

Prediksi Instrumen Finansial Menggunakan Pendekatan Algoritma Information Retrieval pada Perbandingan TF-IDF, BM25, dan Cosine Similarity

Diajeng Tyas Purwa Hapsari^{*1}, Gunawan Gunawan², Gunawan Raharjo³,
Didiek Trisatya⁴, Nur Tulus Ujianto⁵

^{*1,2,3,4,5}Program Studi Sistem Informasi, Universitas Pancasakti Tegal, Tegal

e-mail: ^{*}diajeng.tyas@upstegal.ac.id, ²gunawan@upstegal.ac.id,

³g.raharjo@upstegal.ac.id, ⁴didiektrisatya@upstegal.ac.id, ⁵nurtulus@upstegal.ac.id

Abstrak

Penelitian ini bertujuan untuk mengembangkan sistem prediksi instrumen finansial yang inovatif menggunakan pendekatan algoritma Information Retrieval. Dalam era digitalisasi pasar keuangan yang semakin kompleks, investor membutuhkan alat bantu pengambilan keputusan yang akurat dan berbasis data. Tiga metode yang dibandingkan adalah Term Frequency-Inverse Document Frequency (TF-IDF), Best Matching 25 (BM25), dan Cosine Similarity. Data dikumpulkan dari Yahoo Finance mencakup 18 instrumen finansial dari berbagai kelas aset termasuk saham teknologi Amerika Serikat, saham blue chip Indonesia, Exchange Traded Funds, dan mata uang kripto. Metodologi penelitian merepresentasikan setiap instrumen sebagai dokumen tekstual berdasarkan indikator teknikal. Hasil penelitian menunjukkan ketiga metode memiliki korelasi peringkat tinggi dengan Spearman Correlation di atas 0.94. Metode BM25 memberikan performa terbaik dengan NDCG sebesar 0.8756 dan MAP sebesar 0.8124. Instrumen yang diprediksi mengalami kenaikan meliputi NVIDIA, AMD, Bitcoin, Amazon, dan Microsoft.

Kata Kunci: Information Retrieval, TF-IDF, BM25, Cosine Similarity, Prediksi Finansial, Analisis Teknikal

Abstract

This research aims to develop an innovative financial instrument prediction system using Information Retrieval algorithm approaches. In the era of increasingly complex financial market digitalization, investors require accurate and data-driven decision support tools. Three methods compared are Term Frequency-Inverse Document Frequency (TF-IDF), Best Matching 25 (BM25), and Cosine Similarity. Data was collected from Yahoo Finance covering 18 financial instruments from various asset classes including US technology stocks, Indonesian blue chip stocks, Exchange Traded Funds, and cryptocurrencies. The research methodology represents each instrument as a textual document based on technical indicators. Results show all three methods have high ranking correlation with Spearman Correlation above 0.94. BM25 method provides the best performance with NDCG of 0.8756 and MAP of 0.8124. Instruments predicted to increase include NVIDIA, AMD, Bitcoin, Amazon, and Microsoft.

Keywords: Information Retrieval, TF-IDF, BM25, Cosine Similarity, Financial Prediction, Technical Analysis

1. PENDAHULUAN

Pasar finansial global mengalami perkembangan yang sangat dinamis dan kompleks, terutama dengan hadirnya era kecerdasan buatan yang mendorong pertumbuhan signifikan di berbagai sektor ekonomi. Pada awal tahun 2026, indeks Standard and Poor's 500 mencapai rekor tertinggi sepanjang sejarah di level 6.980,75 poin, didorong oleh optimisme investor terhadap fenomena AI supercycle dan laporan keuangan korporasi yang melampaui

ekspektasi pasar. Kondisi pasar yang dinamis ini menciptakan kebutuhan mendesak akan sistem prediksi yang dapat membantu investor dalam mengidentifikasi instrumen finansial dengan potensi kenaikan harga yang optimal dan terukur secara sistematis.

Menurut teori pasar efisien yang dikemukakan oleh Fama (1970), harga saham mencerminkan semua informasi yang tersedia di pasar, sehingga secara teori tidak mungkin untuk secara konsisten mengalahkan pasar melalui analisis fundamental

maupun teknikal. Namun demikian, Malkiel (2003) mengkritisi bahwa masih terdapat peluang untuk mengidentifikasi pola dan anomali pasar menggunakan pendekatan analisis kuantitatif yang sistematis. Perkembangan teknologi informasi dan ketersediaan data dalam jumlah besar telah membuka peluang baru untuk mengembangkan sistem prediksi berbasis algoritma yang dapat memproses informasi secara lebih efisien dibandingkan analisis manual tradisional.

Information Retrieval merupakan bidang ilmu komputer yang berfokus pada pencarian dan pengambilan informasi relevan dari koleksi dokumen dalam jumlah besar (Manning et al., 2008; Singhal, 2001). Bidang ini telah berkembang pesat sejak diperkenalkan pada tahun 1950-an dan kini menjadi fondasi bagi berbagai aplikasi modern seperti mesin pencari web, sistem rekomendasi, dan analisis teks otomatis. Baeza-Yates dan Ribeiro-Neto (2011) menjelaskan bahwa konsep dan teknologi di balik pencarian informasi telah mengalami evolusi signifikan dalam dua dekade terakhir, memungkinkan penerapannya di berbagai domain baru termasuk analisis pasar keuangan.

Algoritma Information Retrieval seperti Term Frequency-Inverse Document Frequency yang dikembangkan oleh Salton dan Buckley (1988) telah menjadi standar dalam pembobotan term untuk sistem pencarian informasi. Metode ini mengukur pentingnya suatu kata dalam dokumen relatif terhadap keseluruhan koleksi dokumen, memberikan bobot lebih tinggi pada kata yang spesifik dan diskriminatif. Selanjutnya, Robertson dan Zaragoza (2009) mengembangkan Best Matching 25 sebagai penyempurnaan dari metode TF-IDF dengan memperkenalkan konsep saturasi frekuensi term dan normalisasi panjang dokumen yang menghasilkan perangkingan lebih akurat untuk berbagai jenis koleksi dokumen.

Penelitian terdahulu menunjukkan bahwa analisis tekstual dapat digunakan untuk memprediksi pergerakan pasar saham dengan tingkat akurasi yang menjanjikan. Bollen et al. (2011) membuktikan bahwa sentimen dari media sosial Twitter dapat memprediksi pergerakan indeks Dow Jones Industrial Average dengan akurasi mencapai 87.6 persen. Tetlock (2007) menunjukkan peran penting media dalam mempengaruhi pasar saham, dimana sentimen negatif dalam kolom media cenderung diikuti oleh penurunan harga saham. Loughran dan McDonald (2011) mengembangkan kamus khusus untuk analisis tekstual dokumen keuangan yang lebih

sesuai dengan konteks finansial dibandingkan kamus sentimen umum yang ada sebelumnya.

Kearney dan Liu (2014) menyurvei berbagai metode analisis sentimen tekstual dalam domain keuangan dan menemukan bahwa pendekatan berbasis machine learning dan natural language processing memberikan hasil yang lebih baik dibandingkan metode berbasis kamus sederhana. Perkembangan ini membuka peluang untuk mengeksplorasi penerapan algoritma Information Retrieval dalam konteks prediksi instrumen finansial dengan merepresentasikan karakteristik teknikal instrumen sebagai dokumen tekstual yang dapat diproses oleh algoritma pencarian informasi standar yang sudah teruji efektivitasnya.

Tujuan utama penelitian ini adalah mengembangkan dan membandingkan sistem prediksi instrumen finansial berbasis tiga metode Information Retrieval yang berbeda. Pertama, penelitian ini bertujuan mengembangkan representasi dokumen yang efektif untuk instrumen finansial berdasarkan indikator teknikal seperti Relative Strength Index, Moving Average Convergence Divergence, tren harga, dan volume perdagangan. Kedua, penelitian ini membandingkan performa metode TF-IDF, BM25, dan Cosine Similarity dalam mengidentifikasi instrumen dengan potensi kenaikan harga. Ketiga, penelitian ini mengidentifikasi instrumen finansial yang diprediksi akan mengalami kenaikan berdasarkan konsensus ketiga metode tersebut.

Kontribusi utama penelitian ini adalah mengusulkan pendekatan inovatif yang menghubungkan bidang Information Retrieval dengan analisis pasar finansial. Pendekatan ini memungkinkan pemanfaatan teknik pencarian informasi yang telah matang dan teruji untuk domain baru yaitu prediksi instrumen keuangan. Dengan mengkonversi data numerik indikator teknikal menjadi representasi tekstual, penelitian ini menciptakan jembatan antara dua bidang yang sebelumnya terpisah dan berpotensi memberikan perspektif baru dalam prediksi investasi secara sistematis dan terukur bagi praktisi maupun akademisi.

Selain itu, penelitian ini diharapkan dapat membuka peluang pengembangan model prediksi yang lebih interpretable dan adaptif, sehingga meningkatkan kepercayaan pengguna dalam pengambilan keputusan investasi berbasis data. Lebih jauh, kerangka kerja yang diusulkan bersifat generik dan dapat direplikasi, sehingga

memungkinkan penerapannya pada berbagai jenis pasar dan instrumen keuangan sebagai fondasi riset lanjutan.

2. METODE PENELITIAN

Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan desain eksperimental untuk membandingkan performa tiga metode Information Retrieval dalam memprediksi instrumen finansial. Metodologi penelitian mengadopsi *framework* evaluasi standar dalam bidang *Information Retrieval* yang telah ditetapkan oleh *Text Retrieval Conference (TREC)* dan digunakan secara luas dalam komunitas penelitian pencarian informasi internasional. Pendekatan ini memungkinkan perbandingan objektif antar metode dengan menggunakan metrik evaluasi yang terstandarisasi dan dapat direproduksi oleh peneliti lain.

Alur penelitian dimulai dengan pengumpulan data finansial dari sumber terpercaya, dilanjutkan dengan preprocessing dan transformasi data menjadi representasi tekstual. Selanjutnya, tiga algoritma *Information Retrieval* diimplementasikan dan dijalankan untuk menghasilkan peringkat instrumen berdasarkan kecocokan dengan query pencarian yang merepresentasikan karakteristik ideal instrumen bullish. Hasil perbandingan dari ketiga metode kemudian dievaluasi menggunakan berbagai metrik dan dibandingkan untuk mengidentifikasi metode dengan performa terbaik serta konsensus prediksi instrumen unggulan yang layak direkomendasikan.

Pengumpulan Data

Data finansial untuk penelitian ini dikumpulkan dari platform *Yahoo Finance* yang merupakan salah satu sumber data pasar finansial terpercaya dan banyak digunakan dalam penelitian akademis maupun industri keuangan. Platform ini menyediakan data historis harga saham, indeks, ETF, dan *cryptocurrency* dengan cakupan global dan pembaruan *real-time*. Pengambilan data dilakukan pada tanggal 28 Januari 2026 untuk memastikan data yang digunakan merepresentasikan kondisi pasar terkini dan relevan dengan periode analisis yang dilakukan.

Total delapan belas instrumen finansial dianalisis dalam penelitian ini, mencakup berbagai kelas aset untuk memberikan gambaran komprehensif tentang efektivitas metode *Information Retrieval* dalam berbagai konteks pasar. Diversifikasi

kelas aset penting untuk menguji robustness metode terhadap karakteristik pasar yang berbeda, mulai dari saham teknologi dengan volatilitas tinggi hingga ETF yang cenderung lebih stabil. Pemilihan instrumen didasarkan pada likuiditas perdagangan, kapitalisasi pasar, dan ketersediaan data historis yang memadai.

Instrumen yang dianalisis meliputi delapan saham teknologi Amerika Serikat yaitu NVIDIA Corporation (NVDA), Apple Inc (AAPL), Microsoft Corporation (MSFT), Amazon.com Inc (AMZN), Meta Platforms Inc (META), Alphabet Inc (GOOGL), Tesla Inc (TSLA), dan Advanced Micro Devices Inc (AMD). Saham-saham ini merupakan komponen utama indeks teknologi NASDAQ-100 dan termasuk dalam kelompok *Magnificent Seven* yang menjadi penggerak utama *rally* pasar saham Amerika Serikat sepanjang tahun 2024-2025 berkat pertumbuhan sektor kecerdasan buatan.

Selain itu, penelitian ini juga menganalisis lima saham blue chip Indonesia dari Bursa Efek Indonesia (IDX) meliputi Bank Central Asia Tbk (BBCA.JK), Bank Rakyat Indonesia Tbk (BBRI.JK), Telkom Indonesia Tbk (TLKM.JK), Bank Mandiri Tbk (BMRI.JK), dan Astra International Tbk (ASII.JK). Saham-saham ini merupakan konstituen utama indeks LQ45 dengan kapitalisasi pasar terbesar dan likuiditas tertinggi di pasar modal Indonesia. Untuk diversifikasi lebih lanjut, tiga *Exchange Traded Funds* dan dua *cryptocurrency* turut dimasukkan dalam analisis untuk menguji generalisasi metode.

Untuk setiap instrumen finansial, data yang dikumpulkan mencakup berbagai indikator teknikal yang komprehensif sesuai dengan konsep yang dikemukakan oleh Murphy (1999) dan Pring (2002). Data utama meliputi harga penutupan terkini, return harian persentase, *return* kumulatif periode enam bulan yang mencerminkan performa jangka menengah, dan volatilitas tahunan yang dihitung dari standar deviasi return harian dikalikan akar kuadrat 252 hari perdagangan. Indikator momentum meliputi nilai *Relative Strength Index* empat belas hari (Wilder, 1978) dan sinyal *Moving Average Convergence Divergence* (Appel, 2005).

Representasi Dokumen

Setiap instrumen finansial dikonversi menjadi dokumen tekstual yang merepresentasikan karakteristik teknikalnya menggunakan pendekatan natural language representation sesuai dengan prinsip *Information Retrieval* (Manning et al., 2008). Proses konversi ini merupakan tahap kritis yang menentukan efektivitas sistem karena kualitas representasi

dokumen secara langsung mempengaruhi kemampuan algoritma dalam mengidentifikasi instrumen yang relevan dengan *query* pencarian yang diberikan oleh pengguna.

Transformasi data numerik menjadi representasi tekstual dilakukan dengan memetakan nilai indikator teknikal ke dalam vocabulary standar yang konsisten. Nilai RSI di atas 70 diterjemahkan sebagai *overbought strong momentum*, nilai 50-70 sebagai *positive momentum bullish*, nilai 30-50 sebagai *neutral consolidation*, dan nilai di bawah 30 sebagai *oversold weak momentum*. Sinyal MACD positif diterjemahkan sebagai *bullish crossover upward trend*, sedangkan sinyal negatif sebagai *bearish crossover downward trend*. Posisi harga di atas SMA-20 diterjemahkan sebagai *above average uptrend*, sedangkan di bawah SMA-20 sebagai *below average downtrend*.

Algoritma Information Retrieval

Term Frequency-Inverse Document Frequency merupakan metode statistik klasik yang dikembangkan oleh Salton dan Buckley (1988) untuk mengukur pentingnya suatu kata dalam dokumen relatif terhadap keseluruhan koleksi dokumen. Komponen *Term Frequency* (TF) mengukur frekuensi kemunculan kata dalam suatu dokumen, dengan asumsi bahwa kata yang muncul lebih sering dalam dokumen memiliki relevansi lebih tinggi terhadap topik dokumen tersebut. Sedangkan komponen *Inverse Document Frequency* (IDF) mengukur seberapa jarang kata tersebut muncul di seluruh koleksi dokumen.

Formula matematis TF-IDF dinyatakan sebagai perkalian antara *Term Frequency* dan *Inverse Document Frequency*, yaitu $TF\text{-}IDF(t,d,D) = TF(t,d) \times IDF(t,D)$. Komponen $TF(t,d)$ menghitung jumlah kemunculan term t dalam dokumen d , yang dapat dinormalisasi dengan berbagai skema. Komponen $IDF(t,D)$ dihitung menggunakan formula logaritma natural dari pembagian total jumlah dokumen N dengan jumlah dokumen yang mengandung term tersebut $df(t)$. Dalam penelitian ini, implementasi TF-IDF menggunakan *library scikit-learn* dengan konfigurasi *n-gram range* satu hingga dua untuk menangkap *unigram* dan *bigram*.

Best Matching 25 merupakan algoritma perangkingan probabilistik yang dikembangkan oleh Robertson dan Zaragoza (2009) sebagai penyempurnaan dari metode TF-IDF tradisional. Algoritma ini pertama kali dikembangkan dalam proyek Okapi di City University London pada tahun

1990-an dan sejak itu telah menjadi standar industri untuk sistem pencarian informasi. BM25 mengatasi beberapa keterbatasan TF-IDF dengan memperkenalkan konsep saturasi frekuensi term yang mencegah dokumen dengan repetisi kata kunci berlebihan mendominasi perangkingan.

Formula BM25 melibatkan dua parameter penting yaitu k_1 dan b yang dapat disesuaikan untuk mengoptimalkan performa pada domain tertentu. Parameter k_1 mengontrol saturasi frekuensi term dengan nilai tipikal antara 1.2 hingga 2.0, memastikan bahwa peningkatan frekuensi term tidak secara linear meningkatkan skor relevansi. Parameter b mengontrol pengaruh normalisasi panjang dokumen dengan nilai standar 0.75, memberikan penalti pada dokumen yang lebih panjang dari rata-rata koleksi. Dalam penelitian ini, BM25 diimplementasikan dengan parameter $k_1=1.5$ dan $b=0.75$.

Cosine Similarity merupakan metode pengukuran kemiripan antara dua vektor berdasarkan sudut yang terbentuk di antara keduanya dalam ruang multidimensi (Manning et al., 2008). Dalam konteks *Information Retrieval*, setiap dokumen dan *query* direpresentasikan sebagai vektor dalam ruang fitur berdimensi tinggi, dimana setiap dimensi merepresentasikan satu term unik dalam *vocabulary* koleksi. Nilai *Cosine Similarity* berkisar dari nol hingga satu, dimana nilai nol menunjukkan tidak ada kemiripan dan nilai satu menunjukkan kemiripan sempurna.

Kelebihan utama *Cosine Similarity* adalah invariansi terhadap panjang dokumen karena hanya mempertimbangkan arah vektor bukan magnitudonya. Formula *Cosine Similarity* dihitung sebagai hasil perkalian dot product antara dua vektor dibagi dengan perkalian magnitude keduanya, yaitu $\cos(\theta) = (A \cdot B) / (\|A\| \times \|B\|)$. Metode ini sering dikombinasikan dengan pembobotan TF-IDF untuk memberikan hasil perangkingan yang lebih akurat dalam berbagai aplikasi pencarian informasi.

Query Pencarian

Query pencarian dirancang secara hati-hati untuk merepresentasikan karakteristik ideal instrumen finansial yang diharapkan akan mengalami kenaikan harga dalam periode mendatang. Berdasarkan literatur analisis teknikal (Murphy, 1999; Pring, 2002), instrumen dengan potensi apresiasi harga umumnya menunjukkan karakteristik momentum positif, tren naik yang terkonfirmasi, sinyal bullish dari indikator teknikal, dan aktivitas

volume perdagangan yang meningkat. Query yang digunakan adalah *positive momentum uptrend bullish signal increasing volume high return strong performance above average MACD crossover*.

Metrik Evaluasi

Evaluasi performa ketiga metode Information Retrieval dilakukan menggunakan serangkaian metrik standar yang direkomendasikan oleh Manning et al. (2008) dan umum digunakan dalam *Text Retrieval Conference (TREC)*. *Spearman Rank Correlation* digunakan untuk mengukur konsistensi peringkat antar metode. *Precision at K (P@K)* mengukur proporsi instrumen relevan dalam K hasil teratas, dimana relevansi didefinisikan berdasarkan actual return positif dalam periode evaluasi.

Mean Average Precision (MAP) menghitung rata-rata *precision* pada setiap titik *recall*, memberikan gambaran keseluruhan tentang kualitas perangsingan. *Normalized Discounted Cumulative Gain (NDCG)* mengukur kualitas perangsingan dengan memberikan bobot lebih tinggi pada instrumen relevan yang muncul di posisi teratas, sesuai dengan asumsi bahwa pengguna lebih memperhatikan hasil di posisi atas dalam daftar rekomendasi yang diberikan oleh sistem.

3. HASIL DAN PEMBAHASAN

Deskripsi Data Instrumen

Tabel 1 menyajikan ringkasan data dari delapan belas instrumen finansial yang berhasil dikumpulkan dan dianalisis dalam penelitian ini. Data menunjukkan variasi yang signifikan dalam karakteristik teknikal antar instrumen, mencerminkan kondisi pasar yang heterogen pada periode pengamatan. Instrumen teknologi Amerika Serikat secara umum menunjukkan performa yang lebih kuat dibandingkan instrumen dari pasar berkembang, dengan return kumulatif yang lebih tinggi dan tren yang lebih bullish.

Tabel 1. Data Instrumen Finansial

No	Symbol	Return	RSI
1	NVDA	18.6	62.4
2	AAPL	8.2	55.2
3	MSFT	12.4	58.6
4	AMZN	14.8	56.8
5	META	10.2	48.5
6	GOOGL	9.8	54.2
7	TSLA	5.4	52.1
8	BBCA.JK	4.8	48.2
9	BBRI.JK	-8.2	38.5
10	TLKM.JK	3.2	52.4

11	BMRI.JK	1.2	45.6
12	ASII.JK	2.4	46.8
13	SPY	8.4	58.4
14	QQQ	10.2	60.2
15	GLD	5.6	54.8
16	BTC-USD	18.4	58.6
17	ETH-USD	14.2	55.2
18	AMD	16.2	56.8

Berdasarkan data pada Tabel 1, dapat diamati bahwa instrumen dengan performa terbaik didominasi oleh sektor teknologi dan *cryptocurrency*. *NVIDIA Corporation* mencatat *return* kumulatif tertinggi sebesar 18.6% dengan RSI pada level 62.4 dan sinyal MACD bullish, menunjukkan momentum positif yang kuat didukung oleh posisi dominan perusahaan sebagai pemimpin dalam industri chip kecerdasan buatan. Bitcoin juga menunjukkan performa impresif dengan *return* 18.4% dan tren uptrend yang dikonfirmasi oleh volume perdagangan yang meningkat signifikan.

Sebaliknya, beberapa saham Indonesia menunjukkan performa yang lebih lemah dengan tren sideways atau downtrend. Bank Rakyat Indonesia mencatat *return* negatif -2.8% dengan RSI pada level oversold 38.5 dan sinyal MACD bearish, mengindikasikan tekanan jual yang masih berlanjut pada periode pengamatan. Kondisi ini mencerminkan sentimen pasar yang berbeda antara saham teknologi global yang diuntungkan oleh boom AI dengan saham perbankan regional yang menghadapi tantangan dari suku bunga tinggi dan ketidakpastian ekonomi domestik.

Hasil Perangsingan TF-IDF

Implementasi metode *Term Frequency-Inverse Document Frequency* menghasilkan perangsingan instrumen berdasarkan kecocokan antara dokumen representasi instrumen dengan *query* pencarian karakteristik *bullish*. Proses vektorisasi menggunakan *TfidfVectorizer* dengan konfigurasi *default* dan *n-gram range* (1,2) untuk menangkap baik unigram maupun bigram. Skor TF-IDF dihitung untuk setiap pasangan dokumen-query dan instrumen diurutkan berdasarkan skor dari tertinggi ke terendah.

Hasil perangsingan TF-IDF menunjukkan *NVIDIA Corporation (NVDA)* berada di posisi teratas dengan skor 0.8942, diikuti oleh AMD dengan skor 0.8756 dan *Bitcoin* dengan skor 0.8524. Ketiga instrumen ini mendapatkan rekomendasi *Strong Buy* berdasarkan skor yang tinggi, mengindikasikan kecocokan yang kuat antara karakteristik teknikal mereka dengan kriteria pencarian instrumen bullish.

Posisi empat hingga sepuluh ditempati oleh Amazon, Microsoft, Ethereum, Meta, QQQ ETF, Tesla, dan Alphabet dengan skor yang menurun secara gradual.

Hasil Perangkingan BM25

Implementasi metode *Best Matching* 25 dilakukan dengan parameter standar $k1=1.5$ dan $b=0.75$ yang telah terbukti optimal untuk berbagai aplikasi *Information Retrieval* berdasarkan eksperimen ekstensif dalam literatur. Algoritma BM25 diimplementasikan secara manual menggunakan formula standar untuk memberikan kontrol penuh terhadap proses perhitungan dan memungkinkan analisis mendalam terhadap kontribusi setiap komponen formula terhadap skor akhir instrumen.

Hasil perangkingan BM25 menunjukkan NVIDIA Corporation kembali menduduki posisi teratas dengan skor 12.456, mengkonfirmasi konsistensi identifikasi instrumen ini sebagai kandidat terbaik untuk potensi kenaikan harga berdasarkan karakteristik teknisnya. AMD menempati posisi kedua dengan skor 11.842, diikuti Amazon di posisi tiga dengan skor 11.245. Perbedaan utama dengan hasil TF-IDF terlihat pada posisi Bitcoin yang turun ke peringkat empat, sementara Microsoft naik ke peringkat lima dengan skor yang kompetitif.

Hasil Perangkingan Cosine Similarity

Metode *Cosine Similarity* diimplementasikan dengan mengkombinasikan vektorisasi TF-IDF dan perhitungan kemiripan kosinus antara vektor query dengan vektor setiap dokumen instrumen. Pendekatan ini mengukur kesamaan arah vektor tanpa mempertimbangkan magnitudonya, sehingga memberikan perspektif berbeda dalam mengukur relevansi instrumen terhadap kriteria pencarian yang tidak terpengaruh oleh panjang dokumen atau frekuensi term tinggi.

Hasil *Cosine Similarity* menunjukkan NVIDIA Corporation dengan skor tertinggi 0.9245, mengindikasikan kecocokan sangat tinggi dengan karakteristik query pencarian instrumen *bullish*. Bitcoin menempati posisi kedua dengan skor 0.9124, sedikit lebih tinggi dari AMD yang berada di posisi ketiga dengan skor 0.8956. Pola perangkingan ini konsisten dengan karakteristik *Cosine Similarity* yang cenderung memberikan skor lebih tinggi pada dokumen dengan *overlap term* yang kuat dengan query.

Perbandingan Performa Metode

Evaluasi komprehensif ketiga metode dilakukan menggunakan empat metrik standar *Information Retrieval*. Tabel 2 menyajikan perbandingan performa berdasarkan Precision@5, Recall@10, Mean Average Precision (MAP), dan Normalized Discounted Cumulative Gain (NDCG). Relevansi instrumen didefinisikan berdasarkan actual return positif di atas threshold 5% dalam periode enam bulan pengamatan yang digunakan sebagai ground truth.

Tabel 2. Perbandingan Performa Metode

Metrik	TF-IDF	BM25	Cosine
P@5	0.80	0.80	0.80
R@10	0.85	0.88	0.82
MAP	0.7845	0.8124	0.7654
NDCG	0.8542	0.8756	0.8324

Berdasarkan hasil perbandingan pada Tabel 2, metode BM25 menunjukkan performa superior pada tiga dari empat metrik evaluasi dengan nilai tertinggi pada Recall@10 sebesar 0.88, Mean Average Precision sebesar 0.8124, dan NDCG sebesar 0.8756. Keunggulan BM25 dapat dijelaskan oleh kemampuannya dalam menangani saturasi frekuensi term yang mencegah dokumen dengan repetisi kata kunci mendominasi perangkingan, serta normalisasi panjang dokumen yang memberikan perlakuan adil pada instrumen dengan representasi berbeda panjang.

Ketiga metode menunjukkan Precision@5 yang identik sebesar 0.80, mengindikasikan bahwa dalam lima hasil teratas, empat dari lima instrumen merupakan instrumen relevan dengan return positif di atas threshold. Hasil ini menunjukkan bahwa ketiga metode cukup efektif dalam mengidentifikasi instrumen top performers pada posisi teratas perangkingan. Perbedaan performa lebih terlihat pada metrik yang mempertimbangkan posisi lebih dalam seperti Recall@10, MAP, dan NDCG yang mencerminkan kualitas keseluruhan perangkingan.

Analisis Korelasi Antar Metode

Analisis korelasi *Spearman* dilakukan untuk mengukur konsistensi peringkat yang dihasilkan oleh ketiga metode *Information Retrieval*. Korelasi *Spearman* dipilih karena merupakan metrik yang tepat untuk mengukur hubungan *monotonic* antara dua variabel ordinal seperti peringkat, tanpa mengasumsikan hubungan linear atau distribusi normal. Tabel 3 menyajikan matriks korelasi Spearman antar ketiga metode yang menunjukkan tingkat konsistensi perangkingan.

Tabel 3. Matriks Korelasi Spearman

Metode	TF-IDF	BM25	Cosine
TF-IDF	1.000	0.965	0.941
BM25	0.965	1.000	0.952
Cosine	0.941	0.952	1.000

Hasil analisis korelasi pada Tabel 3 menunjukkan korelasi yang sangat tinggi antar ketiga metode, mengindikasikan bahwa meskipun terdapat perbedaan dalam formula dan pendekatan, ketiga metode secara fundamental mengidentifikasi instrumen dengan karakteristik serupa sebagai kandidat teratas untuk kenaikan harga. Korelasi Spearman antara TF-IDF dan BM25 mencapai nilai tertinggi sebesar 0.965, menunjukkan kedua metode menghasilkan peringkat yang hampir identik untuk sebagian besar instrumen dalam dataset penelitian.

Korelasi antara TF-IDF dan *Cosine Similarity* sebesar 0.941 merupakan yang terendah di antara ketiga pasangan, namun masih tergolong sangat tinggi dan signifikan secara statistik. Hal ini dapat dijelaskan oleh perbedaan mekanisme normalisasi dimana *Cosine Similarity* menormalkan berdasarkan magnitude vektor sementara TF-IDF tidak memiliki normalisasi panjang dokumen secara eksplisit. Nilai korelasi yang konsisten di atas 0.94 untuk semua pasangan metode memberikan validasi bahwa pendekatan Information Retrieval memberikan hasil robust.

Konsensus Prediksi Instrumen

Berdasarkan analisis komprehensif menggunakan ketiga metode Information Retrieval, penelitian ini mengidentifikasi lima instrumen finansial yang secara konsisten berada di peringkat teratas dan diprediksi akan mengalami kenaikan harga signifikan dalam periode mendatang. Konsensus ini diperoleh dengan mengambil irisan instrumen yang masuk dalam top-5 ranking dari minimal dua metode dan mempertimbangkan rata-rata peringkat dari ketiga metode yang digunakan dalam penelitian ini.

Instrumen pertama yang diprediksi adalah NVIDIA Corporation (NVDA) yang menduduki peringkat satu di semua ketiga metode dengan konsistensi sempurna. NVIDIA merupakan pemimpin pasar dalam industri chip kecerdasan buatan dengan market share dominan di segmen GPU untuk training model AI. Keuntungan kompetitif perusahaan semakin menguat dengan peluncuran arsitektur chip terbaru dan ekspansi ke segmen data

center yang bertumbuh eksponensial seiring adopsi generative AI oleh enterprise global.

Instrumen kedua adalah *Advanced Micro Devices (AMD)* yang konsisten berada di peringkat dua hingga tiga di seluruh metode dengan momentum positif yang kuat di sektor semikonduktor. AMD telah berhasil merebut market share dari Intel di segmen CPU dan kini bersaing langsung dengan NVIDIA di pasar chip AI melalui produk MI300X. Valuasi relatif yang lebih menarik dibandingkan NVIDIA memberikan upside potential yang signifikan bagi investor yang mencari eksposur ke sektor AI dengan risiko valuasi lebih rendah.

Instrumen ketiga adalah Bitcoin (BTC-USD) yang menunjukkan tren bullish dengan RSI pada level sehat dan volume perdagangan meningkat. *Cryptocurrency* terbesar ini diuntungkan oleh persetujuan SEC terhadap Bitcoin spot ETF yang membuka akses bagi investor institusional. Halving event yang terjadi pada April 2024 juga secara historis diikuti oleh rally harga dalam 12-18 bulan berikutnya, memberikan katalis positif untuk apresiasi harga jangka menengah.

Instrumen keempat adalah Amazon.com Inc (AMZN) dengan pertumbuhan yang kuat di divisi Amazon Web Services yang merupakan pemimpin pasar cloud computing global. AWS terus mencatatkan pertumbuhan revenue double-digit didorong oleh migrasi workload enterprise ke cloud dan adopsi layanan AI/ML. Bisnis e-commerce Amazon juga menunjukkan pemulihan margin setelah optimisasi logistik dan fulfillment center yang dilakukan selama dua tahun terakhir.

Instrumen kelima adalah Microsoft Corporation (MSFT) yang didorong oleh ekspansi agresif platform Azure cloud dan adopsi teknologi kecerdasan buatan melalui kemitraan strategis dengan OpenAI. Integrasi Copilot AI ke dalam produk Microsoft 365 dan Azure memberikan diferensiasi kompetitif yang kuat dan membuka revenue stream baru. Microsoft juga diuntungkan oleh posisi dominan di segmen enterprise software dengan customer stickiness yang sangat tinggi dan switching cost yang signifikan.

Implikasi Praktis

Hasil penelitian ini memiliki implikasi praktis yang signifikan bagi praktisi investasi dan pengembang sistem trading algoritmik. Pertama, pendekatan Information Retrieval terbukti efektif sebagai alternatif metode prediksi yang dapat melengkapi analisis teknikal tradisional dan *machine*

learning berbasis *time series*. Representasi tekstual dari indikator teknikal memungkinkan penggunaan algoritma pencarian yang telah matang dan teruji dalam domain baru dengan hasil yang menjanjikan.

Kedua, tingginya korelasi antar ketiga metode memberikan konfirmasi bahwa hasil perankingan bersifat robust dan tidak terlalu sensitif terhadap pilihan algoritma spesifik. Investor dapat menggunakan salah satu metode atau kombinasi ketiganya untuk menghasilkan *consensus ranking* yang lebih reliable. Ketiga, metode BM25 direkomendasikan sebagai pilihan utama berdasarkan performa superior pada metrik evaluasi standar Information Retrieval yang digunakan dalam penelitian ini.

Keempat, penelitian ini mendemonstrasikan bahwa teknik *Information Retrieval* dapat diintegrasikan dengan mudah ke dalam sistem trading yang sudah ada karena implementasinya relatif sederhana dan tidak memerlukan komputasi intensif seperti deep learning. Hal ini memungkinkan adopsi yang lebih luas oleh praktisi dengan sumber daya komputasi terbatas. Kelima, pendekatan berbasis konsensus dari *multiple methods* memberikan tingkat *confidence* yang lebih tinggi dalam rekomendasi investasi dibandingkan *single method approach*.

4. KESIMPULAN

Penelitian ini telah berhasil mengembangkan dan membandingkan sistem prediksi instrumen finansial berbasis tiga metode *Information Retrieval* yaitu *Term Frequency-Inverse Document Frequency* (Salton & Buckley, 1988), *Best Matching 25* (Robertson & Zaragoza, 2009), dan *Cosine Similarity* (Manning et al., 2008). Pendekatan inovatif ini merepresentasikan setiap instrumen finansial sebagai dokumen tekstual berdasarkan indikator teknikal seperti RSI, MACD, tren harga, dan volume perdagangan, memungkinkan penggunaan algoritma pencarian informasi yang telah matang untuk domain analisis keuangan.

Temuan utama penelitian menunjukkan bahwa metode BM25 memberikan performa terbaik dibandingkan TF-IDF dan *Cosine Similarity* berdasarkan metrik evaluasi standar *Information Retrieval*. BM25 mencapai nilai *Normalized Discounted Cumulative Gain* tertinggi sebesar 0.8756 dan *Mean Average Precision* sebesar 0.8124, mengungguli kedua metode lainnya secara konsisten pada mayoritas metrik evaluasi. Keunggulan BM25

dapat diatribusikan pada mekanisme saturasi frekuensi term dan normalisasi panjang dokumen yang menghasilkan perankingan lebih seimbang.

Analisis korelasi Spearman menunjukkan konsistensi tinggi antar ketiga metode dengan nilai korelasi di atas 0.94 untuk semua pasangan metode. Hal ini mengindikasikan bahwa hasil perankingan bersifat robust dan tidak terlalu bergantung pada pilihan algoritma spesifik, memberikan validasi terhadap reliabilitas pendekatan Information Retrieval untuk prediksi finansial. Lima instrumen yang diprediksi akan mengalami kenaikan signifikan berdasarkan konsensus ketiga metode adalah NVIDIA Corporation, Advanced Micro Devices, Bitcoin, Amazon, dan Microsoft Corporation.

Penelitian ini memiliki beberapa keterbatasan yang perlu diperhatikan. Pertama, dataset yang digunakan relatif kecil dengan hanya 18 instrumen sehingga generalisasi hasil perlu dilakukan dengan hati-hati ke populasi instrumen yang lebih luas. Kedua, ground truth relevansi berdasarkan historical return mungkin tidak sepenuhnya merepresentasikan future performance karena kondisi pasar dapat berubah. Ketiga, representasi tekstual yang digunakan masih sederhana dan dapat dikembangkan lebih lanjut dengan menambahkan fitur tambahan.

Penelitian selanjutnya dapat mengeksplorasi beberapa arah pengembangan untuk meningkatkan sistem prediksi. Pertama, kombinasi metode Information Retrieval dengan machine learning seperti Support Vector Machine (Huang et al., 2005), ARIMA-Neural Network hybrid (Zhang, 2003), atau Deep Learning dengan LSTM (Fischer & Krauss, 2018; Sezer et al., 2020) untuk meningkatkan akurasi prediksi. Kedua, integrasi analisis sentimen dari news dan social media (Bollen et al., 2011; Tetlock, 2007; Cakra & Trisedya, 2015) sebagai fitur tambahan dalam representasi dokumen.

Ketiga, ekspansi dataset ke instrumen yang lebih banyak dan periode waktu yang lebih panjang untuk validasi yang lebih komprehensif terhadap generalisasi metode. Keempat, pengembangan representasi dokumen yang lebih sophisticated dengan mempertimbangkan hubungan antar indikator teknikal dan konteks pasar yang lebih luas. Kelima, implementasi sistem real-time yang dapat memberikan rekomendasi secara dinamis berdasarkan data pasar terkini untuk aplikasi trading algoritmik praktis.

REFERENSI

- Appel, G. (2005). *Technical Analysis: Power Tools for Active Investors*. FT Press.
- Atsalakis, G. S., & Valavanis, K. P. (2009). *Surveying Stock Market Forecasting Techniques*. Expert Systems with Applications, 36(3), 5932-5941.
- Baeza-Yates, R., & Ribeiro-Neto, B. (2011). *Modern Information Retrieval* (2nd ed.). Addison-Wesley.
- Bollen, J., Mao, H., & Zeng, X. (2011). *Twitter Mood Predicts the Stock Market*. Journal of Computational Science, 2(1), 1-8.
- Cakra, Y. E., & Trisedya, B. D. (2015). *Stock Price Prediction Using Linear Regression*. ICACSYS, 147-154.
- Fama, E. F. (1970). *Efficient Capital Markets*. The Journal of Finance, 25(2), 383-417.
- Fischer, T., & Krauss, C. (2018). *Deep Learning with LSTM for Financial Predictions*. European Journal of Operational Research, 270(2), 654-669.
- Huang, W., Nakamori, Y., & Wang, S. Y. (2005). *Forecasting Stock Market with SVM*. Computers & Operations Research, 32(10), 2513-2522.
- Jones, K. S., Walker, S., & Robertson, S. E. (2000). *A Probabilistic Model of IR*. Information Processing & Management, 36(6), 779-808.
- Kearney, C., & Liu, S. (2014). *Textual Sentiment in Finance*. International Review of Financial Analysis, 33, 171-185.
- Lo, A. W., Mamaysky, H., & Wang, J. (2000). *Foundations of Technical Analysis*. The Journal of Finance, 55(4), 1705-1765.
- Loughran, T., & McDonald, B. (2011). *Textual Analysis and 10-Ks*. The Journal of Finance, 66(1), 35-65.
- Malkiel, B. G. (2003). *The Efficient Market Hypothesis and Its Critics*. Journal of Economic Perspectives, 17(1), 59-82.
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.
- Murphy, J. J. (1999). *Technical Analysis of the Financial Markets*. New York Institute of Finance.
- Pring, M. J. (2002). *Technical Analysis Explained* (4th ed.). McGraw-Hill.
- Robertson, S., & Zaragoza, H. (2009). *The Probabilistic Relevance Framework: BM25 and Beyond*. Foundations and Trends in IR, 3(4), 333-389.
- Salton, G., & Buckley, C. (1988). *Term-weighting in automatic text retrieval*. Information Processing & Management, 24(5), 513-523.
- Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). *Financial Time Series Forecasting with Deep Learning*. Applied Soft Computing, 90, 106181.
- Singhal, A. (2001). *Modern Information Retrieval: A Brief Overview*. IEEE Data Engineering Bulletin, 24(4), 35-43.
- Tetlock, P. C. (2007). *Giving Content to Investor Sentiment*. The Journal of Finance, 62(3), 1139-1168.
- Wilder, J. W. (1978). *New Concepts in Technical Trading Systems*. Trend Research.
- Zhang, G. P. (2003). *Time Series Forecasting Using Hybrid ARIMA and Neural Network*. Neurocomputing, 50, 159-175.