses sebelumnya, Setelah menampilkan hasil *Case Folding & Remove Punctuation*. Untuk ke tahap ini user klik menu *stopword*, kemudian sistem akan menampilkan tabel hasil *stopword*

Gambar 4. 32 Tampilan Proses Stemming

Gambar 4.27 merupakan lanjutan dari proses sebelumnya, Setelah menampilkan hasil *Stopword*. Untuk ke tahap ini *user* klik menu *setmming*, kemudian sistem akan menampilkan tabel hasil *stemming*

Gambar 4. 33 Tampilan Hasil Akurasi

Gambar 4.28 merupakan lanjutan dari proses sebelumnya, setelah *user* selesai proses *preprocesing data*, menu selanjutnya adalah *modeling*, pertama *user* klik menu *modeling* kemudian Sistem akan memproses dan menampilkan hasil akurasi pemodelan dan menampilkan diagram hasil akurasi *modeling*. Pada tahap ini ditampilkan hasil akurasi pemodelan dengan beberapa split data yaitu *modeling* 10 % (90:10), *modeling* 20% (80:20) dan *modeling* 30% (70:30)



	text	hasil_prediksi
0	alhamdulih ya aloh moga bang ipul kenal lg rejeki amin	1
1	alhamdulila bang ipul bebas moga lancar jalan depan lancar	1
2	alhamdulilah bahagia bebas bang ipul	1
3	alhamdulilah bang ifull moga sukses	1
4	alhamdulilah bang ipul bebas semangat ya bang somoga sukses lg bng ipul	1
5	alhamdulilah bang ipul dah bebas moga sehat kedepannyalebih baikaamiin	1
6	homo	-1
7	homo aja gegayaan pake istilah lgbtyg ngakuin ada mrk cmn ikut iblis	-1
8	homo arak kek dah asa hidup negeri sodom cc gak malu lu pul ipulmenjijikkan	-1
9	he has no shame really its so disgusting to see his fuckin facerot in hell you freakboikot saiful tandatanga	-1

Gambar 4. 34 Tampilan Prediksi CSV Baru

Pada Gambar 4.29 merupakan tampilan prediksi CSV baru, pengguna dapat melakukan prediksi CSV baru. Pertama pengguna memilih menu *predict CSV Baru*, tampil halaman direktori *folder*, pilih *dataset*, klik *submit* kemudian sistem akan menampilkan hasil prediksi

Gambar 4. 35 Tampilan Prediksi Text Baru

Pada gambar 4.30 pengguna dapat melakukan prediksi *text* baru. Pertama pengguna memilih menu *predict new text*, sistem akan menampilkan *input text form*, lalu ketik teks atau opini yang di inginkan, kemudian klik tombol *submit*, lalu sistem akan menampilkan prediksi sentimen teks atau opini.

Gambar 4. 36Tampilan menu Visualisasi

Pada Gambar 4.31 Pada Tahap ini, menjelaskan proses visualisasi, pertama pengguna memilih menu visualisasi kemudian sistem akan menampilkan hasil visualisasi data prediksi sentimen *cyberbullying* saipul jamil

BAB V
KESIMPULAN

5.1 Kesimpulan

Berdasarkan uraian-uraian yang telah dikemukakan pada bab-bab sebelumnya, maka dapat disimpulkan sebagai berikut :

1. Evaluasi kinerja *naïve bayes* dengan mengkombinasikan dengan fitur seleksi (*chi square vs Information Gain)* dapat meningkatkan nilai akurasi pada proses analisis sentimen diatas 76%. Dengan hasil akurasi NB tanpa fitur seleksi 86%, NB + *Chi Square* 90%, NB + *Information Gain* 89% dan dari hasil akurasi ini ada kenaikan sebesar 4% dibandingkan dengan *naïve bayes* tanpa fitur seleksi.

5.2 Saran

Dalam penelitian ini masih terdapat kekurangan sehingga ke depannya diperlukan pengembangan dan saran-saran yang di usulkan sebagai berikut:

1. Penelitian ini dapat dikembangkan dengan menambahkan konversi emoticon untuk meningkatkan nilai hasil akurasi
2. Penelitian ini dapat dikembangkan dengan memperbandingkan metode data mining lainnya atau memperbandingkan *feature extraction*
3. Penelitian selanjutnya bisa mengembangkan bukan hanya bahasa indonesia, tapi dapat menggunakan bahasa daerah

DAFTAR PUSTAKA

A., V. And Sonawane, S. S. (2016) 'Sentiment Analysis Of Twitter Data: A Survey Of Techniques', *International Journal Of Computer Applications*, 139(11), Pp. 5–15. Doi: 10.5120/Ijca2016908625.

A, V. K. And Aghila, G. (2010) 'A Survey Of Naive Bayes Machine Learning Approach In Text Document Classification', 7(2), Pp. 206–211. Available At: Http://Arxiv.Org/Abs/1003.1795.

Abirami, A. M. And Gayathri, V. (2017) 'A Survey On Sentiment Analysis Methods And Approach', *2016 8th International Conference On Advanced Computing, Icoac 2016*, Pp. 72–76. Doi: 10.1109/Icoac.2017.7951748.

Ahmad, I. S., Bakar, A. A. And Yaakub, M. R. (2019) 'A Review Of Feature Selection In Sentiment Analysis Using Information Gain And Domain Specific Ontology', *International Journal Of Advanced Computer Research*, 9(44), Pp. 283–292. Doi: 10.19101/Ijacr.Pid90.

Alan Dennis, Barbara Haley Wixom, D. T. And David Tegarden, Barbara Haley Wixom, A. D. (2012) *Systems Analysis And Design With Uml, 4th Edition*, *Design*. Available At: Https://Www.Oreilly.Com/Library/View/Systems-Analysis-And/9781118037423/#Toc-Start.

Alan Dennis, Barbara Haley Wixom, R. R. (2012) *System Analysis And Design Fifth Edition*.

Alanazi, I. And Alves-Foss, J. (2020) 'Cyber Bullying And Machine Learning : A Survey', *International Journal If Computer Science And Information Security*, 18(10), Pp. 1–8. Available At: Https://Doi.Org/10.5281/Zenodo.4249340.

Allahyari, M. Pouriyeh S, Assefi M. (2017) 'A Brief Survey Of Text Mining: Classification, Clustering And Extraction Techniques'. Available At: Http://Arxiv.Org/Abs/1707.02919.

Ariona, R. (2016) 'Belajar Html Dan Css Teori Fundamental Dalam Mempelajari Html Dan Css', *Ariona.Net Team*, P. 58.

Astamal, R. (2006) 'Mastering Kode', *Mastering Kode Html*, P. 77.

Azevedo, A. And Santos, M. F. (2008) 'Kdd , Semma And Crisp-Dm : A Parallel Overview Ana Azevedo And M . F . Santos', *Iadis European Conference Data Mining*, Pp. 182–185. Available At: Http://Recipp.Ipp.Pt/Handle/10400.22/136%0ahttp://Recipp.Ipp.Pt/Bitstream/10400.22/136/3/Kdd-Crisp-Semma.Pdf.

Bustami (2010) 'Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah', *Techsi: Jurnal Penelitian Teknik Informatika*, 4, Pp. 127–146.

Dalvi, R. R., Baliram Chavan, S. And Halbe, A. (2020) 'Detecting A Twitter Cyberbullying Using Machine Learning', *Proceedings Of The International Conference On Intelligent Computing And Control Systems, Iciccs 2020*, (Iciccs), Pp. 297–301. Doi: 10.1109/Iciccs48265.2020.9120893.

Giachanou, A. And Crestani, F. (2016) 'Like It Or Not: A Survey Of Twitter Sentiment Analysis Methods', *Acm Computing Surveys*, 49(2). Doi: 10.1145/2938640.

Hadna, M. S., Santosa, P. I. And Winarno, W. W. (2016) 'Studi Literatur Tentang Perbandingan Metode Untuk Proses Analisis Sentimen Di Twitter', *Seminar*

Nasional Teknologi Informasi Dan Komunikasi, 2016(Sentika), Pp. 57–64. Available At: Https://Fti.Uajy.Ac.Id/Sentika/Publikasi/Makalah/2016/95.Pdf.

Hasan, N. F. (2021) 'Deteksi Cyberbullying Pada Facebook Menggunakan Algoritma K-Nearest Neighbor', *Journal Of Smart System*, 1(1), Pp. 35–44. Doi: 10.36728/Jss.V1i1.1605.

Hidayati, A. Kristanto S. Edwin P. (2019) '2019 2nd International Conference Of Computer And Informatics Engineering (Ic2ie) : Proceedings : "Artificial Intelligence Roles In Industrial Revolution 4.0" : 10-11 September 2019, Banyuwangi, East Java, Indonesia', *2019 2nd International Conference Of Computer And Informatics Engineering (Ic2ie)*, Pp. 24–28.

Hung, L. P., Alfred, R. And Hijazi, M. H. A. (2015) 'A Review On Feature Selection Methods For Sentiment Analysis', *Advanced Science Letters*, 21(10), Pp. 2952–2956. Doi: 10.1166/Asl.2015.6475.

Jiang, L, Wang, D, Cai, Z (2007) 'Survey Of Improving Naive Bayes For Classification', *Lecture Notes In Computer Science (Including Subseries Lecture Notes In Artificial Intelligence And Lecture Notes In Bioinformatics)*, 4632 Lnai, Pp. 134–145. Doi: 10.1007/978-3-540-73871-8_14.

Kane, S. N., Mishra, A. And Dutta, A. K. (2016) 'Preface: International Conference On Recent Trends In Physics (Icrtp 2016)', *Journal Of Physics: Conference Series*, 755(1), Pp. 3–9. Doi: 10.1088/1742-6596/755/1/011001.

Khairati, A. Adlina, A, Hertono, F. (2019) 'Kajian Indeks Validitas Pada Algoritma K-Means Enhanced Dan K-Means Mmca', *Prosiding Seminar Nasional Matematika*, 2, Pp. 161–170.

Kpai (2020) *Sejumlah-Kasus-Bullying-Sudah-Warnai-Catatan-Masalah-Anak-Di Awal-2020-Begini-Kata-Komisioner-Kpai @ Www.Kpai.Go.Id, 10 Februari 2020*. Available At: Https://Www.Kpai.Go.Id/Publikasi/Sejumlah-Kasus-Bullying-Sudah-Warnai-Catatan-Masalah-Anak-Di-Awal-2020-Begini-Kata-Komisioner-Kpai (Accessed: 23 April 2021).

Kuhlman, D. (2013) 'A Python Book', *A Python Book*, Pp. 1–227.

Li, G. And Liu, F. (2010) 'A Clustering-Based Approach On Sentiment Analysis', *Proceedings Of 2010 Ieee International Conference On Intelligent Systems And Knowledge Engineering, Iske 2010*, Pp. 331–337. Doi: 10.1109/Iske.2010.5680859.

Lusiana, Gemini, H. And Efendi, Y. (2018) 'Filtering Impolite Words In Social Network Using Naïve Bayes Classifier', *Proceedings Of The 3rd International Conference On Informatics And Computing, Icic 2018*, Pp. 1–5. Doi: 10.1109/Iac.2018.8780464.

Martinez-Plumed, F. and Contreras, O. (2019) 'Crisp-Dm Twenty Years Later: From Data Mining Processes To Data Science Trajectories', *Ieee Transactions On Knowledge And Data Engineering*, 33(8), Pp. 3048–3061. Doi: 10.1109/Tkde.2019.2962680.

Maulana, F. A. Ernawati, I. (2020) 'Analisa Sentimen Cyberbullying Di Jejaring Sosial Twitter Dengan Algoritma Naïve Bayes', *Seminar Nasional Mahasiswa Ilmu Komputer Dan Aplikasinya (Senamika*, Pp. 529–538. Available At: Https://Conference.Upnvj.Ac.Id/Index.Php/Senamika/Article/View/619.

Mehta, J. (2021) 'Cyber Bullying Detection Using Machine Learning', *International Journal For Research In Applied Science And Engineering Technology*, 9(9), Pp. 144–151. Doi: 10.22214/Ijraset.2021.37934.

Microsoft (2021) ‘Civility 69%’, (February), P. 2021. Available At: Https://Query.Prod.Cms.Rt.Microsoft.Com/Cms/Api/Am/Binary/Re4mm8l.

Mufid, M. R. Basofi, A, Rasyid A. (2019) ‘Design An Mvc Model Using Pytho For Flask Framework Development’, *Ies 2019 - International Electronics Symposium: The Role Of Techno-Intelligence In Creating An Open Energy System Towards Energy Democracy, Proceedings*, (Mvc), Pp. 214–219. Doi: 10.1109/Elecsym.2019.8901656.

Nihan, S. T. (2020) ‘Karl Pearsons Chi-Square Tests’, *Educational Research And Reviews*, 15(9), Pp. 575–580. Doi: 10.5897/Err2019.3817.

Nisa, A., Darwiyanto, E. And Asror, I. (2019) ‘Analisis Sentimen Menggunakan Naive Bayes Classifier Dengan Chi-Square Feature Selection Terhadap Penyedia Layanan Telekomunikasi’, *E-Proceeding Of Engineering*, 6(2), Pp. 8650–8659.

Nixon, C. (2014) ‘Current Perspectives: The Impact Of Cyberbullying On Adolescent Health’, *Adolescent Health, Medicine And Therapeutics*, P. 143. Doi: 10.2147/Ahmt.S36456.

Nurhayati, Putra, L. Wardhani. (2019) ‘Chi-Square Feature Selection Effect On Naive Bayes Classifier Algorithm Performance For Sentiment Analysis Document’, 2019 7th International Conference On Cyber And It Service Management, Citsm 2019, (November). Doi: 10.1109/Citsm47753.2019.8965332.

Nuryadi, Tutut, D. Endang S. (2017) *Dasar-Dasar Statistika Penelitian*. Available At:Http://Lppm.Mercubuana-Yogya.Ac.Id/Wp-Content/Uploads/2017/05/Buku-Ajar_Dasar-Dasar-Statistik-Penelitian.Pdf.

Putri, W. S. R., Nurwati, N. And S., M. B. (2016) ‘Pengaruh Media Sosial Terhadap Perilaku Remaja’, *Prosiding Penelitian Dan Pengabdian Kepada Masyarakat*, 3(1). Doi: 10.24198/Jppm.V3i1.13625.

Rahadi, D. R. (2017) ‘Perilaku Pengguna Dan Informasi Hoax Di Media Sosial’, *Jurnal Manajemen Dan Kewirausahaan*, 5(1), Pp. 58–70. Doi: 10.26905/Jmdk.V5i1.1342.

Rossum, G. And Drake, F. L. (2003) *An Introduction To Python : Release 2.2.2*.

Van Rossum, G. And Python Development Team, The (2018) ‘Python Tutorial Release 3.7.0 Guido Van Rossum And The Python Development Team’.

Rumbaugh, J. (2013) *The Unified Modeling Language Reference Manual, Journal Of Chemical Information And Modeling*.

Harrish S., B. And Revandippa B., M. (2017) ‘A Comprehensive Survey On Various Feature Selection Methods To Categorize Text Documents’, *International Journal Of Computer Applications*, 164(8), Pp. 1–7. Doi: 10.5120/Ijca2017913711.

Schröer, C., Kruse, F. And Gómez, J. M. (2021) ‘A Systematic Literature Review On Applying Crisp-Dm Process Model’, *Procedia Computer Science*, 181(2019), Pp. 526–534. Doi: 10.1016/J.Procs.2021.01.199.

Sharma, S. And Singh, D. (2020) ‘Cyber-Bullying Detection Using Naive Bayes And N-Gram’, *International Journal Of Management (Ijm)*, 11(8), Pp. 2090–2104. Doi: 10.34218/Ijm.11.8.2020.184.

Singhal, R. And Rana, R. (2015) ‘Chi-Square Test And Its Application In Hypothesis Testing’, *Journal Of The Practice Of Cardiovascular Sciences*, 1(1), P. 69. Doi: 10.4103/2395-5414.157577.

Sintaha, M. And Mostakim, M. (2019) 'An Empirical Study And Analysis Of The Machine Learning Algorithms Used In Detecting Cyberbullying In Social Media', *2018 21st International Conference Of Computer And Information Technology, Iccit 2018*, Pp. 1–6. Doi: 10.1109/Iccitechn.2018.8631958.

Sulastri, K. (2020) 'Klasifikasi Naïve Bayes Pada Analisis Sentimen Atas Penolakan Dibukanya Larangan Ekspor Benih Lobster', 1(2), Pp. 68–75.

Talpur, B. A. And O'sullivan, D. (2020) 'Cyberbullying Severity Detection: A Machine Learning Approach', *Plos One*, 15(10 October), Pp. 1–19. Doi: 10.1371/Journal.Pone.0240924.

Venugopalan, M. And Gupta, D. (2015) 'Exploring Sentiment Analysis On Twitter Data', *2015 8th International Conference On Contemporary Computing, Ic3 2015*, Pp. 241–247. Doi: 10.1109/Ic3.2015.7346686.

Visa Sofia, D. (2011) 'Confusion Matrix-Based Feature Selection Sofia Visa', *Confusionmatrix-Based Feature Selection Sofia*, 710(January), P. 8.

Watie, E. D. S. (2016) 'Komunikasi Dan Media Sosial (Communications And Social Media)', *Jurnal The Messenger*, 3(2), P. 69. Doi: 10.26623/Themessenger.V3i2.270.

Watori, J. Aryanti, A. Junaidi. (2020) 'Penggunaan Algoritma Klasifikasi Terhadap Analisa Sentimen Pemindahan Ibukota Dengan Pelabelan Otomatis', *Jurnal Informatika*, 7(1), Pp. 85–90. Doi: 10.31311/Ji.V7i1.7528.

Webb, G. I. (2016) 'Encyclopedia Of Machine Learning And Data Mining', *Encyclopedia Of Machine Learning And Data Mining*, (January 2016). Doi: 10.1007/978-1-4899-7502-7.

Widiastuti, Rosarita Niken (2018) *Memaksimalkan Penggunaan Media Sosial Dalam Lembaga Pemerinta*. 1st Edn. Available At: Https://Drive.Google.Com/File/D/1hs6dayxz0cq5i479uc9aaqylbnc9itth/View

Wu, G. And Xu, J. (2016) 'Optimized Approach Of Feature Selection Based On Information Gain', *Proceedings - 2015 International Conference On Computer Science And Mechanical Automation, Csma 2015*, (Mi), Pp. 157–161. Doi: 10.1109/Csma.2015.38.

Zuhri, K., Adha, N. And Saputri, O. (2020) 'Analisis Sentimen Masyarakat Terhadap Pilpres 2019 Berdasarkan Opini Dari Twitter Menggunakan Metode Naive Bayes Classifier', Pp. 259–269.

Zukhrufillah, I. (2018) 'Gejala Media Sosial Twitter Sebagai Media Sosial Alternatif', *Al-I'lam: Jurnal Komunikasi Dan Penyiaran Islam*, 1(2), P. 102. Doi: 10.31764/Jail.V1i2.235.