BAB II

TINJAUAN PUSTAKA

BAB II
TINJAUAN PUSTAKA

II.1 Penelitian Terkait

Studi literatur dalam penelitian ini digunakan untuk mendapatkan gambaran yang menyeluruh tentang apa yang sudah dikerjakan peneliti lain dan bagaimana peneliti mengerjakannya, kemudian mengidentifikasi celah penelitian sebagai dasar penelitian yang akan dilakukan. Penggalian informasi dari literatur dilakukan sebagai upaya memperjelas permasalahan yang memiliki kaitan dengan penelitian ini, sekaligus untuk membedakan antara penelitian ini dengan penelitian sebelumnya.

Dengan adanya berita kenaikan harga bahan bakar, sebagian besar masyarakat menyampaikan sentimen terkait kenaikan harga bahan bakar di salah satu media sosial, Twitter.

Penelitian ini bertujuan untuk membedakan sentimen yang diberikan oleh masyarakat, baik positif, negatif, maupun netral, dengan menggunakan algoritma Naïve Bayes Classifier dan Decision Tree. Hasil analisis menunjukkan bahwa model Algoritma Naïve Bayes, khususnya Bernoulli Naïve Bayes, mencapai akurasi tertinggi sebesar 65,60%, dengan presisi sebesar 68%, recall sebesar 60,30%, dan f1-score sebesar 59,33%.(Muhammad Fadli et al., 2024.)

Setiap metode dibagi menjadi tiga skenario penelitian. Pada Skenario 1 dengan pembagian data 80:20, metode Decision Tree memperoleh akurasi sejumlah 74,71%, presisi sejumlah 57%, recall sejumlah 57%, dan F1- score sejumlah 57%, sedangkan Naïve Bayes memperoleh akurasi sejumlah 73,96%, presisi sejumlah 58%, recall sebesar 34%, dan F1-score sejumlah 29%. Pada Skenario 2, dengan pembagian data 70:30, metode Decision Tree mencapai akurasi sejumlah 73,27%, sedangkan Naïve Bayes mencapai akurasi sebesar 73,99%. Pada Skenario 3, dengan pembagian data 60:40, metode Decision Tree mencapai akurasi sejumlah 71,78%, sedangkan Naïve Bayes mencapai akurasi sejumlah 74,02%. Hasil evaluasi menunjukkan bahwa metode Decision Tree dengan pembagian data 80:20 memiliki akurasi yang lebih unggul dibandingkan dengan metode Naïve Bayes.(Putu Febri Armaeni et al., 2024.)

Setelah mendapatkan analisis sentimen menggunakan lexicon-based, selanjutnya kata-kata tersebut dibobot menggunakan TF-IDF lalu diklasifikasikan dan dievaluasi. Penelitian ini menggunakan algoritma naïve Bayes dan pohon keputusan. Sehingga hasil penelitian perbandingan akurasi algoritma naïve Bayes dan pohon keputusan dengan algoritma pohon keputusan memiliki tingkat akurasi paling tinggi yaitu 99%. Sehingga dapat disimpulkan bahwa menggunakan algoritma pohon keputusan lebih baik dalam mengklasifikasi data analisis sentimen komentar mahasiswa.(Ferat Kristanto et al., 2024.)

Dari hasil analisis didapatkan bahwa ketiga algoritma tersebut menghasilkan perbandingan akurasi data klasifikasi rating film dan acara TV pada Netflix dengan nilai yang berbeda-beda. Berdasarkan matriks konfusi yaitu Accuracy, Precision, dan Recall didapatkan bahwa algoritma Naive Bayes memiliki akurasi tertinggi yaitu 72%, algoritma Decision Tree sebesar 70% dan algoritma KNN memiliki akurasi terendah.

Akurasinya sebesar 61%. Dari hasil tersebut dapat dikatakan bahwa algoritma Naive Bayes dapat mengklasifikasikan data rating film dan acara TV di Netflix lebih baik dibandingkan dengan dua algoritma lainnya.(Zulkarnain et al., 2024.)

Penelitian ini berkontribusi pada pemahaman yang lebih baik tentang sentimen publik terhadap isu-isu politik di Indonesia dan menyoroti pentingnya pemilihan algoritma yang tepat dalam analisis sentimen media sosial.

Saran pengembangan meliputi eksplorasi teknik deep learning, integrasi analisis multimodal, penyeimbangan data (oversampling atau undersampling) dan peningkatan pra-pemrosesan sehingga model lebih mampu menangkap konteks negatif. Hasil penelitian menunjukkan kinerja yang sangat baik dari algoritma Naive Bayes Classifier dan Decision Tree dengan akurasi di atas 95%. Decision Tree unggul dengan akurasi 99%, sedangkan Naïve Bayes Classifier berkinerja lebih baik dengan akurasi 96%. Hasil dengan uji Confusion Matrix adalah precision 0,98, recall 1,00, dan F1-Score 0,99.(Putri Wahyuni et al., 2024.)

Penelitian ini menggunakan naive Bayes classifier dan fungsi ekstraksi TF-IDF untuk menambahkan bobot pada teks. Gunakan pustaka Python scikit-

learn untuk membantu menentukan polaritas kelas sentimen negatif dan positif dalam dataset Anda. Dataset yang digunakan adalah dataset Twitter yang diperoleh dari Oktober hingga Desember 2022, dengan total 15.000 dataset. Skenario pengujian terbaik yang diperoleh dengan pemisahan data uji dan data latih adalah 70% data uji dan 30% data latih, dengan akurasi tertinggi dihasilkan dari dataset Ganjar sebesar 95%. Dengan menggunakan data uji Anies, Ganjar, dan Prabowo, skor mood positif masing-masing kandidat adalah 833, 77, dan 524, sedangkan skor mood negatif masing- masing adalah 637, 1423, dan 976. Pengujian dilakukan dengan menggunakan matriks konfusi dan k-fold cross-validation, dan hasil terbaik diperoleh pada set data Ganjar. Yaitu matriks konfusi sebesar 94,93% dan k-fold cross-validation sebesar 94,46%. Skor f1 terendah untuk kelas positif adalah 67% untuk set data Anies dan 27% untuk kelas negatif untuk set data Ganjar.(Asno Azzawagama Firdaus et al., 2024.)

II.2 Analisis Sentimen

Analisis sentimen merupakan cabang dari *Natural Language Processing (NLP)* yang bertujuan untuk mengidentifikasi dan mengekstrak opini, sentimen, atau emosi dari teks. Analisis sentimen telah berkembang pesat dalam beberapa tahun terakhir, didorong oleh peningkatan jumlah data teks yang tersedia secara online dan kemajuan dalam teknik pemrosesan bahasa alami. Tujuan utama analisis sentimen adalah untuk memahami sikap, opini, atau perasaan yang diungkapkan dalam teks terhadap suatu entitas, peristiwa, atau topik tertentu. Hasil analisis sentimen dapat berupa klasifikasi sentimen (positif, negatif, atau netral), pengukuran intensitas sentimen, atau identifikasi aspek-aspek tertentu yang menjadi fokus sentimen.

Terdapat berbagai pendekatan dalam analisis sentimen, termasuk pendekatan berbasis leksikon, pendekatan berbasis pembelajaran mesin, dan pendekatan hibrida. Pendekatan berbasis leksikon menggunakan kamus atau leksikon yang berisi kata-kata dan frasa yang telah diberi label sentimen. Pendekatan ini relatif sederhana, namun memiliki keterbatasan dalam menangani konteks, sarkasme, dan ambiguitas bahasa. Pendekatan berbasis pembelajaran

mesin, di sisi lain, menggunakan algoritma pembelajaran mesin untuk mempelajari pola dalam data teks dan memprediksi sentimen. Pendekatan ini lebih fleksibel dan mampu menangani data yang lebih kompleks, tetapi membutuhkan data pelatihan yang cukup besar. Pendekatan hibrida menggabungkan keunggulan kedua pendekatan tersebut.

Analisis sentimen komentar YouTube, terutama dalam konteks penutupan TikTok Shop, dapat dilakukan secara efektif menggunakan metode pembelajaran mesin seperti Naïve Bayes dan Decision Trees. Metode ini dikenal luas karena penerapannya dalam tugas klasifikasi teks, termasuk analisis sentimen, yang bertujuan untuk mengklasifikasikan komentar sebagai positif, negatif, atau netral berdasarkan sentimen yang diungkapkan.(Putu Febri Armaeni et al., 2024.)

II.3 Algoritma Decision Tree

Algoritma *Decision Tree* merupakan algoritma pembelajaran mesin yang membangun model prediksi dalam bentuk pohon keputusan. Pohon keputusan terdiri dari node, cabang, dan daun. Node mewakili fitur atau atribut, cabang mewakili aturan keputusan berdasarkan fitur tersebut, dan daun mewakili hasil prediksi. Algoritma *Decision Tree* bekerja dengan secara rekursif membagi dataset berdasarkan fitur yang paling informatif, sehingga menghasilkan pohon yang memaksimalkan pemisahan kelas. Algoritma ini relatif mudah diinterpretasi karena modelnya dapat divisualisasikan dalam bentuk pohon, sehingga aturan keputusan dapat dengan mudah dipahami. Namun, *Decision Tree* rentan terhadap *overfitting*, terutama jika pohonnya terlalu dalam dan kompleks. Teknik *pruning* sering digunakan untuk mengatasi masalah *overfitting* ini.

Algoritma C4.5 merupakan algoritma yang digunakan untuk membentuk pohon keputusan (*Decision Tree*). Pohon keputusan merupakan metode klasifikasi dan prediksi yang terkenal. Pohon keputusan berguna untuk mengekspolari data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan sebuah variabel target. Banyak algoritma yang dapat dipakai dalam pembentukan pohon keputusan, antara lain : ID3, CART, dan C4.5 (Larose, 2005).

Algoritma C4.5 merupakan pengembangan dari algoritma ID3 (Larose, 2005).

Proses pada pohon keputusan adalah mengubah bentuk data (tabel) menjadi model pohon, mengubah model pohon menjadi rule, dan menyederhanakan rule (Basuki & Syarif, 2003).

Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut :

1. Pilih atribut sebagai akar.
2. Buat cabang untuk tiap-tiap nilai.
3. Bagi kasus dalam cabang.
4. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Untuk memilih atribut sebagai akar, didasarkan pada nilai gain tertinggi dari atribut-atribut yang ada. Untuk menghitung *gain* digunakan Persamaan : …… II.1

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy\,(S_i) \quad \text{..... (II.1)}$$

Keterangan :

S : himpunan kasus

A : atribut

n : jumlah partisi atribut A

$|S_i|$: jumlah kasus pada partisi ke-i

|S| : jumlah kasus dalam S

Untuk perhitungan nilai entropi sbb :

$$Entropy(S) = \sum_{i=1}^{n} - pi * log_2\, pi \quad \text{.....(II.2)}$$

Keterangan :

S : himpunan kasus.

A : fitur.

n : jumlah partisi S.

pi : proporsi dari S_i terhadap S

II.4 Algoritma Naive Bayes

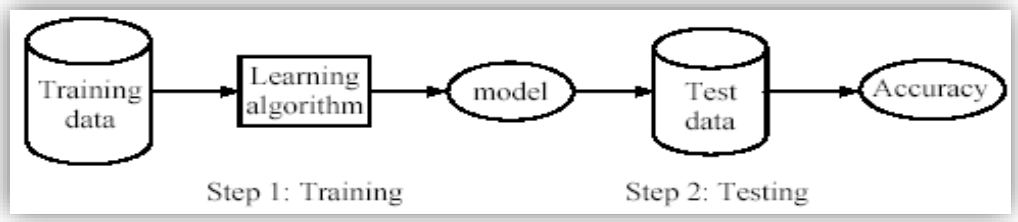
Algoritma *Naive Bayes* merupakan algoritma klasifikasi probabilistik yang didasarkan pada teorema Bayes. Algoritma ini mengasumsikan bahwa fitur-fitur

dalam dataset saling independen secara bersyarat, meskipun asumsi ini seringkali tidak realistis dalam praktiknya. *Naive Bayes* menghitung probabilitas posterior dari setiap kelas dengan menggunakan probabilitas prior dan probabilitas bersyarat dari fitur-fitur yang diberikan. Algoritma ini dikenal karena kecepatan dan efisiensinya, terutama pada dataset yang besar. Meskipun asumsi independensi fitur yang sederhana, *Naive Bayes* seringkali memberikan hasil yang akurat dalam berbagai aplikasi klasifikasi, termasuk analisis sentimen. Berikut ini dua tahapan klasipikasi

- *Learning* (*training*):

  Pembelajaran menggunakan data training (untuk *Naïve Bayesian Classifier*, nilai probabilitas dihitung dalam proses pembelajaran)
- Testing:

  Menguji model menggunakan data testing



II. 1 Gambar Cara Kerja Algoritma

$$\text{Akurasi} = \frac{\text{Jumlah Prediksi Yang Benar}}{\text{Jumlah Prediksi Keseluruhan}} \quad \text{(II.3)}$$

$$\text{Erro} = \frac{\text{Jumlah Prediksi Yang Salah}}{\text{Jumlah Prediksi Keseluruhan}} \quad \text{(II.4)}$$

Dasar Teorema Bayes

1. Diketahui X merupakan sample data.
   a. Dalam bayes X disebut “evidence” atau fakta
   b. Label class tidak diketahui
   c. Umumnya X merupakan record data yang disusun dari n atribut
2. H merupakan suatu hypothesis bahwa X termasuk dari kelas C

3. Classification adalah untuk menentukan P(H|X), peluang hipotesis dari data sample X
    a. Dengan kata lain, dicari peluang bahwa record X termasuk kelas C, dengan diketahui atribut yang menjelaskan X.
    b. Atau, peluang keluarnya hasil H jika diketahui nilai X tertentu.

II.5 Manajemen Krisis Reputasi

Manajemen krisis reputasi merupakan proses strategis yang bertujuan untuk meminimalkan dampak negatif dari krisis terhadap citra dan reputasi suatu organisasi. Krisis reputasi dapat disebabkan oleh berbagai faktor, termasuk skandal, kecelakaan, atau kegagalan produk. Respon yang efektif terhadap krisis reputasi sangat penting untuk mempertahankan kepercayaan publik dan menjaga keberlanjutan bisnis. Strategi manajemen krisis reputasi yang efektif melibatkan identifikasi cepat krisis, pengembangan rencana komunikasi krisis, dan implementasi strategi komunikasi yang transparan dan responsif. Analisis sentimen memainkan peran penting dalam proses ini dengan memberikan wawasan tentang persepsi publik terhadap krisis dan efektivitas strategi komunikasi yang diterapkan. Perusahaan perlu memantau sentimen publik secara real-time untuk mengidentifikasi potensi krisis dan merespon dengan cepat dan efektif.

II.6 Perbandingan Decision Tree dan Naive Bayes dalam Analisis Sentimen

Baik *Decision Tree* maupun *Naive Bayes* telah digunakan secara luas dalam analisis sentimen. *Decision Tree* menawarkan interpretasi yang mudah dipahami, tetapi dapat rentan terhadap *overfitting*. *Naive Bayes*, di sisi lain, lebih efisien komputasi dan seringkali memberikan hasil yang akurat, tetapi asumsi independensi fitur dapat membatasi keakuratannya dalam beberapa kasus. Pilihan antara kedua algoritma ini bergantung pada karakteristik dataset dan kebutuhan analisis. Dalam konteks penelitian ini, kedua algoritma akan dibandingkan untuk menentukan algoritma mana yang lebih efektif dalam menganalisis sentimen masyarakat terhadap pejabat Pertamina pasca kasus Pertamax Oplosan.

II.7 Data Mining

Data Mining merupakan suatu proses analisis untuk mencari pengetahuan dalam suatu basis data. Pengetahuan tersebut dapat diartikan sebagai pola data atau hubungan antar data valid yang belum diketahui sebelumnya .

Analisis sentimen biasanya dilakukan untuk mencari opini publik atau pelanggan terhadap suatu produk atau layanan yang dimiliki oleh suatu perusahaan, organisasi, atau entitas. Analisis sentimen juga dapat diartikan sebagai mempelajari suatu opini, masalah, perasaan, atau emosi dari seseorang atau publik dalam menanggapi sesuatu yang berupa teks atau tulisan. Dalam menentukan suatu sentimen, dilakukan dengan cara menghitung beberapa kata yang terdapat dalam kalimat, dokumen, atau teks.(Nico et al. 2021).

II.8 Mesin Learning Orange

Orange adalah perangkat lunak *open-source* yang digunakan untuk analisis data dan pembelajaran mesin. Orange memiliki antarmuka berbasis visual yang memudahkan pengguna dalam melakukan eksplorasi data, pelatihan model, dan evaluasi performa model tanpa perlu menulis banyak kode pemrograman. Beberapa fitur utama dari Orange adalah:

1. Visualisasi Data: Menyediakan berbagai metode visualisasi seperti diagram batang, scatter plot, dan heatmap.
2. Analisis Data**:** Dapat digunakan untuk pembersihan data, reduksi dimensi, dan transformasi data.
3. Pembelajaran Mesin: Mendukung berbagai algoritma seperti Naïve Bayes, Decision Tree, dan Neural Networks.
4. Evaluasi Model: Memiliki alat untuk mengevaluasi model menggunakan metrik seperti akurasi, presisi, dan recall.

II.9 Web Scraping

Pengumpulan data dilakukan menggunakan metode web crawling atau web scraping, merupakan teknik otomatisasi untuk mengekstraksi data dari situs web (Moaiad Ahmad Khder, 2021), teknik ini menjadi satu-satunya solusi untuk memperoleh data secara langsung tanpa melalui komunikasi antarmuka

pemrograman aplikasi (A O Savelev, et al, 2021). Data diambil secara otomatis melalui Application Programming Interface (API) yang bersumber dari media sosial media.

Tabel Penelian Terdahulu

Tabel I. 1 Penelitian Terdahulu

No	Judul Penelitian	Nama Penulis	Preprocessing Data			Data set	Klasifikasi	Hasil
			Pembersihan	Pelebelan	Pembobotan		Algoritma	
1	Analisis Sentimen Komentar YouTube tentang Penutupan Toko TikTok Menggunakan Perbandingan Metode Naïve	Putu Pebri Armaeni et al., 2024	case folding, Tokenisasi, Stopwords, Stemming	kamus leksikon kata-kata positif dan negatif dalam format	TF-IDF (Term Frequency Inverse Document Frequency)	Crawling data YouTube Menggunakan 7000 data	Naïve Bayes dan Decision Tree	*akurasi sebesar 74,71%, presisi sebesar 57%, recall sebesar 57%, dan F1-score sebesar 57%, sedangkan Naïve Bayes memperoleh akurasi sebesar 73,96%, presisi*

	Bayes dan Decision Tree							*sebesar 58%, recall sebesar 34%, dan F1-score sebesar 29%*
2	Analisis Sentimen tentang Pengacara Indonesia Program Televisi Klub Menggunakan K-Nearest Tetangga, Pengklasifikas	Nico Nathanael Wilim et al., 2021	menghilangkan data duplikat, URL dan simbol, angka, dan tanda baca yang tidak diperlukan dari teks sepert, Case Folding, Tokenisasi, Stopwords,	Tidak dituliskan	TF-IDF (Term Frequency Inverse Document Frequency)	Pengumpulan data menggunakan Python untuk mengambil data di Twitter tahun 2018 dan 2019,sebanyak 30 tweet per bulan masing	K-Nearest Tetangga Naïve Bayes dan Decision Tree	Model berhasil mendapatkan akurasi 66 %

	i Naïve Bayes, dan Pohon Keputusan		Stemming dan filtering.			- masing		
3	Perbandingan Algoritma Naïve Bayes dan Pohon Keputusan Tentang Analisis Sentimen Data Komentar Siswa Di Era Digital Aplikasi Penilai	Ferat Kristanto et al., 2024	proses eliminasi noise agar proses analisis sentimen menjadi lebih akurat	klasifikasi data menggunakan leksikon.	TF-IDF (Term Frequency Inverse Document Frequency)	pengumpulan data mahasiswa dan komentar pada aplikasi DITA.	Naïve Bayes dan Decision Tree	tingkat akurasi sebesar 84,08 %, metode K-NN sebesar 83,38 % dan pohon keputusan sebesar 81,09 %. Hasil penelitian ini

	an Guru (DITA)							dapat menunjukkan bahwa metode naïve bayes mempunyai tingkat akurasi yang lebih tinggi dibandingkan metode lain yang digunakan dengan tingkat akurasi 84,08 %.

4	Analisis Sentimen pada Aplikasi Media Sosial Twitter di Kenaikan Harga Bahan Bakar Minyak Menggunakan Naïve Bayes dan Decision Tree Algoritma	Muhammad Fadli et al., 2024	data cleaning, case folding, tokenizing, stopword removal/filtering, dan data stemming.	Sentimen polariti	Analisis Sentimen Tampilan Barchart	Crawling data Twitter 500 data	Naïve Bayes dan Decision Tree	akurasi tertinggi sebesar 65,60 %, dengan presisi sebesar 68%, recall sebesar 60,30 %, dan f1-score sebesar 59,33 %
5	Analisis Sentimen Komen	Ziyao Zhang	pemrosesan data sampah yang disebutka			data komentar Twitter di	Naïve Bayes	Akurasi model naive Bayes

	tar Twitter Menggunakan Naive Bayes		n di atas dan menggabungkannya dengan penandaan part-of-speech di perpustakaan NLP untuk menyelesaikan normalisasi bagian dari data			Kaggle 25000 data twitt		0,06 poin lebih tinggi daripada pelatihan logika dalam pengaturan pemrosesan yang identik, dan tingkat ingatan 0,05 poin lebih tinggi.

6	Analisis Perbandingan Metode Penambangan Data untuk Menganalisis Pinjaman Pribadi menggunakan Pohon Keputusan dan Klasifikasi Naïve Bayes	Menuka Maharjan	Tidah disebutkan dalam jurnal	CART. Kategori	Indeks Gini diukur untuk menemukan ketidakmurnian D, partisi data atau kumpulan tupel pelatihan, sebagai Gini(D)	Dataset pinjaman Jerman terdiri dari 1000	Naïve Bayes dan Decision Tree	Model berhasil mendapatkan akurasi 95 % pada kategori ke 8

7	Analisis Sentimen Diskusi Twitter tentang Rafael Alun: Multinomial Naïve Bayes dan Pendekatan Pohon Keputusan	Iqbal Sabilirrasyad et al., 2023	Case Transform, Punctuation Cleansing, Mengulangi teks Pembersihan, Pembersihan URL, Data Angka Pembersihan, Penerapan Stemming dan Lemmatizing (Kamus Indonesia), Tokenizing, Menghapus Stopword	Pelabelan manual	Data sudah diberi label	Crawling data Twitter berbasis Python 1021 data	Naïve Bayes dan Decision Tree	Model naïve bayes akurasi 77 % dan decision tree akurasi 72 %

8	Perbandingan Algoritma Multinomial Naive Bayes dan Pohon Keputusan dengan Aplikasi AdaBoost dalam Sentimen Ulasan	Cecep Muhammad Sidik Ramdani et al., 2022	pembersihan data, pelipatan huruf, tokenisasi, pemeriksaan ejaan, penghilangan huruf berhenti, stemming dan penerjemahan.	library vaderSentiment dengan ketentuan jika data memiliki skor > 0 maka labelnya positif dan jika skornya < 0 maka labelnya	CountVectorizer dan TF-IDF. CountVectorizer merupakan fitur untuk mengubah teks menjadi representasi vektor dan juga untuk menghitung	data yang digunakan berasal dari review pengguna di Google Play Store pada aplikasi PeduliLindungi 8305 data.	Algoritma Multinomial Naive Bayes dan Pohon Keputusan dengan Aplikasi AdaBoost	Algoritma Multinomial Naïve Bayes sesudah implementasi AdaBoost memiliki nilai akurasi rata-rata sebesar

	Analisis Aplikasi PeduliLindungi			negatif	jumlah kata dalam sebuah dokumen.			88,8% dan nilai akurasi tertinggi sebesar 90,1%. Untuk hasil kinerja algoritma Decision Tree sesudah implementasi AdaBoost memiliki nilai akurasi rata-

								rata sebesar 84,1% dan nilai akurasi tertinggi sebesar 86,15 %.
9	Perbandingan algoritma decision tree dan naive bayes pada analisis sentimen masyarakat terhadap	Mubarak, 2025	menghilangkan data duplikat, URL dan simbol, angka, dan tanda baca yang tidak diperlukan dari teks sepert, Case Folding,	Sentimen analisis	TF-IDF (Term Frequency Inverse Document Frequency)	Crawling data Twitter berbasis Python 3945	Naïve Bayes dan Decision Tree	

	pejabat pertamina pasca kasus pertamax oplosan		Tokenisasi, Stopwords, Stemming dan filtering.					