

PENDEKATAN MULTIVARIAT DALAM MENDETEKSI



BERITA PALSU DENGAN DATA NUMERIK

Final Project Multivariate Analysis

KELOMPOK

D*** P***

(xxxxxx)

N*** V***

(xxxxxx)

Ririn Ayuning

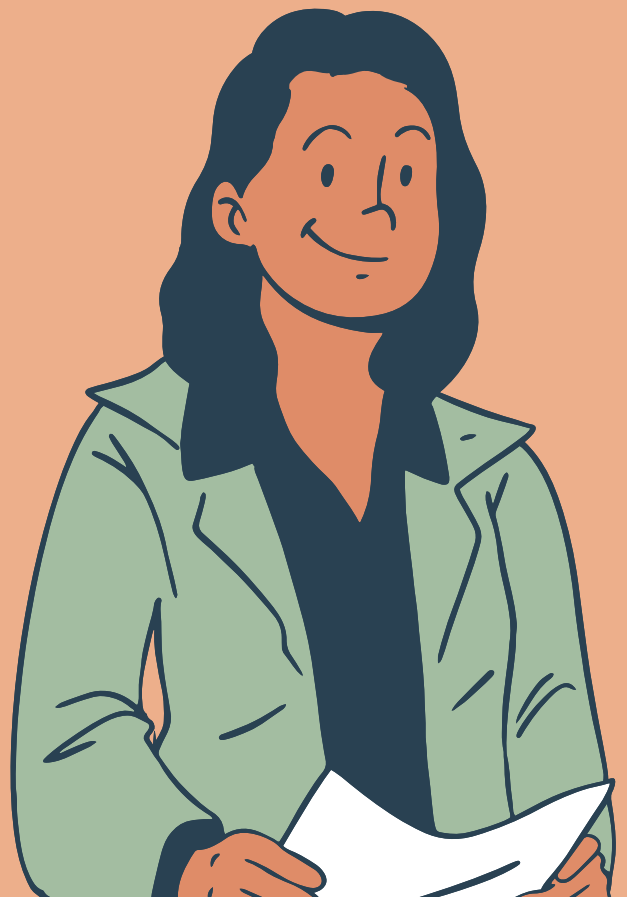
(xxxxxx)

S*** K***

(xxxxxx)



PENDAHULUAN



LATAR BELAKANG

Di era digital yang berkembang pesat, informasi dapat disebar dengan sangat cepat melalui media sosial, situs web, dan aplikasi pesan instan.

Tantangan Baru

- Penyebaran berita palsu (*fake news*)
- Berita palsu dibuat untuk menyesatkan atau menipu pembaca
- Berita palsu mempengaruhi opini publik, menciptakan sensasi, atau keuntungan finansial.

Faktor Penyebaran

- Tingkat penetrasi pengguna internet tinggi (221 juta pengguna di Indonesia pada 2024, menurut APJII).
- Berita palsu mudah tersebar di linimasa media sosial.
- Judul provokatif dan bombastis mempengaruhi pengguna untuk menyebarkannya.

LATAR BELAKANG

Upaya deteksi berita palsu penting dalam hal menjaga integritas informasi yang diterima oleh masyarakat.

Pendekatan Penelitian

- Pendekatan multivariat dengan data numerik.
- Menggunakan bahasa pemrograman R.
- Menggunakan dataset berisi statistik ringkasan artikel berita

Tujuan Penelitian

- Mengembangkan metode deteksi berita palsu yang efektif dan efisien.
- Meningkatkan kesadaran masyarakat akan verifikasi informasi.

CONTOH FAKE NEWS

Dwi Hartanto si Jenius 'Penerus BJ Habibie' Ternyata Pembohong

"Saya sangat berharap dibukakan pintu maaf yang sebesar-besarnya," tulis Dwi dalam surat pernyataan.

Reza Gunadha | [Suara.Com](#)

Senin, 09 Oktober 2017 | 08:49 WIB



Sumber:
[Suara.Com](#)

METHODOLOGI



METODOLOGY

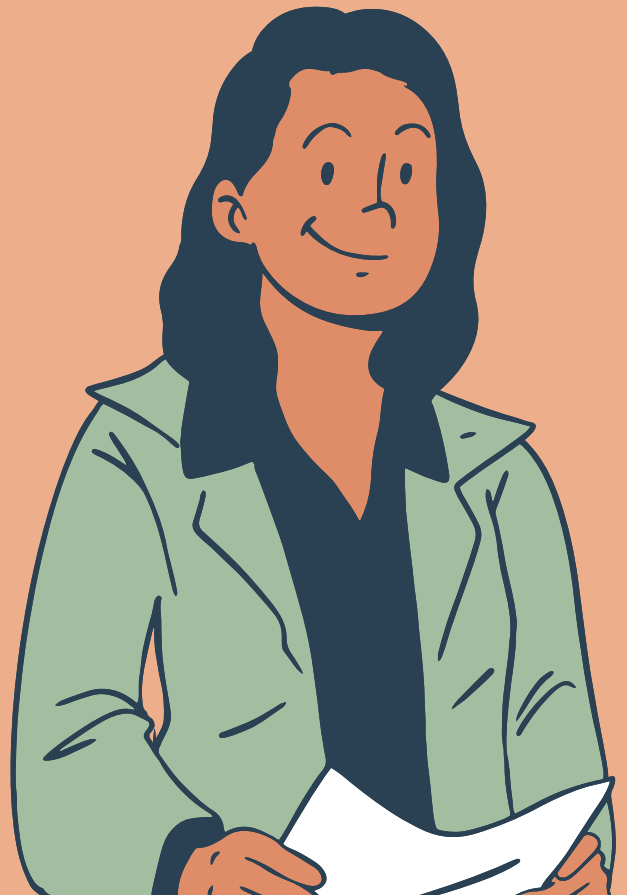
Sumber: Kaggle


Metode: Clustering Kmeans

Langkah Metodologi:

1	2	3	4	5	6	7	8	9
Pengump ulan Data	Eksplorasi dan Persiapan Data	Seleksi dan Normali sasi Fitur	Analisis Cluster	Inisialisas i dan Iterasi Pengelom pokan	Konvergensi	Penambaha n Informasi Cluster ke Data Asli	Visualisa si dan Analisis Hasil Cluster	Interpret asi Hasil

DATA





DATA

Data Overview

Data : Kumpulan artikel berita

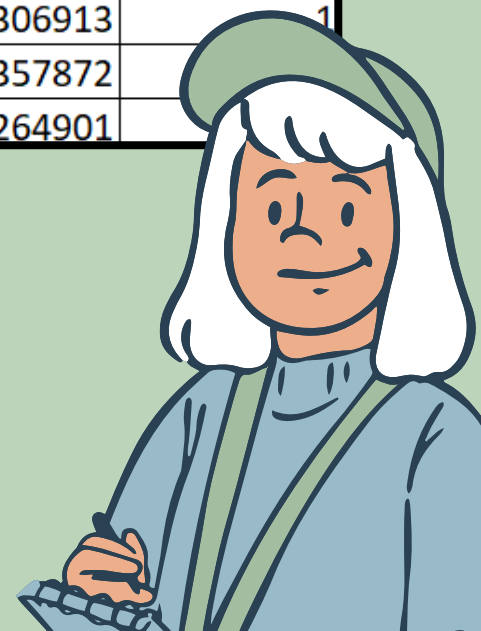
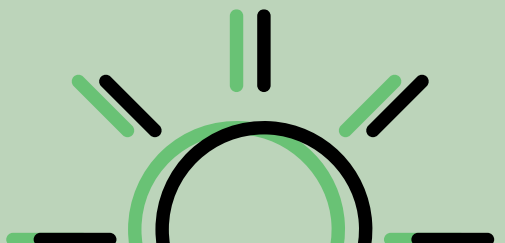
Fitur (6) :

- ID unik (4500)
- Jumlah kata dalam artikel (Word_Count)
- Jumlah kalimat artikel (Number_of_Sentence)
- Jumlah kata unik (Unique_Words)
- Panjang rata-rata kata (Average_Word_Length)
- label: nyata (1) atau palsu (0)

Gambaran data asli :

ID	Word_Count	Number_of_Sentence	Unique_Words	Average_Word_Length	Label
1606	10	4	24	6.176749717	1
3718	10	8	25	5.826769957	1
2634	10	7	18	4.619039519	1
5560	10	6	18	4.961423797	1
7494	10	4	21	4.114323585	1
3159	39	5	16	4.582873768	1
7232	11	4	21	5.756046145	1
7509	11	5	21	4.502689886	1
1509	11	6	24	3.943671991	1
1657	11	8	19	5.41906807	1
4128	11	7	18	5.354000533	1
5537	11	6	21	6.423073953	1
7550	11	4	18	4.092476922	1
7617	11	9	24	5.636306913	1
4515	69	7	15	5.701357872	1
7118	12	7	22	3.173264901	1

Data Source : Kaggle



DATA

DATA PREPARATION

- **Cleaning Data:** Menghilangkan data tidak masuk akal yaitu Unique_Words lebih besar dari Word_Count

- **Jumlah** Data Setelah Cleaning: **3.793**

- **Kode yang digunakan :**

```
df1 <- df %>% filter(Word_Count >= Unique_Words)
df1
```

Gambaran data setelah cleaning :

	ID	Word_Count	Number_of_Sentence	Unique_Words	Average_Word_Length	Label
1	3159	39	5	16	4,582873768	1
2	4515	69	7	15	5,701357872	1
3	6937	85	5	25	4,170042259	1
4	7726	94	5	25	6,956861797	1
5	5221	27	5	23	4,194186258	1
6	8510	28	5	22	6,647107331	1
7	1581	15	8	15	4,674833191	1
8	6974	42	4	19	6,223445894	1
9	2481	15	8	15	5,491922321	1
10	9659	38	7	20	4,856272067	1
11	5549	32	8	25	5,660075699	1
12	5804	36	9	16	4,314407419	1
13	6749	62	7	18	3,951185584	1
14	7631	24	6	20	5,089488698	1
15	9328	36	9	17	5,159520423	1
16	4431	24	6	16	3,941799914	1
17	6092	36	9	25	4,119540872	1
18	1797	39	7	18	6,092643063	1
19	8080	16	8	15	3,52340583	1
20	8382	55	7	20	5,284813457	1



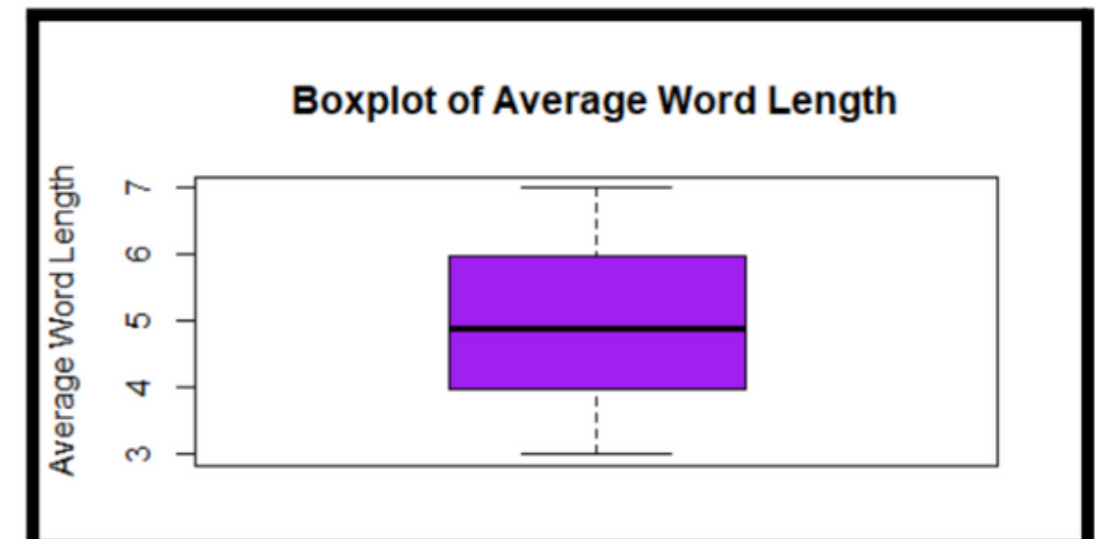
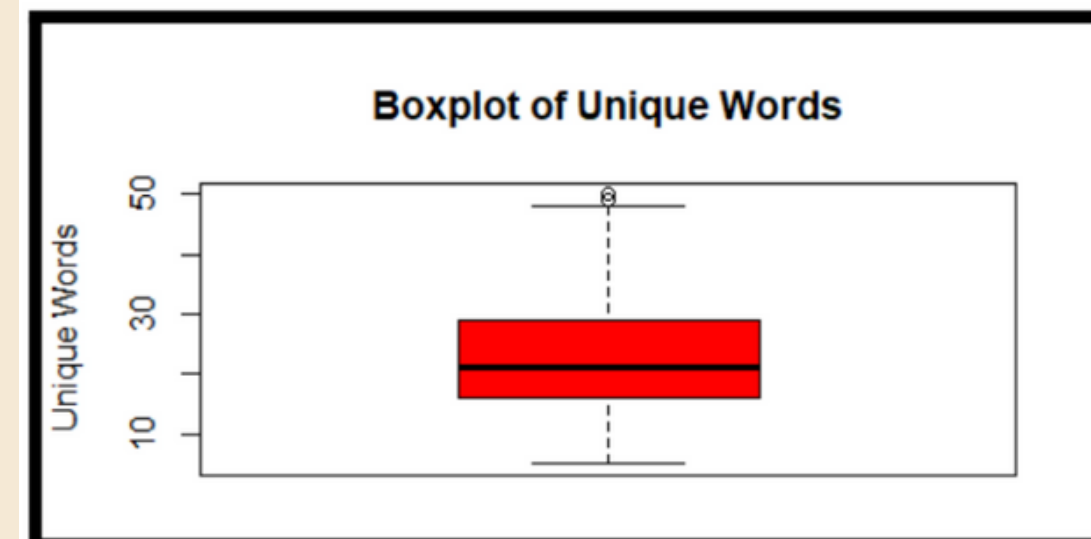
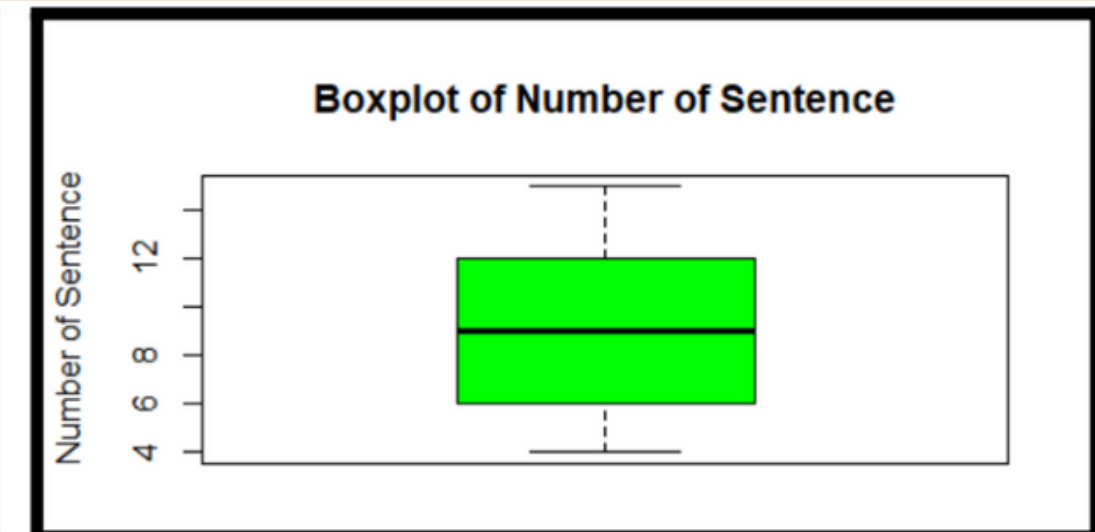
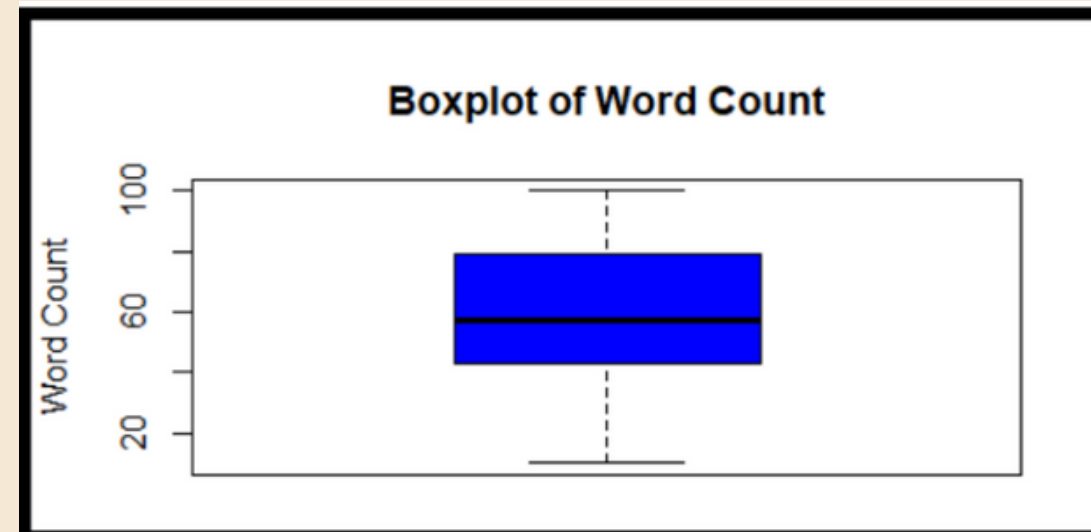


DATA

Descriptive Analytics

- Terdapat outlier pada kolom Unique_Words
- Outlier tidak signifikan mempengaruhi hasil analisis
- Outlier tidak jauh berbeda dengan data lainnya
- **Keputusan** : Data outlier tetap digunakan dalam analisis karena dianggap penting

ID	Word_Count	Number_of_Sentence	Unique_Words
Min. :1002	Min. : 10.00	Min. : 4.000	Min. : 5.00
1st Qu.:3201	1st Qu.: 43.00	1st Qu.: 6.000	1st Qu.:16.00
Median :5411	Median : 57.00	Median : 9.000	Median :21.00
Mean :5443	Mean : 59.91	Mean : 8.962	Mean :23.19
3rd Qu.:7655	3rd Qu.: 79.00	3rd Qu.:12.000	3rd Qu.:29.00
Max. :9999	Max. :100.00	Max. :15.000	Max. :50.00
Average_Word_Length		Label	
Length:3793		Min. :0.0000	
Class :character		1st Qu.:0.0000	
Mode :character		Median :0.0000	
		Mean :0.3599	
		3rd Qu.:1.0000	
		Max. :1.0000	



ANALISIS HASIL





CLUSTERING

- Pemilihan Fitur : hanya variabel-variabel yang berkontribusi signifikan terhadap pembentukan cluster yang digunakan

```
# Select only the relevant features for clustering  
clustering_data <- df1 %>%  
  select(Word_Count, Number_of_Sentence, Unique_Words, Average_Word_Length)
```

- Normalisasi data : memastikan bahwa setiap fitur memiliki skala yang sama dan tidak ada fitur yang mendominasi hasil clustering.

```
# Normalize the data  
clustering_data_scaled <- scale(clustering_data)
```

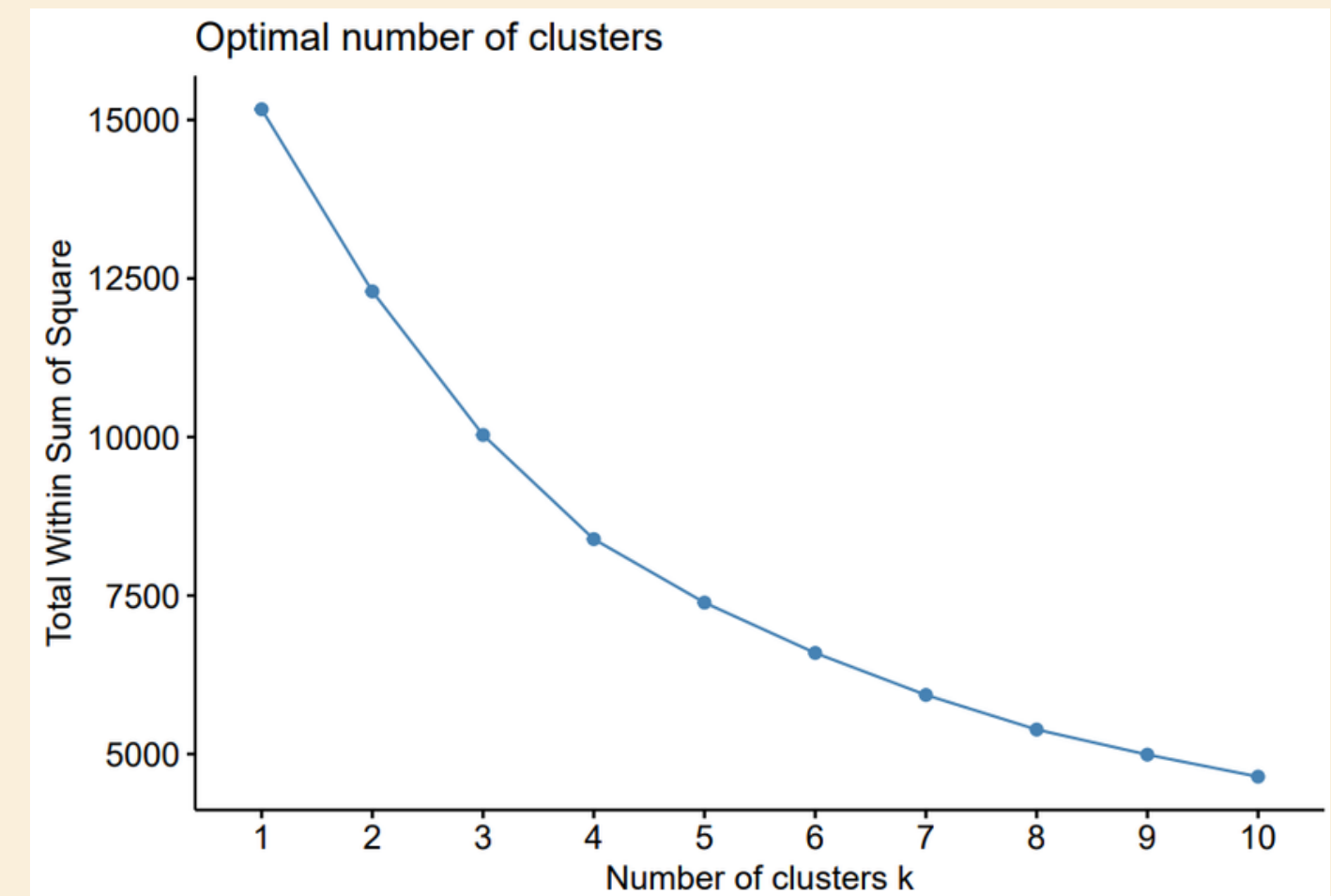
ELBOW METHOD

Digunakan untuk menentukan jumlah cluster optimal.



Determine the optimal number of clusters using the Elbow method
`fviz_nbclust(clustering_data_scaled, kmeans, method = "wss")`

- Evaluasi bagaimana variasi dalam data berkurang dengan menambah jumlah cluster.
- Grafik menunjukkan titik "elbow" di mana penambahan cluster tidak lagi mengurangi variasi signifikan.



Jumlah cluster optimal adalah 3, karena menambah lebih banyak cluster setelah titik ini tidak banyak mengurangi WSS.



K-MEANS

- Tiga cluster dipilih sebagai jumlah optimal

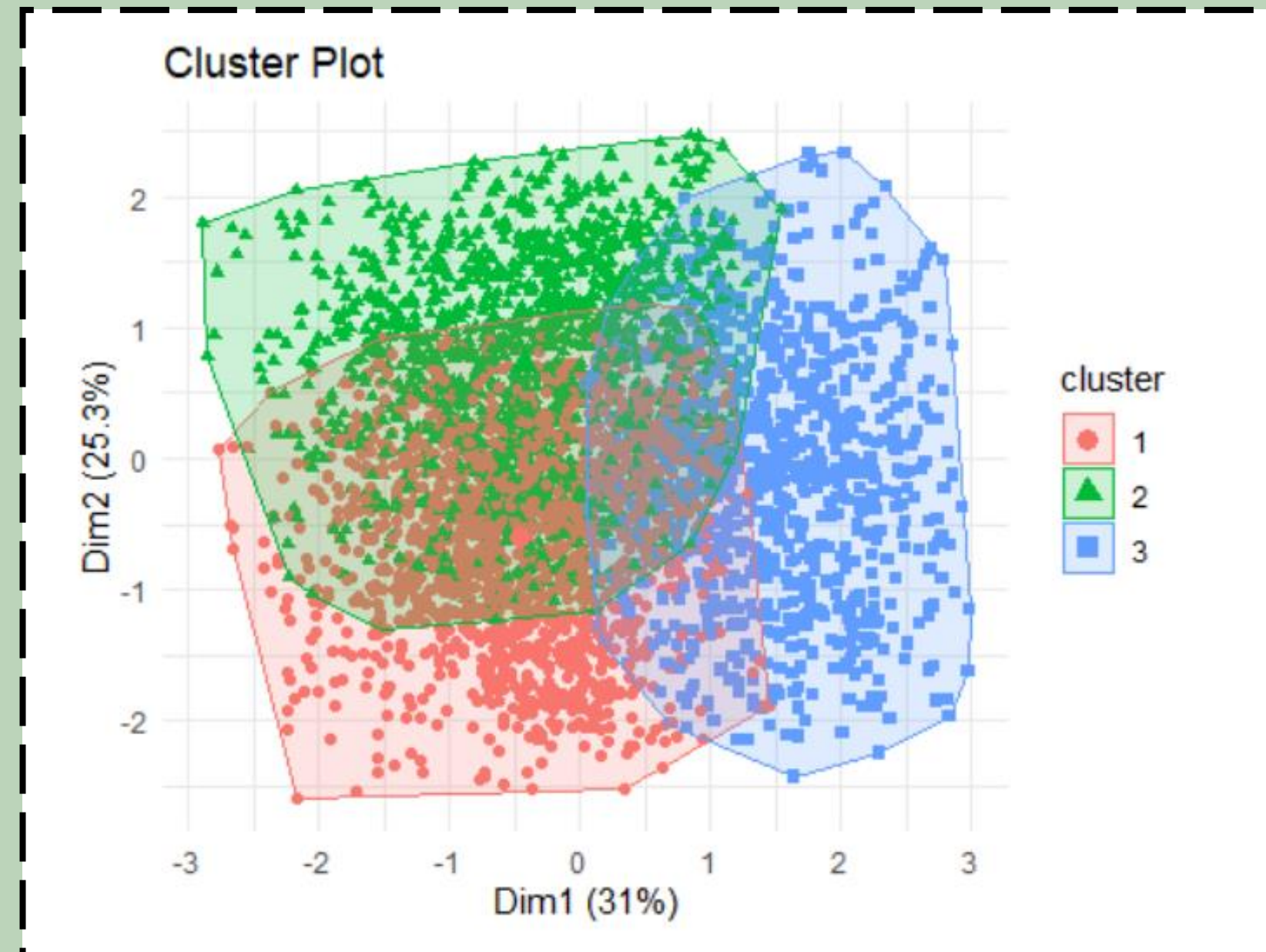
```
# From the plot, choose an optimal number of clusters, say 3
set.seed(123)
kmeans_result <- kmeans(clustering_data_scaled, centers = 3, nstart = 25)
```

```
# Add the cluster assignments to the original data
df1$Cluster <- kmeans_result$cluster
```

- Menambahkan kolom baru bernama "Cluster" ke data asli df1
- Memungkinkan pemahaman yang lebih baik tentang bagaimana data terbagi secara alami berdasarkan fitur-fitur yang dianalisis

VISUALISASI CLUSTER

Melihat hasil visualisasi dari clustering yang sudah dibuat. Hal ini membantu memahami bagaimana data terbagi menjadi kelompok-kelompok berbeda berdasarkan fitur-fitur yang dipilih



CLUSTER SUMMARY

Memberikan gambaran umum tentang karakteristik rata-rata dari setiap cluster. Berguna untuk memahami perbedaan dan kesamaan antar cluster, serta untuk mengevaluasi kualitas dan interpretasi cluster yang terbentuk.

```
## # A tibble: 6 x 7
##   Cluster Label Word_Count Number_of_Sentence Unique_Words Average_Word_Length
##   <int> <int>      <dbl>          <dbl>          <dbl>          <dbl>
## 1       1       0      56.3            9.32           16.8           3.99
## 2       1       1      54.5            8.15           19.8           3.98
## 3       2       0      57.5            9.21           17.6           6.05
## 4       2       1      55.8            7.95           20.0           6.01
## 5       3       0      72.7            9.85           39.4           4.90
## 6       3       1      92.6            7.14           24.7           4.83
## # i 1 more variable: Count <int>
```

DISTRIBUTION TABLE

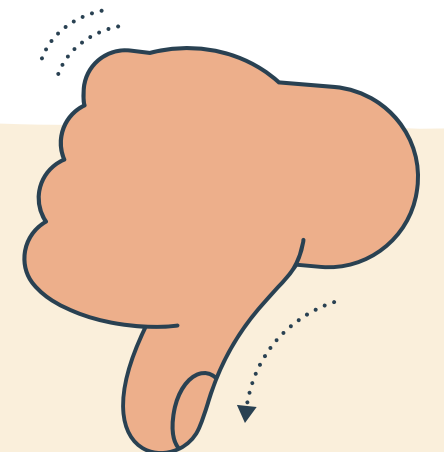
Untuk memahami lebih dalam mengenai bagaimana distribusi berita palsu (0) dan berita asli (1) dalam cluster-cluster yang terbentuk

	0	1
1	819	696
2	749	662
3	860	7

Cluster 1 dan cluster 2 memiliki distribusi yang **cukup seimbang** antara berita palsu dan berita fakta, meskipun masih terdapat sedikit **lebih banyak berita palsu** di masing-masing cluster.



Cluster 3 **didominasi** oleh **berita palsu** dengan distribusi data yang termasuk berita palsu **jauh lebih banyak** dibanding data yang termasuk kategori berita fakta..



KORELASI

	Word_Count	Number_of_Sentence	Unique_Words	Average_Word_Length	Label
Word_Count	1.000000000	0.001592169	0.236197820	0.002657383	-0.15382714
Number_of_Sentence	0.001592169	1.000000000	0.046540708	-0.010757768	-0.20187216
Unique_Words	0.236197820	0.046540708	1.000000000	-0.007189879	-0.22562818
Average_Word_Length	0.002657383	-0.010757768	-0.007189879	1.000000000	0.01049961
Label	-0.153827143	-0.201872155	-0.225628177	0.010499610	1.000000000

- **Word_Count** memiliki korelasi negatif dengan label (-0.1538) menunjukkan artikel dengan lebih banyak kata sedikit cenderung asli.
- **Number_of_Sentence** berkorelasi negatif juga dengan label (-0.2019) yang mana berarti artikel dengan lebih banyak kalimat sedikit cenderung asli.
- **Unique_Words** berkorelasi negatif juga dengan label (-0.2256) berarti artikel dengan lebih banyak kata unik sedikit cenderung asli.
- **Average_Word_Length** memiliki korelasi sangat rendah dengan label (0.0105), walau positif tapi hampir tidak ada hubungan dengan keaslian artikel.

NOTES: KORELASI ANTARA FITUR-FITUR DAN LABEL CENDERUNG RENDAH. FITUR-FITUR INI SENDIRI MUNGKIN TIDAK CUKUP KUAT UNTUK MEMPREDIKSI APAKAH SEBUAH ARTIKEL ADALAH ASLI ATAU PALSU.

KESIMPULAN



KESIMPULAN

- Pendekatan multivariat dengan metode clustering K-Means berhasil mengelompokkan artikel berita berdasarkan kesamaan fitur numerik.
- Proses clustering menunjukkan bahwa tiga cluster adalah jumlah optimal.
- Cluster 1 dan Cluster 2: Distribusi yang seimbang antara berita palsu dan asli, meskipun sedikit lebih banyak berita palsu.
- Cluster 3: Didominasi oleh berita palsu.
- Word_Count: Memiliki hubungan positif dengan Unique_Words.
- Unique_Words dan Number_of_Sentence: Memiliki korelasi negatif yang lemah dengan Label, mengindikasikan bahwa artikel dengan lebih banyak kata unik dan kalimat cenderung lebih mudah diidentifikasi sebagai berita palsu.
- Artikel berita palsu cenderung memiliki jumlah kata unik yang lebih tinggi dan struktur yang lebih kompleks dibandingkan dengan artikel
- Model clustering K-Means terbukti efektif dalam mengelompokkan artikel berita ke dalam cluster yang berbeda berdasarkan fitur numerik yang relevan.
- Memberikan wawasan berharga mengenai pola dan karakteristik umum dari berita palsu dan asli.
- Pendekatan multivariat menggunakan metode clustering K-Means menunjukkan potensi besar dalam mengidentifikasi dan mengelompokkan artikel berita berdasarkan karakteristik numeriknya.
- Diharapkan memberikan kontribusi dalam pengembangan metode deteksi berita palsu yang lebih canggih dan efektif.
- Meningkatkan kesadaran masyarakat akan pentingnya verifikasi informasi sebelum mempercayai dan menyebarkannya.



DAFTAR PUSTAKA

Afrizal Maulana, M., Sandi Yuda, M., & Yulianti, E. (2022). *SYNERGY Jurnal Ilmiah Multidisiplin KEPERCAYAAN MASYARAKAT TERHADAP BERITA PALSU/HOAX DI FACEBOOK PADA PILPRES (Studi Fenomenologis Pada Masyarakat Kota Sukabumi)*. 1(1), 293–300. <https://e-journal.naurendigiton.com/index.php/sjim>

Eric Kunto Aribowo - *Menelusuri Jejak Hoax.pdf*. (n.d.).

Fake News Detection Data (April 2024).

<https://www.kaggle.com/datasets/tasnimmiger/fake-news-detection-data>

Reza Gunadha. (2017, October 9). *Dwi Hartanto si Jenius “Penerus BJ Habibie” Ternyata Pembohong*. Suara.Com.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, 19. <https://doi.org/10.1145/3137597.3137600>





Terima Kasih