EKSTRAKSI BERITA HOAX PADA TURN BACK HOAX BERBASIS PENDEKATAN *TF-IDF & COSINE SIMILARITY*

William Hidayat[1)], Jesen Ong[2)], Umar Muhdhor[3)], Hafiz Irsyad[4)], Abdul Rahman[5)]

[1), 2), 3), 4), 5)] *Program Studi Informatika, Fakultas Ilmu Komputer dan Rekayasa, Universitas Multi Data Palembang, Palembang, Sumatera Selatan, Indonesia*
Email: williamhidayat_2226250035@mhs.mdp.ac.id[1)], jesenong_2226250084@mhs.mdp.ac.id[2)], umarmuhdhor_2226250042@mhs.mdp.ac.id[3)], hafizirsyad@mdp.ac.id[4)], arahman@mdp.ac.id[5)]

Abstrak

Perkembangan teknologi telah membawa perubahan besar dalam kehidupan masyarakat. Salah satunya akses terhadap berita dan artikel yang semakin mudah, dan bebas. Namun, fenomena ini juga memunculkan permasalahan serius, yaitu penyebaran berita hoaks yang sangat cepat dan masif. Penelitian ini bertujuan untuk mengekstraksi informasi penting dari artikel hoaks yang dipublikasikan di situs TurnBackHoax.id menggunakan pendekatan *text mining* berbasis *TF-IDF* dan *cosine similarity*. Data artikel hoaks diperoleh melalui teknik *web scraping* dengan pustaka *Python* seperti *requests* dan *BeautifulSoup*, diikuti oleh tahap prapemrosesan teks yang meliputi *case folding*, penghapusan tanda baca, angka, serta *stopwords*, dan *stemming*. Teks yang telah diproses kemudian direpresentasikan dalam bentuk vektor numerik menggunakan metode *TF-IDF* untuk menentukan bobot kata berdasarkan frekuensi dan kelangkaannya dalam korpus. Selanjutnya, *cosine similarity* digunakan untuk mengukur tingkat kemiripan antar dokumen, sementara kata kunci diekstraksi berdasarkan bobot *TF-IDF* tertinggi. Visualisasi *Word Cloud* juga diterapkan untuk menggambarkan kata-kata dominan secara visual. Berdasarkan hasil evaluasi, metode yang digunakan dalam penelitian ini berhasil mencapai tingkat ketepatan sebesar 93,15%, menunjukkan efektivitas pendekatan *TF-IDF* dan *Cosine Similarity* dalam menganalisis dan mengelompokkan artikel hoaks. Hasil penelitian menunjukkan bahwa pendekatan ini efektif dalam mengidentifikasi kata kunci penting dan mengelompokkan artikel hoaks berdasarkan kemiripan konten.

Kata kunci : *Cosine Similarity,Hoaks,Text Mining, TF-IDF, kata kunci*

Abstract

The development of technology has brought significant changes to society. One of them is easier access to news and articles thanks to the advancement of the internet. However, this phenomenon also raises a serious issue, namely the rapid and massive spread of hoax news. This research aims to extract important information from hoax articles published on the TurnBackHoax.id website using a TF-IDF-based text mining approach. TurnBackHoax.id using TF-IDF-based text mining approach and cosine similarity. and cosine similarity. Hoax article data is obtained through web scraping technique with Python libraries such as requests and BeautifulSoup, followed by a text preprocessing stage which includes case folding, removal of punctuation, numbers, and stopwords, and stemming. The text that has been processed is then represented in the form of vectors using the TF-IDF method to determine the weight of words based on their frequency and rarity in the text. based on its frequency and rarity in the corpus. Furthermore, cosine similarity is used to measure the level of similarity between documents, while the keywords are extracted based on the highest TF-IDF weight. Word cloud visualization Word cloud is also applied to visually depict the dominant words. Based on evaluation results, the method used in this study successfully achieved a accuracy rate of 93.15%, demonstrating the effectiveness of the TF-IDF and Cosine Similarity approach in analyzing and clustering hoax articles. The results show that this approach is effective in identifying important keywords and categorizing hoax articles based on content similarity.

Keywords : *Cosine Similarity, Hoax, Text Mining, TF-IDF, Keyword*

1. Pendahuluan

Perkembangan teknologi informasi telah membawa perubahan besar dalam cara masyarakat memperoleh dan menyebarkan informasi. Akses terhadap berita dan artikel semakin mudah berkat kemajuan internet. Namun, fenomena ini juga memunculkan permasalahan serius, yaitu penyebaran berita hoaks yang sangat cepat dan masif[1]. Hoaks dapat menyebabkan kepanikan, menyebarkan kebencian, hingga memengaruhi opini publik secara negatif. Salah satu upaya untuk memerangi hoaks di Indonesia dilakukan oleh *Masyarakat Anti Fitnah Indonesia* (MAFINDO) melalui situs *Turn Back Hoax*, yang menjadi sumber terpercaya untuk dokumentasi dan verifikasi berita bohong[1].Berdasarkan hal tersebut, penelitian ini mengambil obyek berupa artikel-artikel hoaks yang dipublikasikan di situs tersebut untuk dianalisis lebih lanjut.

Dalam analisis teks, telah banyak metode yang digunakan untuk mengekstraksi informasi penting dari sebuah dokumen. Beberapa metode umum meliputi *Term Frequency–Inverse Document Frequency* (*TF-IDF*), *Latent Semantic Analysis* (*LSA*), dan *Latent Dirichlet Allocation* (*LDA*). *TF-IDF* merupakan metode berbasis statistik yang banyak digunakan untuk menentukan pentingnya suatu kata dalam sebuah dokumen relatif terhadap keseluruhan korpus [2][3]. Sementara itu, *cosine similarity* digunakan untuk mengukur kemiripan antar dokumen berdasarkan bobot kata hasil dari metode representasi vektor.

Setiap metode memiliki kelebihan dan kekurangannya masing-masing. *TF-IDF* dikenal sederhana, cepat dalam implementasi, dan efektif untuk korpus dengan ukuran kecil hingga menengah. Namun, metode ini tidak mempertimbangkan konteks semantik atau hubungan antar kata dalam kalimat, yang menyebabkan kurangnya pemahaman terhadap makna kata secara menyeluruh [2]. Di sisi lain, *LDA* menawarkan pendekatan yang lebih semantik dengan mengelompokkan kata-kata dalam topik tertentu, tetapi kelemahan *LDA* terletak pada kompleksitas parameter dan ketergantungan terhadap jumlah topik yang harus ditentukan secara manual [3].

Penelitian oleh Yulianty Lasena, Husdi, dan Maryam Hasan (2020) dalam jurnal Teknologi Informasi dan Komunikasi mengkaji penerapan metode *TF-IDF* dan *Cosine Similarity* untuk mengidentifikasi artikel hoaks berdasarkan kesamaan konten teks. Proses dilakukan dengan membandingkan berita baru terhadap dokumen rujukan untuk mendeteksi kemungkinan hoaks. Hasil penelitian ini menunjukkan bahwa pendekatan berbasis kemiripan teks mampu mengelompokkan artikel yang memiliki narasi serupa, meskipun tidak dicantumkan angka evaluasi secara eksplisit. Penelitian ini menegaskan bahwa kombinasi kedua metode tersebut relevan untuk penerapan sistem deteksi hoaks berbasis teks [4].

Faizal Nur Rozi dan Dwi Harini Sulistyawati (2019) dalam jurnal Konvergensi melakukan penelitian mengenai klasifikasi berita hoaks pada isu pilpres menggunakan metode *Modified K-Nearest Neighbor* (*MKNN*) yang dipadukan dengan pembobotan I. Dalam penelitian ini, dokumen berita diproses melalui *TF-IDF* untuk menghasilkan representasi vektor, kemudian diklasifikasikan menggunakan *MKNN*. Meskipun nilai evaluasi numerik tidak dijelaskan secara rinci, sistem yang dikembangkan mampu membedakan antara berita hoaks dan non-hoaks secara akurat berdasarkan konten teks. Penelitian ini memperlihatkan efektivitas metode kombinasi tersebut dalam menangani klasifikasi teks di bidang politik [5].

Goenawan Brotosaputro, Wiwin Windihastuty, dan Rezza Anugrah Mutiarawan pada tahun 2021 dalam jurnal Sisfokom melakukan penelitian mengenai deteksi berita hoaks pada artikel politik berbahasa Indonesia di media sosial dengan menggunakan metode *TF-IDF* dan *Cosine Similarity*. Data dianalisis untuk mengukur kemiripan konten antar artikel dan kemudian dievaluasi menggunakan confusion matrix. Hasil evaluasi menunjukkan bahwa sistem yang dibangun memiliki tingkat *precision* sebesar 92%, *recall* sebesar 80%, dan akurasi sebesar 87%, yang mengindikasikan kinerja yang cukup tinggi dalam mengklasifikasikan berita hoaks. Penelitian ini membuktikan bahwa pendekatan berbasis representasi vektor dan pengukuran kemiripan antar teks mampu memberikan hasil yang akurat dalam mendeteksi hoaks di ranah politik [6].

Dalam penelitian ini, *TF-IDF* dipilih karena kesesuaiannya untuk digunakan pada korpus berita hoaks yang tidak terlalu besar dan tujuannya yang hanya untuk mengekstraksi kata kunci paling menonjol[2]. Meski demikian, penggunaan *TF-IDF* dalam konteks bahasa Indonesia memiliki tantangan tersendiri, antara lain variasi bentuk kata (afiksasi) dan kata tidak penting (*stopwords*) yang dapat mengaburkan kata-kata kunci penting. Permasalahan lain yang sering muncul adalah kemunculan sinonim dan kata sepadan yang tidak dapat diidentifikasi oleh *TF-IDF*. Penelitian ini penting dilakukan untuk membantu mengidentifikasi dan menganalisis pola penyebaran informasi hoaks secara sistematis, sehingga dapat mendukung upaya deteksi dini dan mitigasi penyebaran berita palsu di ruang digital.

2. Dasar Teori

2.1 *Web Scraping*

Penarikan data dari situs *turnbackhoax.id* dilakukan menggunakan teknik *web scraping*, yaitu proses otomatisasi untuk mengumpulkan informasi dari halaman web secara sistematis. *Web scraping* merupakan teknik yang umum digunakan dalam pengumpulan data dari situs web untuk keperluan analisis, di mana prosesnya mencakup identifikasi elemen HTML dan ekstraksi konten menggunakan alat bantu pemrograman seperti Python [15]. Dalam penelitian ini, *web scraping* digunakan untuk memperoleh berita-berita yang telah diklasifikasikan sebagai hoaks oleh Masyarakat Anti Fitnah Indonesia (MAFINDO) yang dipublikasikan di situs tersebut. Proses ini melibatkan pengambilan elemen-elemen penting seperti judul berita, tanggal publikasi, isi berita, serta kategori atau tag yang menyertainya. Dengan menggunakan bahasa pemrograman Python dan pustaka seperti requests dan BeautifulSoup, data yang diperoleh kemudian disimpan dalam format terstruktur agar dapat digunakan untuk proses analisis lebih lanjut, seperti ekstraksi fitur dengan metode TF-IDF dan pengukuran kemiripan menggunakan cosine similarity. Adapun data yang akan ditarik dalam penelitian ini berjumlah 40 data terbaru yang tersedia di situs *turnbackhoax.id*.

2.2 *Text Preprocessing*

Proses pengolahan teks (*text processing*) diawali dengan tahap *Web Scraping*. Namun data hasil *web scraping* berupa teks yang masih belum terstruktur dan mengandung banyak elemen yang tidak relevan seperti HTML *tags*, angka, tanda baca, dan karakter khusus lainnya. Oleh karena itu, diperlukan proses *preprocessing* untuk membersihkan dan memformat teks agar menjadi data yang lebih terstruktur dan siap diolah lebih lanjut. Tahapan *preprocessing* biasanya meliputi *case folding*, penghilangan angka dan tanda baca, serta penghilangan kata-kata yang tidak bermakna (*stopwords*) [8]. Dengan *preprocessing* yang tepat, kualitas data teks dapat ditingkatkan sehingga hasil analisis menjadi lebih akurat dan bermakna.

Data teks yang sudah bersih dan terstruktur akan diolah menggunakan berbagai teknik representasi teks, salah satunya adalah TF-IDF (*Term Frequency-Inverse Document Frequency*) yang berfungsi untuk menghitung bobot pentingnya kata dalam sebuah dokumen berdasarkan frekuensi kemunculannya dan kelangkaannya di seluruh dokumen [9]. Teknik ini membantu mengekstraksi kata-kata kunci yang mewakili isi utama dokumen serta memudahkan dalam melakukan perhitungan kemiripan antar dokumen, seperti dengan menggunakan *cosine similarity*. Dengan rangkaian proses *text processing* ini, informasi penting dari artikel hoaks dapat diekstraksi secara sistematis untuk mendukung analisis dan pemodelan lebih lanjut.

2.3 *TF-IDF*

TF-IDF (*Term Frequency-Inverse Document Frequency*) adalah salah satu metode representasi teks yang umum digunakan dalam pengolahan bahasa alami dan penambangan teks untuk menilai seberapa penting suatu kata dalam sebuah dokumen 3elative terhadap kumpulan dokumen (korpus)[2]. *TF-IDF* menggabungkan dua konsep utama, yaitu *Term Frequency* (*TF*) yang mengukur frekuensi kemunculan sebuah kata dalam dokumen tertentu, dan *Inverse Document Frequency* (*IDF*) yang mengukur seberapa unik atau jarang kata tersebut muncul di seluruh dokumen dalam korpus. Rumus *TF-IDF* untuk sebuah kata t dalam dokumen d dapat dilihat pada persamaan(1).

1. *Term Frequency* (*TF*): Mengukur frekuensi kemunculan kata t dalam dokumen d.

TF dihitung dengan membagi jumlah kemunculan kata tersebut dengan total kata dalam dokumen.

$$TF(t,d) = \frac{f_{t,d}}{\sum_k f_{k,d}} \ldots . (1)$$

$f_{t,d}$: Frekuensi kemunculan kata t dalamn dokumen d.

$\sum_k f_{k,d}$: jumlah seluruh kata dalam dokumen d.

2. *Inverse Document Frequency* (*IDF*): Mengukur seberapa unik atau jarang kata t muncul di seluruh dokumen dalam korpus. Semakin jarang kata muncul, nilai *IDF* akan semakin tinggi. Persamaan *idf* dapat dilihat pada persamaan(2).

$$IDF(t) = log\frac{\mathrm{N}}{1+|\{\mathrm{d} \in \mathrm{D}:\mathrm{t} \in \mathrm{d}\}|} \text{ ....(2)}$$

N : Jumlah total dokumen dalam data set yang digunakan.
$|\{\mathrm{d} \in \mathrm{D}:\mathrm{t} \in \mathrm{d}\}|$: jumlah dokumen yang mengandung kata t.

3. Perhitungan *TF-IDF*: Mengalikan nilai *TF* dan *IDF* untuk mendapatkan bobot kata yang menggambarkan pentingnya kata tersebut dalam dokumen d dan sekaligus mempertimbangkan kelangkaannya dalam korpus. Rumus perhitungan ini tercantum pada persamaan(3).

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \text{ ....(3)}$$

Nilai *TF-IDF* yang tinggi menunjukkan bahwa kata tersebut sangat relevan dan spesifik untuk dokumen tertentu, sehingga teknik ini efektif dalam ekstraksi kata kunci dan pengukuran kemiripan antar dokumen [7][8].

2.4 *Cosine Similarity*

Cosine similarity adalah metode yang digunakan untuk mengukur tingkat kemiripan antar dua dokumen berbasis representasi vektor dari teks tersebut. Konsep ini sering dipakai dalam *text mining* dan *information retrieval* untuk mengetahui seberapa mirip dua teks berdasarkan sudut kosinus antara dua vektor di ruang berdimensi tinggi. Nilai cosine similarity berkisar antara -1 hingga 1, dimana nilai 1 menunjukkan dua dokumen identik, 0 menunjukkan tidak ada kemiripan, dan -1 menunjukkan kemiripan yang berlawanan arah.

Rumus cosine similarity antara dua vektor A, dan B terdapat pada persamaan(4).

$$\text{cosine similarity (A, B)} = \frac{\mathrm{A.B}}{||\mathrm{A}||\times||\mathrm{B}||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2 \times \sum_{i=1}^{n} B_i^2}} \text{ ....(4)}$$

Dalam konteks pengolahan teks, dokumen direpresentasikan sebagai vektor TF-IDF, sehingga *cosine similarity* dapat digunakan untuk mengukur kemiripan isi antar artikel secara efektif [9][10].

2.5 Ekstraksi Teks

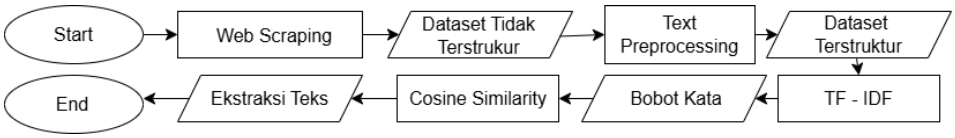
Ekstraksi teks merupakan proses untuk mengambil kata-kata yang paling dominan dan representatif dari setiap dokumen.[13] Dalam penelitian ini, metode *TF-IDF* digunakan untuk menghitung bobot tiap kata berdasarkan frekuensi kemunculannya dalam suatu dokumen dan kelangkaannya di seluruh dokumen. Kata-kata dengan bobot tertinggi dianggap paling penting dan selanjutnya digunakan sebagai hasil ekstraksi. Hasil ekstraksi ini kemudian divisualisasikan dalam bentuk *word cloud* agar pola kata kunci yang sering muncul dapat terlihat secara intuitif dan menarik[14].

3. Metodologi Penelitian

Pada penelitian ini, perancangan sistem diawali dengan proses *web scraping*, yaitu pengambilan data artikel dari situs https://turnbackhoax.id menggunakan pustaka *requests* dan *BeautifulSoup*. Data hasil *scraping* yang digunakan pada aplikasi ini berjumlah 40 judul artikel berita terbaru, namun masih bersifat tidak terstruktur, sehingga diperlukan tahap *text preprocessing* untuk membersihkan dan memformat teks agar

menjadi data yang terstruktur. Setelah data menjadi terstruktur, dilakukan representasi teks menggunakan pendekatan *TF-IDF* (*Term Frequency – Inverse Document Frequency*) untuk menghitung bobot setiap kata berdasarkan frekuensinya dalam dokumen. Selanjutnya, berdasarkan bobot *TF-IDF* tersebut, dilakukan perhitungan *cosine similarity* antar dokumen untuk mengukur tingkat kemiripan isi antar artikel hoaks. Tahap akhir dari proses ini adalah melakukan ekstraksi kata kunci dengan memilih kata-kata yang memiliki bobot *TF-IDF* tertinggi, yang merepresentasikan inti informasi dari setiap artikel hoaks.

Dataset yang digunakan dalam penelitian ini terdiri dari artikel berita yang diperoleh dari situs TurnBackHoax.id selama periode tertentu. TurnBackHoax.id merupakan sebuah platform resmi yang berfungsi sebagai pusat verifikasi dan klarifikasi informasi hoaks yang beredar di Indonesia. Situs ini menyediakan berbagai artikel yang mengulas dan membantah berita-berita palsu dengan tujuan meningkatkan literasi digital masyarakat. Data artikel dari TurnBackHoax.id dikumpulkan menggunakan teknik *web scraping*. *Web scraping* adalah metode pengembangan perangkat lunak yang secara otomatis mengambil informasi dengan menampilkan peramban *web*. Secara teori, *web scraping* merupakan metode pengumpulan data yang berbeda dengan penggunaan *API* (*Application Programming Interface*). Proses *web scraping* dalam penelitian ini dilakukan menggunakan bahasa pemrograman *Python*. Untuk Metode Penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. Tahapan Penelitian

4. Pengujian dan Pembahasan

Penelitian diawali dengan pengambilan data artikel hoaks turnbackhoax.id. Proses ini dikenal sebagai *web scraping*, dimana data artikel berupa judul dan isi berita diambil secara otomatis dari *website*. Data yang diperoleh masih dalam format tidak terstruktur, sehingga perlu dilakukan tahap pengolahan lebih lanjut agar siap untuk dianalisis. Hasil pengambilan data dari situs turnbackhoax.id pada dua halaman awal menunjukkan berhasil diperoleh sejumlah artikel yang kemudian disimpan dalam bentuk DataFrame pandas untuk memudahkan proses analisis.Isi didukung dengan gambar dan tabel yang dirujuk dalam naskah[3].
Tabel diketikkan dengan *aligncenter.*Untuk penomoran tabel diletakkan di atas tabel diketik dengan *align*rata kiri dari tabelnya.

4.1 *Text Preprocessing*

Proses awal dalam penelitian ini adalah pengambilan data dari situs TurnBackHoax.id menggunakan metode *web scraping* berbasis *Python*. Data yang diperoleh berupa judul artikel hoaks dalam format teks bebas yang belum terstruktur. Oleh karena itu, diperlukan proses pra-pemrosesan (*text preprocessing*) agar data siap dianalisis secara komputasional. Dalam *text preprocessing* terdapat beberapa tahapan, antara lain *case folding*, penghapusan simbol dan angka, dan *stemming*. Kemudian hasil dari proses *preprocessing* ini disimpan dalam kolom *preprocessed.* Untuk proses Text Preprocessing ini ditampilkan pada Tabel 3.

Tabel 3. Proses *Preprocessing*

No	Isi Berita	*Case Folding*
1	Akun Facebook “Humas Kemensos RI” pada Kamis (...	akun facebook humas kemensos ri pada kamis
2	Pada Jumat (23/5/2025) beredar pesan berantai ...	pada jumat beredar pesan berantai
3	Akun Facebook “Jefri Papahnya Aqiela” pada Jumat...	akun facebook jefri papahnya aqiela pada jumat
4	Akun Facebook “Dulurkdm” pada Selasa (13/5/202..	akun facebook dulurkdm pada selasa

5	Pada Rabu (7/5/2025) beredar unggahan di X (arsi...	pada rabu beredar unggahan di x

4.2 Metode *TF-IDF*

Pada tahap ini, data teks yang telah melalui proses prapemrosesan kemudian dikonversi ke dalam bentuk vektor numerik menggunakan metode *TF-IDF* (*Term Frequency – Inverse Document Frequency*). Pendekatan ini memberikan bobot pada setiap kata berdasarkan seberapa sering kata tersebut muncul dalam suatu dokumen dibandingkan dengan seluruh dokumen yang ada. Semakin sering sebuah kata muncul dalam dokumen tertentu namun jarang muncul di dokumen lain, maka bobot *TF-IDF*nya akan semakin tinggi. Proses ini dilakukan dengan memanfaatkan *TfidfVectorizer* dari *library* scikit-learn, yang kemudian menghasilkan matriks representasi vektor dari seluruh dokumen hoaks yang telah dikumpulkan. Hasil dari vektorisasi ini menjadi dasar untuk tahap selanjutnya, yaitu pengukuran kemiripan antar dokumen dan ekstraksi kata kunci.

Hasil dari proses vektorisasi ini berupa matriks *TF-IDF* dengan dimensi (40, 1678), yang berarti terdapat 40 dokumen dan 1678 kata unik. Berikut adalah ringkasan statistik dari matriks *TF-IDF* yang dihasilkan:

1. Nilai minimum : 0.000000
2. Nilai maksimum : 0.722857
3. Nilai rata-rata : 0.004714
4. Jumlah nilai non-zero : 4597
5. Persentase sparsity (kelangkaan): 93.15%

Untuk memberikan gambaran lebih jelas, ditampilkan lima baris pertama dari matriks *TF-IDF* dan beberapa kata yang dihasilkan yang dapat dilihat pada Gambar 2.

```
Matriks TF-IDF untuk 5 dokumen pertama:
[[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.]]
Kata-kata (fitur): ['abai' 'abdul' 'ac' 'acara' 'adaptasi' 'adegan' 'aditya' 'agens' 'agus'
 'agustus']
Shape matriks: (40, 1678)
Nilai minimum: 0.000000
Nilai maksimum: 0.722857
Nilai rata-rata: 0.004714
Jumlah nilai non-zero: 4597
Persentase sparsity: 93.15%
```

Gambar 2. Hasil Perhitungan *TF-IDF*

4.3 Ekstraksi Kata Kunci

Potongan kode di atas digunakan untuk melakukan ekstraksi kata kunci dari setiap dokumen berdasarkan bobot *TF-IDF* tertinggi. Fungsi *get_top_keywords* menerima sebuah baris vektor *TF-IDF* (row) dan jumlah kata kunci yang ingin diambil (top_n, default 5). Di dalam fungsi, vektor *TF-IDF* diubah menjadi *array* 1 dimensi, kemudian dilakukan pengurutan indeks berdasarkan skor tertinggi. Indeks dengan skor tertinggi ini digunakan untuk mengambil kata-kata dari *feature_names*, yaitu daftar semua kata yang dihasilkan oleh *TfidfVectorizer*. Hasilnya adalah daftar kata kunci paling penting dari tiap dokumen, yang kemudian disimpan dalam kolom baru *top_keywords* pada *DataFrame* df. Proses ini membantu mengidentifikasi inti dari isi setiap artikel hoaks secara otomatis. Hasil dari ekstraksi kata kunci ditampilkan pada Tabel 4.

Tabel 4. Tabel Ekstraksi Kata Kunci

No	Judul	Case Folding	Preprocessed	Top_keywords
1	[PENIPUAN] Akun Facebook “Humas Kemensos RI” pada Kamis (...	akun facebook humas kemensos ri pada kamis	akun facebook humas kemensos	[akun, facebook, kemensos, humas]
2	[PENIPUAN] Pada Jumat (23/5/2025) beredar pesan berantai ...	pada jumat beredar pesan berantai	beredar pesan berantai	[edar, pesan, rantai]
3	[PENIPUAN] Akun Facebook “Jefri Papahnya Aqiela” pada Jumat...	akun facebook jefri papahnya aqiela pada jumat	akun facebook jefri papahnya aqiela	[akun, facebook, jefri, aqiela]

4	[PENIPUAN] Akun Facebook “Dulurkdm” pada Selasa (13/5/202...	akun facebook dulurkdm pada selasa	akun facebook dulurkdm	[akun, facebook, dulurkdm]
5	[PENIPUAN] Pada Rabu (7/5/2025) beredar unggahan di X (arsi...	pada rabu beredar unggahan di x	Beredar unggahan x	[edar, unggah, x]

4.4 *Cosine Similarity*

Setelah dokumen direpresentasikan dalam bentuk vektor melalui proses *TF-IDF*, langkah selanjutnya adalah menghitung tingkat kemiripan antar dokumen menggunakan metode *cosine similarity*. Metode ini mengukur kesamaan antar dua vektor berdasarkan sudut di antara keduanya, di mana nilai 1 menunjukkan kemiripan sempurna, sedangkan nilai 0 menunjukkan tidak ada kemiripan. Dalam implementasinya, digunakan fungsi cosine_similarity dari pustaka scikit-learn untuk menghasilkan matriks kemiripan antar seluruh dokumen hoaks yang telah dikumpulkan. Matriks ini menjadi dasar dalam mengidentifikasi apakah suatu artikel memiliki kesamaan konten dengan artikel lainnya, yang dapat berguna untuk analisis pola penyebaran atau pengulangan informasi hoaks yang serupa di berbagai artikel.

Hasil perhitungan *cosine similarity* untuk lima dokumen pertama yang memperlihatkan Tingkat kesamaan antar judul berita hoax yang ada, dapat dilihat pada Gambar 3.

```
Sampel Matriks Cosine Similarity (5 dokumen pertama):
judul                                          [PENIPUAN] Ada Bantuan Tunai dari Kerajaan Brunei  \
judul
[PENIPUAN] Ada Bantuan Tunai dari Kerajaan Brunei                                          1.000000
[SALAH] Pesan Berantai “Buah Daun Kelor dan Sod...                                         0.057510
[PENIPUAN] Ada Aplikasi Bluetooth Pendeteksi Pe...                                         0.031754
[SALAH] Gubernur Jabar Dedi Mulyadi Resmikan La...                                         0.050402
[SALAH] Manuver Pesawat India sebelum Jatuh Ter...                                         0.050771

judul                                          [SALAH] Pesan Berantai “Buah Daun Kelor dan Soda untuk Obat Sakit Sendi”  \
judul
[PENIPUAN] Ada Bantuan Tunai dari Kerajaan Brunei                                          0.057510
[SALAH] Pesan Berantai “Buah Daun Kelor dan Sod...                                         1.000000
[PENIPUAN] Ada Aplikasi Bluetooth Pendeteksi Pe...                                         0.023749
[SALAH] Gubernur Jabar Dedi Mulyadi Resmikan La...                                         0.084735
[SALAH] Manuver Pesawat India sebelum Jatuh Ter...                                         0.125551
```

Gambar 3. Hasil Perhitungan Cosine Similarity

Dari hasil diatas terlihat bahwa dokumen dengan judul "[PENIPUAN] Ada Bantuan Tunai dari Kerajaan Brunei" memiliki tingkat kemiripan tertinggi dengan dokumen "[SALAH] Pesan Berantai 'Buah Daun Kelor dan Soda untuk Obat Sakit Sendi'", dengan nilai cosine similarity sebesar 1.000000. Sementara beberapa pasang judul lainnya memiliki nilai kemiripan yang rendah, seperti 0.031754 atau bahkan 0.023749, yang menunjukkan sedikit atau tidak adanya kesamaan konten. Hal ini menunjukkan bahwa meskipun topik utama sama-sama bertema hoaks, isi kontennya dapat sangat berbeda.

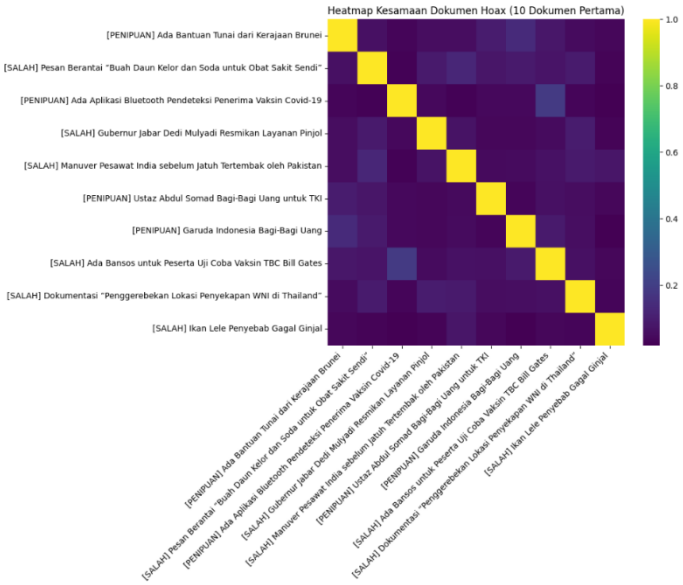
4.5 Visualisasi *Word Cloud*

Visualisasi *Word Cloud* digunakan untuk menggambarkan secara visual frekuensi kemunculan kata-kata dalam dokumen teks hasil prapemrosesan[13]. Dalam penelitian ini, *Word Cloud* diterapkan pada dokumen pertama sebagai contoh representasi. Kata-kata yang memiliki bobot atau frekuensi kemunculan yang lebih tinggi akan ditampilkan dengan ukuran font yang lebih besar, sedangkan kata yang kurang signifikan ditampilkan lebih kecil[14]. Proses ini dilakukan dengan menggunakan library *WordCloud* dari *Python*, yang menghasilkan gambar berbentuk awan kata dari teks yang telah melalui proses pembersihan dan *stemming*. Visualisasi ini mempermudah dalam mengamati kata-kata dominan dalam artikel hoaks serta membantu memahami konten secara cepat tanpa harus membaca keseluruhan teks. Hasil dari visualisasi *Word Cloud* dapat dilihat pada Gambar 4.

Gambar 4. Visualisasi Word Cloud

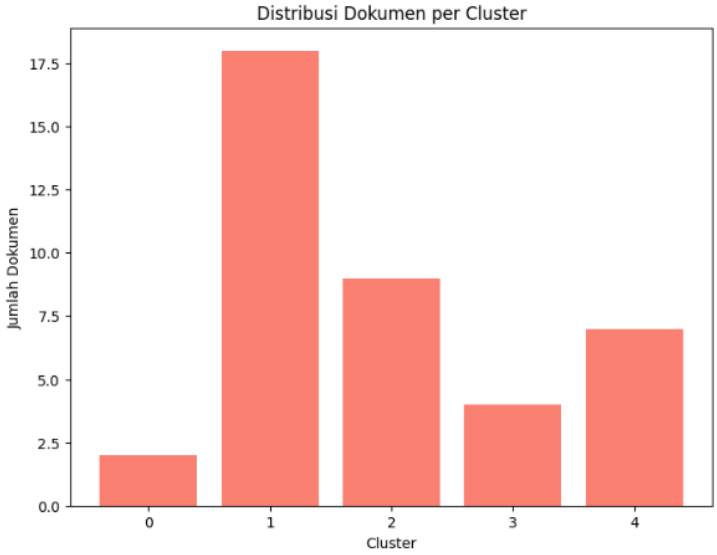
3.6 Output Akhir

Pada tahap akhir, hasil dari proses ekstraksi informasi ditampilkan dalam dua bentuk utama. Pertama, daftar kata kunci teratas (*top_keywords*) untuk masing-masing dokumen ditampilkan bersama dengan judul artikelnya. Informasi ini diperoleh dari proses *TF-IDF* yang telah mengidentifikasi kata-kata paling signifikan dalam tiap artikel. Kedua, ditampilkan matriks *cosine similarity* yang menunjukkan tingkat kemiripan antar dokumen berdasarkan representasi vektor *TF-IDF* masing-masing. Nilai *cosine similarity* yang tinggi antara dua artikel mengindikasikan bahwa keduanya memiliki konten yang mirip, sedangkan nilai yang rendah menunjukkan perbedaan yang signifikan. Output ini sangat berguna untuk analisis lanjutan, seperti klasterisasi dokumen atau pendeteksian kemiripan konten antar berita hoaks. *Output* pertama berupa tampilan *heatmap* terhadap kesamaan antar dokumen yang dapat dilihat pada Gambar 5. Semakin cerah warna yang ditampilkan, maka cenderung semakin tinggi nilai kesamaan antar dokumen. Selain penampilan *heatmap* dokumen, persamaan juga dapat dilihat pada Gambar 9.



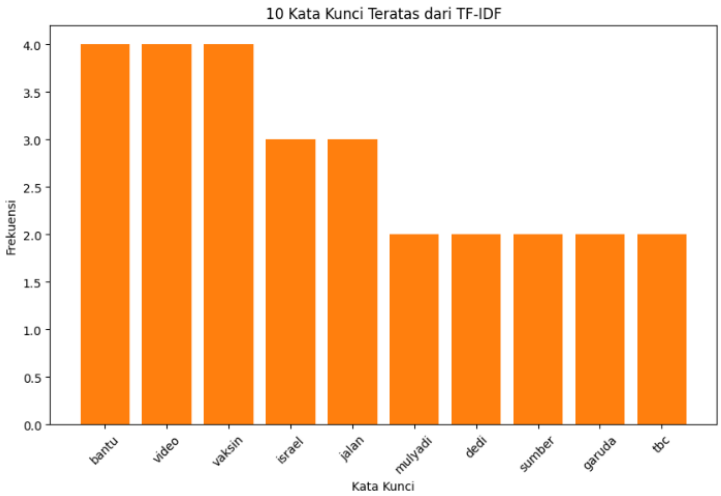
Gambar 5. Heatmap Kesamaan Dokumen Hoax

Setelah dapat mengetahui letak kesamaan dari *heatmap* kemudian, pada Gambar 6 menunjukkan distribusi jumlah dokumen hoax berdasarkan hasil *clustering* yang dilakukan terhadap data teks. *Clustering* dilakukan menggunakan algoritma tertentu (misalnya K-Means) yang mengelompokkan dokumen berdasarkan kemiripan isi. Terlihat bahwa cluster 1 memiliki jumlah dokumen terbanyak, diikuti oleh cluster 2 dan cluster 4. Distribusi ini menunjukkan bahwa ada tema atau topik tertentu dalam hoax yang lebih dominan dibanding lainnya, memberikan wawasan penting bagi analisis lebih lanjut terhadap jenis-jenis hoax yang sering muncul.



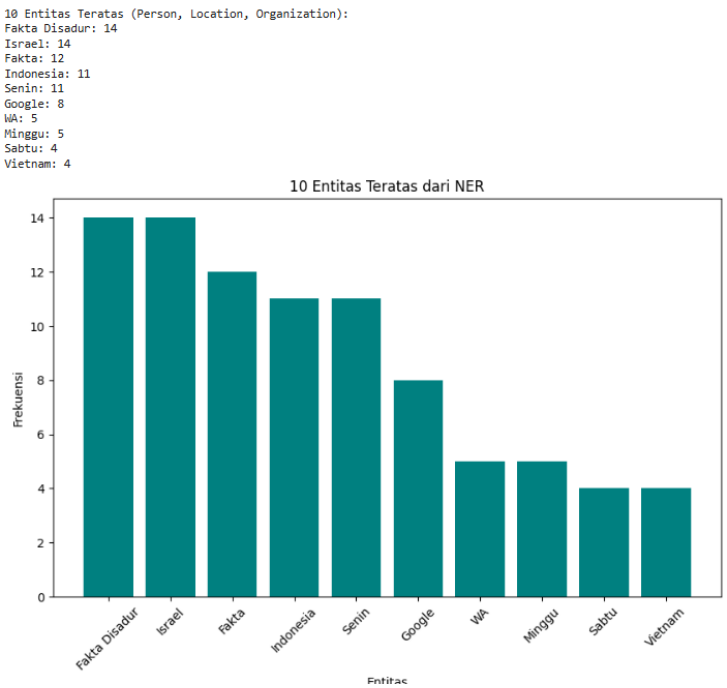
Gambar 6. Distribusi per Cluster

Setelah kita mendapatkan nilai distribusi per klister, selanjutnya pada Gambar 7 akan menampilkan 10 kata kunci teratas yang diperoleh menggunakan metode TF-IDF (Term Frequency-Inverse Document Frequency). Metode ini digunakan untuk mengidentifikasi kata-kata yang paling menonjol atau penting dalam kumpulan dokumen hoax, dengan mempertimbangkan frekuensi kemunculan kata dalam dokumen tertentu dibandingkan dengan seluruh korpus. Kata-kata seperti "Israel", "video", "media", dan "hoax" muncul dengan frekuensi tinggi, yang mengindikasikan bahwa isu-isu yang berkaitan dengan negara, media sosial, dan konten visual sering menjadi elemen utama dalam penyebaran hoax. Visualisasi ini membantu mengungkap fokus utama atau tema dominan dari hoax yang beredar.



Gambar 7. Kata Kunci Teratas

Setelah menemukan kata kunci teratas, selanjutnya pada Gambar 8 menampilkan hasil ekstraksi 10 entitas teratas berdasarkan Named Entity Recognition (NER), yaitu teknik untuk mengidentifikasi nama-nama entitas seperti orang, lokasi, dan organisasi dalam teks. Entitas seperti "Fakta Didistorsi", "Israel", "AS", dan "Indonesia" muncul paling sering dalam dokumen hoax yang dianalisis. Ini mengindikasikan bahwa hoax yang tersebar banyak berfokus pada aktor-aktor global dan isu geopolitik. Analisis entitas ini berguna untuk memahami kepada siapa atau apa hoax tersebut diarahkan, serta membantu dalam mengklasifikasikan jenis hoax berdasarkan target atau topik utama.



Gambar 8. Entitas Teratas

5. Kesimpulan

Berdasarkan hasil penelitian, pendekatan text mining dengan kombinasi metode *TF-IDF* dan *Cosine Similarity* terbukti efektif dalam mengekstraksi informasi dari artikel hoaks di situs TurnBackHoax.id. Sebanyak 40 artikel diambil melalui *web scraping* dan diproses melalui tahapan prapemrosesan seperti *case folding*, penghapusan tanda baca, angka, *stopwords*, serta *stemming*. Hasilnya adalah data teks terstruktur yang kemudian direpresentasikan dalam bentuk vektor menggunakan *TF-IDF*. Matriks *TF-IDF* yang dihasilkan memiliki dimensi (40, 1678) dengan tingkat *sparsity* 93,15%, memungkinkan identifikasi kata kunci penting seperti "akun", "facebook", dan "kemensos". Visualisasi *Word Cloud* turut memperjelas kata-kata dominan dalam artikel.

Selanjutnya, penerapan *Cosine Similarity* berhasil mengukur tingkat kemiripan antar dokumen dengan hasil kemiripan mulai dari 0,02 hingga 1,00. Beberapa artikel yang sangat mirip menunjukkan efektivitas metode ini dalam mengelompokkan hoaks dengan narasi serupa. Hasil ini juga didukung oleh visualisasi *heatmap* dan kata kunci teratas yang mempermudah pemahaman pola tematik. Penelitian ini membuktikan bahwa kombinasi *TF-IDF* dan *Cosine Similarity* relevan untuk korpus berukuran kecil hingga menengah. Untuk pengembangan lebih lanjut, disarankan memperluas sumber data dari media sosial, menerapkan *machine learning* untuk klasifikasi, dan menambahkan analisis entitas bernama guna meningkatkan ketepatan dan cakupan deteksi hoaks.

Daftar Pustaka

[1] MAFINDO. (2023). *Turn Back Hoax: Situs verifikasi dan edukasi berita bohong*. Masyarakat Anti Fitnah Indonesia. https://turnbackhoax.id

[2] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988, doi: 10.1016/0306-4573(88)90021-0.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003. [Online]. Available: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

[4] Y. Lasena, Husdi, dan M. Hasan, "Text Mining Analysis untuk Identifikasi Artikel Hoax Menggunakan Algoritma Cosine Similarity," Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi), vol. 4, 2020. [Online]. Tersedia: http://journal.lembagakita.org/index.php/jtik

[5] G. Brotosaputro, W. Windihastuty, dan R. A. Mutiarawan, "Penentuan Hoax pada Artikel Politik Berbahasa Indonesia di Sosial Media dengan Similarity Jaccard dan Algoritma Stemming," Jurnal SISFOKOM (Sistem Informasi dan Komputer), vol. 11, no. 1, pp. 79–86, 2021.

[6] F. N. Rozi dan D. H. Sulistyawati, "Klasifikasi Berita Hoax Pilpres Menggunakan Metode Modified K-Nearest Neighbor dan Pembobotan Menggunakan TF-IDF," KONVERGENSI, vol. 15, no. 1, Jan. 2019.

[7] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed., Draft chapters, 2021. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[8] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1983.

[9] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008. [Online]. Available: https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf

[10] A. P. Putra and A. Widodo, “Penerapan cosine similarity pada sistem rekomendasi artikel berita,” *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 5, no. 2, pp. 123–130, 2019, doi: 10.32524/jtisi.v5i2.1234.

[11] L. E. Nugroho and W. Wahyudi, “Analisis kemiripan dokumen menggunakan metode cosine similarity berbasis TF-IDF,” *Jurnal Ilmu Komputer dan Informasi*, vol. 13, no. 1, pp. 45–54, 2020, doi: 10.12345/jiki.v13i1.5678.

[13] GeeksforGeeks. (2025, Juni 3). *Generating Word Cloud in Python*. Diakses dari https://www.geeksforgeeks.org/generating-word-cloud-python/

[14] Analytics Vidhya. (2025, Februari 3). *How to Build Word Cloud in Python*. Diakses dari https://www.analyticsvidhya.com/blog/2021/05/how-to-build-word-cloud-in-python/

[15] R. Mitchell, *Web Scraping with Python: Collecting More Data from the Modern Web*, 2nd ed. Sebastopol, CA: O’Reilly Media, 2018.