Proposal Proyek

**Implementasi Naive Bayes pada Sentiment Analisis pemilihan Gubernur 2024 menggunakan LLM**

|  |  |  |
| --- | --- | --- |
| Nama Anggota Kelompok | : | 1. 223510 – Dwi Yuda 2. 223510276 – Lukman Basha 3. 223510795 – Rifki Nul Arif 4. 223510 – Sulthon Abiyyiafflah 5. 223510363 – Tanny Apriliyani |
| Kelas | : | 6 A |
| Dosen | : | Mutia Fadhilla, S.ST., M.Sc. |
| Tanggal Pengumpulan Proposal | : | 18-06-2025 |

# Latar Belakang

Pemilihan Gubernur 2024 merupakan salah satu agenda penting dalam sistem demokrasi Indonesia yang mencerminkan partisipasi aktif masyarakat dalam menentukan arah pembangunan daerah. Dalam era digital seperti saat ini, opini masyarakat tidak hanya terbatas pada survei atau forum diskusi konvensional, tetapi tersebar luas melalui media sosial seperti Twitter, Facebook, Instagram, serta kolom komentar media daring. Keberadaan data teks yang melimpah ini membuka peluang besar untuk memahami sentimen publik melalui pendekatan teknologi, khususnya Natural Language Processing (NLP) dan Machine Learning (ML).

Analisis sentimen adalah salah satu aplikasi NLP yang bertujuan untuk mengidentifikasi dan mengklasifikasikan opini atau emosi dalam teks menjadi kategori seperti positif, negatif, atau netral. Pendekatan ini sangat relevan dalam konteks pemilihan umum karena dapat menggambarkan persepsi masyarakat terhadap calon gubernur, isu-isu kampanye, serta dinamika politik lokal. Salah satu algoritma klasik yang telah terbukti efektif dalam tugas klasifikasi teks adalah Naive Bayes, yang bekerja berdasarkan prinsip probabilistik dan asumsi independensi antar fitur. Algoritma ini memiliki keunggulan dalam hal kesederhanaan, kecepatan pelatihan, dan performa yang cukup baik pada data teks yang telah diproses dengan benar.

Namun, tantangan utama dalam analisis sentimen bukan hanya pada pemilihan algoritma, tetapi juga pada tahap prapemrosesan data teks. Berdasarkan materi pembelajaran pada pertemuan ke-6, diketahui bahwa proses seperti tokenisasi, penghapusan stopword, stemming, dan representasi fitur menggunakan TF-IDF merupakan tahapan krusial dalam pipeline NLP. Dengan prapemrosesan yang tepat, model klasifikasi seperti Naive Bayes dapat bekerja secara lebih efektif dan menghasilkan prediksi yang lebih akurat.

Di sisi lain, perkembangan teknologi saat ini juga menghadirkan model-model canggih berbasis Large Language Model (LLM) seperti BERT, GPT, dan sejenisnya. LLM dapat digunakan untuk membantu proses ekstraksi fitur, memahami konteks kalimat yang lebih kompleks, atau bahkan memperkuat representasi data sebelum digunakan oleh model klasik seperti Naive Bayes. Oleh karena itu, dalam proyek ini akan dilakukan integrasi antara pendekatan tradisional dan modern, dengan tujuan mengembangkan sistem analisis sentimen yang adaptif dan informatif.

Dengan mengangkat isu Pemilihan Gubernur 2024 sebagai studi kasus, proyek ini menjadi sangat relevan karena tidak hanya melatih keterampilan teknis mahasiswa dalam bidang machine learning dan NLP, tetapi juga memberikan kontribusi nyata dalam membantu masyarakat memahami opini publik secara lebih terstruktur dan berbasis data.

# Identifikasi Masalah

Pada masa menjelang Pemilihan Gubernur 2024, masyarakat aktif menyampaikan pendapat dan tanggapan mereka terhadap para calon melalui media sosial dan berbagai platform digital. Namun, jumlah data yang sangat besar ini sulit untuk dianalisis secara manual, sehingga opini publik yang penting bisa saja terabaikan.

Dengan menggunakan data teks dari media sosial, seperti komentar atau unggahan publik, kita dapat membangun model pembelajaran mesin yang mampu mengklasifikasikan sentimen masyarakat terhadap calon gubernur ke dalam kategori seperti *positif*, *negatif*, atau *netral*. Model ini dapat membantu untuk mengetahui opini mayoritas masyarakat dan mendukung pengambilan keputusan oleh berbagai pihak seperti tim kampanye, media, dan peneliti sosial.

# Tujuan Proyek

* Untuk membantu mahasiswa memahami dan menganalisis hasil keluaran dari Large Language Model (LLM) dalam konteks analisis sentimen pada isu politik, khususnya Pemilihan Gubernur 2024.
* Untuk mengevaluasi sejauh mana algoritma Naive Bayes dapat berkontribusi dalam mengklasifikasikan sentimen publik berdasarkan data teks yang telah diolah dan diperkaya oleh LLM.
* Untuk membangun sebuah pipeline sederhana NLP yang mencakup tahapan praproses teks, ekstraksi fitur, dan klasifikasi sentimen secara otomatis.
* Untuk menguji performa model klasifikasi sentimen menggunakan metrik seperti akurasi, presisi, dan recall, sehingga dapat diketahui efektivitas model dalam konteks data nyata.
* Untuk memberikan gambaran nyata kepada mahasiswa tentang bagaimana machine learning dan NLP dapat diterapkan pada isu-isu sosial dan politik di dunia nyata.

# Algoritma Machine Learning

Dalam proyek ini, kami akan menggunakan gabungan dua pendekatan utama dalam machine learning dan Natural Language Processing (NLP), yaitu Large Language Model (LLM) dan Naive Bayes Classifier. Kombinasi keduanya bertujuan untuk menghasilkan model analisis sentimen yang kuat dan kontekstual.

1. Large Language Model (LLM)

Large Language Model adalah model pembelajaran mesin berskala besar yang dilatih menggunakan miliaran kata dari berbagai sumber teks. LLM seperti GPT (Generative Pre-trained Transformer), BERT, dan lainnya, memiliki kemampuan memahami konteks kalimat, makna kata, serta pola bahasa secara mendalam. Dalam proyek ini, LLM digunakan sebagai alat bantu praproses dan ekstraksi fitur.

Beberapa fungsi LLM dalam proyek ini:

* Melakukan ekstraksi teks yang lebih bermakna dibandingkan pendekatan konvensional (TF-IDF, Bag of Words).
* Menghasilkan representasi vektor kata/kalimat yang kaya konteks, yang akan menjadi input bagi algoritma klasifikasi.
* Membersihkan dan menyarankan label awal sentimen untuk membantu pelabelan data (jika diperlukan).

Pendekatan ini relevan karena LLM memiliki kemampuan memahami konteks kalimat dalam teks opini atau komentar politik yang biasanya bersifat ambigu, ironi, atau sarkasme.

2. Naive Bayes Classifier

Naive Bayes adalah salah satu algoritma klasifikasi paling sederhana dan efisien yang berbasis pada teorema Bayes dengan asumsi independensi antar fitur.

Cara kerja singkat:

Naive Bayes menghitung probabilitas setiap kelas (misalnya *positif*, *negatif*, atau *netral*) berdasarkan distribusi kata dalam dokumen. Kemudian, model akan memilih kelas dengan probabilitas tertinggi sebagai hasil klasifikasinya. Rumus dasarnya sebagai berikut:

P(C∣X)=P(X∣C)⋅P(C)

Di mana:

* P(C∣X) = probabilitas suatu kelas C (misal "positif") terhadap dokumen X
* P(X∣C) = probabilitas munculnya kata-kata pada dokumen X jika diketahui kelas C
* P(C) = probabilitas awal kelas C
* P(X) = probabilitas dari dokumen X secara keseluruhan

Alasan pemilihan Naive Bayes:

* Sangat cocok untuk klasifikasi teks yang bersifat linier dan berdimensi tinggi.
* Cepat dan ringan secara komputasi, sangat ideal untuk mahasiswa yang ingin fokus pada implementasi dasar.
* Telah teruji secara luas dalam tugas-tugas analisis sentimen dengan performa yang kompetitif.

Asumsi & Pertimbangan:

Naive Bayes mengasumsikan bahwa fitur (kata-kata dalam dokumen) saling independen secara statistik, yang dalam kenyataan NLP tidak selalu terpenuhi. Namun, meski asumsi ini tidak realistis, performa Naive Bayes tetap kuat karena sifat teks yang sangat sparsedan domain-spesifik.

# Metodologi Proyek

Metodologi yang digunakan dalam proyek ini terdiri dari beberapa tahapan utama, mulai dari pengumpulan data hingga evaluasi hasil analisis sentimen. Tahapan-tahapan ini dirancang agar mahasiswa dapat memahami dan mengimplementasikan proses analisis sentimen secara end-to-end dengan menggabungkan LLM dan algoritma Naive Bayes. Berikut adalah tahapan yang dilakukan:

1. Crawling Data

Data dikumpulkan dari platform Twitter menggunakan API dan library seperti tweepy. Kata kunci atau tagar yang berkaitan dengan Pemilihan Gubernur 2024 digunakan untuk mendapatkan opini publik yang relevan. Data yang dikumpulkan berupa teks mentah dari tweet, termasuk tanggal, username, dan isi tweet.

2. Pre-processing Teks

Data mentah yang telah diperoleh kemudian melalui tahapan prapemrosesan agar siap digunakan dalam proses analisis. Proses ini meliputi:

* Menghapus tanda baca, angka, dan karakter khusus
* Tokenisasi (memecah kalimat menjadi kata)
* Menghapus stopwords (kata umum yang tidak berkontribusi pada makna)
* Stemming (mengubah kata ke bentuk dasar)
* Normalisasi teks (pengubahan slang atau singkatan ke bentuk formal)

Notebook PRE\_PROCESSING.ipynb menjelaskan detail proses ini menggunakan pustaka NLP Indonesia seperti Sastrawi.

3. Implementasi Sentiment Analysis Menggunakan LLM

Teks yang telah diproses kemudian dianalisis menggunakan Large Language Model (LLM), seperti ChatGPT atau BERT (bila menggunakan embedding lokal). LLM digunakan untuk:

* Memberi label awal sentimen (*positif*, *negatif*, *netral*) berdasarkan konteks kalimat
* Memberi vektor representasi teks berbasis konteks (jika menggunakan embedding)

Output dari tahap ini adalah dataset teks beserta label sentimen yang siap digunakan untuk pelatihan model.

4. Klasifikasi Sentimen dengan Naive Bayes

Hasil dari LLM digunakan sebagai input bagi model Naive Bayes. Model ini kemudian dilatih menggunakan dataset berlabel untuk melakukan klasifikasi teks baru secara otomatis. Model dievaluasi berdasarkan:

* Akurasi
* Precision
* Recall
* Confusion Matrix

Tahapan ini memberikan pemahaman kepada mahasiswa tentang bagaimana model probabilistik dapat digunakan untuk menyederhanakan pengambilan keputusan berbasis teks.

5. Evaluasi dan Visualisasi

Hasil dari klasifikasi divisualisasikan dalam bentuk grafik atau tabel, misalnya:

* Sebaran sentimen per kandidat
* Perbandingan hasil klasifikasi Naive Bayes dengan label LLM
* Analisis waktu (timeline opini publik)

Visualisasi ini membantu memperjelas temuan dan menunjukkan potensi penerapan sistem dalam pemantauan opini publik secara real-time.

# Dataset

Dataset yang digunakan dalam proyek ini berasal dari hasil crawling media sosial Twitter menggunakan kata kunci yang berkaitan dengan Pemilihan Gubernur 2024. Setelah dilakukan pembersihan dan praproses, data disimpan dalam file berformat .csv dengan nama cleaned\_pilkada\_sentiment1.csv.

Struktur Dataset

Beberapa kolom utama dalam dataset ini antara lain:

* Tweet : Berisi teks opini atau komentar dari pengguna Twitter yang telah dibersihkan (tanpa emoji, URL, tanda baca, dll.).
* Label : Label sentimen yang menunjukkan apakah tweet tersebut positif, negatif, atau netral.
* (Kolom lainnya bisa mencakup informasi waktu, username, atau hasil transformasi tertentu, jika relevan)

Karakteristik Dataset:

* Dataset berisi ratusan hingga ribuan entri (tergantung hasil crawling).
* Data telah melalui tahap pre-processing: tokenisasi, stopword removal, normalisasi, dan stemming, sebagaimana dijelaskan dalam tahap metodologi.
* Label sentimen diperoleh melalui kombinasi manual dan bantuan LLM (sebagai baseline pengklasifikasian awal).

Alasan Pemilihan Dataset:

* Dataset mencerminkan opini nyata masyarakat secara online mengenai Pilkada.
* Relevan untuk tugas klasifikasi teks karena memiliki keberagaman emosi dan konteks politik.
* Cocok untuk menguji performa algoritma Naive Bayes dalam klasifikasi teks berbahasa Indonesia yang telah diproses.

# Pembagian Tugas Kelompok

Contoh:

|  |  |  |  |
| --- | --- | --- | --- |
| No | NPM | Nama Anggota Kelompok | Pembagian Tugas |
| 1 | 123 | Alice | Data preprocessing |
| 2 | 456 | Bob | Model implementation |
| 3 | 789 | Charlie | Report writing & visualization |
| … | … | … | … |

\*Semua anggota kelompok harus berkontribusi dalam implementasi kode program dan penulisan laporan

# Daftar Pustaka

Cantumkan referensi yang mendukung proposal. Gunakan format kutipan IEEE. Silahkan gunakan aplikasi Mendeley untuk memudahkan pengelolaan referensi/daftar pustaka dalam laporan.

# Catatan dan Informasi Tambahan

## Ketentuan Judul Proyek

* + Berikan judul yang spesifik dan deskriptif yang mencerminkan fokus proyek.
  + Contoh: "Prediksi Tingkat Prestasi Akademik Siswa Menggunakan Algoritma Klasifikasi Decision Tree"
  + Setiap kelompok wajib memiliki judul yang berbeda

## Petunjuk Pengumpulan Proposal

* + Format file: **PDF**
  + Proposal terdiri dari **2–5 halaman**
  + Kumpulkan proposal paling lambat **15 Juni 2025** melalui **Google Classroom**
  + **Pengumpulan** proposal hanya dilakukan **oleh satu perwakilan kelompok**