LAPORAN AKHIR PROJECT TEXT PROCESSING

"Analisis dan Klasifikasi Berita Clickbait: Pendekatan Random Forest, Similaritas Judul-Konten, dan Pengelompokan Topik Hasil Klasifikasi"

Dosen pengampu : Riskyana Dewi Intan P, M.Kom.

Disusun oleh :

Muhammad Faiz Munif Billah	NIM. 23031554028
Ibrahim Frosly Alesandro	NIM. 23031554021
Gesang Nur Zamroji	NIM. 23031554145

PROGRAM STUDI SAINS DATA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

DAFTAR ISI

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Seiring dengan meningkatnya penggunaan internet di masyarakat, berita online semakin menjadi pilihan utama dalam mengakses informasi. Berita online yang mudah diakses dan disebarluaskan melalui berbagai perangkat menawarkan kecepatan dan kemudahan bagi pembaca. Namun, kemudahan ini juga dimanfaatkan oleh beberapa produsen berita untuk menarik perhatian dengan cara yang tidak etis, seperti memproduksi berita clickbait. Berita clickbait, yang biasanya ditandai dengan judul sensasional yang tidak relevan atau tidak sesuai dengan isi, bertujuan untuk meningkatkan jumlah pembaca atau klik semata. Fenomena ini berdampak signifikan terhadap opini publik, kredibilitas media, serta kepercayaan masyarakat terhadap sumber informasi.

Secara sosial, berita clickbait dapat menyesatkan pembaca dan membentuk persepsi yang salah tentang suatu isu, sehingga berpotensi memperkeruh suasana masyarakat. Dalam konteks politik, berita semacam ini sering kali digunakan untuk menyebarkan propaganda atau memanipulasi opini pemilih, yang pada akhirnya mempengaruhi stabilitas demokrasi. Dari segi ekonomi, berita clickbait dapat mengalihkan perhatian pembaca dari sumber informasi berkualitas tinggi ke konten yang kurang bermutu, menurunkan standar jurnalistik, dan menyebabkan hilangnya kepercayaan pada platform berita yang kredibel.

Untuk mengatasi tantangan ini, berbagai solusi berbasis teknologi telah mulai diterapkan. Salah satunya adalah penggunaan model random forest dan text similarity untuk mengklasifikasikan berita clickbait dan non clickbait secara otomatis. Model dapat mempelajari pola-pola tertentu dalam struktur judul berita, penggunaan kata kunci, serta elemen emosional yang sering ditemui pada berita clickbait. Sistem ini memungkinkan identifikasi cepat dan akurat, sehingga membantu platform berita maupun pembaca untuk menghindari berita yang menyesatkan. Selain itu, sistem verifikasi berita juga dapat diterapkan untuk mengevaluasi kredibilitas dan validitas informasi sebelum disebarluaskan.

Selain solusi teknologi, peningkatan literasi digital di kalangan masyarakat juga menjadi langkah penting. Dengan literasi digital yang baik, masyarakat dapat lebih kritis dalam mengevaluasi informasi yang mereka konsumsi dan menghindari jebakan berita clickbait. Kombinasi antara solusi teknologi dan edukasi publik ini diharapkan dapat menciptakan ekosistem informasi yang lebih sehat dan terpercaya di era digital.

1.2 Tujuan

1. Klasifikasi berita clickbait dan non clickbait
2. Membandingkan random forest classifier dan teknik similarity untuk kasus klasifikasi
3. Clustering untuk menentukan topik pada berita clickbait dan non clickbait

1.3 Manfaat

1. Memberikan metode yang efektif untuk membedakan berita *clickbait* dan *non clickbait* sehingga membantu pembaca untuk mengakses informasi yang lebih relevan dan kredibel.
2. Memberikan wawasan tambahan terkait kinerja model klasifikasi Random Forest dan Similarity, serta menambah cakupan dengan melakukan clustering untuk tahu topik dominan berita pada kategori *clickbait* atau *non clickbait.*

BAB 2

LANDASAN TEORI

2.1 Exploratory Data Analysis

EDA (Exploratory Data Analysis) adalah langkah awal dalam investigasi data yang bertujuan untuk memahami dan merangkum karakteristik utama dataset. Dalam analisis teks, proses ini mencakup pemeriksaan terhadap isi, struktur, dan kualitas data teks yang akan dianalisis. EDA memungkinkan pengguna untuk memperoleh gambaran yang jelas tentang data, yang kemudian digunakan sebagai dasar untuk pengembangan model machine learning atau analisis statistik (Khasanah, 2020).

Melalui EDA, analis dapat menghemat waktu sekaligus mengurangi risiko kesalahan selama proses analisis. Dengan memahami dataset secara menyeluruh, mereka dapat menentukan langkah berikutnya dengan lebih tepat dan strategis. Selain itu, EDA membantu memastikan bahwa data yang digunakan relevan dan valid sesuai dengan tujuan yang ingin dicapai (*Exploratory-Data Analysis*, 2022).

Secara keseluruhan, EDA merupakan tahap penting dalam analisis teks yang memberikan wawasan awal yang berharga, sehingga mempersiapkan peneliti atau praktisi untuk melanjutkan ke analisis yang lebih mendalam dan kompleks.

2.2 Pre-Processing

Text preprocessing digunakan untuk menyiapkan teks sebelum menggunakannya dalam pengujian atau pelatihan untuk tujuan mengurangi noise dalam data (Indraloka & Santosa, 2017). Pre-processing yang dilakukan yaitu dengan:

1. Case folding mengonversi semua kalimat yang memiliki huruf menjadi huruf kecil dan menghapus tidak valid pada karakter, termasuk angka, simbol, tanda baca, dan URL (Uniform Resource Locators).
2. Tokenizing melibatkan pemotongan kalimat menurut kata-kata yang menyusunnya. Tokenisasi memberikan gambaran proses pembagian teks menjadi kata-kata dengan menggunakan spasi sebagai pembatas dengan tujuan agar berdiri sendiri untuk setiap kata tanpa adanya hubungan dengan kata lain.

3. Filtering disebut juga dengan menghilangkan stopwords, yaitu proses kata-kata yang dihilangkan dan dianggap tidak relevan atau tidak menggambarkan makna isi kalimat. Dalam sebuah kalimat seringkali terdapat makna yang tidak lagi memiliki kaitan pada, seperti “ini”, “itu”, dll. Oleh karena itu, kata kata tersebut dihapus tidak berdampak besari.
4. Stemming melibatkan konversi kata-kata dengan imbuhan yang berbeda ke kata dasarnya, langkah ini sering dilakukan untuk teks berbahasa Inggris, karena struktur afiks pada bahasa inggris cenderung stabil. Pada penelitian ini stemming yang dilakukan menggunakan modul sastrawi untuk bahasa Indonesia. Kata yang diproses menjadi bentuk dasarnya dengan menghapus imbuhan yang melekat pada kata tersebut, seperti “in-”, “-nya” dan lain-lain..

2.3 TF-IDF (Term Frequency-Inverse Document Frequency)

TF (Term Frequency) adalah frekuensi kemunculan suatu kata dalam setiap dokumen. Dari TF kita mendapatkan DF (Document Frequency), yaitu jumlah kata yang terkandung dalam dokumen tersebut. TF-IDF adalah nilai untuk menghitung dokumen dengan bobot kata yang telah ditemukan. TF-IDF diperoleh dengan mengalikan TF dan IDF, dimana IDF adalah kebalikan dari DF (Sammut & Webb, 2011).Perhitungannya dapat ditulis sebagai berikut:

$$\mathrm{TF}\left(t,\ \ d\right) = \frac{(\text{Number of occurrences of term } t \text{ in document } d)}{(\text{Total number of terms in the document } d)}$$

$$\mathrm{IDF}\left(t\right) = \log_e \frac{(\text{Total number of documents in the corpus})}{(\text{Number of documents with term } t \text{ in them })}$$

$$tfidf_{t,d} = tf_{t,d} \times idf_d$$

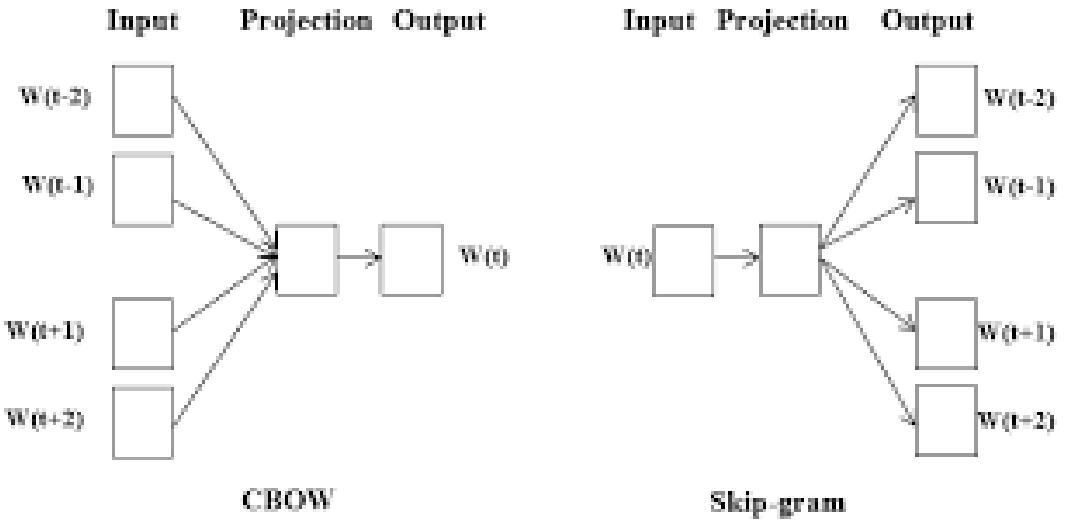
2.4 Word2Vec

Word2Vec adalah model *shallow neural network* yang merubah representasi kata yang merupakan kombinasi dari karakter *alphanumeric* menjadi *vector* (Girsang, 2020). Dalam proses training, representasi vector memiliki hubungan satu sama lain. Ada dua metode untuk mempelajari representasi kata::

1. Continuous bag-of-words model: menentukan kata tengah berdasarkan kata-kata konteks di sekitarnya. Konteks terdiri dari beberapa kata sebelum dan sesudah kata

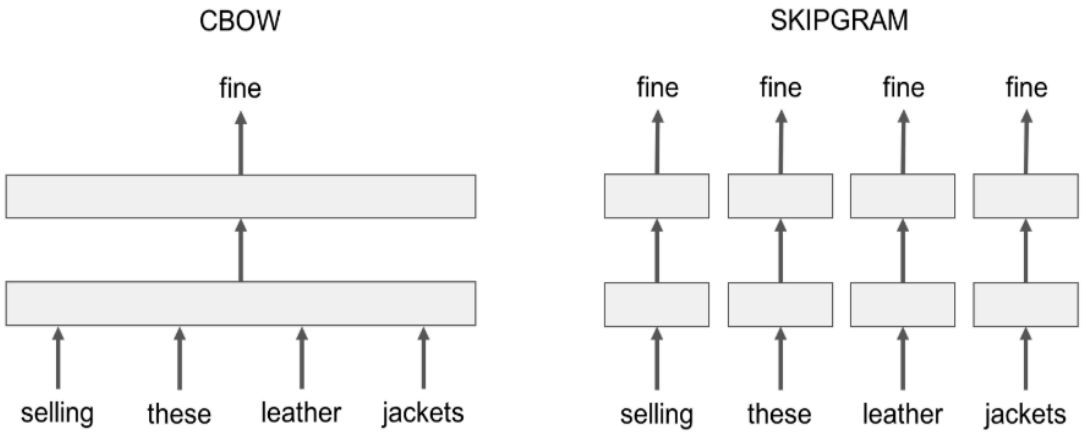
saat ini (tengah). Arsitektur ini disebut model kantong kata karena urutan kata dalam konteks tidak penting.

2. Continuous skip-gram model: menentukan kata pada suatu rentang yang ditentukan sebelum dan sesudah kata saat ini dalam kalimat yang sama.



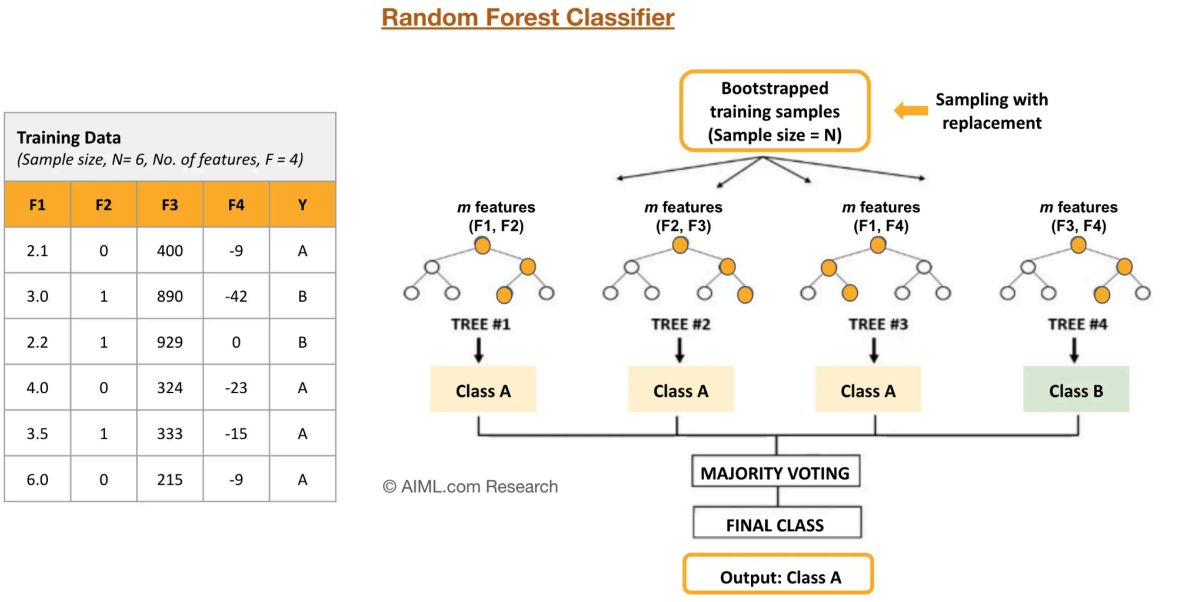
2.5 FasText

Berbeda dari Word2Vec, FastText tidak memakai hanya satu kata secara utuh untuk diproses, tapi FastText menggunakan n-gram. Contoh implementasi n-gram pada kata “pintar” dengan trigram (n=3) berupa “pin”, “int”, “nta”, “tar”. Kelebihan FastText adalah waktu proses yang relatif cepat. Berbeda dengan Word2Vec, FastText dapat menangani kata yang tidak pernah muncul di kamus (vocabulary), yang mana dalam Word2Vec hal seperti ini akan menghasilkan error (Girsang, 2021). FastText menyediakan pilihan untuk menggunakan salah satu dari dua algoritma utama FastText, yakni Continous Bag of Words (CBOW) dan Skip-gram.



2.6 Random Forest Classifier

Random Forest adalah algoritma *ensemble learning* yang terdiri dari banyak *decision trees* untuk meningkatkan akurasi prediksi (Breiman, 2001). Dengan menggunakan metode bagging, random forest akan menggambar beberapa kumpulan sampel pelatihan yang berbeda satu sama lain. Setiap kumpulan sampel membuat *Decision Tree* dengan atribut yang dipilih secara acak. Metode ini ditandai unggul dengan kemampuannya yang baik untuk menahan kinerja yang berat dalam kemampuan klasifikasi.



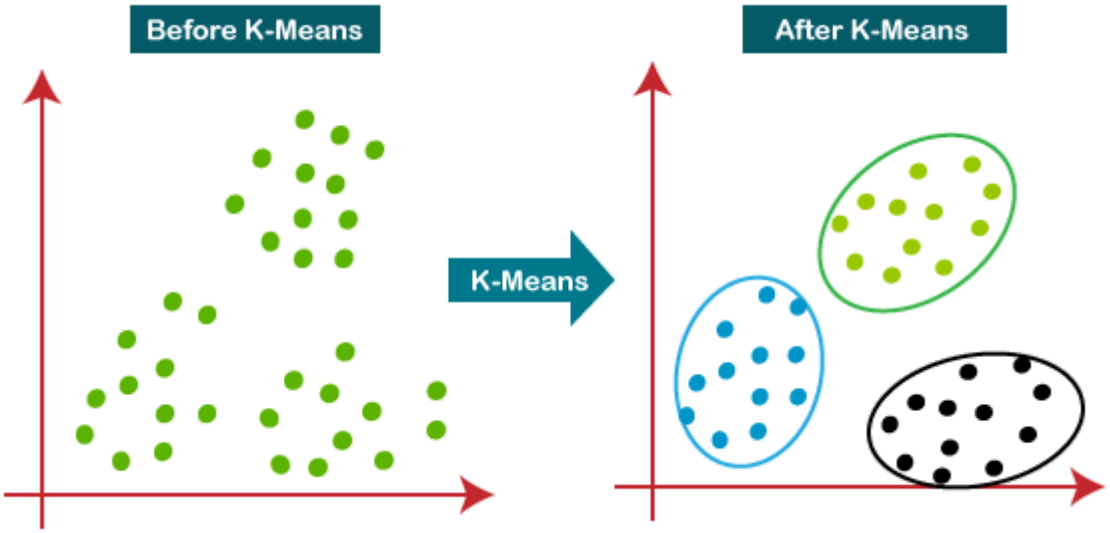
(*Explain the Copcept and Working of Random Forest Model*, 2023)

2.7 K-Means

K-Means clustering bekerja dengan membagi data ke dalam k *cluster* berdasarkan kesamaannya, kemudian setiap data akan dimasukkan ke dalam *cluster* dengan centroid terdekat (Lloyd, 1982). Cara kerja K-Means dibagi menjadi beberapa langkah antara lain:

1. Menentukan Jumlah Kluster
2. Inisialisasi Centroid
3. Data point ditugaskan ke cluster terdekat
4. Memperbarui Centroid
5. Iterasi pint 3 dan 4

Tujuan untuk mengecilkan jarak total dari data point dan centroid biasanya disebut inertia. Semakin kecil nilai inertia maka semakin bagus model dalam clustering, yang menandakan bahwa data point dalam satu cluster lebih mirip satu sama lain.



2.8 Data Scaling dan PCA

Data Scaling adalah teknik untuk mengubah skala data agar semua fitur berada dalam rentang yang sama agar distribusi data serupa. *StandardScaler* untuk menyamakan skala data dengan mengubah nilai fitur agar memiliki *mean* 0 dan standar deviasi 1 (Han et al., 2012). Sedangkan, *PCA* adalah teknik mereduksi dimensi data dengan mengurangi jumlah fitur tanpa kehilangan banyak informasi penting, dengan mengubah fitur asli ke dalam bentuk *principal components*. Penerapan *Scaling* dan *PCA* dapat mengurangi noise dan mempercepat proses komputasi.

2.9 Similarity Text

Similarity Text memiliki peran yang semakin penting dalam penelitian dan aplikasi terkait teks seperti pengambilan informasi dan klasifikasi teks termasuk clickbait atau non clickbait.Kata-kata bisa serupa dalam dua cara secara leksikal dan semantik. Kata-kata serupa secara leksikal jika memiliki urutan karakter yang serupa. Kata-kata serupa secara semantik jika memiliki hal yang sama, berlawanan satu sama lain, digunakan dengan cara yang sama, digunakan dalam konteks yang sama dan yang satu adalah jenis yang lain.

2.9.1 Cosine Similarity

Cosine Similarity adalah ukuran untuk menentukan tingkat kesamaan antara dua vektor dengan mengukur kosinus sudut antara dua vektor tanpa terpengaruh ukuran vectornya. Semakin kecil sudutnya, semakin mirip kedua vektor. dan semakin besar nilai cosin-nya .

Cosine Similarity juga bisa dikatakan metrik yang diimplementasikan secara luas dalam pencarian informasi dan studi terkait. Metrik ini memodelkan dokumen teks sebagai vektor. Similarity antara dua dokumen/dua item dapat diperoleh dengan menghitung nilai kosinus antara vektor dari dua dokumen/dua item tersebut.

$$cos(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$$

$$cos(\mathbf{u}, \mathbf{v}) = \frac{\sum_{i=1}^{V} u_i v_i}{\sqrt{\sum_{i=1}^{V} u_i^2}\sqrt{\sum_{i=1}^{V} v_i^2}}$$

2.9.1 Jaccard Similarity

Jaccard Similarity adalah metode yang digunakan untuk menganalisis kesamaan antara dua sampel data/dokumen. Untuk dua himpunan hingga A dan B, jaccard sendiri juga bisa dibilang *intersection over union* dimana skor kesamaan jaccard berkisar dari 0 hingga 1 dimana 1 mewakili yang paling relevan dan 0 mewakili yang kurang adanya kemiripan antara data/dokumen.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \qquad J(A,B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

2.10 Silhouette Score

Skor Silhouette adalah metrik yang digunakan untuk mengevaluasi kualitas clustering dalam analisis data. Metrik ini mengukur seberapa mirip objek dengan klusternya sendiri dibandingkan dengan cluster lain, memberikan nilai antara -1 dan 1 (Shitao et al., 2018).

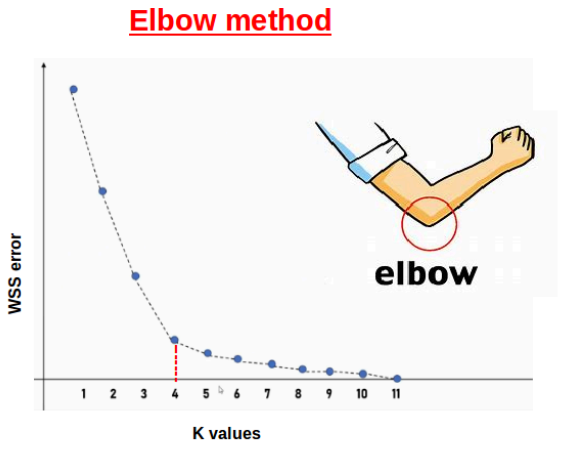
$$s = \frac{b - a}{\max(a, b)}$$

Cohesion (a): Mean jarak antara data dengan data lain dalam kluster yang sama.
Separation (b): Mean jarak antara data dengan data lain di kluster terdekat lainnya.
Hasil dari s dari -1 hingga 1:

- Nilai mendekati 1: Data dikelompokkan dengan baik/sesuai
- Nilai mendekati 0: Data berada di antara dua kluster atau kluster tersebut mungkin saling tumpang tindih.
- Nilai negatif: Data mungkin dikelompokkan di kluster yang salah (elfanmauludi, 2023).

2.11 Elbow Method

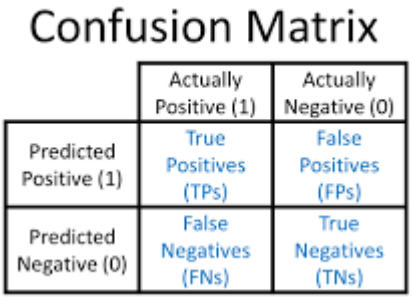
Metode siku adalah teknik untuk menentukan jumlah cluster optimal salah satunya dalam pengelompokan k-mean. Ini mengidentifikasi titik di mana jumlah jarak kuadrat antara titik dan sentroid cluster menunjukkan penurunan minimal, sering dinilai secara visual (Matsuga dan Sheremet, 2023).



Jika dilihat dari grafik, K dapat ditentukan melalui titik yang berbentuk seperti siku.

2.12 Evaluation Matrix

2.12.1 Confusion Matrix



TP (True Positive): Model benar mendeteksi sebagai "positif".

TN (True Negative): Model benar mendeteksi sebagai “negatif”.

FP (False Positive): Model mengklasifiksikan sebagai positif yang seharusnya negatif. FP juga dikenal dengan *Type I error*.

FN (False Negative): Model mengklasifiksikan sebagai positif yang seharusnya negatif. FN juga dikenal dengan *Type II error* (*Confusion Matrix*, n.d.).

2.12.2 Accuracy

Merepresentasikan rasio data yang diklasifikasikan benar (TP+TN) terhadap jumlah total data (TP+TN+FP+FN).

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)}$$

2.12.3 Precision

Merepresentaasikan rasio data yang diklasifikasikan benar (*TP*) terhadap jumlah total data yang diprediksi (*TP* + *FP*).

$$Precision = \frac{TP}{TP + FP}$$

2.12.4 Recall

Didefinisikan sebagai rasio data yang diklasifikasikan dengan benar *(TP)* dibagi dengan jumlah total data yang benar diprediksi *(TP + FN)*

$$\text{Re}call = \frac{TP}{TP + FN}$$

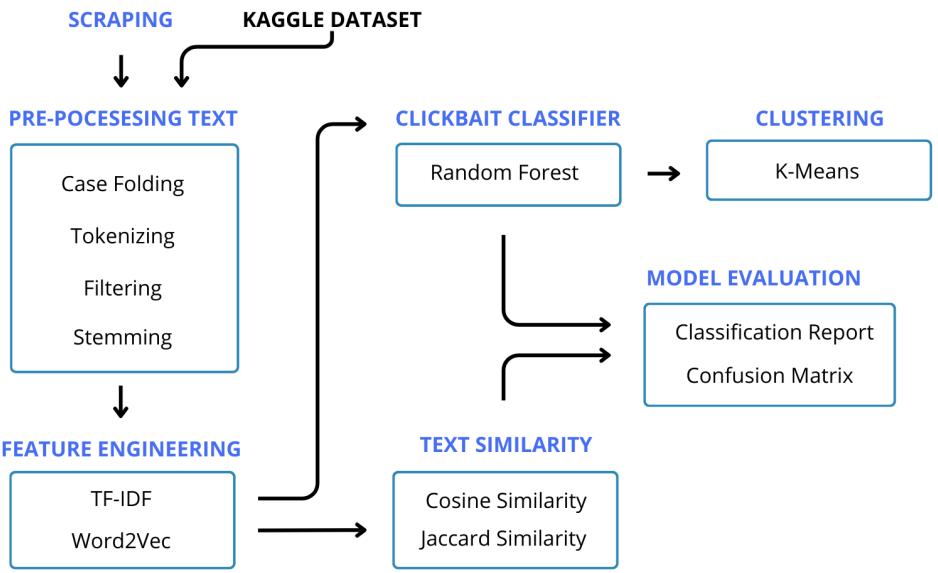
2.12.5 F1-Score

Menyatakan keseimbangan antara presisi dan recall

$$F1Score = \frac{2 * precision * recall}{precision + recall}$$

BAB 3

METODOLOGI PENELITIAN



3.1 Scraping

Project ini menggunakan tiga website berita yang nantinya akan dijadikan menjadi satu file.csv untuk dilakukan tahap ke selanjutnya. Kami menggunakan library pada umumnya yaitu request dan BeautifulSoup untuk melakukan akses dan mengambil struktur HTML website berita tersebut.

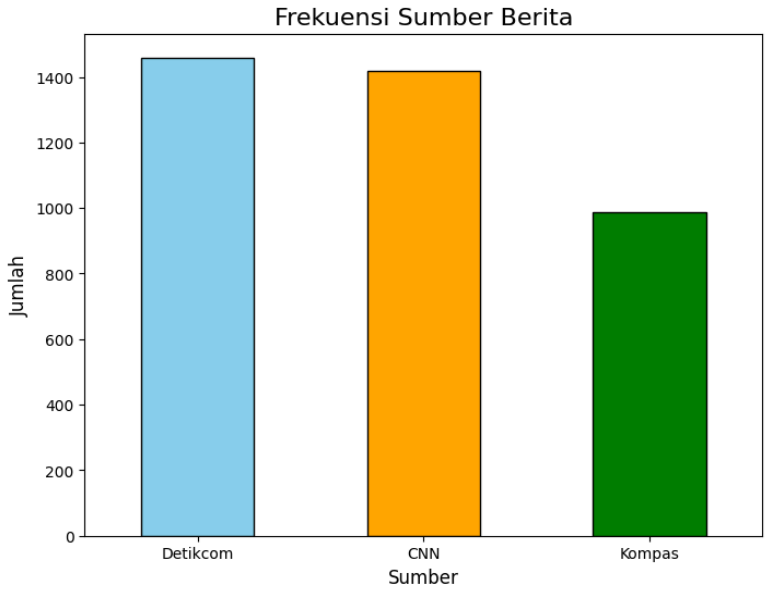
- Detik.com

  Kami mengambil data dari detik.com sebanyak 1459 data dengan beberapa informasi didalamnya.
- CNN Indonesia

  Kami mengambil data dari website CNN Indonesia sebanyak 1461 data yang akan dilakukan pemrosesan ke tahap selanjutnya.
- Kompas.com

  Kami melakukan scraping dari website Kompas.com sebanyak 987 data untuk dilakukan pemrosesan lebih lanjut.

Dari ketiga website diatas kami melakukan penggabungan data hasil scraping untuk dijadikan data testing ke dalam model yang kami gunakan nantinya.



3.2 Exploratory Data Analysis

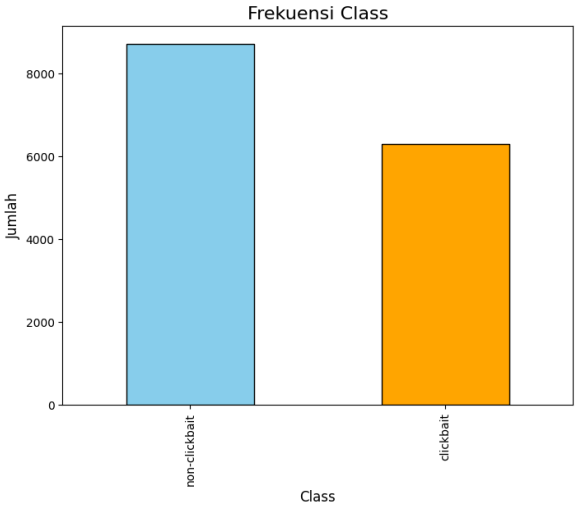
3.2.1 Data Train Kaggle

Pada data ini kami melakukan Exploratory Data Analysis (EDA) bertujuan untuk mengidentifikasi pola, menemukan anomali, dan memeriksa asumsi.

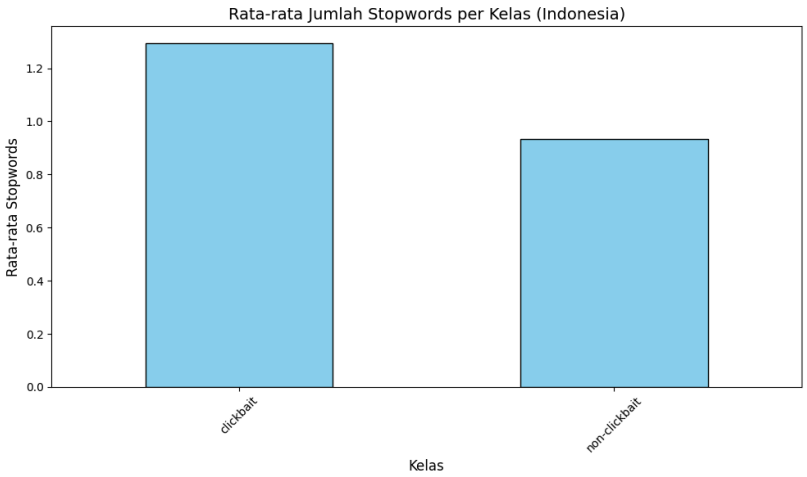
a. Judul yang memiliki nilai null sehingga kami menghapusnya

	0
title	1
label	0
label_score	0

b. Perbandingan banyaknya Clickbait dan Non-Clickbait



c. Perbandingan jumlah banyaknya Clickbait dan Non-Clickbait berdasarkan rata-rata stopwords per judul



3.2.2 Data Scraping

a. Mengecek data null/Nan

Ditemukan beberapa data null pada kolom title dan konten sebagai berikut

```
data_srap.isna().sum()
```

	0
date	0
title	2
content	44
sumber	0

Setelah kami cek ternyata data null tersebut memang pada judulnya menyebutkan “FOTO” yang hanya menampilkan FOTO.Kami memutuskan untuk menghapusnya saja

FOTO: Ngeri Rusuh Peru sampai Polisi Dibakar H...	NaN	CNN
FOTO: Kedai Kuno Berusia Nyaris 5.000 Tahun Di...	NaN	CNN
FOTO: Israel Gunakan Robot untuk Percepat Bela...	NaN	CNN
FOTO: Robot Anjing Curi Perhatian Pengunjung M...	NaN	CNN
FOTO: India Rayakan Festival Holi dengan 'Pera...	NaN	CNN

b. Menyamakan format dari kolom date

dalam kolom date format tanggal hasil scraping memiliki format tanggal yang berbeda sehingga kami menyamakan formatnya

	date	date_convert
0	31/12/2022	2022-12-31
1	31/12/2022	2022-12-31
2	31/12/2022	2022-12-31
3	1/1/2023	2023-01-01
4	1/1/2023	2023-01-01
...	...	...
3903	2023/12/30	2023-12-30

c. Setelah pre-processing

	title	content	sumber	date_convert	title_clean
0	Aturan Baru Sri Mulyani: Gaji Minimal Rp 5 Jut...	KOMPAS.com -Pemerintah dan DPR mengubah batas ...	Kompas	2022-12-31	atur baru sri mulyani gaji minimal rp juta ken...
1	Paus Benediktus XVI Meninggal Dunia di Usia 95...	VATIKAN, KOMPAS.com -MantanPaus Benediktus XVI...	Kompas	2022-12-31	paus benediktus xvi tinggal dunia di usia tahun
2	Kondisi Indra Bekti Menurun, Penglihatan Kabur...	JAKARTA, KOMPAS.com -Adik presenterIndra Bekti...	Kompas	2022-12-31	kondisi indra bekti turun lihat kabur dan tens...
3	Aldila Jelita Dikritik Usai Buka Donasi untuk ...	JAKARTA, KOMPAS.com- Istri presenterIndra Bekt...	Kompas	2023-01-01	aldila jelita kritik usai buka donasi untuk bi...
4	Banjir Semarang Telan Korban Jiwa, Dua Mahasis...	KOMPAS.com- Dua mahasiswa tewas akibat terseng...	Kompas	2023-01-01	banjir semarang telan korban jiwa dua mahasisw...
...	...	...	...	...	...
3858	15 Prajurit TNI Ditahan Usai Diduga Aniaya Rel...	Sebanyak 15 prajurit TNI ditahan buntut dugaan...	CNN	2023-12-30	prajurit tni tahan usai duga aniaya rawan ganj...
3859	Istri Kesal Calon Anak Diberi Nama Seperti Kuc...	Nama adalah doa. Tak ayal semua orangtua pasti...	CNN	2023-12-31	istri kesal calon anak beri nama seperti kucin...
3860	RSUD Sumedang Retak Imbas Gempa M 4,8 di Malam...	Gempa bumi yang mengguncang Sumedang, Jawa Bar...	CNN	2023-12-31	rsud sumedang retak imbas gempa m di malam tah...
3861	Arsenal di Akhir 2023: Mimpi Posisi Satu, Terd...	Harapan Arsenal mengakhiri tahun 2023 di posis...	CNN	2023-12-31	arsenal di akhir mimpi posisi satu dampar di p...
3862	Gelombang Besar Hantam Pantai California, Warg...	Gelombang besar melanda pantai California, Ame...	CNN	2023-12-31	gelombang besar hantam pantai california warga...

3863 rows × 5 columns

3.3 Pre-Processing

Pada tahap penelitian melakukan pre-processing untuk data train dan data hasil scraping yang nanti akan dilakukan untuk analisis lebih lanjut. Sebelum masuk pada teks pre-processing, penelitian melakukan penghapusan data yang memiliki nilai null pada data scraping. Pre-processing yang dilakukan yaitu dengan: Case folding, Tokenizing Filtering, Stemming.

Penelitian melakukan 2 pre-processing yaitu dengan **mempertahankan stopwords** dan **menghapus stopwords**. Dari kedua pre-processing tersebut kami membandingkan hasil yang lebih optimal.persiapan data teks yang akan digunakan untuk kemungkinan pengolahan pada tahap selanjutnya.

3.4 Fitur Engineering dan Word Embedding

1. TF-IDF (Term Frequency-Inverse Document Frequency)

TF - IDF diterapkan pada data kaggle dan data scraping ,yakni melakukan teknik pembobotan kata dan mengukur relevansi kata dalam dokumen berita milik kami terhadap sekumpulan dokumen yang menjadi data train ataupun test. TF - IDF ini juga berguna disini untuk mengatasi keseimbangan data, mengurangi dimensi data, dan untuk keakuratan akurasi dalam model. Namun pada **data kaggle menambahkan fitur** tambahan antara lain *'title_length'*, *'exclamation_count'*, *'capital_ratio'*, *'clickbait_word_count'*, *'has_number'* untuk bahan analisis dan perbandingan ketika *feature engineering* dimasukkan ke dalam model dan melihat korelasi fiturnya terhadap label data kami.

img1.tf-idf data kaggle

	aa	aachen	aaji	aaliyah	aap	aaron	abad	abadi	abah	abai	...	zulkifli	zumba	zun	zylgwyn	zylvechia	title_length	exclamation_count	capital_ratio	clickbait_word_count	has_number
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	14	0	0.159574	0	1
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	11	0	0.166667	0	0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	10	0	0.121622	0	0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	10	0	0.171875	0	0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	10	0	0.144928	0	0

img2.tf-idf data scrap

	aa	aachen	aaji	aaliyah	aap	aaron	abad	abadi	abah	abai	...	zuhur	zul	zulgywyn	zulham	zulhas	zulkifli	zumba	zun	zylgwyn	zylvechia
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
14994	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14995	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14996	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14997	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14998	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

2. Word2vec

Word2vec CBOW diterapkan pada data Kaggle dan data scraping, yakni dengan merepresentasikan kata dalam bentuk vektor dan menangkap hubungan semantik antar kata. Teknik ini memungkinkan untuk memahami konteks kata dalam dokumen berita kami terhadap sekumpulan dokumen yang menjadi data train ataupun test

3.5 Klasifikasi Clickbait Random Forest

Data train kaggle dilatih yang telah dilatih sebelumnya digunakan untuk klasifikasi pada data scraping. Proses pelatihan menggunakan data stopwords dan yang tidak menggunakan stopwords dari hasil TF-IDF dan Word2vec dengan parameter yang sama.

Fitur Engineering	TF-IDF		Fitur Kata		WORD2VEC (CBOW)	
Remove Stopwords	With	Without	With	Without	With	Without
Accuracy	0.76	0.79	0.62	0.71	0.63	0.67

hasil training menunjukan bahwa model random forest TF-IDF tanpa menghapus stopword menghasilkan akurasi tertinggi dengan akurasi 0.79. Selanjutnya model tersebut disimpan dan digunakan untuk melabeli data hasil scraping.

3.6 Similarity Text

Dalam kasus *clickbait* dan *non clickbait* dari sebuah berita, proses preprocessing pada judul dan isi berita diperlukan untuk mempermudah program mengeksekusi tahapan selanjutnya(3.3 Pre-processing).Berita *clickbait* dan *non clickbait* seringkali berhubungan dengan stopwords, maka dari itu data clean dengan remove stopwords dan tanpa remove stopwords akan dilakukan analisis. *Feature engineering* yang digunakan diantaranya ***TF-IDF*** dan ***Word2Vec*** untuk merepresentasikan vektor numerik yang nantinya diimplementasikan pada ***cosine similarity*** ataupun ***jaccard similarity***. Untuk fitur yang mungkin bisa dibilang sangat tinggi atau saling beriringan nilainya sehingga diperlukan threshold yang sesuai distribusi rata-rata pada data, maka akan dilakukan distribusi threshold untuk menentukan threshold yang sesuai dengan *similarity score*.

3.6.1 Cosine Similarity

Dalam kasus ini, memakai dua kemungkinan karena stopwords bisa berpengaruh terhadap judul dan isi berita sehingga nantinya akan

menggunakan remove stopwords dan no remove stopwords. Selanjutnya akan menggunakan teknik pendekatan *feature engineering TF-IDF* dan *Word2Vec* yang dimana kedua pendekatan tersebut nantinya akan merepresentasikan bentuk vektor numerik untuk bisa dilakukan pemodelan selanjutnya sehingga bisa diambil beberapa informasi dari hasil similaritynya.

$$cos(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$$

$$cos(\mathbf{u}, \mathbf{v}) = \frac{\sum_{i=1}^{V} u_i v_i}{\sqrt{\sum_{i=1}^{V} u_i^2}\sqrt{\sum_{i=1}^{V} v_i^2}}$$

Dari rumus diatas dapat diketahui dimana pembilang merupakan dot product dari dua vektor u dan v, kemudian penyebutnya ada panjang norm dari u dan v. Selanjutnya, akan diambil threshold θ yang menjadi batas keputusan. Dalam project ini nantinya keputusan diambil jika hasil *similarity score* < θ maka termasuk dalam berita *clickbait* dan jika hasil *similarity score* > θ maka termasuk berita *non clickbait*. Untuk thresholdnya sendiri ditentukan melalui distribusi rata rata similarity score dari keseluruhan data. Semakin tinggi nilai similarity antara vektor dokumen/item akan semakin banyak relevansi antara dokumen/item.

	No Remove Stpwrds	Remove Stpwrds	TF IDF	Word2Vec	Score 1	Score 2	Score 3
Cosine Similarity	☑	☐	☑	☐	0.801519	0.799078	0.787411
Cosine Similarity	☐	☑	☑	☐	0.852405	0.835246	0.834535
Cosine Similarity	☑	☐	☐	☑	0.982513	0.979994	0.978387
Cosine Similarity	☐	☑	☐	☑	0.997667	0.997043	0.996987

3.6.1 Jaccard Similarity

Jaccard Similarity adalah metode yang digunakan untuk menganalisis kesamaan antara dua sampel data/dokumen. Untuk dua himpunan hingga A dan B, jaccard sendiri juga bisa dibilang *intersection over union* dimana skor kesamaan jaccard berkisar dari 0 hingga 1 dimana 1 mewakili yang paling relevan dan 0 mewakili yang kurang adanya kemiripan antara data/dokumen.

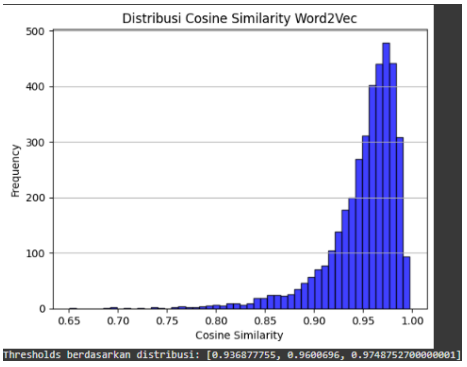
$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \qquad J(A,B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

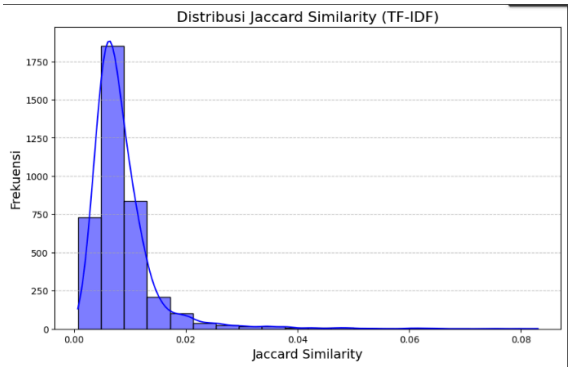
Dari rumus diatas pembilangnya berupa interception himpunan A dan B. Sedangkan bagian penyebut berupa union A dan B. Data clean yang

digunakan terdapat dua tipe yaitu menggunakan stopwords dan tidak menggunakan stopwords. Sebelum memasukkan input data ke dalam *jaccard similarity,* teknik pendekatan *feature engineering* diperlukan. *TF-IDF* dan *Word2Vec* digunakan untuk merepresentasikan bentuk numerik suatu data/dokumen yang kemudian nantinya akan menjadi bahan untuk perhitungan kesamaan dalam project ini. Hasilnya nanti *clickbait* jika skor similarity nya rendah dan jika skor similarity tinggi kemungkinan besar berita tersebut *non clickbait.*

	No Remove Stpwrds	Remove Stpwrds	TF IDF	Word2Vec	Score 1	Score 2	Score 3
Jaccard Similarity	☑	☐	☑	☐	0.054878	0.052464	0.047472
Jaccard Similarity	☐	☑	☑	☐	0.083034	0.079439	0.075669
Jaccard Similarity	☑	☐	☐	☑	0.6	0.526316	0.473684
Jaccard Similarity	☐	☑	☐	☑	0.583333	0.555556	0.545455

Distribusi Threshold





3.7 Clustering

Analisis clustering dalam penelitian ini dengan mengelompokkan judul berita **dengan menghapus stopwords** menjadi dua kategori utama, yaitu non-clickbait (label 0) dan clickbait (label 1). Alasan menghapus stopwords karena judul berita bisa lebih dikenali jika tanpa stopwords. Pendekatan yang dilakukan dengan menggunakan representasi teks yaitu, TF-IDF, Word2Vec, dan FastText untuk dianalisis pada setiap kategori. Setelah data teks direpresentasikan dalam pembobotan kata atau bentuk vektor, dilakukan *StandardScaler* untuk menstandarkan data. Setelah itu dengan teknik Principal Component Analysis (PCA) untuk mempertahankan informasi yang signifikan secara lebih efisien. Apabila beberapa komponen utama pertama menjelaskan lebih dari 85% hingga 95% variasi data asli, maka informasi dalam komponen utama ini sudah mencukupi (Kodinariya & Makwana, 2013). Proses PCA hanya dilakukan pada TF-IDF karena dimensi yang sangat tinggi sedangkan Word2Vec dan FastText berdimensi kecil.

Proses clustering menggunakan algoritma K-Means, untuk penentuan jumlah cluster optimal dilakukan berdasarkan metode Elbow dan evaluasi menggunakan Silhouette Score sebagai ukuran tambahan. Tujuannya untuk memperoleh pemahaman bagaimana data teks di setiap kategori (non-clickbait dan clickbait) dikelompokkan dan menemukan pola-pola khusus yang membedakan kedua kategori tersebut.

BAB 4

HASIL DAN ANALISIS

4.1 Model Klasifikasi

4.1.1 Random Forest TF-IDF

1. **Random forest final_features (without stopwords)**

Classification Report:

	precision	recall	f1-score	support
0	0.75	0.88	0.81	1729
1	0.79	0.60	0.68	1271
accuracy			0.76	3000
macro avg	0.77	0.74	0.74	3000
weighted avg	0.76	0.76	0.75	3000

Akurasi: 0.76

Precision dan Recall:

- Non-clickbait: Precision **0.75**, Recall **0.88**
- Clickbait: Precision **0.79**, Recall **0.60**

F1-Score:

- Non-clickbait: **0.81**
- Clickbait: **0.68**

Kesimpulan:

a. Model cukup baik dalam mendeteksi berita **non-clickbait** (Recall tinggi: **0.88**), tetapi kurang optimal dalam mendeteksi berita **clickbait** (Recall rendah: **0.60**).

b. Fitur tanpa stopwords mengurangi beberapa informasi konteks, terutama untuk pola clickbait.

2. Random forest final_features2 (with stopwords)

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

           0       0.78      0.89      0.83      1729
           1       0.81      0.65      0.72      1271

    accuracy                           0.79      3000
   macro avg       0.79      0.77      0.78      3000
weighted avg       0.79      0.79      0.78      3000
```

Akurasi: 0.79

Precision dan Recall:

- Non-clickbait: Precision **0.78**, Recall **0.89**
- Clickbait: Precision **0.81**, Recall **0.65**

F1-Score:

- Non-clickbait: **0.83**
- Clickbait: **0.72**

Kesimpulan:

a. Akurasi meningkat dibandingkan tanpa hapus stopwords, dengan performa yang lebih seimbang.

b. Peningkatan Recall untuk clickbait (dari **0.60** ke **0.65**) menunjukkan bahwa model mampu mengenali lebih banyak pola clickbait dengan mempertahankan kata-kata penghubung.

➢ **Random forest with fitur kata kata, remove stopwords**

	precision	recall	f1-score	support
Non-Clickbait	0.61	0.94	0.74	1753
Clickbait	0.66	0.17	0.27	1247
accuracy			0.62	3000
macro avg	0.64	0.56	0.51	3000
weighted avg	0.63	0.62	0.55	3000

➢ **Random forest with fitur kata kata, no remove stopwords**

	precision	recall	f1-score	support
Non-Clickbait	0.74	0.79	0.76	1753
Clickbait	0.67	0.60	0.63	1247
accuracy			0.71	3000
macro avg	0.70	0.70	0.70	3000
weighted avg	0.71	0.71	0.71	3000

4.1.2 Random Forest Word2Vec

1. **Random forest Word2vec without remove stopwords**

	precision	recall	f1-score	support
0	0.69	0.80	0.74	1729
1	0.65	0.50	0.57	1271
accuracy			0.67	3000
macro avg	0.67	0.65	0.65	3000
weighted avg	0.67	0.67	0.67	3000

2. **Random forest Word2vec with remove stopwords**

	precision	recall	f1-score	support
0	0.66	0.76	0.70	1729
1	0.58	0.46	0.51	1271
accuracy			0.63	3000
macro avg	0.62	0.61	0.61	3000
weighted avg	0.62	0.63	0.62	3000

3. **Perbandingan**

Menyertakan stopwords justru menurunkan performa pada Word2Vec, baik untuk non-clickbait maupun clickbait. Word2Vec lebih sensitif terhadap kata-kata yang berlebihan atau kurang relevan, sehingga penghapusan stopwords lebih disarankan.

4.1.3 Analisis Word2Vec dan Random Forest dalam klasifikasi

Fitur kata (TF-IDF atau bag-of-words) lebih cocok untuk mendeteksi clickbait karena pola clickbait sering kali eksplisit dan berdasarkan frekuensi

kata tertentu. Word2Vec, yang bergantung pada representasi semantik, kurang efektif dalam kasus ini karena clickbait cenderung menggunakan pola yang eksplisit dan bukan hubungan semantik yang mendalam. Hal ini menunjukkan bahwa pola sederhana dalam clickbait lebih mudah ditangkap oleh model berbasis frekuensi kata dibandingkan representasi semantik (Word2Vec).

4.1.4 Random forest try data test

Model random forest dengan akurasi tertinggi digunakan untuk labeling data scraping.

	title	prediksi_label_text
0	Aturan Baru Sri Mulyani: Gaji Minimal Rp 5 Jut...	non-clickbait
1	Paus Benediktus XVI Meninggal Dunia di Usia 95...	non-clickbait
2	Kondisi Indra Bekti Menurun, Penglihatan Kabur...	non-clickbait
3	Aldila Jelita Dikritik Usai Buka Donasi untuk ...	clickbait
4	Banjir Semarang Telan Korban Jiwa, Dua Mahasis...	clickbait
...	...	...
3858	15 Prajurit TNI Ditahan Usai Diduga Aniaya Rel...	non-clickbait
3859	Istri Kesal Calon Anak Diberi Nama Seperti Kuc...	clickbait
3860	RSUD Sumedang Retak Imbas Gempa M 4,8 di Malam...	non-clickbait
3861	Arsenal di Akhir 2023: Mimpi Posisi Satu, Terd...	non-clickbait
3862	Gelombang Besar Hantam Pantai California, Warg...	non-clickbait

3863 rows × 2 columns

No	Clickbait	Non-Clickbait
1	Korban Pembunuhan Berantai di Cianjur Sudah Lama Hilang, Kenapa Baru Terungkap?	Aturan Baru Sri Mulyani: Gaji Minimal Rp 5 Juta Kena Pajak 5 Persen
2	Identitas Prajurit Gadungan yang Ajak Wanita Foto Studio Terungkap, TNI: Domisili Bandung	Beda Perlakuan Polisi terhadap Kasus Kecelakaan yang Tewaskan Mahasiswa UI, Hasya dan Annisa

No	Clickbait	Non-Clickbait
1	Korban Pembunuhan Berantai di Cianjur Sudah Lama Hilang, Kenapa Baru Terungkap?	Aturan Baru Sri Mulyani: Gaji Minimal Rp 5 Juta Kena Pajak 5 Persen
3	Terungkap Alasan Pria Buka Pintu Darurat Asiana Airlines saat Penerbangan	Kapolri Sebut Tidak Ada Orang yang Disandera di Kasus Pesawat Susi Air
4	Warganet Mengeluh Nomor Pribadinya Dijadikan Kontak Darurat Pinjol, Ini Saran OJK	Bareskrim Tetapkan 9 Tersangka Kasus "Robot Trading" Net89, 1 Meninggal, 2 DPO
5	Terungkap, Bayi Tertukar di Bogor karena Gelang Dipasangkan Suster Rumah Sakit Dobel	Dukung Anies Capres, Amien Rais Doakan Prabowo jadi Presiden
6	Bocah di Ponorogo 3 Hari Tinggal dengan Jenazah Ibu, Terungkap Saat Tetangga Terima WA	Profil Lengkap Rafael Trisambodo, PNS Pajak Berharta Rp 56 Miliar
7	Terungkap! Dekan FMIPA Unila Pakai APBN untuk Beri THR Eks Rektor Karomani	Kemenkeu Tolak Pengunduran Diri Rafael Alun dari ASN Ditjen Pajak
8	Anies Baswedan: Kalau di Survei Nomor 3 Buat Apa Dijegal?	AG, Pacar Mario, Ditetapkan sebagai Pelaku Kasus Penganiayaan D
9	Terungkap! Fredrich Yunadi Sudah Bebas dari Penjara	Surya Paloh Pastikan Anies Baswedan Lanjutkan Pembangunan Era Jokowi jika Terpilih Jadi Presiden
10	Siksa Kubur Rampung Syuting, Aksi Reza Rahadian Terungkap	Buntut Kasus Rafael Alun Trisambodo, Kepala Bea dan Cukai Makassar Diperiksa Kemenkeu, Miliki Harta Rp 13,7 Miliar

4.2 Model Similarity

➢ **Cosine Similarity**

1. TF IDF

- Remove stopwords

I	N	O	P
similarity_remove_stopwor	label_threshold_0.3	label_threshold_0.5	label_threshold_0.7_
0.610108973	0	0	1
0.751533811	0	0	0
0.548712203	0	0	1
0.601666954	0	0	1
0.433577907	0	1	1
0.613933617	0	0	1
0.466132673	0	1	1
0.486914628	0	1	1
0.403784951	0	1	1
0.574702729	0	0	1

- No remove stopwords

J	K	L	M
similarity_no_remove	label_threshold_0.3	label_threshold_0.5	label_threshold_0.7
0.514275389	0	0	1
0.696443198	0	0	0
0.561119597	0	0	1
0.554580654	0	0	1
0.35363449	0	1	1
0.48129771	0	0	1
0.295914928	0	1	1
0.459366499	0	1	1
0.207166423	0	1	1
0.545828303	0	0	1

2. Word2Vec

- No remove stopwords

U	V	W	X
similarity_word2vec_no_remove_stopv	label_word2vec_threshold_0.8631725	label_word2vec_threshold_0.8981749	label_word2vec_threshold_0.9257667
0.9100472	0	0	1
0.7762558	1	1	1
0.922637	0	0	1
0.9178746	0	0	1
0.90493816	0	0	1
0.9269005	0	0	0
0.9511925	0	0	0
0.8123075	1	1	1
0.83958465	1	1	1
0.9296339	0	0	0

- Remove stopwords

AC	AD	AE	AF
similarity_word2vec_remove_	label_word2vec_remove_threshold_0.936877	label_word2vec_remove_threshold_0.9600696	label_word2vec_remove_threshold_0.97487527
0.98588246	0	0	0
0.83791447	1	1	1
0.91204596	1	1	1
0.9402889	0	1	1
0.9741797	0	0	1
0.9596745	0	1	1
0.97507095	0	0	0
0.81373835	1	1	1
0.9706267	0	0	1
0.94814634	0	1	1

➢ **Jaccard Similarity**

1. TF IDF
   - No remove stopwords

I	J	K	L
similarity_jaccard_tfid	label_jaccard_threshold_0.002	label_jaccard_threshold_0.003	label_jaccard_threshold_0.004
0.003275639	0	0	1
0.006570219	0	0	0
0.004749114	0	0	0
0.003722018	0	0	1
0.002675155	0	1	1
0.003487665	0	0	1
0.001896891	1	1	1
0.003146346	0	0	1

   - Remove stopwords

M	N	O	P
similarity_jaccard_	label_jaccard_threshold_0.0022_(	label_jaccard_threshold_0.0072_	label_jaccard_threshold_0.0087_
0.00802775	0	0	1
0.019270098	0	0	0
0.009610358	0	0	0
0.007713679	0	0	1
0.005475612	0	1	1
0.008185782	0	0	1
0.005483914	0	1	1
0.005728407	0	1	1

2. Word2Vec
   - No remove stopwords

Y	Z	AA	AB
similarity_jacc	label_jaccard_threshold_0.0293_	label_jaccard_threshold_0.0654_	label_jaccard_threshold_0.1183
0.040816327	0	1	1
0.060150376	0	1	1
0.071428571	0	0	1
0.055555556	0	1	1
0.05952381	0	1	1
0.074534161	0	0	1
0.045226131	0	1	1
0.053571429	0	1	1
0.021621622	1	1	1
0.095238095	0	0	1

   - Remove stopwords

S	T	U	V
similarity_jaccard_word2vec_remove_stop	label_jaccard_threshold_0.0293	label_jaccard_threshold_0.0654	label_jaccard_threshold_0.1183
0.048951049	0	1	1
0.083333333	0	0	1
0.090909091	0	0	1
0.067164179	0	0	1
0.057377049	0	1	1
0.060869565	0	1	1
0.039735099	0	1	1
0.056	0	1	1
0.037878788	0	1	1
0.107526882	0	0	1

Ringkasan

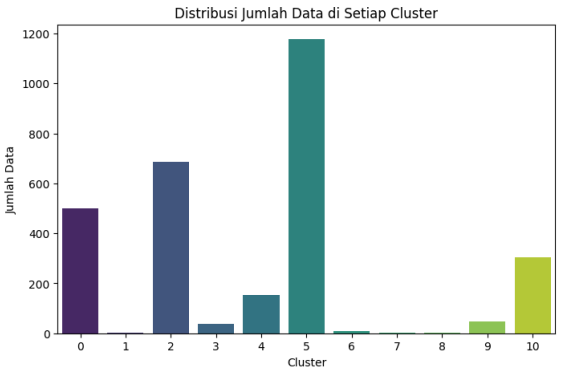
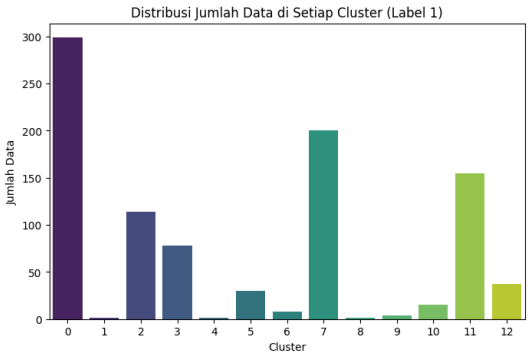
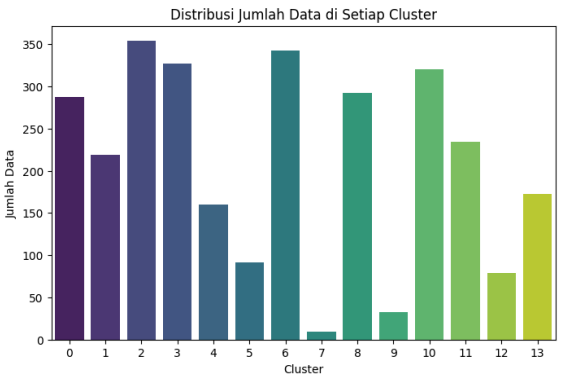
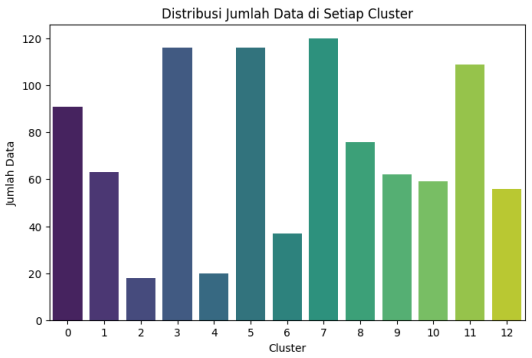
	No Remove Stpwrds	Remove Stpwrds	TF IDF	Word2Vec	Score 1	Score 2	Score 3
Cosine Similarity	☑	☐	☑	☐	0.801519	0.799078	0.787411
Cosine Similarity	☐	☑	☑	☐	0.852405	0.835246	0.834535
Cosine Similarity	☑	☐	☐	☑	0.982513	0.979994	0.978387
Cosine Similarity	☐	☑	☐	☑	0.997667	0.997043	0.996987

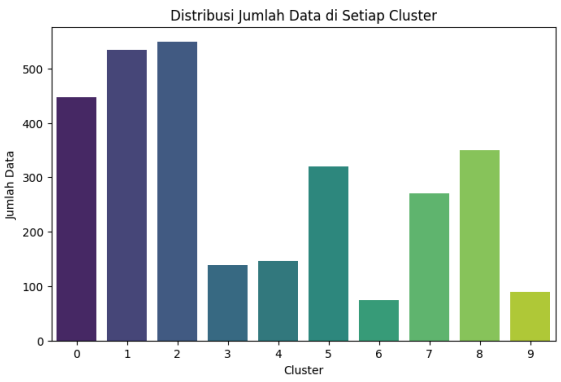
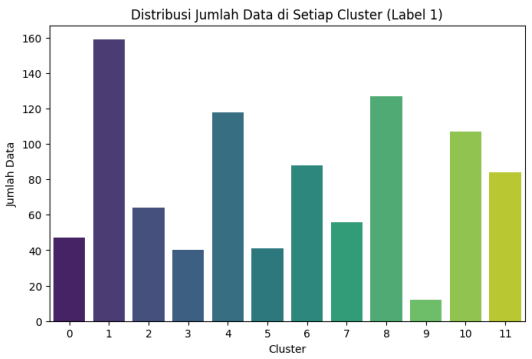
	No Remove Stpwrds	Remove Stpwrds	TF IDF	Word2Vec	Score 1	Score 2	Score 3
Jaccard Similarity	☑	☐	☑	☐	0.054878	0.052464	0.047472
Jaccard Similarity	☐	☑	☑	☐	0.083034	0.079439	0.075669
Jaccard Similarity	☑	☐	☐	☑	0.6	0.526316	0.473684
Jaccard Similarity	☐	☑	☐	☑	0.583333	0.555556	0.545455

4.3 Analisis Clustering

		Label 0 (Non-Clickbait)			Label 1 (Clickbaut)		
Fitur Engineering		TF-IDF	Word2Vec	FastText	TF-IDF	Word2Vec	FastText
Jumlah Komponen PCA		1832	-	-	692	-	-
Pemilihan K	Shiloutte Score Optimal	0.43	0.049	0.21	0.40	0.013	0.028
	Elbow Method Optimal	k=11	k=10	k=14	k=13	k=12	k=13

Hasil clustering akan diambil dengan evaluasi elbow method terlebih dahulu, lalu dipertimbangkan dengan silhouette score yang lebih mendekati 1 .

	Label 0 (Non-Clickbait)	**Label 1 (Clickbait)**
TF-IDF		
FastText		

Word2Vec		

Untuk klasifikasi topiknya berdasarkan visualisasi di atas dapat ditentukan bahwa

1. Pada **TF-IDF** kemungkinan kata-kata lebih spesifik untuk menghasilkan cluster..
2. **Word2Vec** dan **FastText** cenderung menghasilkan cluster berbasis semantik (makna kalimat).
3. Berdasarkan hasil visualisasi, **word2vec** menghasilkan cluster yang optimal dengan persebaran kata top 10:
    a. Label 0

Cluster	Top Words
0	rp (35), soal (31), tak (25), polisi (24), jokowi (22), jadi (21), bakal (20), warga (19), sebut (17), minta (17)
1	polisi (55), soal (51), kasus (45), tak (44), warga (39), jadi (37), duga (35), indonesia (35), kpk (34), usai (30)
2	hasil (25), sebut (16), bakar (13), gelar (13), juara (11), air (11), rp (11), mahfud (11), temu (11), jakarta (11)
3	jadi (59), soal (36), cawapres (34), prabowo (33), tak (29), ganjar (28), gibran (23), usai (22), dukung (20), sebut (19)
4	tewas (61), orang (47), polisi (39), korban (24), warga (24), anak (21), bogor (21), usai (15), israel (14), bunuh (13)
5	soal (62), jadi (56), prabowo (55), tak (52), ganjar (42), jokowi (41), indonesia (40), gibran (37), sebut (34), dukung (33)
6	gempa (75), m (63), guncang (41), jadi (27), barat (14), malu (13), sulut (10), magnitudo (10), tsunami (9), potensi (8)
7	hasil (53), indonesia (33), liga (33), vs (26), usai (22), inggris (17), polisi (15), dunia (14), gol (14), man (13)
8	the (10), s (6), film (6), mu (6), motogp (6), lihat (5), lebih (5), baik (5), baru (5), pakai (5)
9	u (71), indonesia (62), piala (58), dunia (36), timnas (31), usai (16), aff (13), asia (12), vs (10), main (9)

b. Label 1

Cluster	Top Words
0	sebut (9), alas (7), tak (6), bogor (6), lapor (6), temu (5), anak (5), bikin (5), lebih (4), foto (4)
1	rp (30), polisi (30), usai (26), viral (26), juta (17), cerita (14), anak (14), soal (13), minta (13), duga (13)
2	prabowo (25), anak (17), tak (17), usai (13), anies (13), mau (12), ganjar (11), soal (10), dukung (10), pilih (8)
3	detik (16), sambo (4), vonis (4), mahfud (4), soal (4), gera (3), tanah (3), milu (3), indonesia (3), teman (3)
4	viral (13), mobil (8), bakar (7), polisi (7), tinggal (7), masuk (7), fakta (6), warga (6), prabowo (6), tewas (5)
5	jokowi (26), temu (13), video (7), sebut (6), ganjar (4), soal (4), kata (4), viral (3), pdi (3), p (3)
6	ungkap (25), tak (23), alas (10), malam (10), hasil (7), diri (7), bunuh (6), viral (6), tni (6), jalan (6)
7	cawapres (29), jadi (21), soal (19), alas (13), anies (13), capres (12), prabowo (10), ungkap (9), sebut (9), gibran (9)
8	hasil (11), indonesia (7), putri (6), final (6), bikin (6), all (5), siapa (5), juara (5), respons (5), soal (5)
9	kata (21), usai (8), indonesia (5), u (3), soal (2), juara (2), dunia (2), latih (2), baliho (1), gambar (1)
10	tiba (13), soal (11), anies (10), kata (9), gibran (8), hingga (8), apa (8), tewas (7), hukum (7), baru (7)
11	jadi (30), tahun (17), cerita (10), cara (7), ungkap (7), orang (6), hingga (6), sangka (5), polisi (4), tutup (4)

4. Topik ditentukan berdasarkan judul pada tiap cluster:
   a. Label 0

Cluster	Titles
0	Aturan Baru Sri Mulyani: Gaji Minimal Rp 5 Juta Kena Pajak 5 Persen,Kronologi Tukang Becak Cairkan Uang Rp 320 Juta di Surabaya Versi Keluarga Pemilik Rekening,Buntut Kasus Rafael Alun Trisambodo, Kepala Bea dan Cukai Makassar Diperiksa Kemenkeu, Miliki Harta Rp 13,7 Miliar
1	Polisi di Lampung Dikepung Warga Saat Tangkap Bandar Narkoba, Mobil Digulingkan Massa,Tembakkan Gas Air Mata ke Suporter PSIS, Polisi Dinilai Tak Belajar dari Tragedi Kanjuruhan,Pekerjaan Pelaku Mutilasi Wanita di Sleman Terungkap, Polisi: Mengurus Tenda
2	Hasil India Open 2023, Monster Ganda Campuran China Tersingkir,Hasil Proliga 2023: Bandung bjb Tandamata Raih Tiket Grand Final,Hasil Proliga 2023: Taklukkan Pertamina 3-1, Tandamata Juara Putaran Kedua
3	Balita Tewas di Pasar Rebo, Diduga Jadi Jaminan Utang Ibunya Sebesar Rp 300.000,Jadi Bakal Capres Favorit Partai Ummat, Anies Diundang ke Rakernas,Dukung Anies Capres, Amien Rais Doakan Prabowo jadi Presiden
4	Banjir Semarang Telan Korban Jiwa, Dua Mahasiswa Tewas Tersengat Listrik Jelang Tahun Baru 2023,Beda Perlakuan Polisi terhadap Kasus Kecelakaan yang Tewaskan Mahasiswa UI, Hasya dan Annisa,Tertembak Senapan Angin Dimainkan Anak, Perempuan di Semarang Tewas
5	Soal Teguran Komisi I DPR, Panglima TNI Akan Sampaikan ke Dudung,Pemkab Deli Serdang soal Jalan Umum Dijual Rp 1,6 Miliar: Tak Salahi Aturan,Soal Pilpres 2024, Jokowi: Saya Enggak Mau Dibawa Ke Sana-sini...
6	Gempa Terkini M 5,6 Guncang Pacitan, Tidak Berpotensi Tsunami,Gempa M 6,4 Guncang Padang Sidempuan, Warga Panik Lari ke Luar Rumah,Gempa M 6,4 Guncang Padang Sidempuan Sumut, Tak Berpotensi Tsunami
7	Hasil Malaysia Open 2023: Axelsen Juara dalam Tempo 40 Menit,Hasil India Open 2023: Jonatan Christie Sukses Revans, Lolos 16 Besar,Hasil Sidang Etik, Richard Eliezer Dipertahankan Polri
8	Ahli Unsoed: Jembatan The Geong Banyumas Gunakan Kaca Bekas,7 Film Dewasa 'Hot' di Netflix, Salah Satunya The Next 365 Days,AHY ke Anies: You Are The Superstar, You Will Lead Us All
9	Hasil Timnas U20 Indonesia Vs Irak 0-2: Garuda Kalah Lawan 10 Pemain,Jadwal Siaran Langsung Indonesia Vs Uzbekistan di Piala Asia U20, Laga Penentuan Garuda,Pemerintah Lobi FIFA soal Partisipasi Timnas Israel di Piala Dunia U20

b. Label 1

Cluster	Filtered Titles
0	Sang Ibu Sebut Nono Bocah Juara Sempoa Dunia Tolak Hadiah Mobil dari Astra, Ini Alasannya,Keluarga TikToker Bima: Gubernur Lampung Sebut Orangtua Bima Tak Bisa Didik Anak,Disebut Memaki Inara Saat Diminta Jadi Penengah, Ibunda Virgoun: Dia Nangis Masih Saya Peluk
1	LSM Minta Uang Damai Rp 200 Juta Kasus Pemerkosaan di Brebes, Keluarga Korban Hanya Diberi Rp 30 Juta,Cerita di Balik Tukang Becak di Surabaya Cairkan Uang Rp 320 Juta dari Rekening Bukan Miliknya,Pelanggan Mohon Tiang Listrik Dipindah Malah Diminta Rp 4,3 Juta, Ini Penjelasan PLN
2	Daftar 6 Nama Capres-Cawapres 2024 yang Muncul pada Musra Jateng, Ada Ganjar hingga Prabowo,Jokowi Kumpulkan Ketum Parpol, PPP Ungkap Potensi Wujudkan Dukungan untuk Ganjar-Prabowo,Prabowo Pamer Foto Bareng Jokowi Naik Maung di Hadapan Massa Kader Gerindra
3	Detik-detik Terakhir Sebelum Meninggal, Nani Wijaya Tersenyum dan Melambaikan Tangan,Detik-detik Wajah Pesulap Limbad Terbakar Saat Atraksi Sembur Api di Madiun,Detik-detik KDRT Panca di Jagakarsa: Awalnya Sisiri Rambut Istri, Tiba-tiba Emosi lalu Menganiaya
4	Video Viral CVT Honda PCX 160 Jebol, Ini Penjelasannya,Viral, Unggahan Ratusan Calon Maba UB Disebut Mengundurkan Diri, Ini Kata Pihak Kampus,Istri Polisi di Probolinggo yang Viral Karena Bentak Siswi Magang Kini Diperiksa
5	Ketika Jokowi Menyebut Gibran dengan Panggilan "Pak Wali",Bupati Grobogan Cukur Rambut 2 Kades Gondrong yang Terekam di Video Viral Sentil Nama Jokowi,GASPOL! Hari Ini: PDI-P Tolak Israel di Piala Dunia U-20, Takut Jokowi Dimakzulkan?
6	Korban Pembunuhan Berantai di Cianjur Sudah Lama Hilang, Kenapa Baru Terungkap?,Ketua RW Ungkap Sosok Bripka Madih yang Ngaku Diperas Polisi: Dia Suka Bikin Onar,Identitas Prajurit Gadungan yang Ajak Wanita Foto Studio Terungkap, TNI: Domisili Bandung
7	PAN Ungkap Alasan Dukung Ganjar-Erick Jadi Capres-Cawapres 2024,Jokowi Minta Prabowo Jadi Cawapres Ganjar saat Pertemuan di Solo? Ini Kata Gerindra,Tolak Tawaran PKS Jadi Cawapres Anies, Mahfud: Kalau Diajak, Malah Merusak Demokrasi
8	Hasil Lengkap Malaysia Open 2023 Hari Ini: Minions Gugur, 5 Wakil Indonesia Melaju,Hasil Lengkap Indonesia Masters 2023: All Indonesian Final Bersejarah!,Hasil Final Indonesia Masters 2023: Gebuk Chico dengan Smes 364 Km/Jam, Jonatan Juara!
9	Soal Baliho Bergambar Kaesang dengan Jokowi, Ini Kata Giring PSI,Bakal Dukung Anies di Pilpres 2024? Ini Kata Veronica Tan,Kata-kata STY Usai Timnas Indonesia U-20 Tumbang dari Selandia Baru
10	Isak Tangis Iringi Jenazah Syabda Perkasa Belawa dan Ibunya Saat Tiba di Rumah Duka,Tiba-tiba Menangis Saat Tanggapi Kedekatan Gading Marten dengan Pacar Baru, Gisella Anastasia: Perasaan yang Rumit,Bus Rombongan Mudik Gratis Pemkot Medan Masuk Jurang, Mesin Tiba-tiba Mati
11	Kronologi Aiptu AR "Jual" Istri ke Sesama Polisi di Pamekasan, Terjadi Sejak 2015, Konsumsi Narkoba Sebelum Beraksi,Visa Habis, WNI Mantan Pemetik Buah di Inggris Jadi Imigran Gelap hingga Cari Suaka,Bandara Internasional Bakal Dikurangi Jadi 15, Erick Thohir: Yang Lain Hanya Boleh Layani Haji dan Umrah

4.4 Analisis Perbandingan Hasil Klasifikasi dan Similarity

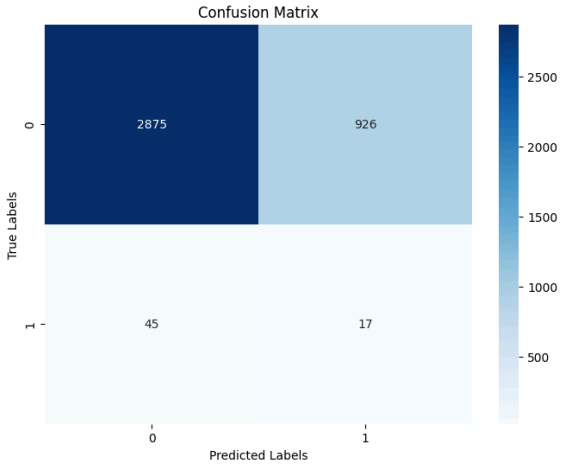
1. Perbandingan akurasi Cosine dan Jaccard Similarity

Penelitian ini mengasumsikan data hasil pelabelan Random forest sebagai acuan untuk membandingkan dua pendekatan similarity yaitu Cosine similarity dan Jaccard similarity.

Similarity	Fitur Engineering	Stopword	Tresshold	Accuracy
Cosine Similarity	TF-IDF	Remove	0.3	0.71
			0.5	0.46
			0.7	0.26
		No Remove	0.3	0.70
			0.5	0.46
			0.7	0.26
	Word2Vec	Remove	0.94	0.59
			0.96	0.43
			0.97	0.31
		No Remove	0.87	0.64
			0.90	0.31
			0.93	0.38
Jaccard Similarity	TF-IDF	Remove	0.002	0.70
			0.003	0.58
			0.004	0.46
		No Remove	0.0022	0.75
			0.0072	0.51
			0.0087	0.43
	Word2Vec	Remove	0.00293	0.74
			0.00654	0.50
			0.1183	0.28
		No Remove	0.0293	0.74
			0.0654	0.40
			0.1183	0.27

Berdasarkan tabel, similarity Jaccard dengan *feature engineering* TF-IDF tanpa menghapus stopword menggunakan threshold 0,002 menghasilkan akurasi tertinggi diantara percobaan lainnya ,dengan akurasi 0.75 dan rincian classification report dan confusion matrix sebagai berikut :

Confusion Matrix



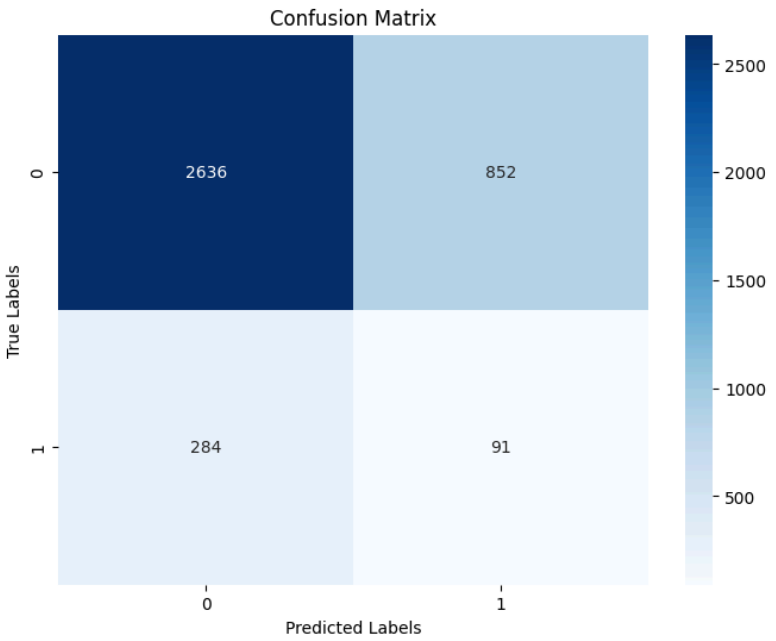
Hasil confusion matrix menunjukan bahwa dengan akurasi 0.75, model dapat memprediksi label non clickbait(0) dengan cukup baik dengan hanya terdapat 926 kesalahan dan 2875 benar. Namun kurang optimal untuk label clickbait karena hanya 17 data yang predik benar sedangkan 45 lainnya salah.

Classification Report:

	precision	recall	f1-score	support
0	0.98	0.76	0.86	3801
1	0.02	0.27	0.03	62
accuracy			0.75	3863
macro avg	0.50	0.52	0.44	3863
weighted avg	0.97	0.75	0.84	3863

Peneliti juga menganalisis hasil Cosine similarity tertinggi dengan akurasi 0.71. rincian classification report dan confusion matrix sebagai berikut :

Confusion matriks



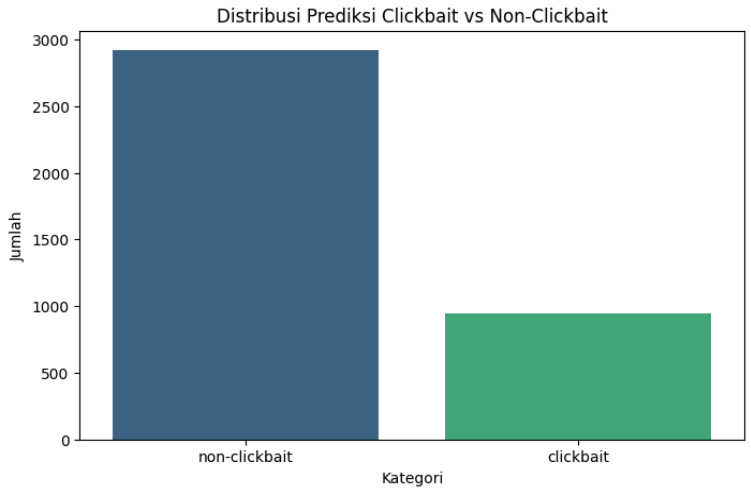
Hasil confusion matrix menunjukan bahwa dengan akurasi 0.71, model dapat memprediksi label non clickbait(0) dengan cukup baik dengan terdapat 852 kesalahan dan 2636 benar. Namun kurang optimal untuk label clickbait karena hanya 91 data yang predik benar sedangkan 284 lainnya salah.Namun 91 data yang di predik benar tersebut lebih besar dari jaccard similarity yang hanya memprediksi benar 17 data.

Classification Report:

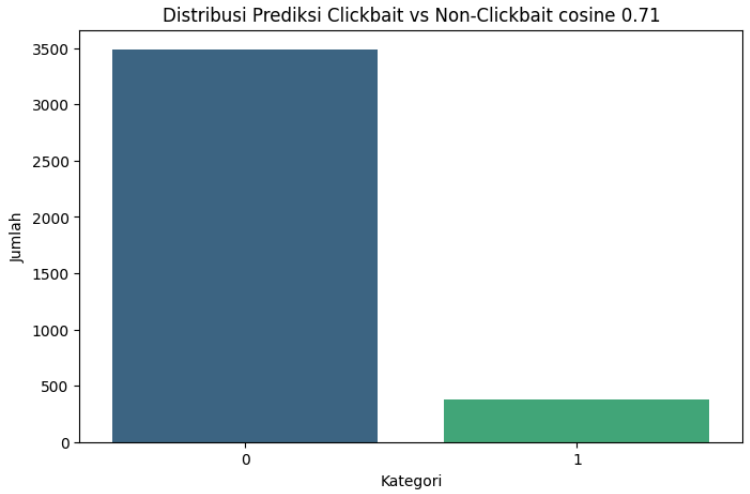
	precision	recall	f1-score	support
0	0.90	0.76	0.82	3488
1	0.10	0.24	0.14	375
accuracy			0.71	3863
macro avg	0.50	0.50	0.48	3863
weighted avg	0.82	0.71	0.76	3863

2. Perbandingan hasil pelabelan

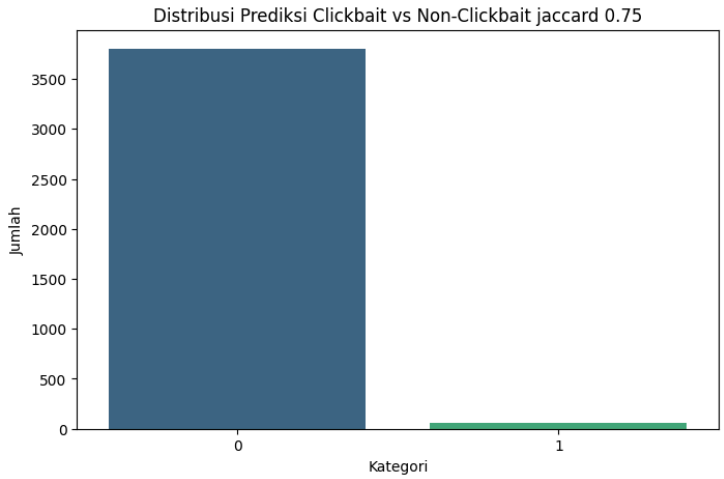
- **Random Forest**



- **Cosine Similarity 0.71**



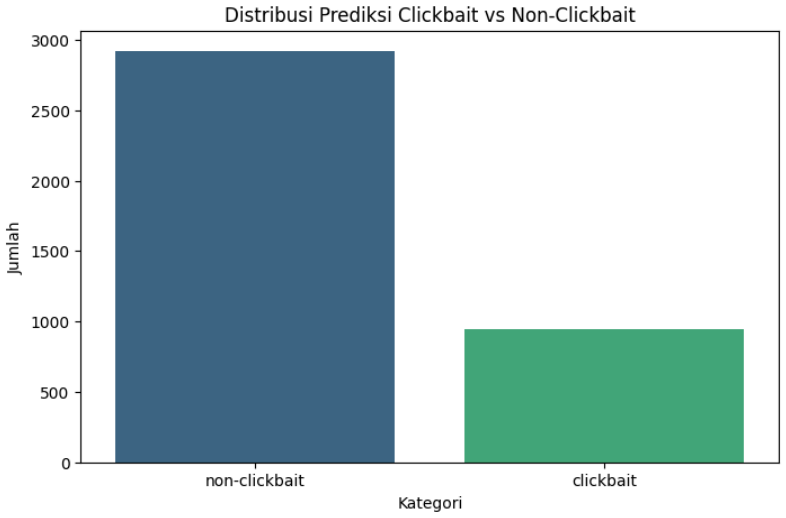
- **Jaccard Similarity 0.75**



Distribusi hasil similarity class non clickbait menggunakan cosine dan Jaccard menunjukkan sekitar 3.500 data. Sementara itu, pada hasil pelabelan menggunakan Random Forest, hanya terdapat sekitar 3.000 data untuk kelas non-clickbait. Sebaliknya, untuk kelas clickbait, Random Forest menghasilkan sekitar 1.000 data, sedangkan cosine dan Jaccard hanya menghasilkan kurang dari 500 data.

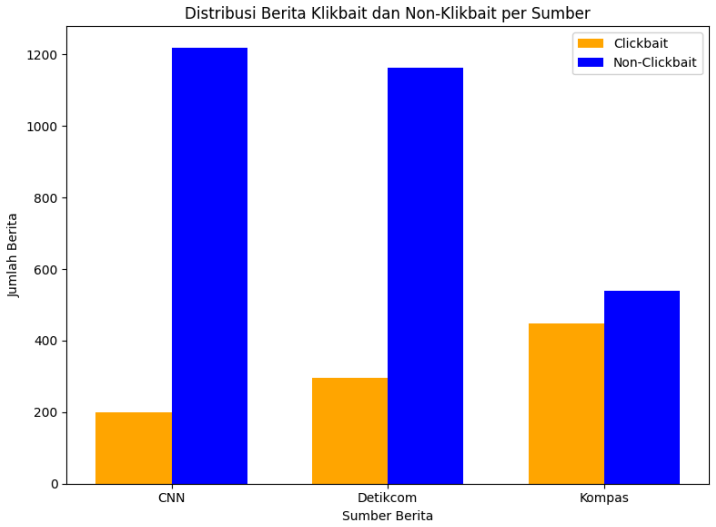
4.5 Analysis Insight

Dari histogram di bawah, terlihat bahwa jumlah berita non-clickbait lebih dominan dibandingkan berita clickbait dalam keseluruhan data scraping. Berita non-clickbait tercatat sebanyak sekitar 3.000 berita, sedangkan berita clickbait hanya mencapai sekitar 1.000 berita

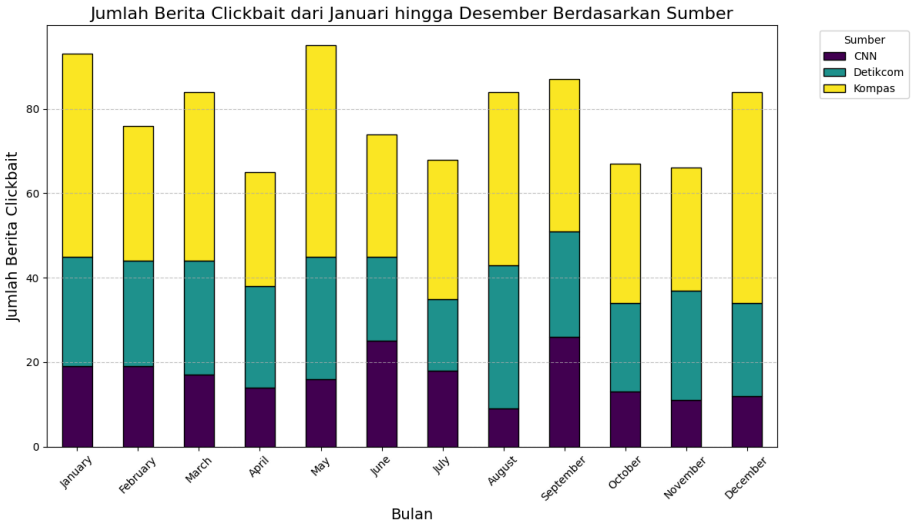


Berdasarkan grafik di bawah ini, Kompas menempati posisi pertama dengan jumlah berita clickbait terbanyak, yaitu lebih dari 400 berita. Detikcom berada di

peringkat kedua dengan jumlah berita yang sedikit lebih rendah, sedangkan CNN berada di peringkat terakhir dengan sekitar 200 berita clickbait.



Tren bulanan sepanjang tahun 2023 menunjukkan bahwa Kompas secara konsisten mencatat jumlah berita terbanyak setiap bulan dibandingkan dengan platform lainnya. Hal ini mengindikasikan dominasi Kompas dalam jumlah publikasi berita selama periode tersebut.



BAB 5

PENUTUP

KESIMPULAN

Berdasarkan beberapa analisis yang telah dilakukan untuk kasus klasifikasi berita *clickbait* dan *non clickbait*, dari awal tahap preprocessing data seperti penghapusan stopwords dan tanpa penghapusan stopwords memberikan pengaruh yang berbeda pada hasil klasifikasi, variasi hasil masing-masing pendekatan tersebut menunjukkan bahwa langkah preprocessing menentukan kualitas fitur teks yang dihasilkan.

Penggunaan *feature engineering* seperti TF-IDF, Word2Vec, dan FastText juga menghasilkan kualitas fitur yang berbeda dimana masing-masing teknik tersebut memiliki karakteristik tersendiri.

Random forest TF IDF tanpa remove stopwords menunjukkan hasil akurasi terbaik dalam kasus klasifikasi ini. Dengan mencoba beberapa teknik pendekatan fitur, perbedaan tiap tiap pendekatan terlihat signifikan dalam menentukan label *clickbait* maupun *non clickbait.* Hasil masing-masing label akan di-clustering untuk menentukan topik berdasarkan judul.

Pendekatan similarity dalam kasus ini menggunakan *cosine similarity* dan *jaccard similarity* yang dimana pendekatan ini memberikan hasil klasifikasi yang sederhana dibandingkan random forest. Selanjutnya hasil similarity dan hasil random forest (*supervised learning*) dilakukan analisis untuk melihat lebih *worth it* mana antara *cosine* dan *jaccard.*

Secara keseluruhan, kombinasi preprocessing, teknik representasi teks, dan modeling terbukti menjadi faktor penentu untuk hasil terbaik dalam kasus klasifikasi berita *clickbait* atau *non clickbait*.

KENDALA DAN SARAN

Kendala dalam kasus ini, distribusi dan kredibilitas dataset perlu divalidasi lagi karena sekilas terlihat kurangnya kelengkapan data karena ada beberapa folder namun hanya dilakukan eksekusi terhadap beberapa kolom dalam file folder tersebut. Selain itu, fasilitas collab yang kurang memadai sehingga menimbulkan berkali-kali reconnect yang mengakibatkan run code ulang secara keseluruhan.

Untuk kedepannya, analisis bisa dilakukan dengan memastikan atau memining data lebih dalam lagi dan lebih detail untuk mempermudah jalur *step by step* dalam penyelesaian kasus project serta berlangganan collab pro demi kenyamanan penggunaan tools yang baik dan nyaman tanpa ada kendala teknis.

BAB 6

DAFTAR PUSTAKA

Kodinariya, T. M., & Makwana, D. R. (2013). Review on determining number of cluster in K-Means clustering. International Journal of Advance Research in Computer Science and Management Studies, 90-95.

Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2), 129–137.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.

Indraloka, D. S., & Santosa, B. (2017). Penerapan Text Mining untuk Melakukan Clustering Data Tweet Shopee Indonesia. JURNAL SAINS DAN SENI ITS, 51-56.

Ahmadi, H. A., & Chowanda, A. (2023). Clickbait Classification Model on Online News with Semantic Similarity Calculation Between News Title and Content. *Building of Informatics, Technology and Science (BITS)*, *Volume 4*(4), 1990. DOI 10.47065/bits.v4i4.3030. Retrieved 01 Rabu, 2025, from DOI 10.47065/bits.v4i4.3030

exploratory-data analysis. (2022, Maret 1). Algoritma. Retrieved January 1, 2025, from https://algorit.ma/blog/exploratory-data-analysis-2022/

Girsang, A. S. (2021, December 31). *Word Embedding dengan FastText*. mti.binus.ac.id. Retrieved January 1, 2025, from https://mti.binus.ac.id/2021/12/31/word-embedding-dengan-fasttext/

Gomma, W. H., & Fahmy, A. A. (2013). A Survey of Text Similarity Approaches. *International Journal of Computer Applications*, *Volume 68*(13), 14.

https://www.researchgate.net/publication/259181798_A_Survey_of_Text_Similarity_Approaches. Retrieved 01 Rabu, 2025, from https://www.researchgate.net/publication/259181798_A_Survey_of_Text_Similarity_Approaches

Khasanah, L. U. (2020, September 2). *Exploratory Data Analysis : Pahami Lebih Dalam untuk Siap Hadapi Industri Data*. DQLab. Retrieved January 1, 2025, from https://dqlab.id/data-analisis-machine-learning-untuk-proses-pengolahan-data

K-Means Clustering Algorithm. (n.d.). Javatpoint. Retrieved January 1, 2025, from https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

Mahendra, D. S., Rahmat, B., & Mumpuni, R. (2024, Agustus). Implementasi Metode Multinomial Naive Bayes dalam Klasifikasi Judul Berita Clickbait. *Neptunus : Jurnal Ilmu Komputer Dan Teknologi Informasi*, *Volume 2*(3), 304. https://doi.org/10.61132/neptunus.v2i3.249

Winland, V. (2024, June 26). *What is k-means clustering?* IBM. Retrieved January 1, 2025, from https://www.ibm.com/think/topics/k-means-clustering

word2vec | Text. (2024, July 19). TensorFlow. Retrieved January 1, 2025, from https://www.tensorflow.org/text/tutorials/word2vec

Girsang, A. S. (2021, December 31). *Word Embedding dengan Word2vec*. mti.binus.ac.id. Retrieved January 1, 2025, from https://mti.binus.ac.id/2020/11/17/word-embedding-dengan-word2vec/

Word representations · fastText. (n.d.). fastText. Retrieved January 1, 2025, from https://fasttext.cc/docs/en/unsupervised-tutorial.html

Explain the copcept and working of random forest model. (2023, October 8). AIML.com. Retrieved January 1, 2025, from https://aiml.com/what-is-random-forest-2/

Shitao, Zhao., Jianqiang, Sun., Kentaro, Shimizu., Koji, Kadota. (2018). Silhouette Scores for Arbitrary Defined Groups in Gene Expression Data and Insights into Differential Expression Results. Biological Procedures Online, 20(1):5-5. doi: 10.1186/S12575-018-0067-8

elfanmauludi. (2023, December 30). *Memahami Metode Silhouette Score dalam Analisis Klastering*. penelitian.id. Retrieved January 1, 2025, from https://www.penelitian.id/2023/12/memahami-metode-silhouette-score-dalam.html

O., Matsuga., V., S., Sheremet. (2023). Оцінювання оптимальної кількості кластерів для методу k-середніх на основі кусково-лінійної регресії з одним вузлом. Aktual′nì problemi avtomatizacìï ta ìnformacìjnih tehnologìj, doi: 10.15421/432302

Confusion Matrix. (n.d.). ScienceDirect. https://www.sciencedirect.com/topics/engineering/confusion-matrix#:~:text=A%20confusion%20matrix%20is%20a,performance%20of%20a%20classification%20algorithm.

Kontribusi

1. Muhammad Faiz Munif Billah NIM. 23031554028

   Scraping dan pre processing data website kompas.com, Feature Engineering, Modeling, Similarity, Laporan.

2. Ibrahim Frosly Alesandro NIM. 23031554021

   Scraping dan pre processing data website detik.com, Modeling, Feature Engineering, Laporan, Try data test, Clustering.

3. Gesang Nur Zamroji NIM. 23031554145

   Scraping dan pre processing data website CNN Indonesia, Modeling, Feature Engineering, Laporan, PPT, analisis dan visualisasi hasil

link presentasi : https://youtu.be/fBMINF8uveQ