TUGAS PROJECT MACHINE LEARNING (BULLYING)

Nama Ketua : Randy Ardiansyah (20230801131)

Nama Anggota :
Briyan Jonathan (20230801036)
Sandi (20230801064)
Muhammad Febriyan Putrahariska (20230801043)
Arie awijaya (20230801120)
Nico saputra (20230801206)

FAKULTAS ILMU KOMPUTER

PROGRAM STUDI S1-TERAPAN TEKNIK INFORMATIKA JAKARTA

2024

Studi kasus: Analisis Sentiment Terhadap Kasus Perundungan (Bullying) di masyarakat dengan Dataset Twitter berbasis Hybrid Model Machine Learning

TUGAS:

1) Anda harus membuat konsep aplikasi dengan jelas, Jelaskan dalam
Project Overview !
2) Lakukan Crawling Data Twitter Minimal 6.000 Dataset dengan String Value / Kata Kunci Pencarian sesuai studi kasus masing
masing!
3) Buatlah Flowchart Tahapan Pre-processing Data Dalam Sebuah Sentiment Analysis!
4) Buatlah Arsitektur Model Hybrid Machine Learning !

KONSEP

Project Overview : Analisis Sentimen Terhadap Kasus Perundungan (Bullying) di Masyarakat Berbasis Hybrid Model Machine Learning

Latar Belakang :
Kasus perundungan (bullying) di masyarakat merupakan masalah sosial yang serius dan kompleks. Fenomena ini tidak hanya berdampak pada korban secara fisik dan psikologis, tetapi juga dapat mempengaruhi lingkungan sosial secara keseluruhan. Untuk menangani dan mencegah kasus perundungan, penting untuk memahami pola dan sentimen masyarakat terhadap perundungan. Salah satu cara untuk mengidentifikasi dan menganalisis sentimen ini adalah melalui analisis data media sosial, di mana banyak orang mengungkapkan pandangan dan pengalaman mereka.

Tujuan Proyek :
Proyek ini bertujuan untuk mengembangkan sebuah aplikasi yang mampu menganalisis sentimen masyarakat terhadap kasus perundungan menggunakan pendekatan hybrid model machine learning. Aplikasi ini akan mengumpulkan data dari berbagai platform media sosial, menganalisis sentimen publik, dan memberikan wawasan yang dapat digunakan oleh pemangku kepentingan seperti lembaga pendidikan, pemerintah, dan organisasi non-profit untuk mengembangkan strategi penanganan dan pencegahan perundungan yang lebih efektif.

Studi kasus : penggunaan learning machine sebagai pilar utama untuk mengenelusuri kasus bullying yang marak di lingkungan masyarakat atau pun sekolah, agar lebih optimal kami menyajikan 6000 data kasus bullying, yang kami ambil dalam salah satu aplikasi jejaring media sosial yaitu X atau twitter.

Problem :

- Marak nya kasus bullying
- Identifikasi kasus bullying
- Dampak kasus bullying terhadap masyarakat

Goals :

- Meningkatkan experience baru terhadap learning machine
- Menangani cyber-bullying dengan efektif
- Meningkatkan partisipasi dan tanggung jawab masyarakat
- Menghargai berbagai perbedaan dan budaya lingkungan

Deliverables :

- Pengumpulan data dengan data crawling
- Memproses data kedalam machine learning
- Evaluasi
- Implementasi dan hasil kerja

Risk or Obstacles :

- Pengumpulan data : terdapat limit perhari yang menjadi salah satu faktor masalah dalam pengumpulan data
- Masalah kualitas data : Data yang tidak lengkap, akurat, dan terhadapat berita hoax. Memastikan kualitas melalui pembersihan, validasi

Manfaat Proyek

1. **Pemahaman yang Lebih Baik**: Memahami sentimen masyarakat terhadap perundungan membantu dalam menyusun kebijakan dan program intervensi yang lebih efektif.

2. **Deteksi Dini: Identifikasi** tren dan pola perundungan secara real-time memungkinkan tindakan preventif yang lebih cepat.

3. **Peningkatan Kesadaran**: Informasi yang disajikan melalui aplikasi ini dapat digunakan untuk meningkatkan kesadaran masyarakat tentang dampak perundungan dan pentingnya pencegahan.

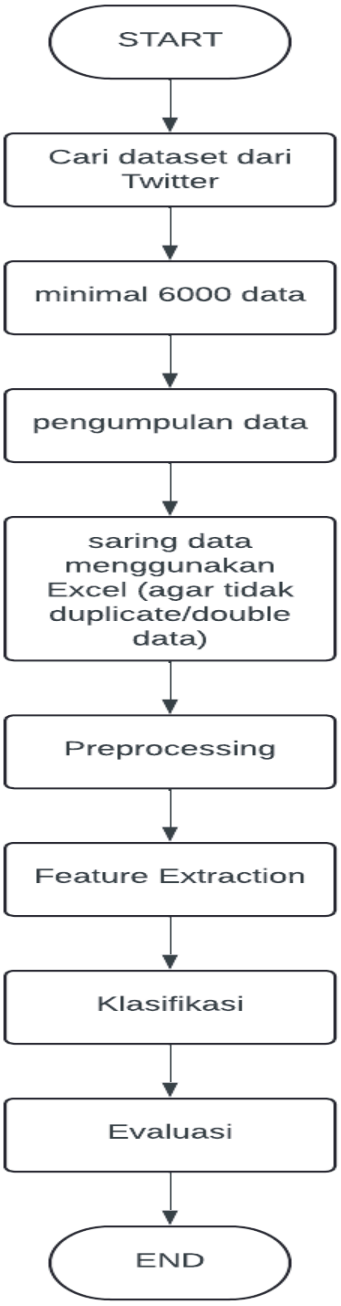
Kesimpulan :

Proyek analisis sentimen terhadap kasus perundungan berbasis hybrid model machine learning ini merupakan langkah inovatif dalam memahami dan mengatasi masalah perundungan di masyarakat. Dengan memanfaatkan teknologi machine learning dan analisis data, diharapkan aplikasi ini dapat memberikan kontribusi nyata dalam menciptakan lingkungan sosial yang lebih aman dan inklusif.

2. DATASET (6550) FROM TWITTER

: https://sg.docworkspace.com/d/sIFiIjo2zAaWlsrMG?sa=cl

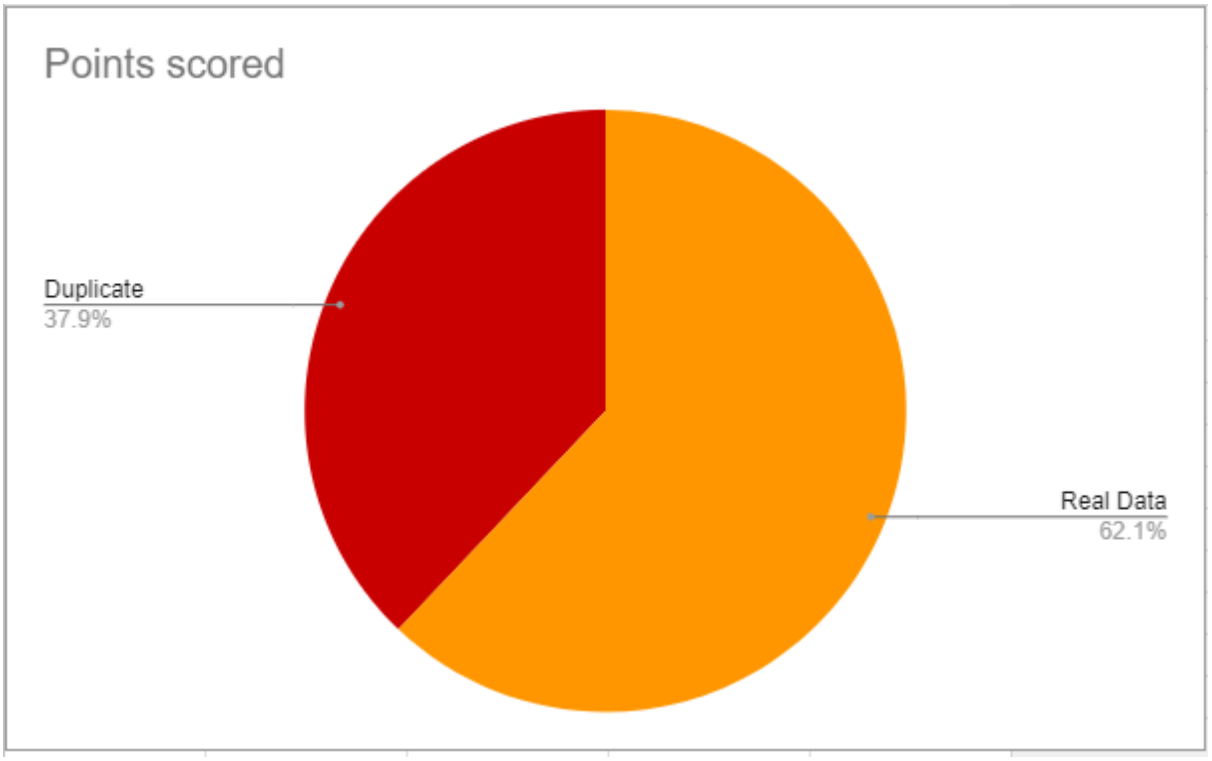
FLOWCHART (Preprocessing data)



Deskripsi Flowchart

1. cari dataset dari twitter.
2. minimal 6000 data , yang terkumpul (6550).
3. Penumpulan data, Twitter menjadi sumber data pada penelitian ini. Proses pengumpulan data dilakukan dengan memanfaatkan Twitter
   .Dalam proses penggunaan Twitter (auth token), sebagai akses untuk python untuk mencari data tweet. berupa Consumer Key Access Auth Token. Data

yang diambil merupakan data cuitan berbahasa Indonesia. Cuitan yang terkumpul untuk membangun model pada penelitian ini adalah 10.550 cuitan seperti yang terlihat pada gambar 2 dengan cacah masing masing yaitu 6550 (Real) cuitan bullying dan 4000 (duplicate) cuitan bullying.



4. Saring data menggunakan Excel agar tidak duplicate
5. Pre-prosessing adalah teknik untuk menyiapkan data agar lebih siap untuk dilakukan lebih lanjut dalam rangka ekstraksi pengetahuan.
6. Feature Extraction: Menggunakan berbagai teknik untuk mengubah teks menjadi fitur yang bisa digunakan oleh model machine learning.
7. Klasifikasi, Bagian ini menjabarkan tentang proses klasifikasi cuitan dengan menggunakan beberapa algoritma Machine Learning yaitu Multinomial Naïve Bayes, Random Forest, Decision tree, dan CNN. Algoritma tersebut dipilih karena terbukti memiliki akurasi yang sangat baik dalam klasifikas data dalam bentuk teks.
8. Evaluasi.

4. Buatlah Arsitektur Model Hybrid Machine Learning !

Kita mengambil model:

Machine learning:
- Random forest
- naive Bayes

Deep learning :
- CNN

Pembahasan:

Model hybrid menggabungkan kekuatan beberapa algoritma machine learning untuk meningkatkan akurasi dan kinerja model. Dalam konteks deteksi bullying, kita dapat menggabungkan metode machine learning tradisional dengan teknik deep learning. Berikut adalah arsitektur untuk model hybrid yang menggabungkan Random Forest, Naive Bayes, dan Convolutional Neural Network (CNN).

1. Data Preparation and Preprocessing

Langkah-langkah:

- Pembersihan Data: Menghapus data yang tidak lengkap atau tidak relevan. Mengatasi teks yang tidak sesuai, seperti menghapus karakter khusus dan melakukan normalisasi teks.
- Tokenisasi: Memecah teks menjadi kata-kata atau frasa.
- Stop Words Removal: Menghapus kata-kata umum yang tidak informatif (misalnya "dan", "atau").
- Stemming/Lemmatization: Mengubah kata-kata ke bentuk dasarnya.
- Feature Extraction: Bag of Words (BoW): Menghitung frekuensi setiap kata dalam teks.
- TF-IDF: Menghitung frekuensi term-terbalik dokumen untuk setiap kata.
- Word Embedding: Mengubah kata-kata menjadi vektor densitas menggunakan metode seperti Word2Vec atau GloVe.

2. Model Building

A. Machine Learning Models:

Random Forest:

Membangun beberapa pohon keputusan dengan subset acak dari data dan fitur. Melatih model Random Forest dengan data fitur dari BoW atau TF-IDF.

Naive Bayes:

Melatih model Naive Bayes dengan data fitur dari BoW atau TF-IDF.

B. Deep Learning Model:

Convolutional Neural Network (CNN):

- Embedding Layer: Mengubah kata-kata menjadi vektor densitas.
- Convolutional Layer: Menggunakan filter untuk mengekstraksi fitur spasial dari teks.
- Pooling Layer: Mengurangi dimensionalitas fitur dan mempertahankan informasi penting.
- Fully Connected Layer: Menggabungkan fitur untuk membuat keputusan klasifikasi.
- Output Layer: Menghasilkan probabilitas untuk setiap kelas (bullying atau non-bullying).

3. Model Fusion and Meta-Classifier

Langkah-langkah:

- Training Individual Models: Melatih model Random Forest, Naive Bayes, dan CNN secara terpisah menggunakan data training yang sama.
- Generating Predictions: Menghasilkan prediksi dari masing-masing model untuk data validation/test.
- Meta-Classifier: Menggunakan model machine learning lain (misalnya Logistic Regression) sebagai meta-classifier untuk menggabungkan prediksi dari Random Forest, Naive Bayes, dan CNN. Menggabungkan prediksi

model-model individu sebagai fitur input untuk meta-classifier. Melatih meta-classifier dengan data prediksi dari model-model individu.

4. Evaluation

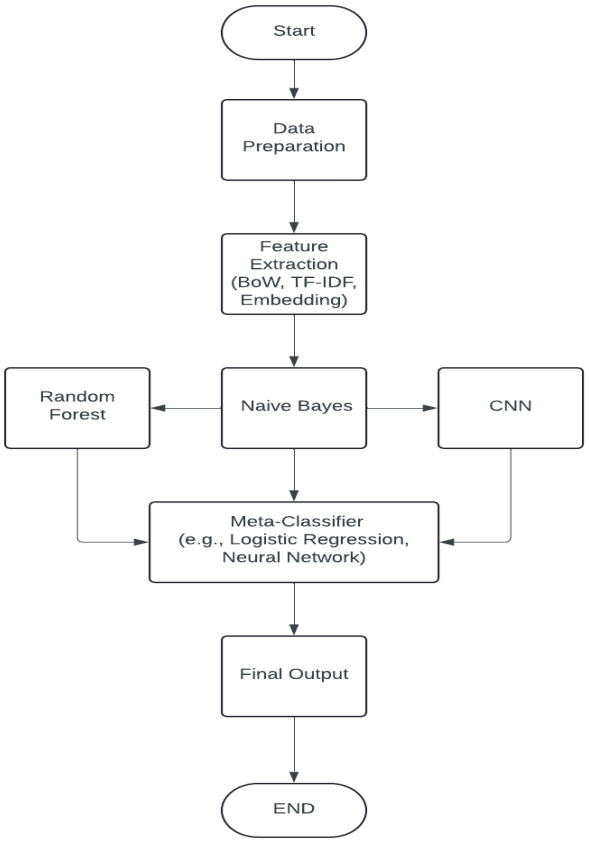
Langkah-langkah:

Validation: Menggunakan teknik cross-validation untuk memastikan tidak ada overfitting dan model bekerja dengan baik pada data validation.

Performance Metrics: Mengukur performa model hybrid menggunakan metrik seperti akurasi, presisi, recall, dan F1-score.

Comparison: Membandingkan performa model hybrid dengan model individu untuk memastikan peningkatan kinerja.

Diagram Arsitektur Hybrid Model



Penjelasan Arsitektur

- Data Preparation: Proses pembersihan dan normalisasi teks sebelum digunakan oleh model.
- Feature Extraction: Menggunakan berbagai teknik untuk mengubah teks menjadi fitur yang bisa digunakan oleh model machine learning.
- Model Building: Membangun dan melatih model individual (Random Forest,Naive Bayes, CNN).
- Meta-Classifier: Menggabungkan prediksi dari model-model individual menggunakan model tambahan untuk meningkatkan akurasi dan robustnes prediksi.
- Evaluation: Mengevaluasi kinerja model hybrid dengan berbagai metrik untuk memastikan kinerjanya.

Kesimpulan nya:

- Random Forest adalah model yang baik untuk interpretabilitas dan dapat menangani data dengan fitur yang kompleks.
- Naive Bayes adalah model yang cepat dan efisien untuk data teks dengan asumsi independensi fitur.
- CNN adalah model deep learning yang kuat untuk deteksi bullying dalam teks melalui pemrosesan fitur spasial, meskipun memerlukan lebih banyak data dan daya komputasi.

Arsitektur hybrid ini mengkombinasikan kekuatan dari berbagai model untuk mencapai kinerja yang lebih baik dalam mendeteksi bullying. Model ini memanfaatkan kekuatan interpretabilitas dari model tradisional dan kemampuan fitur spasial dari model deep learning.