Jurnal Teknik Informatika (JUTIF)
Vol. 5, No. 2, April 2024, pp. 539-550

DOI: https://doi.org/10.52436/1.jutif.2024.5.2.1635
p-ISSN: 2723-3863
e-ISSN: 2723-3871

COMBINATION K-MEANS AND LSTM FOR SOCIAL MEDIA BLACK CAMPAIGN DETECTION OF INDONESIA PRESIDENTIAL CANDIDATES 2024

Wisnu Priambodo[*1], Eri Zuliarso[2]

[1,2]Master of Information Technology, Faculty of Information Technology and Industry, Universitas Stikubank, Indonesia
Email: [1]wisnupriambodo0016@mhs.unisbank.ac.id, [2]eri299@edu.unisbank.ac.id

(Article received: December 24, 2023; Revision: January 15, 2023; published: April 15, 2024)

Abstract

Social media has become the main platform for the public and political figures to voice opinions and run political campaigns. Despite its positive impact, social media also has negative impacts, particularly in the spread of Black Campaigns. This phenomenon has become critical, especially about the 2024 elections in Indonesia that target presidential candidates. Black campaigns can trigger conflict and damage the image of presidential candidates in the eyes of the public. Therefore, it is important to detect black campaigns against presidential candidates. This research develops a Black Campaign detection model using the K-means clustering algorithm and the Long Short-Term Memory (LSTM) approach. K-means is implemented to cluster text data on Twitter social media, while LSTM is used to learn word order patterns and detect text. The result is that K-means can effectively prepare the data, and classification using LSTM shows an accuracy of 90.28%. The comparison with Ensemble Learning classification model achieved an accuracy of 94.31%. Evaluation involved accuracy, precision, recall, and F1-score, with the result that Ensemble Learning was slightly superior in the evaluation matrix. However, compared to Ensemble Learning, LSTM has an advantage in understanding word order, which can be achieved by utilizing the advantages of Deep Learning Recurrent Neural Network architecture. Testing on sample data shows the similarity between LSTM and Ensemble Learning models in detecting Black Campaigns on Twitter social media post text data.

Keywords: *Black Campaign, Detection, K-Means, LSTM.*

KOMBINASI K-MEANS DAN LSTM UNTUK DETEKSI BLACK CAMPAIGN DI MEDIA SOSIAL PADA CALON PRESIDEN INDONESIA 2024

Abstrak

Media sosial telah menjadi *platform* utama bagi publik dan tokoh politik untuk menyuarakan opini dan kampanye politik. Meskipun memiliki dampak positif, media sosial juga membawa dampak negatif, khususnya dalam penyebaran Kampanye Hitam. Fenomena ini menjadi kritis, terutama terkait pemilu 2024 di Indonesia yang menargetkan calon presiden. Kampanye Hitam dapat memicu konflik dan merusak citra calon presiden di mata masyarakat. Oleh karena itu, deteksi kampanye hitam terhadap calon presiden menjadi penting untuk dilakukan. Penelitian ini mengembangkan model deteksi Kampanye Hitam menggunakan pendekatan algoritma *clustering K-means* dan *Long Short-Term Memory (LSTM)*. *K-means* diimplementasikan untuk mengelompokkan data teks di media sosial *Twitter*, sedangkan *LSTM* digunakan untuk mempelajari pola urutan kata dan mendeteksi teks. Hasilnya melalui *K-means* dapat mempersiapkan data secara efektif dan klasifikasi menggunakan *LSTM* menunjukkan akurasi sebesar 90.28%. Perbandingan dengan Model klasifikasi *Ensemble Learning* mencapai akurasi 94.31%. Evaluasi melibatkan akurasi, presisi, *recall*, dan *F1-score*, dengan hasil *Ensemble Learning* sedikit unggul dalam matrik evaluasi. Meski demikian, dibandingkan *Ensemble Learning*, *LSTM* memiliki keunggulan dalam memahami urutan kata, hal ini dapat dicapai dengan memanfaatkan kelebihan arsitektur *Deep Learning Recurrent Neural Network*. Pengujian pada sampel data menunjukkan kemiripan antara model *LSTM* dan *Ensemble Learning* dalam mendeteksi Kampanye Hitam pada data teks postingan sosial media *Twitter*.

Kata kunci: *Deteksi, K-Means, Kampanye Hitam, LSTM.*

1. PENDAHULUAN

Media sosial yang fleksibel telah menjadi fasilitas publik, tokoh politik, dan institusi dalam

menyampaikan pandangan politik dan informasi. Dengan kemampuan untuk mencapai audiens yang lebih besar dan tidak dibatasi oleh status sosial, ekonomi dan politik yang ada di masyarakat, penyebaran informasi politik di media sosial dapat membentuk opini publik dan menjadi strategi penting dalam kampanye politik modern [1], [2]. Pemilihan Presiden Indonesia yang dijadwalkan akan dilaksanakan pada tahun 2024 menjadi momen penting dalam proses demokrasi di indonesia [3], [4]. Meskipun demikian, dalam konteks politik modern, muncul fenomena yang dikenal sebagai "*black campaign*" atau kampanye hitam dapat menjadi ancaman dalam proses demokrasi.

Fenomena *black campaign* menjadi perhatian serius, mengingat penyebaran berita palsu atau fitnah dapat merusak reputasi. Selain dapat melanggar norma, perilaku penghinaan dan pelecehan dapat mempengaruhi citra calon presiden [5], [6], [7], [8]. Oleh karena itu, analisis *Black campaign* saat ini telah menjadi suatu kebutuhan di masyarakat, karena tersebarnya fitnah memiliki potensi untuk menimbulkan kegaduhan dan kekhawatiran di kalangan masyarakat [9], [10], [11]. Dalam hal ini, *Black campaign* dapat dengan mudah menyebar dan mempengaruhi proses demokrasi. Oleh karena itu, deteksi *Black campaign* dapat menjadi salah satu cara untuk mencegah hal tersebut. Namun, deteksi teks secara manual kurang efektif untuk diimplementasikan [12]. Sehingga dalam penelitian ini diusulkan model deteksi dengan mengombinasikan teknik pembelajaran mesin.

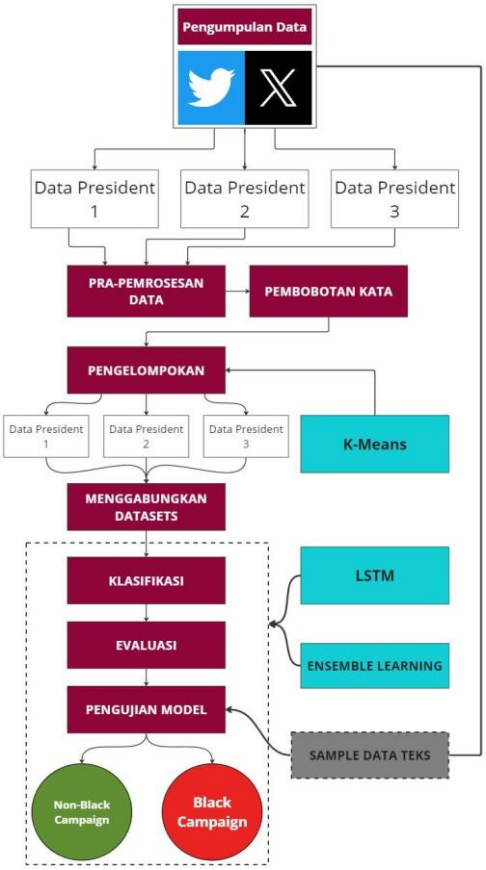
Salah satu teknik pembelajaran mesin dalam penelitian ini adalah Algoritma *K-means*, yang digunakan dalam analisis sosial media karena sederhana dan efisien [13]. *K-Means* digunakan untuk memilah data teks dari media sosial *Twitter* atau *X* terkait dengan Calon Presiden Indonesia 2024. Selain *K-means*, penelitian ini juga menggunakan model berbasis jaringan saraf yang semakin populer dan efektif [14]. Seperti yang ditunjukkan oleh beberapa penelitian, Salah satu model berbasis jaringan saraf yang populer adalah *Long Short-Term Memory (LSTM)*, yang memiliki kemampuan untuk memahami variasi dan kompleksitas dalam data teks [15], [16]. Oleh karena itu, *LSTM* dapat digunakan untuk mendeteksi konten kampanye hitam.

Studi sebelumnya oleh Wahyuni et al., 2023, mengeksplorasi penggunaan algoritma *K-Means* untuk meramalkan kelompok peserta pemilu dan mengidentifikasi daerah dengan tingkat abstain tinggi. Pada penelitian tersebut mengonfirmasi bahwa algoritma *K-Means* efektif untuk pengelompokan data [17].

Studi-studi lainnya juga mengeksplorasi penggunaan teknik-teknik analisis data yang inovatif. Muhariya et al., 2022, menggunakan algoritma *K-Means* untuk mengelompokkan data *cyberbullying* di Sosial Media, Hasil pengelompokan kata-kata *cyberbullying* di *Instagram* menghasilkan akurasi tertinggi, yaitu 67.38% [18]. Bisht et al., 2020, mengusulkan penggunaan *LSTM* dalam mendeteksi ujaran kebencian di *Twitter*, hasil klasifikasi mencapai akurasi sebesar 86%, menunjukkan *LSTM* mampu dalam memahami urutan kata yang kompleks [19]. Terakhir, Chang dan Masterson, 2019, menggunakan *LSTM* untuk mengklasifikasikan dokumen teks di bidang ilmu politik. *LSTM* digunakan untuk menangkap urutan antar kata dalam dokumen, dan dapat meningkatkan akurasi klasifikasi [20].

Berdasarkan beberapa penelitian terdahulu, algoritma *K-means* dipilih dan diimplementasikan untuk mengelompokkan teks serupa di media sosial. [18], [21], [22], dan *LSTM* dapat memberikan solusi yang efektif untuk mendeteksi pola dan struktur kata yang kompleks pada postingan di media sosial [20], [19], [23]. Dalam penelitian ini, perbandingan dengan metode klasifikasi *Ensemble Learning* dilakukan untuk mengevaluasi apakah penggabungan beberapa model klasifikasi memberikan peningkatan kinerja.

2. METODE PENELITIAN



Gambar 1. Langkah-langkah Metode Penelitian

Merujuk pada teknik yang digunakan untuk mencapai tujuan penelitian, yang ditunjukkan pada Gambar 1, penelitian ini dilakukan untuk menguji akurasi dan melihat kemampuan model dalam mendeteksi kampanye hitam terhadap calon presiden indonesia di sosial media *Twitter* berbasis kombinasi model *K-menas* dan *LSTM*. Adapun tahapan dalam

mencapai tujuan dari penelitian, adalah sebagai berikut.

2.1. Pengumpulan Data

Penelitian ini dimulai dengan pengumpulan data percakapan di media sosial *Twitter* terkait Pemilihan Presiden 2024, yang melibatkan tiga calon presiden yang sedang bersaing. Pengumpulan data dilakukan secara terpisah untuk setiap calon presiden, agar memudahkan proses pengelompokan dan penentuan label. Tahap awal menghubungkan ke *platform* sosial media dan penentuan kata kunci yang relevan untuk setiap calon presiden. Metode *Crawling* berbasis *Python* digunakan dalam mengumpulkan data dengan memanfaatkan *tool Tweet-Harvest*.

Pada proses *crawling* data, dilakukan secara aktif mencari, menyaring, dan mengumpulkan percakapan relevan dari media sosial *Twitter* berdasarkan topik pemilihan presiden di indonesia. Data yang terkumpul disimpan dalam format *Comma Separated Values (CSV)* untuk setiap calon presiden. Pemisahan data ke dalam tiga file *datasets* memberikan manfaat dalam memberikan hasil yang lebih akurat proses pengelompokan dan mengungkapkan pola yang lebih spesifik terkait dengan kemungkinan kampanye hitam dalam percakapan tersebut.

2.2. Pra-Pemrosesan Data

Tahap pra-pemrosesan bertujuan untuk membersihkan data dari gangguan atau *noise* sehingga tidak mempengaruhi hasil pengelompokan data menggunakan *K-means*. Setiap datasets yang diperoleh akan dilakukan tahapan preprocessing, tahapan ini dilakukan agar proses pengelompokan data dapat dilakukan dengan lebih efektif dan efisien [24], [25]. Terdapat beberapa tahapan dalam proses preprocessing data, yaitu:

1. *Case Folding*, Merupakan proses ini dilakukan dengan mengubah seluruh kalimat dalam data menjadi huruf kecil agar tidak terjadi perbedaan ketika proses pengolahan data [26], [27].
2. *Removal Punctuation*, merupakan langkah untuk menghapus mention,url,angka dan tanda baca yang ada pada teks postingan sosial media.
3. *Tokenizing*, Merupakan tahap yang dilakukan dengan memisahkan setiap kata dalam kalimat dan menghapus tanda baca, sehingga hanya tersisa kata-kata saja [26], [27].
4. *Stopword*, tahap ini untuk menghilangkan kata-kata yang dianggap tidak penting dalam data, seperti kata depan atau kata penghubung [9], [26].
5. *Normalization*, Adalah tahap yang dilakukan untuk mengubah kata-kata yang tidak standar menjadi lebih standar agar memudahkan proses analisis [28].
6. *Stemming*, Proses ini digunakan untuk mengubah kata-kata menjadi bentuk dasarnya agar lebih mudah dalam proses analisis [28].

2.3. Pembobotan Kata

Dalam penelitian ini melibatkan *TF-IDF (Term Frequency-Inverse Document Frequency)* untuk mengekstrak fitur-fitur penting dari data tekstual yang diperoleh dari *Twitter*. Metode ini digunakan untuk menghitung nilai dari *Term Frequency (TF)* yaitu seberapa sering kata tersebut muncul dalam dokumen tertentu dan *Inverse Document Frequency (IDF)* yaitu seberapa jarang kata tersebut muncul dalam seluruh dokumen yang ada [23], [29], [30].

TF-IDF memiliki keunggulan dalam model pembelajaran mesin dan Pemrosesan Bahasa Alami *(NLP)* untuk analisis teks, terutama ketika perhatian diberikan pada frekuensi kata-kata yang muncul [31]. Proses ini memfasilitasi identifikasi dan klasifikasi yang efisien atas informasi yang berpotensi menyesatkan. Dengan demikian, rumus untuk menghitung *TF-IDF* dari term *t* yang ada dalam dokumen *d* dijelaskan dalam Persamaan 1.

$$tf - idf\,(d,t) = tf(t) * idf(d,t) \qquad (1)$$

Secara keseluruhan Algoritma *TF-IDF* digunakan untuk menghasilkan vektor kalimat berdasarkan frekuensi kata yang muncul di dalam datasets setelah pra-pemrosesan data.

2.4. Pengelompokan

Komponen selanjutnya melibatkan pemanfaatan *K-means clustering* dalam pengelompokan data teks. *K-mean*s adalah algoritma machine learning tanpa pengawasan atau unsupervised learning yang mengelompokkan titik-titik data yang serupa ke dalam kelompok-kelompok berdasarkan karakteristiknya [17], [18], [30]. *K-means Clustering* mengelompokkan titik-titik data yang mirip, menjadikannya langkah kunci dalam pengenalan pola dan pengambilan keputusan [21]. Tahapan dalam penggunaan algoritma k-means adalah sebagai berikut:

1. Inisialisasi: Menentukan jumlah cluster (k) sebagai centroid awal.
2. *Assignment*: Hitung jarak antara setiap data dengan centroid menggunakan rumus *Euclidean Distance*. Setiap data akan di-assign ke cluster dengan centroid terdekat, Seperti yang ditunjukkan dalam Persamaan 2.

$$(x,y) = \sqrt{\Sigma_{i=1}^{n}(x_i - y_i)^2} \qquad (2)$$

x dan *y* adalah dua titik dalam ruang n-dimensi, dan *n* adalah jumlah dimensi.

3. *Update Centroid*: Setelah semua data di-assign ke cluster, dihitung rata-rata dari setiap cluster untuk mendapatkan centroid baru. Persamaan 3 merupakan Rumus *Update Centroid*.

$$(C) = \frac{1}{|C|}\Sigma_{x\epsilon C^2} \quad (3)$$

C adalah kluster, |*C*| adalah jumlah data point dalam kluster *C*, dan *x* adalah data point dalam kluster *C*. Dengan Metode k-means sebagai pendekatan dalam pengelompokan data teks dengan memanfaatkan pembobotan *vektor TF-IDF*. Proses *clustering* ini dapat ditentukan dua kelompok data untuk dijadikan label untuk komponen penghubung pada proses klasifikasi, tiga label tersebut antara lain *Non-Black Campaign*, dan *Black Campaign*, dimana dua label tersebut digunakan pada tahapan selanjutnya yaitu kasifikasi yang memerlukan data label untuk proses *train*, *validation* dan *test*.

2.5. Penggabungan Datasets

Tahapan ini melibatkan penggabungan tiga datasets pada tiap-tiap calon presiden hasil dari proses sebelumnya yaitu Pra-Pemrosesan Data, Pembobotan Kata, dan Pengelompokan menggunakan *K-means* ke dalam satu dataset yang lebih besar. Didalam proses ini juga dilakukan penggabungan duplikasi baris data dan penghapusan baris data kosong.

2.6. Klasifikasi

Klasifikasi data teks kampanye hitam dilakukan dengan menerapkan beberapa proses diantaranya.

2.6.1. Proses *Encoding*

Tahap awal dalam proses klasifikasi melibatkan representasi teks menggunakan metode *One-Hot Encoding*. Melalui pendekatan *One-Hot Encoding*, setiap kata dalam teks diwakili oleh vektor biner, setiap elemen menunjukkan keberadaan atau ketiadaan kata dalam dokumen teks [32].

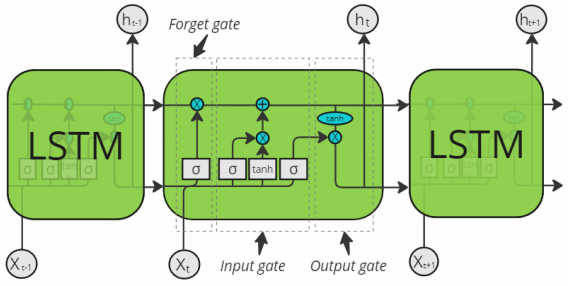
2.6.2. Distribusi Data

Proses distribusi data secara seimbang digunakan untuk memastikan model yang dilatih memiliki akurasi yang baik dan tidak bias [33], data dibagi menjadi data *Validation*, *Training*, dan *Testing*.

2.6.3. *LSTM*

Selanjutnya, proses pelatihan model menggunakan algoritma *Long Short-term Memory Networks (LSTM)*. *LSTM* adalah sebuah jenis jaringan syaraf tiruan yang dikenal dengan kemampuannya untuk menganalisis data berurutan [19], [20]. Selain itu, *LSTM* merupakan metode *Recurrent Neural Network* (*RNN)* yang menambahkan sel memori agar dapat menyimpan informasi untuk waktu yang lama. Tujuannya adalah untuk mengatasi masalah vanishing gradient saat memproses data urutan yang panjang pada *RNN* [29].

Struktur dari algoritma *LSTM* terdiri dari jaringan saraf dan beberapa unit memori yang berbeda yang disebut sebagai sel. Sel ini memiliki dua bagian penting, yaitu state sel dan hidden state, yang akan diteruskan ke selanjutnya. Gambar 2 menggambarkan ilustrasi dari sel *LSTM*.



Gambar 2. Ilustrasi *LSTM*

Algoritma *LSTM* mengumpulkan informasi data berlabel *Non-Black Campaign*, dan *Black Campaign* yang diperoleh pada proses *clustering K-Means*, kemudian disimpan dalam sel *(cell)*, sementara manipulasi memori dilakukan oleh sebuah komponen yang dikenal sebagai gate. Terdapat tiga jenis gate yang ada dalam algoritma LSTM, yaitu *Forget gate*, *Input gate*, dan *Output gate*. Sehingga dengan *LSTM* ini dapat digunakan sebagai solusi untuk mengatasi *vanishing gradient*. Model *LSTM* dilatih menggunakan data yang telah diberi label hasil dari proses *clustering*.

2.6.4. *Ensemble Learning*

Model *LSTM* yang telah dibangun akan dibandingkan berdasarkan evaluasi dan deteksi sampel dengan algoritma *Ensemble Learning*, *Ensemble Learning* merupakan teknik penggabungan model klasifikasi yang berbeda dalam satu model untuk meningkatkan kinerja dibandingkan dengan pengklasifikasi tunggal [34]. Gabungan algoritma machine learning yang dipakai dalam penelitian ini diantaranya *Naive Bayes*, *Random Forest*, dan *Support Vector Machine (SVM)*, dengan menggunakan teknik *Voting* [35].

2.7. Evaluasi

Evaluasi model bertujuan mengukur kinerja model yang telah dibangun, memberikan pemahaman tentang kemampuan model dalam mendeteksi teks. Hal ini membantu memahami sejauh mana model mampu menggeneralisasi dari data pelatihan ke data yang belum pernah dilihat sebelumnya. Tahapan evaluasi model meliputi:

1. *Confusion Matrix:* Merupakan tabel yang memperlihatkan hasil prediksi model dalam bentuk matriks. Matriks ini memuat empat indikator yaitu *True Positive (TP)*, *False Negative (FN)*, *False Positive (FP)*, dan *True Negative (TN)*. Dengan *confusion matrix*, dapat dilihat sejauh mana model mampu memprediksi dengan benar dan salah untuk setiap kelas [36] [37].

2. *Accuracy* (Akurasi)**:** Akurasi mengukur berapa persen data yang diklasifikasikan dengan benar dari keseluruhan data yang dievaluasi [38], [39]. Akurasi dari sebuah model secara bergantung pada berbagai indikator penting seperti Jumlah *True Positive (TP)* dan *True Negative (TN)* dihitung dalam skala rasio dengan jumlah agregat *TP, TN, False Positive (FP)*, dan *False Negative (FN)*. Seperti yang dinyatakan dalam Persamaan 4.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

3. *Precision* (Presisi): Presisi memberikan informasi tentang seberapa baik model membedakan antara kelas positif dan kelas negatif [38], [39] . Precision dihitung dengan membagi jumlah prediksi benar positif dengan total prediksi positif. Persamaan 5 menunjukkan presisi yang merupakan rasio *TP* terhadap *TP* dan *FP*.

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

4. *Recall* (Sensitivitas): *Recall* dihitung dengan membagi jumlah prediksi benar positif dengan total jumlah contoh positif yang sebenarnya [38], [39]. Ini mengindikasikan jumlah total prediksi yang benar yang dilakukan oleh algoritma. Ini merupakan perbandingan antara *True Positive (TP)* dengan jumlah keseluruhan *True Positive* dan *False Negative (FN)* yang diakumulasikan. Seperti yang ditunjukkan dalam Persamaan 6.

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

5. *F1-score*: Adalah *harmonic mean* (rata-rata harmonik) dari precision dan *recall*. *F1-score* memberikan pengukuran gabungan antara *precision* dan *recall*, dalam mengevaluasi kinerja suatu model klasifikasi. Metrik ini dapat digunakan dalam kelas-kelas yang dievaluasi memiliki distribusi yang tidak seimbang *(imbalanced)*. [38], [39]. Seperti yang dinyatakan dalam Persamaan 7.

$$F1 - score = \frac{2 * TP}{2 * TP + FP + FN} \quad (7)$$

Dalam evaluasi model, *confusion matrix*, akurasi, *precision*, *recall*, dan *F1-score* memberikan informasi kinerja model dalam mengklasifikasikan data. Hal ini membantu memahami kekuatan dan kelemahan model.

2.8. Pengujian Model

Pada tahapan ini model akan diuji untuk mendeteksi teks yang terindikasi terindikasi kampanye hitam. Data diambil dari sosial media Twitter. Data ini belum digunakan untuk membangun model. Tujuan dari pengujian ini adalah untuk menunjukkan hasil analisis yang dilakukan oleh model Klasifikasi *LSTM* dalam memprediksi konten teks baru yang belum dikenali oleh Model,Selain itu pengujian model juga dilakukan dengan model *Ensemble Learning* sebagai pembanding.

3. HASIL DAN PEMBAHASAN

Proses pengambilan data twitter dimulai dari Januari 2023 sampai dengan September 2023 menggunakan *tool tweet-harvest*, dengan jumlah masing-masing data yaitu data calon president-1 berjumlah 2520, data calon president-2 berjumlah 2099, data calon president-3 berjumlah 4001, total keseluruhan adalah 8620 data. Selanjutnya tahapan-tahapan pemrosesan analisis data menggunakan *platform cloud Google Colab* dengan bahasa pemrograman *Python*.

3.1. Pra-Pemrosesan Data

Untuk mempercepat dan menyederhanakan pengelompokan, data *tweet* yang telah diperoleh akan dilakukan pra-pemrosesan data. Beberapa tahapan pra-pemrosesan data ditunjukkan pada Tabel 2.

Table 2. Alur Pra-Pemrosesan Data

Proses	Data Teks
Teks Awal	Gosah sok peduli ganjar, @PDI_Perjuangan kritik food estate krna MENTERI yg di tugaskan pak jokowi tdk BECUS bekerja namanya PRABOWO SUBIANTO. Buat yg paham ajah @adearmando61 bersebrangan dg PDIP krn @psi_id condong ke GERINDRA makanya suka melakukan serangan ke pengusung ganjar
Case Folding	gosah sok peduli ganjar, @pdi_perjuangan kritik food estate krna menteri yg di tugaskan pak jokowi tdk becus bekerja namanya prabowo subianto. buat yg paham ajah @adearmando61 bersebrangan dg pdip krn @psi_id condong ke gerindra makanya suka melakukan serangan ke pengusung ganjar
Removal Punctuation	gosah sok peduli ganjar perjuangan kritik food estate krna menteri yg di tugaskan pak jokowi tdk becus bekerja namanya prabowo subianto buat yg paham ajah bersebrangan dg pdip krn id condong ke gerindra makanya suka melakukan serangan ke pengusung ganjar
Tokenizing	['gosah', 'sok', 'peduli', 'ganjar', 'perjuangan', 'kritik', 'food', 'estate', 'krna', 'menteri', 'yg', 'di', 'tugaskan', 'pak', 'jokowi', 'tdk', 'becus', 'bekerja', 'namanya', 'prabowo', 'subianto', 'buat', 'yg', 'paham', 'ajah', 'bersebrangan', 'dg', 'pdip', 'krn', 'id', 'condong', 'ke', 'gerindra', 'makanya', 'suka', 'melakukan', 'serangan', 'ke', 'pengusung', 'ganjar']
Stopword	['gosah', 'sok', 'peduli', 'ganjar', 'perjuangan', 'kritik', 'food', 'estate', 'krna', 'menteri', 'tugaskan', 'jokowi', 'becus', 'namanya', 'prabowo', 'subianto', 'paham', 'ajah', 'bersebrangan', 'pdip', 'id', 'condong', 'gerindra', 'suka', 'serangan', 'pengusung', 'ganjar']
Normalization	['jangan', 'sok', 'peduli', 'ganjar', 'perjuangan', 'kritik', 'food', 'estate', 'krna', 'menteri', 'tugaskan', 'jokowi', 'becus', 'namanya',

Stemming	'prabowo', 'subianto', 'paham', 'ajah', 'bersebrangan', 'partaidemokrasiIndonesiaperjuangan', 'id', 'condong', 'gerindra', 'suka', 'serangan', 'pengusung', 'ganjar'] ['jangan', 'sok', 'peduli', 'ganjar', 'juang', 'kritik', 'food', 'estate', 'krna', 'menteri', 'tugas', 'jokowi', 'becus', 'nama', 'prabowo', 'subianto', 'paham', 'ajah', 'bersebrangan', 'partaidemokrasiindonesiaperjuangan', 'id', 'condong', 'gerindra', 'suka', 'serang', 'usung', 'ganjar']

3.2. Pembobotan Kata dengan *TF-IDF*

Langkah ini melibatkan penghitungan bobot pada setiap kata menggunakan *TF-IDF* dari *library Scikit-learn*. nilai bobot diberikan kepada kata-kata berdasarkan dua komponen *Term Frequenc (TF)* dan *Inverse Document Frequency (IDF)*. Tabel 3 merupakan hasil sampel proses *TF-IDF* pada tiap-tiap datasets.

Table 3. Hasil Proses *TF-IDF*

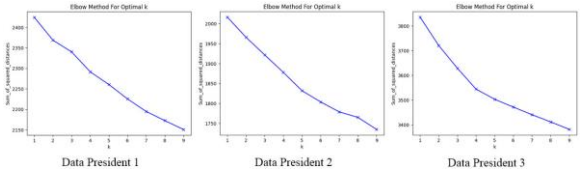
Datasets	Kata	TF	TF-IDF
Data President 1	paspampres	0,041667	0,280558
	anies	0,041667	0,007241
	baret	0,083333	0,460786
Data President 2	isu	0,090909	0,244216
	prabowo	0,090909	0,015037
	gampar	0,090909	0,595509
Data President 3	ngomong	0,125	0,639531
	ngaca	0,125	0,835608
	prestasi	0,125	0,535871

3.3. Mengelompokkan Label Data

Algoritma *K-means* tidak secara otomatis menghasilkan kelompok data yang membedakan antara *black-campaign* dan *nonblack-campaign*. Oleh karena itu, peran manusia diperlukan dalam menentukan label yang tepat. Tahapan dalam menentukan label data diawali dengan penentuan jumlah *cluster* atau kelompok data.

3.3.1. Penentuan jumlah *Cluster*

Pada tahap ini dilakukan analisis menggunakan metode *Elbow* untuk menentukan jumlah *cluster* yang optimal. Meskipun hasil visualisasi elbow menunjukkan 9 *cluster* yang optimal, namun untuk mempermudah analisis, dalam penelitian ini digunakan 6 *cluster*. Visualisasi *Elbow* pada masing-masing data calon presiden ditunjukkan pada Gambar 4.

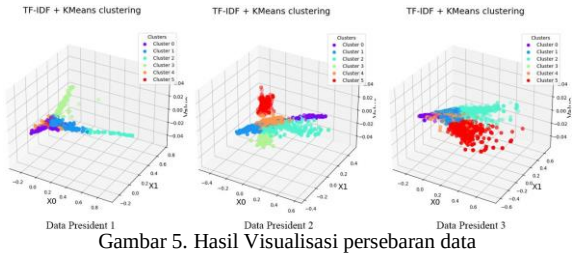


Data President 1 Data President 2 Data President 3

Gambar 4. Hasil *Elbow*

3.3.2. Analisis Persebaran Data

Visualisasi data membantu memahami distribusi data di setiap cluster, dan memperjelas perbedaan karakteristik antara cluster. Proses ini merupakan langkah awal dalam menentukan label data, Hasil visualisasi data *cluster* dengan algoritma *K-means* dapat divisualisasikan pada Gambar 5.

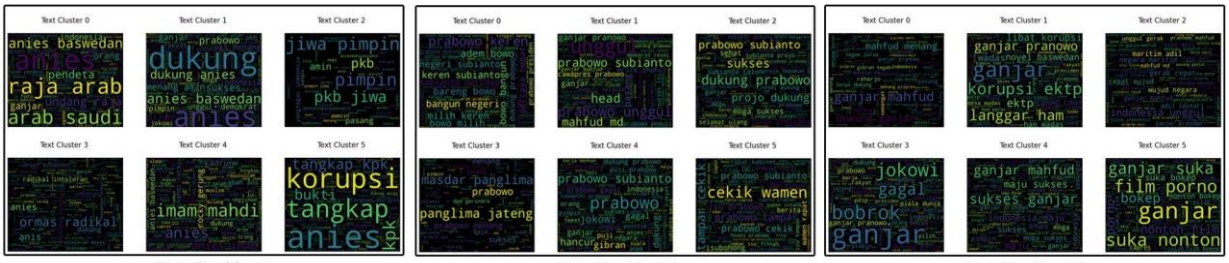


Data President 1 Data President 2 Data President 3

Gambar 5. Hasil Visualisasi persebaran data

3.3.3. Visualisasi Hasil *Cluster*

Visualisasi melalui *WordCloud* dilakukan untuk membuat gambaran visual tentang kata-kata yang sering muncul di setiap *cluster*. Langkah ini dapat membantu dalam memahami topik dan pola data. Proses ini merupakan langkah awal dalam menentukan label data dan membuatnya lebih mudah dalam menentukan label yang sesuai pada setiap *cluster* berdasarkan hasil *WordCloud*. Gambar 6 merupahan hasil proses *WordCloud* pada data sample pada masing-masing calon presiden.

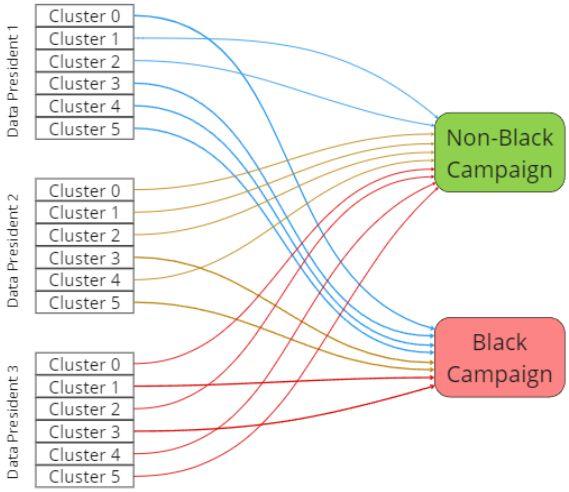


Data President 1 Data President 2 Data President 3

Gambar 6.Hasil Visualisasi *WordCloud*

3.3.4. Penentuan Label

Secara keseluruhan proses penentuan label (*black-campaign* dan *nonblack-campaign*) perlu mempertimbangkan dengan cermat dan berhati-hari agar tidak mempengaruhi hasil klasifikasi. Ilustrasi penentuan label pada tiap-tiap *Cluster* dapat dilihat pada Gambar 7.



Gambar 7. Ilustrasi penentuan label data.

3.4. Penggabungan Datasets

Setelah proses pengelompokan data dengan metode *K-means*, Tiga dataset yang telah dilabeli akan disatukan menjadi satu dataset besar, dilanjutkan menghapus data yang dianggap tidak informatif, termasuk data yang kosong dan duplikat. Sebelum proses penggabungan, terdapat 8620 baris data, dengan total 1594 merupakan duplikat data. Setelah proses penghapusan data duplikat jumlah total data yang digunakan untuk analisis menjadi 7025 baris data, proses ini menjadikannya dataset menjadi bersih dan siap untuk dianalisis.

3.5. Klasifikasi

Dalam tahap klasifikasi menggunakan model *Long Short-Term Memory (LSTM)*, digunakan untuk mempelajari urutan dan hubungan antar kata. Model *LSTM* ini memanfaatkan kelebihan arsitektur *Recurrent Neural Network (RNN)* untuk memproses informasi sekuensial pada data teks. Selain itu, penerapan teknik-teknik pencegahan overfitting, seperti *Dropout* dan *Early Stopping*, digunakan untuk memastikan performa model yang optimal dan menghindari masalah *overfitting*. Selanjutnya, hasil klasifikasi dari model *LSTM* dibandingkan dengan hasil dari model *Ensemble Learning*, yang menggunakan teknik *Voting* dengan gabungan model klasifikasi *Machine Learning* diantaranya *Naive Bayes, Random Forest*, dan *SVM*.

Teknik *ensemble* memungkinkan kombinasi dari beberapa model klasifikasi untuk meningkatkan keakuratan prediksi secara keseluruhan. Tahapan-tahapan dalam proses klasifikasi dijabarkan sebagai berikut:

3.5.1. Implementasi *One-Hot Encoding*

Dalam proses ini, digunakan *library TensorFlow* melalui modul *keras* untuk melakukan tahapan *one-hot encoding* pada dokumen teks. Dimulai dengan mengimpor fungsi *one_hot* dari *library TensorFlow*. Selanjutnya, menetapkan ukuran *vocab* sebanyak 10,000 untuk menentukan jumlah kata unik yang akan direpresentasikan dalam *encoding*, proses ini merubah data teks kedalam bentuk biner. Kemudian, lapisan *LSTM* digunakan untuk memahami urutan dan hubungan antar kata, sehingga kemampuan *LSTM* dalam mendeteksi teks kampanye hitam akan lebih baik.

Berbeda dengan *Ensemble Learning* yang berbasiskan *Machine Learning*, Model *Ensemble Learning* melihat kalimat sebagai kumpulan kata tanpa memahami urutan dan hubungan antar kata.

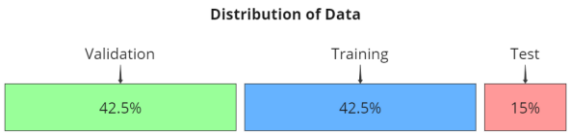
3.5.2. Proses Distribusi Data

Langkah awal dalam distribusi data adalah menyamaratakan jumlah label dalam dataset, untuk memastikan keseimbangan antara kategori *black-campaign* dan *nonblack-campaign*. Dalam proses penyeimbangan data ini didapati data berjumlah 2812 untuk masing-masing label data, yang sebelumnya 4213 data pada label black-campaign dan 2812 data pada label nonblack-campaign. Tabel 4 dibawah merupakan hasil sebelum dan sesudah penyeimbangan data.

Table 4. Hasil perbandingan penyeimbangan data.

Label	Sebelum	Sesudah
black-campaign	4213	2812
nonblack-campaign	2812	2812

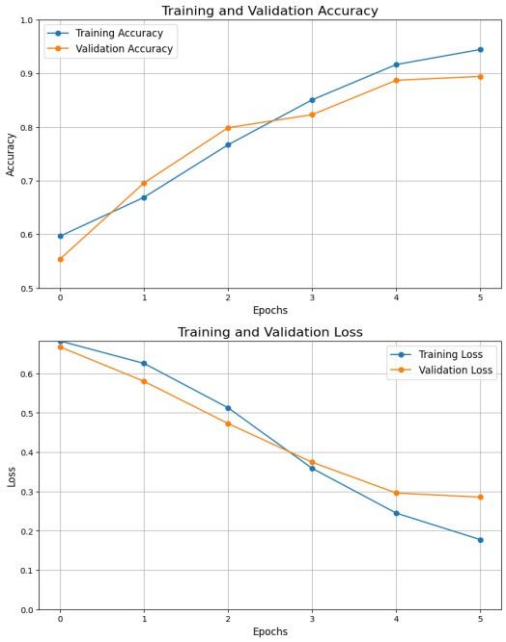
Selanjutnya, distribusi dataset untuk *Training*, *Testing* dan *Validation*. Pada tahap ini, dataset dibagi menjadi data pelatihan, pengujian dan validasi agar distribusi label seimbang di setiap set (train, test, dan validation), ditambahkan modul *StratifiedShuffleSplit* dari *scikit-learn*. Distribusi label yang seimbang pada data *Test*, *Train* dan *Validation* dilakukan untuk mencegah terjadinya *overfitting*. Distribusi label dapat dilihat pada Gambar 8.



Gambar 8. Visualisasi distribusi data

3.5.3. Pelatihan Model

Pada Proses pelatihan model menggunakan algoritma *LSTM* untuk mempelajari urutan kata-kata. Penambahan fungsi *Dropout* sebesar 20%, *learning rate* yang rendah dan *Early Stopping* dilakukan untuk mencegah permasalahan *Overfitting* yang mungkin terjadi selama pelatihan model. Model dilatih dengan jumlah layer 128 dan total 6 epochs, Gambar 9 menunjukkan kinerja model dalam proses *training* dan *validation accuracy*, serta *training* dan *validation loss*.



Gambar 9. Grafik hasil pelatihan model

Berdasarkan hasil pelatihan yang ditunjukkan di Gambar 9. Model mampu melakukan klasifikasi dengan baik. Proses ini bertujuan memastikan bahwa model dapat memberikan prediksi yang akurat dan dapat dilanjutkan pada proses evaluasi.

3.6. Evaluasi

Penelitian ini mengevaluasi dua model klasifikasi *LSTM* dan *Ensemble Learning*. Evaluasi dilakukan dengan berbagai metrik untuk mengukur kinerja keduanya menggunakan *confusion matrix*, perbandingan antara kedua model mengungkap gambaran menarik tentang kemampuan klasifikasi. Hasil Evaluasi dapat dilihat pada Tabel 5 di bawah.

Table 5. Hasil Evaluasi confusion matrix

Model	TP	FP	FN	TN
LSTM	377	37	45	385
Ensemble Learning	396	22	26	400

Pada tabel *confusion matrix*, Secara umum *Ensemble Learning* menunjukkan kinerja yang lebih baik dibandingkan *LSTM* dalam hal akurasi prediksi, dengan menghasilkan lebih banyak prediksi benar.

Selain itu Evaluasi model dinyatakan dalam beberapa metrik kinerja klasifikasi utama, seperti akurasi, presisi, *recall*, *F1-score*. Tabel 6 merupakan hasil evaluasi dari model klasifikasi *LSTM* dan *Ensemble Learning*.

Table 6. Hasil Evaluasi

Model	Akurasi	Presisi	Recal	F-1 Score
LSTM	90.28%	0.90	0.91	0.90
Ensemble Learning	94.31%	0.94	0.95	0.94

Perbandingan antara kedua model ini menunjukkan kedua model memiliki kinerja yang baik, namun dengan karakteristik yang berbeda. *Ensemble Learning* menunjukkan bahwa model ini memberikan hasil yang lebih baik dibandingkan dengan model *LSTM* dalam proses klasifikasi yang dijalankan. Meskipun perbedaan tidak terlalu signifikan, namun hasil ini menunjukkan bahwa penggabungan beberapa model pada *Ensemble Learning* dengan teknik *Voting* dapat meningkatkan performa klasifikasi.

3.7. Hasil Pengujian Model

Pada tahap pengujian ini, diperlihatkan hasil deteksi model klasifikasi dengan menggunakan data yang belum pernah dikenali sebelumnya oleh model yang telah dilatih, baik itu *LSTM* ataupun *Ensemble Learning*. Hasil uji deteksi sampel ditujukkan pada Tabel 7.

Table 7. Hasil Deteksi Model

Tanggal Posting	Data Teks	LSTM	Ensemble Learning
4:54 AM · Nov 4, 2023	TOGOG TOLOLGOBLOG YG PUNYA BOTOL BODOHTOLOLOGI block pak Anies klo mmg korupsi Formula E sdh pasti di tangkap KPK block dmn KPK sampai 16 kali gelar perkara utk mentersangkakan pak Anies krn beliau tdk korupsilah mangkana KPK gak bisa membuktikan beda dg Ganjar	*Black Campaign*	*Black Campaign*
2:44 PM · Nov 6, 2023	Lembaga Survei Indonesia (LSI) mencatat Prabowo Subianto unggul atas dua lawannya, Ganjar Pranowo dan Anies Baswedan, dalam simulasi pilpres dua nama.	*Non-Black Campaign*	*Non-Black Campaign*
9:22 AM · Nov 8, 2023	Terlambat,.....semua orang dah tahu Jateng propinsi termiskin, UMR paling rendah, HAM di Krendeng & Wadas, Ganjar terima 520ribu dolar dari kasus eKTP	*Black Campaign*	*Black Campaign*
4:00 PM · Nov 15, 2023	sudah terlihat sangat jelas bahwa saat ini Elektabilitas Prabowo Gibran makin kokoh , dan sudah meninggalkan jauh Ganjar Mahfud dan Amin tetap semangat yakin pasti menang dalam 1 putaran	*Non-Black Campaign*	*Non-Black Campaign*

5:56 PM · Nov 21, 2023	Waspadalah, krn kpu & bawaslu tdk netral nih, pdahal paslon prabowo-gibran itu cacat hukum, terbukti langgar konstitusi	*Black Campaign*	*Black Campaign*
7:30 AM · Nov 21, 2023	Survei terbaru IPO Anies-Muhaimin sukses menyalip Ganjar-Mahfud.. Selangkah lagi menang satu putaran!!	*Non-Black Campaign*	*Non-Black Campaign*

Berdasarkan hasil dari tabel 7, terlihat bahwa kedua model cenderung memberikan hasil yang serupa dalam mendeteksi data. Meskipun demikian, dengan memanfaatkan kelebihan arsitektur *Recurrent Neural Network* untuk memproses informasi sekuensial data teks, *LSTM* mampu memahami frasa atau urutan dan hubungan antar kata. Sementara, *Ensemble Learning* tidak, dikarenakan, *Ensemble Learning* hanya dapat mempelajari pola data secara keseluruhan.

4. DISKUSI

Penelitian ini menunjukkan bahwa pemrosesan data awal menggunakan *K-means* dapat mempermudah identifikasi label, serta dapat mempersiapkan data untuk proses klasifikasi. Hal ini didukung oleh hasil penelitian sebelumnya oleh Muhariya et al., yang menemukan bahwa pengelompokan *K-means* dengan Preprocessing dan pembobotan kata *TF-IDF* berhasil mengelompokkan data komentar menjadi dua kelompok [18]. Data sosial media yang digunakan dalam penelitian ini berbeda dengan data penelitian sebelumnya. Penelitian ini menerapkan *Removal Punctuation* dalam proses pra-pemrosesan untuk menghapus tanda baca dalam data teks, sehingga dapat menjadi teknik yang efektif untuk meningkatkan akurasi.

Selanjutnya, Proses pelatihan menggunakan model klasifikasi *LSTM* dengan penerapan teknik-teknik pencegahan *overfitting*, Mengacu pada penelitian yang disampaikan oleh Bisht et al., Distribusi data kelas yang tidak seimbang dapat mempengaruhi kinerja model menjadi kurang maksimal [19]. Hal ini disebabkan model cenderung memprediksi kelas yang lebih banyak muncul. Oleh karena itu, ditambahkan modul *StratifiedShuffleSplit* dari *scikit-learn*. Sehingga distribusi label dapat seimbang. Selain itu, teknik teknik seperti *Dropout* dan *Early Stopping* diterapkan untuk mencegah *overfitting*

Confusion matrix yang menggambarkan kemampuan model dalam mengenali dan memprediksi kelas positif dan negatif. Meskipun perbedaan antara keduanya tidak signifikan, namun model *Ensemble Learning* memiliki sedikit peningkatan dalam beberapa metrik evaluasi.

Hasil evaluasi menunjukkan bahwa model *LSTM* memiliki akurasi sebesar 90.28%, sedangkan model *Ensemble Learning* mencapai akurasi 94.31%. Selain itu, Model *LSTM* dan *Ensemble Learning (Voting)* menunjukkan hasil yang baik pada nilai presisi, *recall*, dan *F1-score*.

Pengujian terhadap sampel data menegaskan bahwa model *LSTM* dan *Ensemble Learning* cenderung memberikan hasil yang serupa dalam mendeteksi *Black Campaign* dan *Non-Black Campaign* pada data teks.

Meskipun hasil evaluasi *Ensemble Learning* memiliki akurasi yang lebih tinggi, Namun, *Ensemble Learning* tidak dapat mempelajari pola-pola urutan kata. *Ensemble learning* hanya dapat mempelajari pola-pola data secara keseluruhan. Oleh karena itu, *LSTM* lebih cocok untuk deteksi kampanye hitam.

Hal ini sesuai dengan penelitian yang dilakukan Chang dan Masterson. Dengan memanfaatkan kelebihan arsitektur *Recurrent Neural Network*, *LSTM* dapat memahami urutan kata [20], sehingga kemampuan *LSTM* dalam memahami makna kalimat didalam teks kampanye hitam lebih baik dibanding *Ensemble learning*.

5. KESIMPULAN DAN SARAN

Berdasarkan penelitian ini, terlihat bahwa model deteksi menggunakan metode *K-means*, *LSTM* dapat diaplikasikan secara efektif untuk analisis data teks *Twitter* terkait isu *black campaign* pada pencalonan presiden Indonesia tahun 2024. Hasil penelitian menunjukkan bahwa pemrosesan data awal dengan *K-means* memberikan kontribusi positif dalam persiapan data untuk analisis selanjutnya. Pengembangan model klasifikasi dengan *LSTM* menunjukkan akurasi 90.28%. Sementara itu, model *Ensemble Learning* mencapai akurasi lebih tinggi, yakni 94.31%, Menunjukkan bahwa penggabungan beberapa model dapat meningkatkan kinerja klasifikasi. Meskipun evaluasi menunjukkan bahwa *Ensemble Learning* memberikan hasil yang baik, namun *LSTM* memiliki keunggulan dalam memahami urutan kata. Keunggulan ini dapat diperoleh dengan memanfaatkan struktur *Recurrent Neural Network*.

Penelitian ini dapat menjadi rekomendasi dan landasan bagi penelitian mendatang, dengan fokus pada peningkatan akurasi model melalui *fine-tuning*, dengan cara penyesuaian model yang dilakukan berdasarkan hasil evaluasi. Pengembangan lebih lanjut pada model klasifikasi teks di media sosial diharapkan dapat merespons lebih baik terhadap kompleksitas data teks. Selain itu, penelitian ini juga dapat menjadi eksplorasi pengembangan *Prototype* aplikasi inovatif berbasis *online*, yang bertujuan mendeteksi kampanye hitam secara lebih luas, dan memberikan dampak positif pada aktivitas menyuarakan opini dan pendapat di media sosial.

DAFTAR PUSTAKA

[1] J. Indrawan, R. E. Barzah, and H. Simanihuruk, 'INSTAGRAM SEBAGAI MEDIA KOMUNIKASI POLITIK BAGI GENERASI MILENIAL', *EKSPRESI DAN PERSEPSI: JURNAL ILMU KOMUNIKASI*, vol. 6, no. 1, pp. 170–179, 2023, doi: 10.33822/jep.v6i1.4519.

[2] E. H. Susanto, 'MEDIA SOSIAL SEBAGAI PENDUKUNG JARINGAN KOMUNIKASI POLITIK', *Jurnal ASPIKOM*, vol. 3, no. 3, 2017, doi: 10.24329/aspikom.v3i3.123.

[3] A. A. Munzir, 'Beragam peran media sosial dalam dunia politik di Indonesia', *JPPUMA: Jurnal Ilmu Pemerintahan dan Sosial Politik UMA (Journal of Governance and Political Social UMA)*, vol. 7, no. 2, pp. 173–182, 2019, doi: 10.31289/jppuma.v7i2.2691.

[4] L. Syafirullah, A. S. Prabowo, and R. H. Maharrani, 'THE AHP METHOD IN DETERMINING RI 2024 PRESIDENTIAL CANDIDATES MILENIAL GENERATION CILACAP STATE POLYTECHNIC: METODE AHP DALAM MENENTUKAN BAKAL CALON PRESIDEN RI 2024 GENERASI MILENIAL POLITEKNIK NEGERI CILACAP', *Jurnal Minfo Polgan*, vol. 12, no. 2, pp. 819–829, 2023, doi: 10.33395/jmp.v12i1.12498.

[5] H. Sazali, U. A. R. SM, and R. F. Marta, 'Mapping Hate Speech Relationships Indonesia's Religion and State in Social Media', *Communicatus: Jurnal Ilmu komunikasi*, vol. 6, no. 2, pp. 189–208, 2022, doi: 10.15575/cjik.v6i2.20431.

[6] D. Wiana, 'Analysis of the use of the hate speech in social media In the case of presidential election in 2019', *Journal of Applied Studies in Language*, vol. 3, no. 2, pp. 158–167, 2019, doi: 10.31940/jasl.v3i2.1541.

[7] W. W. Utami and D. Darmaiza, 'Hate Speech, Agama, Dan Kontestasi Politik Di Indonesia', *Indonesian Journal of Religion and Society*, vol. 2, no. 2, pp. 113–128, 2020, doi: 10.36256/ijrs.v2i2.108.

[8] N. F. Octarina and H. Djanggih, 'Legal Implication of Black Campaigns on The Social Media in The General Election Process', *Jurnal Dinamika Hukum*, vol. 19, no. 1, pp. 271–282, 2019, doi: 10.20884/1.jdh.2019.19.1.2115.

[9] H. S. Al-Ash and W. C. Wibowo, 'Fake news identification characteristics using named entity recognition and phrase detection', in *Proceedings of 2018 10th International Conference on Information Technology and Electrical Engineering: Smart Technology for Better Society, ICITEE 2018*, 2018. doi: 10.1109/ICITEED.2018.8534898.

[10] R. Ali, U. Farooq, U. Arshad, W. Shahzad, and M. O. Beg, 'Hate speech detection on Twitter using transfer learning', *Computer Speech & Language*, vol. 74, p. 101365, 2022, doi: 10.1016/j.csl.2022.101365.

[11] F. E. T. Sirait, 'Ujaran Kebencian, Hoax dan Perilaku Memilih (Studi Kasus pada Pemilihan Presiden 2019 di Indonesia)', *Jurnal Penelitian Politik*, vol. 16, no. 2, 2020, doi: 10.14203/jpp.v16i2.806.

[12] S. Long, X. He, and C. Yao, 'Scene text detection and recognition: The deep learning era', *International Journal of Computer Vision*, vol. 129, pp. 161–184, 2021, doi: 10.1007/s11263-020-01369-0.

[13] R. Nainggolan, R. Perangin-angin, E. Simarmata, and A. F. Tarigan, 'Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method', *J. Phys.: Conf. Ser.*, vol. 1361, no. 1, p. 012015, Nov. 2019, doi: 10.1088/1742-6596/1361/1/012015.

[14] J. Cui, Z. Wang, S.-B. Ho, and E. Cambria, 'Survey on sentiment analysis: evolution of research methods and topics', *Artificial Intelligence Review*, pp. 1–42, 2023, doi: 10.1007/s10462-022-10386-z.

[15] Z. Amiri, A. Heidari, N. J. Navimipour, M. Unal, and A. Mousavi, 'Adventures in data analysis: A systematic review of Deep Learning techniques for pattern recognition in cyber-physical-social systems', *Multimedia Tools and Applications*, pp. 1–65, 2023, doi: 10.1007/s11042-023-16382-x.

[16] A. Chatterjee, U. Gupta, M. K. Chinnakotla, R. Srikanth, M. Galley, and P. Agrawal, 'Understanding Emotions in Text Using Deep Learning and Big Data', *Computers in Human Behavior*, vol. 93, pp. 309–317, Apr. 2019, doi: 10.1016/j.chb.2018.12.029.

[17] S. N. Wahyuni, N. N. Khanom, and Y. Astuti, 'K-Means Algorithm Analysis for Election Cluster Prediction', *JOIV: International Journal on Informatics Visualization*, vol. 7, no. 1, pp. 1–6, 2023, doi: 10.30630/joiv.7.1.1107.

[18] A. Muhariya, I. Riadi, and Y. Prayudi, 'Cyberbullying Analysis on Instagram Using K-Means Clustering', *JUITA : Jurnal Informatika*, vol. 10, no. 2, p. 261, Nov. 2022, doi: 10.30595/juita.v10i2.14490.

[19] A. Bisht, A. Singh, H. S. Bhadauria, J. Virmani, and Kriti, 'Detection of Hate Speech and Offensive Language in Twitter Data Using LSTM Model', in *Recent Trends in Image and Signal Processing in Computer*

Vision, vol. 1124, S. Jain and S. Paul, Eds., in Advances in Intelligent Systems and Computing, vol. 1124. , Singapore: Springer Singapore, 2020, pp. 243–264. doi: 10.1007/978-981-15-2740-1_17.

[20] C. Chang and M. Masterson, 'Using Word Order in Political Text Classification with Long Short-term Memory Models', *Political Analysis*, vol. 28, no. 3, pp. 395–411, Jul. 2020, doi: 10.1017/pan.2019.46.

[21] I. Kurniawan and A. Susanto, 'Implementasi Metode K-Means dan NaÃ ve Bayes Classifier untuk Analisis Sentimen Pemilihan Presiden (Pilpres) 2019', *Jurnal Eksplora Informatika*, vol. 9, no. 1, pp. 1–10, 2019, doi: 10.30864/eksplora.v9i1.237.

[22] Y. Saini, V. Bachchas, Y. Kumar, and S. Kumar, 'Abusive Text Examination Using Latent Dirichlet Allocation, Self Organizing Maps and K Means Clustering', in *Proceedings of the International Conference on Intelligent Computing and Control Systems, ICICCS 2020*, 2020. doi: 10.1109/ICICCS48265.2020.9121090.

[23] M. Chiny, M. Chihab, O. Bencharef, and Y. Chihab, 'Lstm, vader and tf-idf based hybrid sentiment analysis model', *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 7, 2021, doi: 10.14569/IJACSA.2021.0120730.

[24] G. A. M. K. Jaluwana, G. M. A. Sasmita, and I. M. A. D. Suarjaya, 'Analysis of Public Sentiment Towards Goverment Efforts to Break the Chain of Covid-19 Transmission in Indonesia Using CNN and Bidirectional LSTM', *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 4, pp. 511–520, 2022, doi: 10.29207/resti.v6i4.4055.

[25] S. Saadah, K. M. Auditama, A. A. Fattahila, F. I. Amorokhman, A. Aditsania, and A. A. Rohmawati, 'Implementation of BERT, IndoBERT, and CNN-LSTM in Classifying Public Opinion about COVID-19 Vaccine in Indonesia', *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 4, pp. 648–655, 2022.

[26] P. F. Muhammad, R. Kusumaningrum, and A. Wibowo, 'Sentiment analysis using Word2vec and long short-term memory (LSTM) for Indonesian hotel reviews', *Procedia Computer Science*, vol. 179, pp. 728–735, 2021, doi: 10.1016/j.procs.2021.01.061.

[27] S. Santi and H. Februariyanti, 'IMPLEMENTATION OF CLUSTERING ON TWEET UPLOADING SIDE EFFECTS OF COVID-19 POST VACCINATION USING K-MEANS ALGORITHM', *Jurnal Teknik Informatika (Jutif)*, vol. 4, no. 4, Art. no. 4, Aug. 2023, doi: 10.52436/1.jutif.2023.4.4.704.

[28] A. Oussous, F.-Z. Benjelloun, A. A. Lahcen, and S. Belfkih, 'ASA: A framework for Arabic sentiment analysis', *Journal of Information Science*, vol. 46, no. 4, pp. 544–559, 2020, doi: 10.1177/0165551519849516.

[29] M. I. Alfarizi, L. Syafaah, and M. Lestandy, 'Emotional Text Classification Using TF-IDF (Term Frequency-Inverse Document Frequency) And LSTM (Long Short-Term Memory)', *JUITA : Jurnal Informatika*, vol. 10, no. 2, p. 225, Nov. 2022, doi: 10.30595/juita.v10i2.13262.

[30] T. A. B. Sembiring and M. S. Hasibuan, 'TEXT CLUSTERING IN KARO LANGUAGE USING TF-IDF WEIGHTING AND K-MEANS CLUSTERING', *Jurnal Teknik Informatika (Jutif)*, vol. 4, no. 5, Art. no. 5, Nov. 2023, doi: 10.52436/1.jutif.2023.4.5.1462.

[31] S. Akuma, T. Lubem, and I. T. Adom, 'Comparing Bag of Words and TF-IDF with different models for hate speech detection from live tweets', *International Journal of Information Technology*, vol. 14, no. 7, pp. 3629–3635, Dec. 2022, doi: 10.1007/s41870-022-01096-4.

[32] A. Gazizullina and M. Mazzara, 'Prediction of twitter message deletion', presented at the 2019 12th International Conference on Developments in eSystems Engineering (DeSE), IEEE, 2019, pp. 117–122. doi: 10.1109/DeSE.2019.00031.

[33] W. Afandi, S. N. Saputro, A. M. Kusumaningrum, H. Adriansyah, M. H. Kafabi, and S. Sudianto, 'Klasifikasi Judul Berita Clickbait menggunakan RNN-LSTM', *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 7, no. 2, Art. no. 2, May 2022, doi: 10.30591/jpit.v7i2.3401.

[34] S. K. Dasari, S. Gorla, and P. R. PVGD, 'A stacking ensemble approach for identification of informative tweets on twitter data', *International Journal of Information Technology*, pp. 1–12, 2023, doi: 10.1007/s41870-023-01316-5.

[35] S. A. Kokatnoor and B. Krishnan, 'Twitter hate speech detection using stacked weighted ensemble (SWE) model', presented at the 2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), IEEE, 2020, pp. 87–92. doi: 10.1109/ICRCICN50933.2020.9296199.

[36] Duhok Polytechnic University, A. A. Salih, A. M. Abdulazeez, and Duhok Polytechnic

University, 'Evaluation of Classification Algorithms for Intrusion Detection System: A Review', *JSCDM*, vol. 02, no. 01, Apr. 2021, doi: 10.30880/jscdm.2021.02.01.004.

[37] H. Zhao, S. Sun, and B. Jin, 'Sequential fault diagnosis based on LSTM neural network', *Ieee Access*, vol. 6, pp. 12929–12939, 2018, doi: 10.1109/ACCESS.2018.2794765.

[38] M. Al Razib, D. Javeed, M. T. Khan, R. Alkanhel, and M. S. A. Muthanna, 'Cyber threats detection in smart environments using SDN-enabled DNN-LSTM hybrid framework', *IEEE Access*, vol. 10, pp. 53015–53026, 2022, doi: 10.1109/ACCESS.2022.3172304.

[39] R. M. AlZoman and M. J. F. Alenazi, 'A Comparative Study of Traffic Classification Techniques for Smart City Networks', *Sensors*, vol. 21, no. 14, p. 4677, Jul. 2021, doi: 10.3390/s21144677..