

Relevansi Berita terhadap Kasus Korupsi Dana Iklan Bank BJB Menggunakan TF-IDF dan Cosine Similarity

Rizki Ambarwati¹⁾, Cindy Meilani²⁾, Hafiz Irsyad³⁾

¹⁾²⁾³⁾Informatika, Fakultas Ilmu Komputer dan Rekayasa, Universitas Multi Data Palembang
Jl Rajawali No 14, Palembang

¹⁾rizkiambarwati_2226250079@mhs.mdp.ac.id

²⁾cindymeilani_2226250107@mhs.mdp.ac.id

³⁾hafizirsyad@mdp.ac.id

Abstrak

Kasus korupsi dana iklan Bank BJB menjadi perhatian publik karena menyangkut penyalahgunaan dana yang melibatkan institusi keuangan milik pemerintah daerah. Banyaknya pemberitaan dari berbagai media membuat informasi yang tersebar sulit untuk disaring berdasarkan relevansinya terhadap inti kasus. Penelitian ini bertujuan untuk menganalisis dan mengukur relevansi berita-berita yang beredar terhadap kasus korupsi dana iklan Bank BJB dengan memanfaatkan metode *Term Frequency-Inverse Document Frequency* (TF-IDF) dan *Cosine Similarity*. Data yang digunakan berupa kumpulan artikel berita daring dari berbagai sumber media yang membahas topik terkait. Tahapan penelitian dimulai dengan preprocessing teks, seperti pembersihan data, tokenisasi, *stopword removal*, dan *stemming*. Dokumen yang memiliki tingkat kemiripan tinggi dianggap relevan terhadap berita acuan utama kasus korupsi ini. Hasil pengujian TF-IDF dan Cosine Similarity menunjukkan tingkat *accuracy* sebesar 71%, *precision* sebesar 100%, *recall* 71 % dan hasil *F1 score* 83 % sehingga dengan metode tersebut dapat mengelompokkan berita secara efektif berdasarkan tingkat relevansinya.

Kata kunci: Bank BJB, Cosine Similarity, Korupsi, TF-IDF.

Abstract

The Bank BJB advertising fund corruption case has attracted public attention because it involves misuse of funds involving financial institutions owned by the local government. The large number of news reports from various media makes it difficult to filter the information circulating based on its relevance to the core of the case. This study aims to analyze and measure the relevance of circulating news to the Bank BJB advertising fund corruption case by utilizing the *Term Frequency-Inverse Document Frequency* (TF-IDF) and *Cosine Similarity* methods. The data used are a collection of online news articles from various media sources that discuss related topics. The research stages begin with text preprocessing, such as data cleaning, tokenization, *stopword removal*, and *stemming*. Documents that have a high level of similarity are considered relevant to the main reference news of this corruption case. The results of the TF-IDF and Cosine Similarity tests show an *accuracy* level of 71%, a *precision* of 100%, a *recall* of 71% and an *F1 score* of 83% so that this method can effectively group news based on its level of relevance.

Keywords: Bank BJB, Cosine Similarity, Corruption, TF-IDF.

1. PENDAHULUAN

Penyebaran teknologi informasi yang semakin meluas di era digital saat ini berlangsung sangat cepat, terutama melalui media online dan media sosial. Internet dan media konvensional telah menjadi dua bentuk media yang paling dominan dalam menyajikan berita dan informasi kepada masyarakat [1]. Informasi yang tersebar di media sosial berpengaruh besar terhadap persepsi masyarakat [2]. Maraknya isu-isu sosial dan politik pada saat ini adalah kasus dugaan

korupsi. Salah satu yang menjadi sorotan dan perbincangan publik saat ini adalah dugaan penggelapan dana yang dilakukan oleh Bank BJB. Bank BJB yang merupakan bagian dari PT Bank Pembangunan Daerah Jawa Barat dan Banten, yang pusatnya terletak di daerah Bandung. Kasus ini tengah diselidiki oleh Komisi Pemberantasan Korupsi (KPK) dengan dugaan bahwa dana iklan yang seharusnya digunakan untuk media promosi tetapi disalahgunakan untuk praktik korupsi yang melibatkan persekongkolan dan penggelapan anggaran. Hingga saat ini, pihak KPK juga telah menetapkan 5 pihak yang terlibat dengan bukti transfer dan dana yang dipakai untuk belanja oleh pihak yang bersangkutan.

Dalam kasus yang terjadi akibat korupsi dana iklan Bank BJB ini, dibutuhkan suatu penyelesaian dengan melakukan pendekatan sistematis guna untuk mengukur sejauh mana penyebaran berita tersebar luas. Dikarenakan sebagian masyarakat Indonesia masih kesulitan dalam membedakan mana yang merupakan berita valid dan hoax terkait kasus korupsi. Hal ini dapat dilihat dari hasil survei Kantor Berita Indonesia ANTARA yang menyatakan bahwa dari 1,116 responden meneruskan berita bohong atau hoax seputar isu sosial politik tanpa memperhatikan ketidakjelasan sumber berita. Sementara 83,20 persen responden memeriksa kebenaran berita hoax dan sekitar 15,9 persen memilih mengabaikan berita tersebut. Namun, dalam survei tersebut juga mengungkapkan bahwa 91,8 persen responden sering menerima hoax dari media sosial seperti Line, WA atau Telegram [3]. Berdasarkan hasil survei Kantor Berita Indonesia ANTARA yang telah dilakukan menunjukkan bahwa sebagian masyarakat masih belum memperhatikan kebenaran berita terhadap isu hoax.

Salah satu pendekatan yang dapat digunakan untuk mengukur an untuk mengukur sejauh mana relevansi berita online adalah dengan metode *Term Frequency - Inverse Document Frequency* (TF-IDF) dan *cosine similarity*. Beberapa penelitian telah menunjukkan keberhasilan menggunakan metode ini. Penelitian sebelumnya melakukan klasifikasi berita online dari Kompas.com menunjukkan bahwa metode ini efektif dalam mengukur kemiripan konten antar dokumen menggunakan TF-IDF dan *cosine similarity* mencapai tingkat akurasi sebesar 91,25% [4]. Penelitian lain menerapkan TF-IDF dan *cosine similarity* dalam sistem klasifikasi dokumen skripsi dan memperoleh akurasi klasifikasi hingga 98% [5]. Hasil penelitian menunjukkan kinerja yang signifikan dengan skor akurasi 93.6% menggunakan metode TF-IDF dan *cosine similarity* serta representasi *bigram* saja [6].

Selain itu, efektivitas metode ini dalam mendeteksi kemiripan dokumen mampu menghasilkan nilai kemiripan dari masing-masing teks pembanding [7]. Penelitian [8] menyatakan bahwa pembobotan TF-IDF dan *cosine similarity* dianggap cocok untuk klasifikasi artikel pada website. Berdasarkan hal tersebut penelitian ini menjadikan media sosial menjadi topik yang menarik untuk diteliti. Saat ini banyak penelitian yang membahas seputar artikel berita. Penelitian ini bertujuan untuk menganalisis relevansi berita daring terhadap kasus korupsi dana iklan Bank BJB menggunakan TF-IDF dan *cosine similarity*. Penelitian ini dilakukan dengan mengumpulkan sejumlah artikel berita dan media yang memuat informasi tentang kasus korupsi dana iklan Bank BJB. Berita-berita yang didapat kemudian diproses mulai dari pembersihan data, analisis menggunakan TF-IDF. TF-IDF digunakan untuk pembobotan kata dalam dokumen, sedangkan *cosine similarity* digunakan untuk mengukur tingkat kemiripan antar dokumen dengan menghitung kedekatan vektor teks secara sistematis. Berdasarkan nilai yang dihasilkan dari proses perhitungan *cosine similarity* akan didapatkan sejumlah berita yang memiliki kemiripan. Hasil dari data tersebut akan menunjukkan berita mana yang paling relevan terhadap dokumen, yaitu berita yang memuat informasi tentang kasus korupsi Bank BJB, dari informasi tersebut dapat diidentifikasi mana berita yang tidak relevan. Dengan pendekatan TF-IDF dan *cosine similarity* yang digunakan memungkinkan untuk mengidentifikasi berita-berita yang relevan, sekaligus menghindari pengaruh informasi yang tidak sesuai.

Oleh karena itu, Penelitian ini diharapkan dapat memberikan gambaran yang jelas mengenai relevansi dan kredibilitas berita daring terutama tentang kasus korupsi Bank BJB, sehingga dapat membantu memperbaiki kualitas informasi yang beredar dan tersebar di ruang publik. Manfaat yang diharapkan dari penelitian ini meliputi peningkatan literasi digital masyarakat, penguatan penyaringan berita oleh media dan lembaga terkait, serta terciptanya

lingkungan informasi yang lebih sehat dan kondusif. Dengan demikian, penelitian ini tidak hanya berkontribusi pada bidang akademis, tetapi juga memiliki nilai praktis dalam meningkatkan kesadaran masyarakat dan mendukung penguatan sistem pemberantasan korupsi di Indonesia.

2. TINJAUAN PUSTAKA

2.1 Relevansi Berita

Relevansi memiliki kata dasar dari relevan. Relevan yang memiliki arti yaitu bersangkut-paut atau berguna secara langsung. Pengertian relevansi adalah hubungan antara dua hal yang saling terikat apabila kedua hal itu dicocokkan satu sama lain lalu memiliki keterkaitan satu dengan yang lainnya [9]. Begitu pula dengan keterkaitan antara relevan dan berita, yang dimana berita adalah sebuah informasi mengenai peristiwa yang terjadi di suatu lokasi yang bisa disajikan dalam bentuk teks maupun visual. Berita bisa ditemukan di berbagai portal berita dan media cetak. Umumnya setiap berita dikelompokkan berdasarkan kategori umum seperti yang meliputi berita terbaru, olahraga, politik, ekonomi, sosial, kesehatan, dan lain-lain [10]. Dalam konteks ini relevansi sangat penting karena untuk memilih berita yang kebenarannya memang benar atau berita itu termasuk hoax, dalam kasus korupsi dana iklan Bank BJB mengacu pada kesesuaian isi berita dengan suatu isu, topik dan peristiwa yang menjadi sorotan utama.

2.2 TF-IDF

Term Frequency –Inverse Document Frequency (TF-IDF) adalah teknik pembobotan berbasis statistik yang sering diterapkan di berbagai permasalahan penggalian informasi. Namun, secara umum TF-IDF tidak banyak dikenal sebagai algoritma untuk peringkasan teks otomatis [11]. Metode TF-IDF ini menggabungkan dua konsep untuk perhitungan bobot, yaitu frekuensi kemunculan sebuah kata di dalam sebuah dokumen tertentu dan inverse frekuensi dokumen yang mengandung kata tersebut [12]. Metode dalam kasus berita korupsi dana iklan Bank BJB ini menghitung frekuensi kemunculan kata dalam dokumen (*term frequency*) dan memperhitungkan seberapa sering kata tersebut muncul di seluruh dokumen teks berita yang ada (*inverse document frequency*). Berikut perhitungan matematis yang digunakan untuk menghitung TF-IDF pada kasus berita dana iklan Bank BJB, dapat dilihat pada persamaan 1:

$$IDF(word) = \log \frac{td}{df} \quad (1)$$

Keterangan:

$IDF(word)$: Nilai IDF dari setiap kata yang di cari

td : Jumlah keseluruhan dokumen yang ada

df :Jumlah kemunculan kata pada semua dokumen

2.3 Cosine Similarity

Metode *cosine similarity* merupakan metode yang digunakan untuk menghitung *similarity* (tingkat kesamaan) antar dua buah objek. Secara umum penghitungan metode ini didasarkan pada *vector space similarity measure* [13]. Kemiripan antar dokumen dihitung menggunakan suatu fungsi ukuran kemiripan (*similarity measure*). Semakin besar hasil fungsi *similarity*, maka kedua objek yang dievaluasi semakin mirip, demikian pula sebaliknya. Ukuran ini memungkinkan perbandingan dokumen sesuai dengan kemiripan atau relevansinya terhadap query. Dalam kasus korupsi dana iklan Bank BJB penggunaan *cosine similarity* digunakan untuk membandingkan 7 dokumen berita yang beredar dengan menghitung judul dan isi dari dokumen. Berikut perhitungan matematika yang digunakan untuk menghitung *cosine similarity* pada kasus berita dana iklan Bank BJB, dapat dilihat pada persamaan 2:

$$\cos\theta = \text{sim}(x, y) = \frac{x \cdot y}{||x|| ||y||} \quad (2)$$

Keterangan:

$\cos\theta$: Nilai cosinus antara dua vektor

$\text{sim}(x, y)$: Fungsi *Similarity*

$||x||$: Norm dari vektor

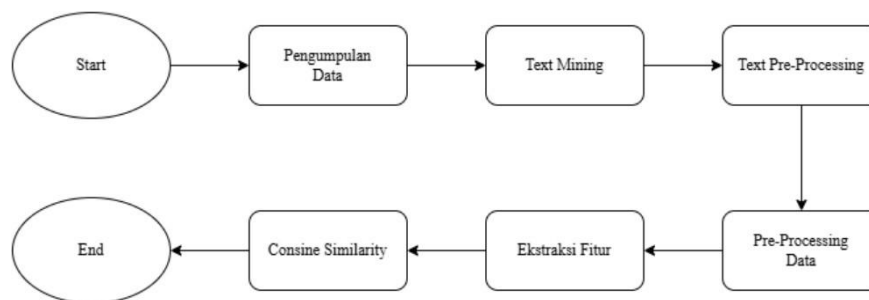
$||y||$: Norm dari vektor

2.4 Program

Penelitian ini menggunakan google collab sebagai platform implementasi kode program untuk membuktikan ke relevansian berita korupsi dana iklan Bank BJB dengan ekstraksi fitur TF-IDF dan *cosine similarity*.

3. METODE PENELITIAN

Penelitian ini menggunakan beberapa tahap, dari tahap termasuk *preprocessing*, menghitung bobot TF-IDF, memasukkan data, menghitung *cosine similarity*, dan hasil evaluasi. Alur penelitian dapat dilihat pada gambar 1.



Gambar 1. Alur Penelitian

3.1 Pengumpulan Data

Pengumpulan data dilakukan dengan mengumpulkan sejumlah berita *online* dari berbagai website yang membahas korupsi iklan Bank BJB. Pengambilan data dari url *website* berita lalu dimasukkan ke dalam *google docs* dengan menuliskan judul dari berita. Dokumen yang digunakan dalam penelitian ini menggunakan dokumen dari beberapa artikel berita karena kebanyakan berita yang ada tidak sesuai fakta yang ada dan untuk tingkat akurasi juga cukup dipertanyakan maka dari itu penelitian ini mengambil data dari artikel berita agar mengetahui berita mana yang relevan dan sesuai dengan keadaan saat ini.

3.2 Text Mining

Text Mining merupakan bagian dari data mining yang sama sama memiliki tujuan untuk menambang data untuk mencari sebuah pola atau keterkaitan unik yang merupakan perwakilan isi atau ciri khas sebuah dokumen teks. Sumber data pada *text mining* adalah kumpulan dokumen yang berupa data teks yang tidak terstruktur maka dari itu *text mining* digunakan untuk mengklasifikasikan apakah judul berita korupsi iklan Bank BJB menghasilkan kebenaran akan relevansi berita tersebut.

3.3 Text Preprocessing

Tahapan selanjutnya adalah preprocessing teks dimana mengolah data awal yang masih bermacam-macam untuk dijadikan sebuah data terstruktur yang siap dipakai [14]. Adapun proses dalam *text preprocessing* adalah sebagai berikut:

1. Case Folding

Case Folding adalah proses perubahan seluruh huruf atau karakter dalam teks menjadi huruf kecil. *Case folding* hanya memproses huruf alphabet dari “a” sampai dengan “z” saja agar karakter selain huruf akan dihapus. Maka dari itu, case folding akan mengkonversi keseluruhan teks dalam dokumen menjadi huruf kecil.

2. Tokenizing

Setelah proses *case folding* selesai selanjutnya yaitu proses tokenizing dimana pada proses ini teks akan dipecah dalam kata. Misalnya, kalimat “Kasus korupsi dana iklan Bank BJB” ini akan dipecah menjadi “kasus” “korupsi”, “dana”, “iklan”, “bank”, “bjb”. Adapun tujuan pada tahap *tokenizing* adalah untuk memecah data mentah menjadi bagian-bagian yang lebih kecil dan lebih mudah untuk dianalisis.

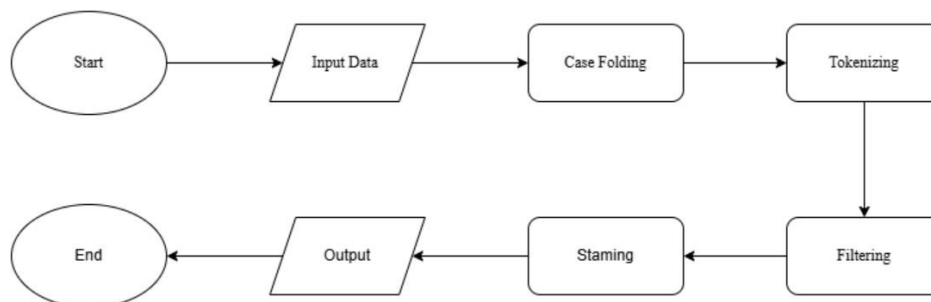
3. Filtering

Pada tahap *filtering* ini kata-kata penting dari hasil *tokenizing* diambil dengan menggunakan *stopword removal*. Pada proses *stopword* ini kata-kata umum yang sering muncul dan kurang penting akan dihapus.

4. Stemming

Proses stemming berguna untuk mengubah kata menjadi bentuk dasarnya, misalnya kata ‘dikorupsi’ akan diubah menjadi ‘korupsi’. Tujuan dari tahap *stemming* ini adalah untuk menghilangkan berbagai variasi kata yang sebenarnya memiliki makna yang sama.

Tahapan *text preprocessing* dapat dilihat pada gambar 2 berikut ini.



Gambar 2. Alur Pre-Processing

Hasil tahap preprocessing teks pada relevansi berita mengenai kasus korupsi iklan Bank BJB dapat dilihat pada tabel 1. Pada gambar menampilkan hasil dari preprocessing teks yang terdapat judul dari berita, isi berita, sumber dan tanggal berita tersebut diterbitkan.

Tabel 1. Hasil *Preprocessing* Teks

No	Judul	Isi	Sumber	Tanggal
0	Kasus Korupsi Dana Iklan Bank BJB, Dari Anggaran...	KPK mengungkap bahwa dari total anggaran Rp 40...	Tempo	2025-03-15
1	KPK Tetapkan 5 Tersangka Korupsi Pengadaan Ikl..	KPK menetapkan lima tersangka dalam kasus koru...	KPK	2025-03-14

No	Judul	Isi	Sumber	Tanggal
----	-------	-----	--------	---------

2	Korupsi Iklan BJB: Anggaran Rp 409 Miliar, Dir..	Dari anggaran Rp 409 miliar untuk iklan, hanya...	Detik	2025-03-13
3	Konstruksi Lengkap Kasus Bank BJB Rugikan Nega...	KPK mengungkap bahwa pengadaan iklan di Bank BJB	CNN Indonesia	2025-03-13

3.4 Preprocessing Data

Setelah tahap *preprocessing* teks selesai, selanjutnya adalah tahap *preprocessing* data. Pada tahap ini, data teks diproses berdasarkan judul dan isi setiap dokumen yang dapat dilihat pada tabel 2.

Tabel 2. Hasil *Preprocessing Data*

No	Judul	Isi	Sumber	Tanggal	Pemrosesan Judul	Pemrosesan Isi
0	Kasus Korupsi Dana Iklan Bank BJB, Dari Anggar...	KPK mengungkap bahwa dari total anggaran Rp 40...	Tempo	2025-03-15	korupsi dana iklan bank bjb anggar rp miliar r...	kpk ungkap total anggar rp miliar iklan bank b...
1	KPK Tetapkan 5 Tersangka Korupsi Pengadaan Ikl..	KPK menetapkan lima tersangka dalam kasus koru...	KPK	2025-03-14	kpk tetap sangka korupsi ada iklan bank bjb	kpk tetap sangka korupsi ada iklan bank bjb di...
2	Korupsi Iklan BJB: Anggaran Rp 409 Miliar, Dir..	Dari anggaran Rp 409 miliar untuk iklan, hanya...	Detik	2025-03-13	korupsi iklan bjb anggar rp miliar realisasi r...	anggar rp miliar iklan rp miliar realisasi ses...
3	Konstruksi Lengkap Kasus Bank BJB Rugikan Nega...	KPK mengungkap bahwa pengadaan iklan di Bank BJB	CNN Indonesia	2025-03-13	konstruksi lengkap bank bjb rugi negara rp miliar	kpk ungkap ada iklan bank bjb libat tunjuk age...

Tanpa *preprocessing* pada teks, hasil dari kemiripan tidak bisa akurat karena kata umum tidak diproses dan dihilangkan. Oleh karena itu, tahap *preprocessing* pada teks diperlukan sehingga bisa menghasilkan hasil kemiripan setiap dokumen teks berita.

3.5 Ekstraksi Fitur

Pada ekstraksi fitur yang dilakukan dan di implementasikan ke dalam kode program, peneliti menggunakan metode TF-IDF yang dimana metode ini digunakan untuk mengubah data teks pada kolom judul dan isi berita menjadi vektor numerik. Seluruh teks yang telah digabungkan berguna untuk membangun *vocabulary* yang kemudian menggunakan *TfidfVectorizer* dengan parameter *unigram*, *biagram*, dan batas minum kemunculan data sebanyak dua kali. Hasil dari TF-IDF berbentuk dataframe agar mudah dalam analisis data dalam tahap ekstraksi fitur yang dapat dilihat pada tabel 3 dimana hasil ekstraksi fitur untuk judul dan tabel 4 merupakan hasil ekstraksi fitur pada isi kasus berita korupsi iklan Bank BJB.

Tabel 3. Hasil Ekstraksi Fitur Judul

No	bank bjb	bjb	bjb anggaran	tetap	tetap sangka	tuju	tuju direktur
0	0.135110	0.126390	0.329806	0.000000	0.000000	0.0	0.0
1	0.143650	0.134379	0.000000	0.311995	0.311995	0.0	0.0
2	0.000000	0.133003	0.347062	0.000000	0.000000	0.0	0.0
3	0.218866	0.204741	0.000000	0.000000	0.000000	0.0	0.0

Tabel 4. Hasil Ekstraksi Fitur Isi

No	bank bjb	bjb	bjb anggaran	tetap	tetap sangka	tuju	tuju direktur
0	0.169574	0.158629	0.000000	0.000000	0.000000	0.206967	0.206967
1	0.085901	0.080357	0.000000	0.186569	0.186569	0.000000	0.000000
2	0.076251	0.071330	0.000000	0.000000	0.000000	0.186130	0.186130
3	0.112101	0.104866	0.000000	0.000000	0.000000	0.000000	0.000000

3.6 Cosine Similarity

Setelah ekstraksi fitur selesai, selanjutnya adalah menghitung nilai *cosine similarity* antara judul dan isi untuk setiap berita. Hasil dari perhitungan *cosine similarity* dapat dilihat pada tabel 5 berikut ini.

Tabel 5. Hasil Cosine Similarity

No	Judul	Isi	Cosine Similarity
0	Kasus Korupsi Dana Iklan Bank BJB, Dari Anggaran...	KPK mengungkap bahwa dari total anggaran Rp 40...	0.513299
1	KPK Tetapkan 5 Tersangka Korupsi Pengadaan Ikl..	KPK menetapkan lima tersangka dalam kasus koru...	0.612098
2	Korupsi Iklan BJB: Anggaran Rp 409 Miliar, Dir..	Dari anggaran Rp 409 miliar untuk iklan, hanya...	0.488527
3	Konstruksi Lengkap Kasus Bank BJB Rugikan Nega...	KPK mengungkap bahwa pengadaan iklan di Bank B...	0.417631

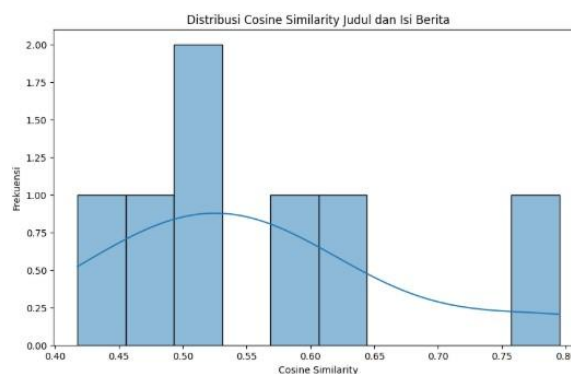
4. PEMBAHASAN

Penelitian ini mengukur relevansi dari 7 berita mengenai kasus korupsi dana iklan Bank BJB, dimana metode yang digunakan adalah TF-IDF dan *cosine similarity*. Setelah dilakukan *pre-processing* dengan tahapan *case folding*, *tokenisasi*, *stopword removal* dan *stemming*, selanjutnya data teks diambil dan kemudian data teks tersebut diubah menjadi vektor menggunakan TF-IDF. Tahapan *pre-processing* selesai, kemudian menghitung nilai *cosine similarity* antara judul dan isi untuk setiap berita. Hasil dari *cosine similarity* yang diklasifikasikan berdasarkan *threshold* ditampilkan hasil analisis dan visualisasi setiap sumber berita serta hasil analisis nilai *cosine similarity*nya adalah 5 relevan dan 2 tidak relevan yang dapat dilihat pada tabel 6.

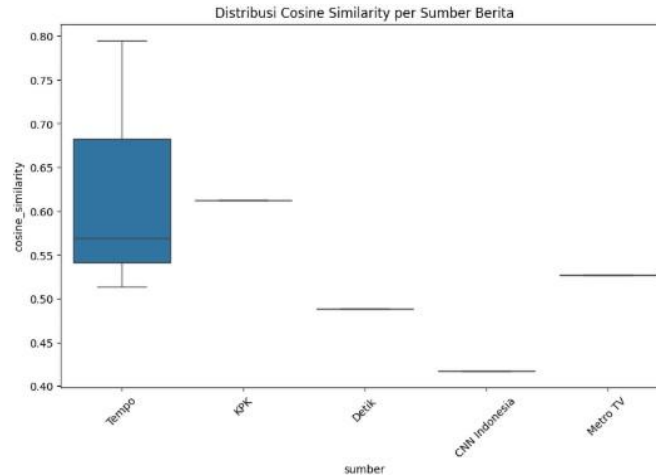
Tabel 6. Hasil Analisis Model

No	<i>Cosine similarity</i>	Relevan
0	0.513299	Ya
1	0.612098	Ya
2	0.488527	Ya
3	0.417631	Tidak
4	0.569312	Tidak
5	0.794973	Ya
6	0.527012	Ya

Berikutnya hasil analisis model dan ke relevansi berita ditampilkan melalui media grafik distribusi *cosine similarity* yang dimana dibagi menjadi dua grafik yaitu grafik distribusi judul dan berita serta grafik per sumber berita yang ditujukan pada gambar 3 dan gambar 4. Pada gambar 3 menunjukkan hasil grafik dari perhitungan distribusi *cosine similarity* yang dihitung dari judul dan isi berita menghasilkan frekuensi kemiripan berita awal 0.50 di pertengahan terdapat kenaikan 0.75 sampai frekuensi akhir mendapatkan 0.25 dari judul dan isi berita yang telah dihitung kemiripan dengan kemiripan tingkat sedang pada berita. Gambar 4 menunjukkan grafik distribusi *Cosine Similarity* per-sumber berita, sumber berita dari tempo menunjukkan variasi dan kemiripan yang paling tinggi, dan sumber berita yang lain memiliki kemiripan yang rendah serta bervariasi. Dari grafik yang ditampilkan dengan menunjukkan nilai *cosine similarity* berkisar 0,40 sampai 0,80 dari nama per-sumber berita.

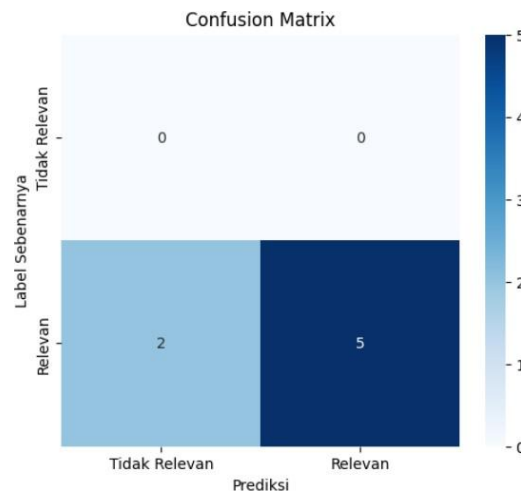


Gambar 3. Grafik Distribusi Judul dan Isi



Gambar 4. Grafik Distribusi Per-sumber Berita

Tahap terakhir yaitu evaluasi model menggunakan *confusion matrix* dimana hasil *accuracy* 71%, *precision* 100%, *recall* 71% dan *F1 score*nya 83%, yang ditampilkan pada gambar 5 dan tabel 7. Dimana pada gambar 5 menampilkan hasil prediksi yang menunjukkan 2 berita yang tidak relevan dan 5 berita yang relevan dari label sebenarnya dengan menggunakan diagram *confusion matrix*.



Gambar 5. Hasil Evaluasi Model

Tabel 7. Hasil Evaluasi Model

Evaluasi	Hasil
<i>Accuracy</i>	0.71
<i>Precision</i>	1.00
<i>Recall</i>	0.71
<i>F1 Score</i>	0.83

5. KESIMPULAN

Berdasarkan penelitian yang dilakukan dengan relevansi berita terhadap kasus korupsi dana iklan Bank BJB dengan penyelesaian menggunakan pendekatan TF-IDF dan *cosine similarity* menghasilkan hasil akurasi sebesar 71 %. Berdasarkan dataset berita yang dikumpulkan menjadi satu ke dalam *google docs* dimana dari 7 data berita tersebut 5 diantaranya termasuk berita yang relevan, sedangkan 2 data berita tidak termasuk ke dalam berita yang relevan.

DAFTAR PUSTAKA

- [1] A. Gani, "Pengaruh Media Sosial Terhadap Perkembangan Anak Remaja," *Jurnal Mitra Manajemen*, vol. 7, no. 2, pp. 32-44, 2015. <https://doi.org/10.35968/jmm.v7i2.533>
- [2] R. Haqqu, "Era Baru Televisi dalam Pandangan Konvergensi Media" *Jurnal Fotografi Televisi Animasi*, vol. 16, no. 1, pp. 15-20, 2020. <https://doi.org/10.24821/rekam.v16i1.3721>
- [3] M. A. Iskandar, "Survei: satu persen responden langsung teruskan hoax," ANTARA News, 2017, [Online]. Tersedia: <https://www.antaranews.com/berita/612245/@JambiAntara> [Diakses 4 Juni 2025].
- [4] B. Herwijayanti, D. E. Ratnawati and L. Muflikhah, "Klasifikasi Berita Online dengan menggunakan pembobotan TF-IDF dan Cosine Similarity" *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 1, pp. 306-312, 2018.
- [5] R. T. Wahyuni, D. Parstiyanto and E. Suprptono, "Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi" *Jurnal Teknik elektro*, vol. 9, no. 1, pp. 18-23, 2017. <https://doi.org/10.15294/jte.v9i1.10955>
- [6] A. Widiyanto, E. Pebriyanto, Fitriyani and Marna, "Document Similarity Using Term Frequency-Inverse Document Frequency Representation and Cosine Similarity," *Journal of Dinda : Data Science, Information Technology, dan Data Analytics*, vol. 4, no. 2, pp. 149–153, 2024. <https://doi.org/10.20895/dinda.v4i2.1589>
- [7] A. Riyani, M. Z. Naf'an and A. Burhanuddin, "Penerapan Cosine Similarity dan Pembobotan TF-IDF untuk Mendeteksi Kemiripan Dokumen," *Jurnal Linguistik Komputasional*, vol. 2, no. 1, 2019.
- [8] Nico, U. Budiyanto and T. Fatimah, "Implementasi Algoritma Pembobotan TF-IDF dan Cosine Similarity untuk Penetapan Kategori Artikel pada Website Universitas Budi Luhur," *Jurnal TICOM: Technology of Information and Communication*, vol. 10, no. 3, pp. 218–223, 2012. <https://doi.org/10.70309/ticom.v10i3.38>
- [9] A. Syatar, "Relevansi Antara Pemidanaan Indonesia dan Sanksi Pidana Islam," *Jurnal Syariah dan Hukum DIKTUM*, vol. 16, no. 1, pp. 1–134, 2018. <https://doi.org/10.35905/diktum.v16i1.525>
- [10] N. T. Hariyanti, T. Rahmawati and A. Wirapraja, "Perancangan Aplikasi Mobile Menggunakan Machine Learning Untuk Menentukan Klasifikasi Kategori Berita," *Teknika*, vol. 13, no. 3, pp. 487–492, 2024. <https://doi.org/10.34148/teknika.v13i3.1093>
- [11] D. H. Wahid and A. SN, "Peringkasan Sentimen Ekstraktif di Twitter Menggunakan Hybrid TF-IDF dan Cosine Similarity," *Indonesian Journal of Computing and Cybernetics Systems*, vol. 10, no. 2, pp. 207–218, 2016. <https://doi.org/10.22146/ijccs.16625>
- [12] H. Sujadi, "Analisis Sentimen Pengguna Media Sosial Twitter Terhadap Wabah Covid-19 Dengan Metode Naive Bayes Classifier dan Support Vector Machine," *Indonesian Journal of Computing and Cybernetics Systems*, vol. 8, no. 1, pp. 22–27, 2022. <https://doi.org/10.31949/infotech.v8i1.1883>
- [13] N. Hikmah, D. Ariyanti and F. A. Pratama, "Implementasi Chatbot Sebagai Virtual Assistant di Universitas Panca Marga Probolinggo menggunakan Metode TF-IDF," *Jurnal*

Teknologi Informasi dan Multimedia, vol. 4, no. 2, pp. 133–148, 2022.

<https://doi.org/10.35746/jtim.v4i2.225>

- [14] R. F. Kurniawan, M. F. Arif “Implementasi Text Mining Menggunakan Metode Cosine Similarity untuk Klasifikasi Konten Berita di Postingan Grup Facebook Info Lantas dan Kriminal Pasuruan,” *Jurnal Ahli Muda Indonesia*, vol. 3, no. 1, pp. 9–17 2022.
<https://doi.org/10.46510/jami.v3i1.41>