**ANALISIS SENTIMEN POSTINGAN YANG MENGANDUNG UNSUR BUNUH DIRI PADA MEDIA SOSIAL TWITTER DENGAN METODE NAÏVE BAYES CLASSIFIER**

**SKRIPSI**

diajukan untuk menempuh ujian sarjana

pada Fakultas Matematika dan Ilmu Pengetahuan Alam

Universitas Padjadjaran

ZAHRAN HANIF FATHANMUBIN

NPM 140810200062



UNIVERSITAS PADJADJARAN

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

PROGRAM STUDI TEKNIK INFORMATIKA

SUMEDANG

2024

**BAB 1**

**PENDAHULUAN**

**1.1. Latar Belakang**

Beberapa waktu yang lalu beredar kabar seorang mahasiswi tewas bunuh diri. Hal ini tentu sangat memprihatinkan. Parahnya, selain ungkapan belasungkawa, ternyata terdapat juga komentar-komentar aneh lainnya. “Kamu hebat ya kak udah berani. Aku belum seberani kakak.” “Tunggu aku ya kak,” “Kamu keren banget kak berani melakukan itu,” Itu adalah contoh komentar-komentar yang muncul di berita orang bunuh diri. Hal ini tentu sangat aneh dan menjadi perhatian.

Tidak hanya di kalangan remaja saja, tetapi juga di kalangan anak-anak. Sempat ramai di media sosial seorang anak menggunakan *profile picture semicolon* yang menunjukan hal terkait *mental health*. Setelah dilihat lebih jauh, ternyata anak tersebut mendiagnosa dirinya sendiri yang di mana hal ini tentu tidak bisa sembarangan dilakukan oleh orang yang bukan ahli. Pun di rumahnya, anak tersebut tidak pernah mendapatkan tekanan.

Menurut WHO, bunuh diri adalah penyebab kematian tertinggi keempat untuk kalangan usia 15 – 29 tahun pada 2019. Jumlahnya pun tidak pernah berkurang. Lantas darimana hal ini berasal? Mengapa di zaman yang serba digital ini, yang di mana mental health terus digaungkan, tetapi malah tidak bisa mengurangi angka bunuh diri?

Jawabannya ada pada sosial media. Bunuh diri seakan menjadi trend. Mudah sekali ditemukan konten-konten yang mengandung unsur bunuh diri di sosial media. Mulai dari curhat sederhana hingga tulisan yang memang ingin dilakukan. Salah satu sosial media yang sering digunakan untuk bercurhat adalah Twitter. Penggunanya yang banyak dan aktif ditakutkan akan dapat menggiring opini karena tersu terpapar konten-konten bunuh diri. Akibatnya, bunuh diri yang dianggap sebagai hal tabu lamba laun menjadi suatu hal yang dinormalisasikan melihat banyak dan mudahnya hal-hal dengan unsur tersebut ditemukan.

Perlu pembatasan terkait hal ini. Pemblokiran konten-konten terkait bisa menjadi salah satu hal yang dapat membantu. Terdapat ribuat bahkan mungkin jutaan konten yang ada. Tidak mungkin kalau harus diseleksi satu-persatu secara manual. Perlu adanya otomatisasi guna mengetahui adanya unsur bunuh diri dalam suatu postingan.

Analisis sentimen bisa menjadi suatu alternatif. Konten-konten tersebut akan dimasukan ke komputer. Lalu komputer akan menilai sendiri apakah konten tersebut mengandung unsur bunuh diri. Dalam penelitian ini, penulis akan mencoba menerapkan hal tersebut. Harapannya output dari penelitian ini bisa menjadi salah satu cara dalam melakukan *filtering* konten di Twitter.

**1.2. Identifikasi Masalah**

Berangkat dari hal-hal yang telah dijelaskan pada bagian latar belakang, terdapat beberapa permasalahan yang ada. Melalui penelitian ini, penulis akan membuatkan solusi dari permasalahan-permasalahan tersebut sebagai berikut :

1. Bagaimana mengklasifikasikan suatu konten mengandung unsur bunuh diri atau tidak?
2. Seberapa tinggi tingkat akurasinya?
3. Bagaimana dan dengan media apa data tersebut ditampilkan?

**1.3. Batasan Masalah**

Berangkat dari permasalahan di atas, terdapat beberapa batasan masalah antara lain sebagai berikut :

1. Sosial media yang akan diambil untuk melakukan training data adalah Twitter.
2. Hasil training data akan mengkategorikan apakah suatu postingan mengandung konten dengan unsur bunuh diri atau tidak.
3. Penelitian ini akan menggunakan konten berbahasa inggris.
4. Penelitian akan menggunakan analisis sentimen dengan metode machine learning.
5. Pengujian tidak hanya dilakukan pada data yang sudah terdapat pada sumber *training* saja, tetapi juga akan mencakup beberapa data dari luar data yang telah dikumpulkan.
6. Penelitian akan menggunakan Excel dan bahasa pemrograman Python.
7. Postingan yang diambil akan memiliki satu hingga dua kalimat.
8. Target minimal akurasi adalah 70%.

**1.4. Maksud dan Tujuan Penelitian**

Penelitian ini bermaksud untuk melihat apakah suatu postingan yang mengandung unsur bunuh diri atau tidak. Menggunakan analisis sentimen dengan metode machine learning yang akan mempermudah penganalisaan. Hasil dari penelitian ini diharapkan dapat menjadi salah satu cara yang dapat dilakukan guna menciptakan lingkungan internet yang lebih sehat. Adapun tujuan dari penelitian ini ada sebagai berikut :

1. Membuat pengkategorian suatu postingan dengan kategori memiliki unsur bunuh diri atau tidak.
2. Memberi akurasi minimal 70% supaya hasil dari penelitian ini dapat dipertanggungjawabkan.
3. Memberi kesimpulan dari seluruh proses dengan tampilan yang mudah untuk dilihat.

**1.5. Manfaat Penelitian**

Manfaat dilakukannya penelitian ini adalah sebagai berikut :

1. Penelitian ini dapat menjadi awal untuk pengembangan suatu teknologi lebih lanjut untuk memberi deteksi dini pada suatu postingan yang mengandung unsur bunuh diri.
2. Penelitian ini dapat digunakan untuk membuat kebijakan media sosial yang lebih ramah pengguna.

**1.6. Metodologi Penelitian**

Metodologi penelitian ini adalah melakukan melalui beberapa langkah utama. Pertama, data dikumpulkan dari Twitter menggunakan API Twitter atau alat scraping lainnya, dengan mengumpulkan tweet yang mengandung kata kunci atau hashtag terkait bunuh diri. Selanjutnya, data tersebut diproses melalui tahap pra-pemrosesan yang meliputi tokenisasi, normalisasi, dan penghapusan stopword untuk membersihkan data dari noise seperti tautan, emoji, dan karakter spesial lainnya. Lalu, setelah data diproses, analisis sentimen dilakukan dengan melabeli tweet menjadi sentimen positif dan negatif. Labeling ini bisa dilakukan secara manual atau dengan menggunakan dataset yang sudah dilabeli sebelumnya. Kemudian, model analisis sentimen dilatih menggunakan algoritma Naive Bayes.

Peneliti kemudian menginterpretasikan hasil analisis untuk memahami tren dan pola sentimen terkait postingan bunuh diri. Berdasarkan hasil analisis tersebut, kesimpulan dapat ditarik dan rekomendasi dapat diberikan untuk tindakan lanjut seperti intervensi atau pencegahan bunuh diri di media sosial. Dengan metodologi ini, analisis sentimen yang mendalam terhadap postingan yang mengandung unsur bunuh diri di media sosial Twitter dapat dilakukan secara sistematis dan menyeluruh.

**1.7. Sistematika Penulisan**

Supaya penelitian ini berjalan secara runtut, diperlukan sistematika penulisan yang berisi mengenai materi apa saja yang akan dibahas pada tiap bab dalam penelitian ini. Adapun rincian sistematika penulisan pada penelitian ini sebagai berikut :

**BAB I PENDAHULUAN**

Bab ini menjelaskan tentang latar belakang dari topik penulisan skripsi, pokok permasalahan yang berupa identifikasi dan batasan masalah, tujuan dan manfaat, metodologi yang digunakan serta sistematika penulisan yang digunakan.

**BAB II TINJAUAN PUSTAKA**

Bab ini menjelaskan tentang seluruh landasan teori yang berhubungan dengan penelitian, yaitu tentang metode perhitungan yang digunakan, penjelasan teoritis *framework* yang digunakan, serta teori lainnya guna memahami permasalahan yang dibahas pada penelitian ini.

**BAB III ANALISIS DAN PERANCANGAN**

Bab ini menjelaskan tentang metode pengembangan aplikasi yang digunakan meliputi data kebutuhan aplikasi, analisis kebutuhan sistem, perancangan aplikasi, diagram pemodelan sistem, dan sebagainya.

**BAB IV HASIL DAN PEMBAHASAN**

Bab ini menjelaskan tentang implementasi website yang telah dibangun, tampilan website, pengujian aplikasi dan pengujian sistem pakar.

**BAB V KESIMPULAN DAN SARAN**

Bab ini merupakan penutup yang berisi kesimpulan dan saran dari penelitian yang sudah dilakukan

**BAB 2**

**LANDASAN TEORI**

**2.1. Analisis Sentimen**

Analisis sentimen atau *opinion mining* sebuah metode untuk menganalisa opini, penilaian, evaluasi, sikap, dan emosi seseorang terhadap suatu topik atau permasalahan. Analisis sentimen akan ditampilkan menjadi tiga jenis kategori yaitu positif, negatif, dan netral(<https://books.google.co.id/books?hl=en&lr=&id=xYhyEAAAQBAJ&oi=fnd&pg=PP1&dq=liu+sentiment+analysis&ots=rlTvPBW8DB&sig=46LZfXP7-3NXP_Zu81pOw5xUoGg&redir_esc=y#v=onepage&q=liu%20sentiment%20analysis&f=false> ). Pengkategorian ini tergantung dari kasus yang diambil. Pengkategorian atau pengkalisifikasian penting supaya data-data yang terkumpul lebih mudah diolah. Tujuan analisis sentimen adalah untuk mengolah data secara otomatis guna mendapatkan informasi yang dibutuhkan (<file:///C:/Users/Zahran/Downloads/2026-4436-1-PB.pdf> ).

Dalam melakukan analisis sentimen, pada garis besarnya dibagi menjadi dua metode, yaitu *lexicon based method* dan *machine learning*. Pendekatan *lexicon based* adalah pendekatan yang menggunakan kamus sebagai dasar untuk menentukan sentimen dari suatu kata. Contohnya kata seperti indah, luar biasa, menakjubkan, bagus, baik akan dimasukan dalam sentimen positif. Sementara kata-kata buruk akan masuk ke dalam sentimen negatif (<https://www.researchgate.net/profile/Bing-Liu-120/publication/228667268_Sentiment_analysis_and_subjectivity/links/5472bbea0cf24bc8ea199f7c/Sentiment-analysis-and-subjectivity.pdf> ).

*Machine learning* sendiri pada dasarnya menciptakan algoritma guna membuat mesin mempelajari data yang disediakan (<https://researchgate.net/profile/Vladimir-Nasteski/publication/328146111_An_overview_of_the_supervised_machine_learning_methods/links/5c1025194585157ac1bba147/An-overview-of-the-supervised-machine-learning-methods.pdf> ). Data tersebut yang nanti akan digunakan sebagai dasar dalam menentukan suatu analisis. Hal yang membuat machine learning lebih baik adalah dalam machine learning algoritma dapat didesain supaya menyesuaikan dengan data yang ada, sementara lexicon based berpacu pada sentimen melalui kamus. Hal ini menyulitkan dalam penentuan kalimat seperti “hari yang indah untuk memukul orang.”

**2.2. Machine Learning**

Machine learning adalah penggunaan komputer untuk mempelajari suatu pola tertentu guna menemukan hasil yang lebih baik terhadap suatu permasalahan (<https://bafflerbach.github.io/test_software_carpentry/files/Module%201%20Lecture%20Notes.pdf> ). Machine learning pada garis besarnya dibagi menjadi dua grup, yaitu supervised dan *unsupervised learning*. Pada supervised, data-data yang disiapkan telah ditentukan sebelumnya. Data-data diberi label sehingga memudahkan proses. Mesin akan melakukan review pada data tersebut untuk kemudian membangun sebuah model pembelajaran.

Pada *unsupervised learning*, algoritma tidak disediakan dengan klasifikasi. Tujuan utamanya adalah pengembangan secara otomatis tabel klasifikasi. Pada algortima ini, kesamaan karakteristik dari suatu data akan dicari. Kesamaan tersebut akan menentukan apakah data tersebut dapat dimasukan ke dalam satu kategori dan menjadi suatu kelompok (<https://www.researchgate.net/profile/Vladimir-Nasteski/publication/328146111_An_overview_of_the_supervised_machine_learning_methods/links/5c1025194585157ac1bba147/An-overview-of-the-supervised-machine-learning-methods.pdf> ).

**2.3. Natural Language Processing (NLP)**

Natural Language Processing (NLP) merupakan area penelitian yang meneliti bagaimana komputer dapat digunakan untuk mengerti dan memanipulasi teks bahasa asli. Tujuan dari NLP adalah untuk mengumpulkan pengetahuan tentang bagaimana cara manusia mengerti dan menggunakan bahasa (<https://pure.strath.ac.uk/ws/portalfiles/portal/131112/strathprints002611.pdf> ). Hal ini dapat berguna untuk pengembangan alat dan teknik yang layak untuk membuat sistem komputer yang dapat memanipulasi dan mengerti bahasa asli guna melakukan tugas yang ada. Penerapan NLP mencakup berbagai jenis bidang studi, seperti mesin penerjemah, pemrosesan dan ringkasan teks, pengenalan suara, sistem pakar, dan lain-lain (<https://pure.strath.ac.uk/ws/portalfiles/portal/131112/strathprints002611.pdf> ).

Secara garis besar Natural Language Processing (NLP) dapat dibagi menjadi dua bagian, yaitu Natural Language Understanding yang merupakan linguistik dan Natural Language Generation yang di dalamnya termasuk tugas untuk mengerti dan menciptakan teks (<https://link.springer.com/content/pdf/10.1007/s11042-022-13428-4.pdf> ) . Natural Language Understanding (NLU) mengizinkan mesin untuk mengerti dan menganalisa natural language seperti konsep, kata kunci, emoticon, dan lain-lain. Natural Languange Generation (NLG) merupakan proses untuk memparafrase kalimat yang berguna untuk kebutuhan proses. Perbedaan dari kedua hal tersebut adalah pada Natural Language Understanding (NLU) akan berfokus pada memahami bahasa manusia secara otomatis. Pada Natural Language Generation, fokusnya adalah untuk menghasilkan suatu teks secara alami berdasarkan data yang ada (<https://link.springer.com/content/pdf/10.1007/s11042-022-13428-4.pdf> ).

**2.4. Python**

Python merupakan salah satu bahasa pemrograman. Pertama kali dibuat oleh Guido Van Rossum pada tahun 1991. Pengembangannya terinspirasi dari pemrograman ABC yang saat itu sedang berkembang. Pengembangan bahasa python tidak hanya melibatkan pihak dari kalangan IT, tapi juga dari pihak-pihak lainnya (<https://tahtamedia.co.id/index.php/issj/article/view/344> ). Python bersifat *open source.*

Termasuk sebagai *high level programming language* ([*https://journal.ijresm.com/index.php/ijresm/article/view/395*](https://journal.ijresm.com/index.php/ijresm/article/view/395)). Python dapat digunakan hampir dalam setiap aspek pemrograman. Menggunakan interpreter di dalamnya sehingga python dapat menerjemahkan kode secara langsung. Python dapat dijalankan di berbagai OS seperti Windows, Linux , dan sebagainya (<https://tahtamedia.co.id/index.php/issj/article/view/344> ).

Hal yang menarik dari python adalah python dapat menjalankan tambahan kode yang diambil dari bahasa lainnya. Hal ini tentu akan memudahkan pekerjaan (<file:///C:/Users/Zahran/Downloads/journalsresaim,+IJRESM_V3_I12_3.pdf> ). Selain itu, kemudahan dari sintaks nya yang mirip dengan bahasa inggris membuat pekerjaan menjadi lebih sederhana. Didukung dengan berbagai macam library sehingga jangkauan penggunaannya sangat luas.

Dalam machine learning python sangat sering digunakan. Python sering menjadi pilihan utama dikarenakan library yang disediakannya. Umumnya, dalam melakukan machine learning, ditemukan operasi linear aljabar serta array multidimensi. Fungsinya adalah untuk merepresentasikan vektor dan matriks (<https://www.mdpi.com/2078-2489/11/4/193> ). Library seperti Pandas, NumPy, Scipy, dan sejenisnya akan sering digunakan. Python menyediakan library tersebut.

**2.5. Tokenisasi**

Tokenisasi adalah suatu proses untuk menciptakan token dari sebuah sumber data yang berisi teks yang bisa dibaca manusia dalam bentuk kode (<file:///C:/Users/Zahran/Downloads/encyclopedia-03-00024.pdf> ). Token sendiri merupakan elemen dalam pemrograman seperti ekspresi perkalian atau kata kunci untuk keluar dari suatu program. Pada tokenisasi, kata yang ditemukan akan dirubah menjadi token. Selanjutnya, token tersebut akan dikelompokan dan diberikan hierarki. Pada akhirnya kata yang telah diubah dapat diproses.

Dalam analisis sentimen, tokenisasi masuk ke dalam tahap *prepocessing*. Pada tahap ini, tanda baca akan dihilangkan (<https://www.ijprems.com/uploadedfiles/paper/issue_4_april_2024/33384/final/fin_ijprems1714118825.pdf> ). Hal ini dilakukan untuk mempermudah penganalisaan karena tanda baca tidak bisa dinilai tanpa adanya kata. Kalimat akan dipecah menjadi kata. Setiap kata akan dinilai dan diberi bobot. Fungsi dari pembobotan adalah penentuan nilai. Nilai tersebut akan mengkategorikan apakah suatu kata masuk kedalam kategori positif, negatif, atau netral. Pada akhirnya penilaian tersebut akan dapat memberi penilaian terhadap suatu kalimat tergantung dari topik yang diambil.

**2.6. Naive Bayes Classifier**

Nave Bayes Classifier merupakan salah satu algoritma dalam melakukan pemrosesan analisis sentimen. Naïve Bayes Classifier sendiri merupakan sebuah algoritma yang bertujuan mencari probabilitas tertinggi untuk mengklasifikasikan data uji pada kategori yang tepat. Metode ini mengasumsikan bahwa setiap atribut merupakan atribut yang bebas. Artinya tidak terpengaruh oleh variabel kelas (<https://e-jurnal.lppmunsera.org/index.php/PROSISKO/article/view/6922/2754> ).  
 Salah satu keunggulan dari metode naïve bayes classifier adalah tingkat akurasinya yang tinggi. Selain itu, naïve bayes classifier terbilang cukup mudah untuk melakukan pengolahan data dikarenakan tidak memerlukan data training yang banyak. Seperti yang telah dijelaskan di atas, setiap atribut merupakan atribut bebas atau *independent*, maka hanya varians dari suatu variabel dalam sebuah kelas yg dibutuhkan untuk memberi klasifikasi (<https://bios.sinergis.org/bios/article/view/15> ). Kegunaan dari metode *naïve bayes classifier* adalah sebagai berikut :

1. Klasifikasi dokumen berupa tulisan.
2. Algoritma ini merupakan algoritma atau metode yang menggunakan basis probabilitas dalam *machine learning*.
3. Pembuatan diagnosis medis otomatis.
4. Alat pendeteksi dan penyaring spam.

Dalam mengitung probabilitas pada metode ini, diperlukan beberapa hal, yaitu :

𝑃(𝐻|𝑋) = 𝑃(𝑋|𝐻)𝑃(𝐻) / 𝑃(𝑋)

Penjelasan:

a. 𝑃(𝐻|𝑋) : posterior probability dari sebuah kategori, kategori H data X.

b. 𝑃(𝐻) : prior probability sebuah kategori H.

c. 𝑃(𝑋|𝐻) : kemungkinan data X pada kategori H.

d. 𝑃(𝑋) : prior probability data X.

Bentuk pengaplikasian teorema Bayes diatas kedalam agoritma Naïve Bayes Classifier adalah menurut algoritma ini, setiap dokumen direpresentasikan dengan atribut 𝑥1, 𝑥2, 𝑥3, 𝑥4 … , 𝑥𝑛. 𝑥1 adalah kata pertama yang muncul, 𝑥2 adalah kata kedua yang muncul dan seterusnya hingga 𝑥𝑛 yang merupakan kata terakhir pada sebuah teks.

Pada algoritma Naïve Bayes Classifier, probabilitas kata di suatu dokumen tertentu akan dibandingkan dengan setiap kategori yang ada.

𝑃(𝐶𝑖 |𝑋) = 𝑃(𝑋|𝐶𝑖 )𝑃(𝐶𝑖) / 𝑃(𝑋)

𝑃(𝑋) = 𝑃(𝑥1, 𝑥2, 𝑥3, 𝑥4… , 𝑥𝑛) adalah konstan maka 𝑃(𝐶𝑖 |𝑋)𝑃(𝐶𝑖) dimaksimalkan. Perubahan persamaan akan berubah menjadi

𝑃(𝐶𝑖 |𝑋) = 𝑃(𝐶𝑖 , |𝑋)𝑃(𝐶𝑖)

𝑃(X |𝐶𝑖 ) = ∏ 𝑃(𝑋𝑘|𝐶𝑖) 𝑛 𝑘=1

𝑃(X |𝐶𝑖 ) = 𝑃(𝑋1 |𝐶𝑖 ) 𝑥 𝑃(𝑋2 |𝐶𝑖 ) 𝑥 𝑃(𝑋3 |𝐶𝑖 ) 𝑥 … 𝑥 𝑃(𝑋𝑛|𝐶𝑖)

Bandingkan ilai probabilitas tertinggi dokumen yang diuji dengan nilai probabilitas tiap dokumen pada tiap kategori.

𝑃(X |𝐶𝑖 )𝑃(𝐶𝑖 ) > 𝑃(X |𝐶𝑗)𝑃(𝐶𝑗) 𝑢𝑛𝑡𝑢𝑘 1 ≤ 𝑗 ≤ 𝑚,𝑗 ≠ i

Sehingga 𝑃(X |𝐶𝑖 )𝑃(𝐶𝑖 ) adalaha maksimum sehingga menghasilkan persamaan

𝑃(𝐶𝑚) = |𝑑𝑜𝑐𝑠 𝑚| / |𝑑𝑜𝑐𝑠|

𝑃(𝑥𝑖 |𝐶𝑚) = 𝑛𝑘 + 1 / 𝑛 + |𝑤𝑜𝑟𝑑𝑠| + 2

Penjelasan :

1. |𝑑𝑜𝑐𝑠 𝑚| : jumlah semua dokumen pada kategori 𝑚.
2. |𝑑𝑜𝑐𝑠 | : jumlah semua dokumen pada data training.
3. 𝑛𝑘 : frekuensi munculnya kata 𝑥𝑖 pada kategori 𝐶𝑚.
4. 𝑛 : frekuensi munculnya kata dari tiap kategori.
5. |𝑤𝑜𝑟𝑑𝑠| : jumlah kata semua kategori pada data training.

**BAB 3**

**ANALISIS DAN PERANCANGAN**

Bab ini akan menjelaskan bagaimana tahapan-tahapan yang akan dilalui melalui penelitian analisis sentimen Postingan Mengandung Unsur Bunuh Diri

**3.1. Fase *Communication***

Fase komunikasi merupakan fase di mana data akan diambil. Penelitian ini akan diambil menggunakan data yang diambil dari Twitter menggunakan Python. Data yang akan digunakan akan berupa komentar postingan di Twitter yang menggunakan bahasa inggris yang telah di himpun pada laman Kaggle. Data tersebut akan diambil untuk dijadikan sebagai sumber pelatihan.

Data yang diambil akan diambil dari rentang waktu… . Pengumpulan data akan menghasilkan … baris data. Data akan dikategorikan menjadi positif dan negatif.

**3.2. Fase *Planning***

Tujuan dari fase perencanaan adalah untuk memprediksi sumber daya apa yang akan digunakan. Fase ini juga dapat menjadi estimasi waktu pengerjaan dari penelitian ini. Gunanya adalaha untuk memastikan target dapat tercapai. Fase ini akan dibagi menjadi tiga, yaitu analisis fungsional, analisis perangkat keras, dan analisis perangkat lunak.

|  |  |  |  |
| --- | --- | --- | --- |
| No. | Kode | Nama Fitur | Fungsi |
| 1 | F001 | Beranda | Menampilkan halaman beranda. |
| 2 | F002 | Komentar | Menyajikan data yang telah diambil. |
| 3 | F003 | Hasil Training | Menyajikan data yang sudah diberi label. |
| 4 | F002 | Analisa | Menampilkan persentase positif, negatif, akurasi, presisi, dan recall seerta menampilkan diagram positif dan negatif |

**3.2.1. Analisis Kebutuhan Fungsional**

Analisi kebutuhan fungsional akan berguna untuk menjelaskan fitur-fitur apa saja yang akan tersedia pada penelitian analisis sentimen terhadap postingan yang mengandung unsur bunuh diri. Fitur-fitur yang tersedia juga menjadi bentuk produk yang dihasilkan dari penelitian ini. Berikut fitur-fitur yang akan tersedia.

**3.2.2. Analisis Kebutuhan Perangkat Keras**

Analasisi Pernagkat keras berfungsi untuk mengetahui perangkat keras yang digunakan dalam penelitian ini. Beriku spesifikasi dari perangkat keras yang digunakan

1. Manufaktur : ASUS TUF Gaming A15
2. Prosessor : AMD Ryzen 7 7735HS with Radeon Graphics, 3201 Mhz, 8 Core(s), 16 Logical Processor(s)
3. RAM : 16 GB
4. Tipe Sistem : 64-bit
5. Sistem Operasi : Microsoft Windows 11 Home Single Language

**3.2.3. Analisis Kebutuhan Perangkat Lunak**

Analisis kebutuhan perangkat lunak dilakukan supaya kebutuhan perangkat lunak yang akan menunjang penelitian ini dapat diketahui. Berikut adalah perangkat lunak yang akan digunakan

1. Microsoft Excel yang akan digunakan untuk memberi label setiap data secara manual. Data yang telah dilabeli akan disimpan dalam format CSV.
2. Google Collab (Python) yang akan digunakan untuk mengerjakan model analisis sentimen. Setelah itu akan dilakukan analisa pada model guna menghasilkan *output* yang diinginkan.
3. Visual Studio Code sebagai teks editor yang berfungsi dalam pembangunan model analisis sentimen
4. Google Chrome sebagai browser yang digunakan untuk mengakses Google Collab secara *online*.

**3.3. Fase *Modeling***

Bertujuan membuat analisa model yang akan dibuat guna memberi gambaran secara garis besar dari penelitian ini. Fase ini akan terdiri dari percangangan UML dan rancangan antar muka. Selain itu, tahapan pembangunan model analisis sentimen terhadap postingan bunuh diri di Twitter juga dijelaskan.

**3.3.1 Pelabelan Data**

Data telah dikumpulkan pada fase sebelumnya, yaitu fase communication. Data yang telah dikumpulkan tersebut belum rapih, artinya masih banyak data yang memiliki missing value dan target dari data juga belum memiliki pelabelan. Data kemudian dipisahkan berdasarkan kategori. Diperlukan fase pelabelan data supaya data dapat diolah oleh mesin dengan lebih mudah.

Data akan diberi dua jenis label, yaitu positif dan negatif. Data mana yang memiliki label positif atau negatif tergantung dari topik penelitiannya. Pada penelitian ini, label positif akan diberikan pada data yang tidak memiliki unsur bunuh diri, sementara label negatif akan diberikan pada data yang memiliki unsur bunuh diri. Selanjutnya data akan diolah.

**3.3.2. Preprocessing Data**

Data yang diambil masih memiliki hal-hal yang sulit untuk diolah seperti emoji, nama akun, *hashtag*, tanda baca, dan sebagainya. Hal ini dapat menyulitkan penelitian karena bisa saja hal-hal ini mempengaruhi hasil dari pembelajaran yang dilakukan mesin. Tahap ini akan menghilangkan hal-hal tersebut sehingga mesin akan hanya mengelola dari kata per kata. Tahapan dari preprocessing adalah sebagai berikut :

1. **Cleansing**

Merupakan tahap di mana hal-hal yang tidak dipergunakan dalam proses pelatihan dihilangkan. Hal-hal ini antara lain menghilangkan tanda baca, nama akun pengguna, simbol-simbol, emoji, hashtag, dan sebagainya. Contoh dapat dilihat pada tabel … .

|  |  |
| --- | --- |
| **Sebelum Cleansing** | **Sesudah Cleansing** |
| @ValenValdez Oh, that's good to hear. But is it over already? Or you'll continue it after the Holy week? | Oh that’s good to hear But is it over already Or you’ll continue it after the Holy week |
| @dizzyhrvy that crap took me forever to put together. iÃ¢Â€Â™m going to go sleep for DAYS | that crap took me forever to put together going to go sleep for DAYS |

1. **Case Folding**

Tahap ini merupakan tahap penyamaan hal-hal yang tidak seragam. Dalam suatu kalimat akan terdiri dari huruf kapital dan huruf kecil. Tahap ini akan menyeragamkan semua tulisan menjadi huruf kecil saja guna mempermudah proses pelatihan data. Contoh dari case folding ada pada tabel 3. .

|  |  |
| --- | --- |
| **Sebelum Case Folding** | **Sesudah Case Folding** |
| Oh that’s good to hear But is it over already Or you’ll continue it after the Holy week | oh that’s good to hear but is it over already or you’ll continue it after the holy week |
| that crap took me forever to put together going to go sleep for DAYS | that crap took me forever to put together going to go sleep for days |

1. **Normalisasi**

Data yang tersedia merupakan data yang diambil dari sosial media. Pada penulisan suatu postingan sosial media terdapat banyak kata yang tidak mengikuti kaidah bahasa yang baik. Kesalahan penulisan, singkatan, *slang*, dan lain-lain akan banyak ditemukan. Contohnya kata “i’m” yang merupakan singkatan dari “i am”. Selanjutnya ada kata “nvm” yang merupakan singkatan dari “*nevermind*.” Pentingnya normalisasi adalah untuk memudahkan pelatihan yang dilakukan oleh mesin. Kata-kata singkatan atau beberapa yang telah disebutkan di atas dapat diartikan berbeda oleh mesin biarpun memiliki maksud yang sama.

Kata “nvm” memiliki arti yang sama dengan “*nevermind”*, tapi karena penulisannya berbeda maka mesin bisa menganggap hal itu dua tulisan dengan arti yang berbeda. Melihat hal tersebut tentu perlu untuk melakukan normalisasi supaya kesalahan pengartian dapat dihindari. Terdapat … kata yang dibuat manual yang diambil dari dataset. Adapun contoh dari tahap normalisasi dapat dilihat pada tabel 3. .

|  |  |
| --- | --- |
| **Sebelum Normalisasi** | **Sesudah Normalisasi** |
| oh that’s good to hear but is it over already or you’ll continue it after the holy week | oh that is good to hear but is it over already or you will continue it after the holy week |
| that crap took me forever to put together going to go sleep for days | that crap took me forever to put together going to go sleep for days |

1. **Tokenisasi**

|  |  |  |
| --- | --- | --- |
| **Setelah Tokenisasi** | | |
| that | crap | took |
| me | forever | to |
| put | together | going |
| to | go | sleep |
| for | days |  |

Merupakan tahap pemecahan. Kalimat terdiri dari kata-kata. Setiap kata akan dipecah satu persatu. Kata akan disimpan dalam bentuk array. Contoh dari tokenisasi ada pada tabel 3. .

|  |
| --- |
| **Sebelum Tokenisasi** |
| that crap took me forever to put together going to go sleep for days |

1. **Stopwords Removal**

|  |  |  |
| --- | --- | --- |
| **Sesudah Stopwords Removal** | | |
| crap | sleep | days |

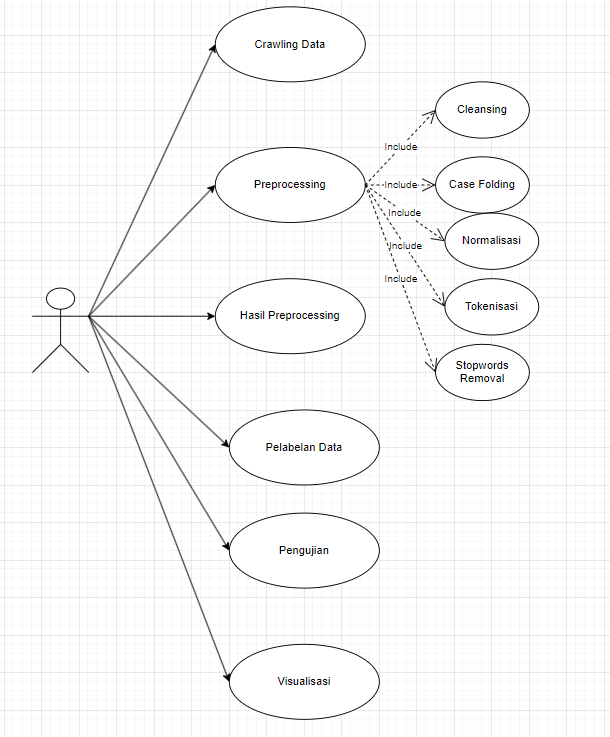
Fase ini merupakan fase penghapusan kata-kata yang tidak dibutuhkan dalam proses pelatihan mesin. Contoh dari *stopwords* ini sendiri adalah “because”, “a”, “an; “I”, to be, dan masih banyak lagi. Sumber stopword diambil dari github. Adapun contoh dari fase stopword removal ada pada tabel 3. .

|  |  |  |
| --- | --- | --- |
| **Sebelum Stopwords Removal** | | |
| that | crap | took |
| me | forever | to |
| put | together | going |
| to | go | sleep |
| for | days |  |

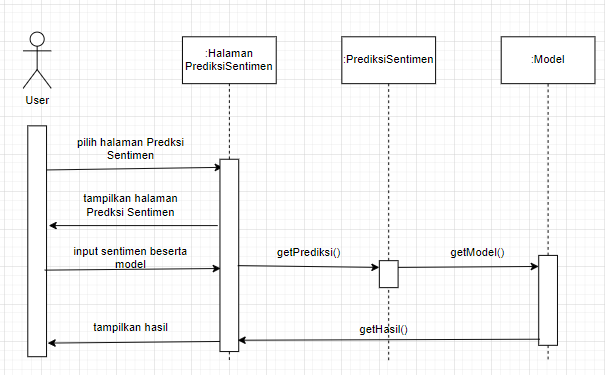
**3.4. Perancangan UML**

1. **Use Case Diagram**

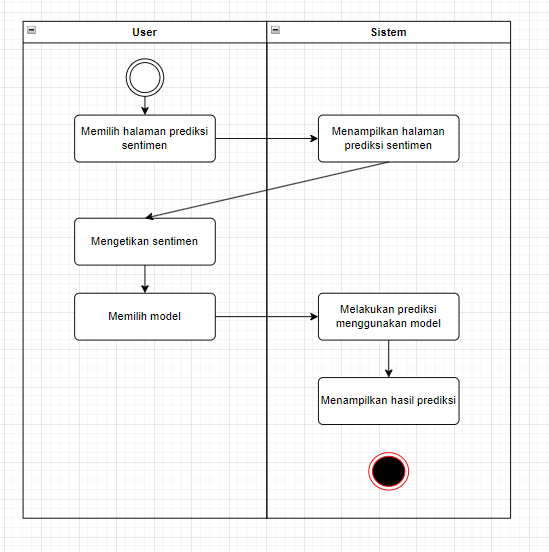
Use case diagram menampilkan alur bagaimana penelitian akan dilakukan. Tahap pertama adalah pengambilan data yang diambil dari media sosial Twitter. Selanjutnya, data yang telah diambil akan masuk ke tahap preprocessing. Dalam preprocessing, data akan melalui lima fase, yaitu *cleansing*, *case folding*, normalisasi, tokenisasi, dan *stopwords removal*. Lalu dari *preprocessing* akan diberi label. Setelahnya data-data tersebut akan diuji dan pada akhirnya dapat divisualisasikan.

****

1. **Sequence Diagram**

****

1. **Activity Diagram**



**3.5. Perancangan Antar Muka**

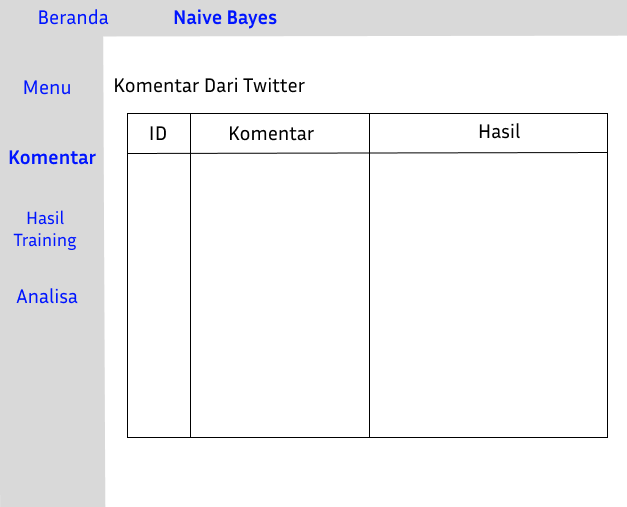
**3.5.1 Beranda**

Saat pertama kali membuka website, user akan masuk ke halaman beranda. Beranda hanya akan berisi judul dari penelitian ini. Selanjutnya user dapat masuk ke halaman Naïve Bayes. Halaman ini akan berisi tiga sub halaman, yaitu komentar, hasil training, dan analisa. Rincian lebih lanjut akan dijelaskan pada bagian selanjutnya.

****

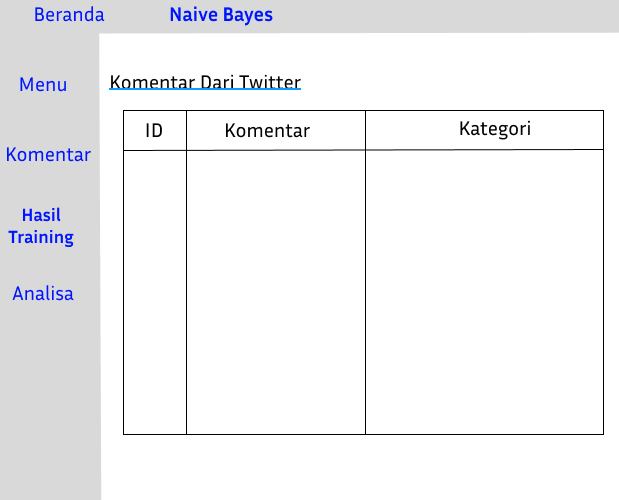
**3.5.2 Komentar**

Pada halaman komentar akan ditampilkan hasil data yang telah diambil. Data-data tersebut sudah dipisahkan dan diberikan ID masing-masing.

****

**3.5.3 Hasil Training**

Halaman hasil training akan berisi hasil dari pelabelan setiap data. Pada penelitian ini tidak ada label netral dikarenakan untuk penentuannya hanya akan ada dua hal, yaitu positif (tidak mengandung unsur bunuh diri) dan negatif (mengandung unsur bunuh diri).

****

**3.5.4 Analisa**

Pada halaman analisa, user dapat melihat jumlah komentar positif maupun negatif. Jumlahnya akan ditaampilkan dalam diagram dan dalam persen. Selanjutnya, user juga dapat melihat tingkat akurasi, presisi, dan recall. Ketiga hal tersebut juga direpresentasikan melalui persentase.

****

**3.6. Perhitungan Naïve Bayes Classification**

Metode Naïve Bayes melakukan klasifikasi melalu training dan testing setelah tahapan preprocessing. Akan ada pembagian kalimat yang dianggap positif maupun negatif. Berikut flowchart proses Naïve Bayes :

**3.6.1 Data Training**

Data akan dibagi menjadi dua kategori. Pertama kategori positif (tidak mengandung unsur bunuh diri) dilambangkan dengan angka “1”. Kedua ada kategori negatif (mengandung unsur bunuh diri) dilambangkan dengan angka “-1”. Pada penelitian ini, data netral atau “0” dihilangkan atau dianggap sebagai data positif. Alasannya adalah pada penelitian ini hanya akan menentukan apakah postingan tersebut mengandung unsur bunuh diri atau tidak. Asal tidak masuk kategori negatif “-1” maka dianggap tidak memiliki unsur bunuh diri.

|  |  |
| --- | --- |
| **Bobot** | **Komentar** |
| 1 | [‘oh’ ‘that’ ‘is’ ‘good’ ‘to’ ‘hear’ ‘but’ ‘is’ ‘it’ ‘over’ ‘already’ ‘or’ ‘you’ ‘will’ ‘continue’ ‘it’ ‘after’ ‘the’ ‘holy’ ‘week’] |
| -1 | [‘that’ ‘crap’ ‘took’ ‘me’ ‘forever’ ‘to’ ‘put’ ‘together’ ‘going’ ‘to’ ‘go’ ‘sleep’ ‘for’ ‘days’] |

**3.7. Fase *Construction***

Merupakan fase untuk merubah rancangan pada fase sebelumnya, yaitu fase *modeling* menjadi menjadi bahasa yang dapat dimengerti oleh mesin. Sistem ini akan menggunakan bahasa pemrograman python sebagai alat pengembangan sistem. Penelitian akan disajikan dalam bentuk *website* yang akan menggunakan framework dari python, yaitu *framework flask*. Pada nantinya *website* akan dapat diakses oleh pengguna dari berbagai lokasi secara daring. Adapun halaman yang akan tersedia pada website antara lain : (1) Beranda, (2) Search analisis sentimen.

**3.8. Fase *Deployment***

Merupakan fase di mana hasil dari peneletian berupa sistem yang sudah selesai akan diunggah ke media online. Sistem yang diunggah akan berupa website. Kegunaan fase ini adalah supaya sistem yang telah selesai dapat diakses dari berbagai media seperti PC, laptop, dan handphone. Setelah berhasil dideploy, akan dilakukan maintenance atau pemeliharaan sistem. Fungsi dari maintenance adalah untuk terus merawat sistem apabila terjadi *bug* atau *error*.