

Sistem Temu Balik Informasi Berita Multitema Berbahasa Inggris Menggunakan BM25 dan Optimasi Semantik Kata

Reva Marlita¹, Noer Fotin Octavia², Husna Maulida³, Afra Rihadatul Aisy⁴,
Aulia Fahrizha Febriana⁵, Yoga Faris Setiadi⁶

^{1,2,3,4,5} Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Amikom Purwokerto, Jawa Tengah, Indonesia

⁶ Program Studi Sistem Informasi, STMIK Bina Patria, Indonesia, Jawa Tengah, Indonesia

surel: ¹revamarlita54@gmail.com, ²22sa11a147@mhs.amikompurwokerto.ac.id, ³husdil250502@gmail.com, ⁴afraaisy11@gmail.com,
⁵febrianafahrizha@gmail.com, ⁶yogafariss107@gmail.com

Info Artikel

Sejarah artikel:

Diterima 02-04-2026

Revisi 06-26-2026

Diterima 26-08-2026

Kata kunci:

Sistem Temu Balik Informasi
BM25
Semantic Query Expansion
WordNet
Berita Multitema

ABSTRAK

Pertumbuhan berita daring berbahasa Inggris yang semakin pesat dan mencakup berbagai tema menimbulkan tantangan dalam menemukan informasi yang relevan. Salah satu permasalahan utama dalam pencarian berita adalah perbedaan kosakata antara kata kunci yang dimasukkan pengguna dan istilah yang digunakan dalam dokumen berita. Penelitian ini bertujuan untuk merancang sistem temu balik informasi berita multitema berbahasa Inggris yang mampu meningkatkan relevansi hasil pencarian secara kontekstual. Pendekatan yang digunakan mengombinasikan metode leksikal melalui algoritma BM25 dengan pendekatan semantik menggunakan teknik *semantic query expansion* berbasis WordNet. Dataset yang digunakan berasal dari BBC News dengan total 42.115 dokumen, di mana judul dan deskripsi berita digabungkan untuk memperkaya konteks dokumen. Tahapan penelitian meliputi *preprocessing* teks, pembentukan model BM25, perluasan *query* menggunakan sinonim WordNet, serta pemeringkatan dokumen berdasarkan skor relevansi. Hasil pengujian menunjukkan bahwa penerapan ekspansi semantik memungkinkan sistem menemukan dokumen yang relevan meskipun tidak mengandung kata kunci yang sama secara langsung. Dengan demikian, integrasi BM25 dan optimasi semantik terbukti mampu meningkatkan kualitas pencarian berita berdasarkan makna, bukan sekadar kecocokan kata. Penelitian ini diharapkan dapat menjadi dasar pengembangan sistem pencarian berita yang lebih efektif dan kontekstual.

Penulis yang sesuai:

Reva Marlita

Program Studi Informatika Fakultas Ilmu Komputer Universitas Amikom Purwokerto

Email: revamarlita54@gmail.com

1. PENDAHULUAN

Perkembangan pesat data digital, khususnya berita daring berbahasa Inggris, menghasilkan koleksi dokumen teks yang besar dan bersifat multitema. Kondisi ini meningkatkan kebutuhan akan sistem temu balik informasi (*Information Retrieval*, IR) yang mampu menyajikan dokumen relevan secara efektif sesuai dengan kebutuhan pengguna. Pada domain berita, variasi gaya bahasa, topik, dan istilah yang digunakan menyebabkan proses pencarian informasi menjadi semakin kompleks, sehingga sistem IR dituntut tidak hanya cepat, tetapi juga mampu memahami konteks dan makna dari *query* yang diajukan pengguna [1].

Salah satu metode pemeringkatan dokumen yang masih banyak digunakan dalam sistem IR adalah algoritma BM25. Metode ini mampu menghitung tingkat relevansi dokumen berdasarkan frekuensi kemunculan istilah dan panjang dokumen, sehingga efektif diterapkan pada koleksi dokumen teks berskala besar. BM25 juga dikenal memiliki performa yang stabil dan efisien sehingga sering dijadikan sebagai metode dasar (*baseline*) dalam pengembangan sistem temu balik informasi, termasuk pada domain berita dan dokumen teks lainnya [2]. Namun demikian, BM25 masih memiliki keterbatasan karena berfokus pada pencocokan leksikal antara *query* dan dokumen.

Ketergantungan pada kecocokan kata secara literal menyebabkan permasalahan *vocabulary mismatch*, yaitu kondisi ketika dokumen yang relevan secara makna tidak terambil karena menggunakan sinonim atau variasi istilah yang berbeda dari *query* pengguna. Permasalahan ini umum terjadi pada pencarian berita, di mana satu konsep dapat direpresentasikan dengan berbagai istilah yang berbeda [3]. Kondisi tersebut mendorong perlunya pendekatan yang mampu memperluas representasi *query* agar mencakup variasi makna yang lebih luas.

Untuk mengatasi permasalahan tersebut, berbagai penelitian mengembangkan teknik *query expansion*, yaitu memperluas *query* dengan istilah lain yang memiliki keterkaitan makna. Teknik *query expansion* tradisional seperti *pseudo-relevance feedback* (PRF) memanfaatkan dokumen hasil pencarian awal, sedangkan pendekatan berbasis sumber leksikal seperti WordNet memanfaatkan relasi sinonim antar kata. Penelitian Hutapea dkk. mengombinasikan BM25 dengan *query expansion* berbasis Word2Vec dan menunjukkan peningkatan kinerja sistem temu balik informasi [2]. Pendekatan yang mengintegrasikan informasi semantik konseptual terbukti mampu meningkatkan kualitas hasil pencarian dengan menangkap hubungan makna yang tidak dapat direpresentasikan melalui pencocokan leksikal semata [4].

Penelitian Baihaqi dkk. menerapkan algoritma BM25 dengan seleksi fitur *term frequency* pada sistem temu balik informasi berita [5]. Hasil penelitian tersebut menunjukkan peningkatan kinerja pencarian, khususnya pada metrik *precision@k*. Namun, pendekatan yang digunakan masih sepenuhnya mengandalkan pencocokan leksikal dan belum mempertimbangkan hubungan makna antar istilah dalam proses pemrosesan *query*. Sebaliknya, penelitian Pan dkk. mengusulkan integrasi pendekatan semantik melalui *semantic pseudo-relevance feedback* untuk memperkaya representasi *query*. Pendekatan tersebut terbukti mampu meningkatkan relevansi hasil pencarian dengan memanfaatkan relasi semantik, tetapi di sisi lain menambah kompleksitas sistem dan berpotensi menimbulkan *query drift* akibat perluasan makna yang kurang terkontrol [6].

Berdasarkan kajian tersebut, dapat diidentifikasi adanya kesenjangan metodologis dalam penelitian IR, yaitu belum banyak penelitian yang mengkaji integrasi pendekatan semantik yang bersifat sederhana, terkontrol, dan efisien dengan metode pemeringkatan leksikal yang telah terbukti kuat seperti BM25, khususnya pada domain berita multitema berbahasa Inggris. Oleh karena itu, penelitian ini bertujuan untuk merancang dan mengevaluasi sistem temu balik informasi berita multitema berbahasa Inggris menggunakan algoritma BM25 yang dioptimasi dengan *semantic query expansion* berbasis WordNet pada dataset BBC News. Penelitian ini diharapkan mampu memberikan kontribusi berupa pendekatan optimasi semantik kata yang efektif dalam meningkatkan relevansi hasil pencarian berdasarkan makna, tanpa menambah kompleksitas sistem secara signifikan.

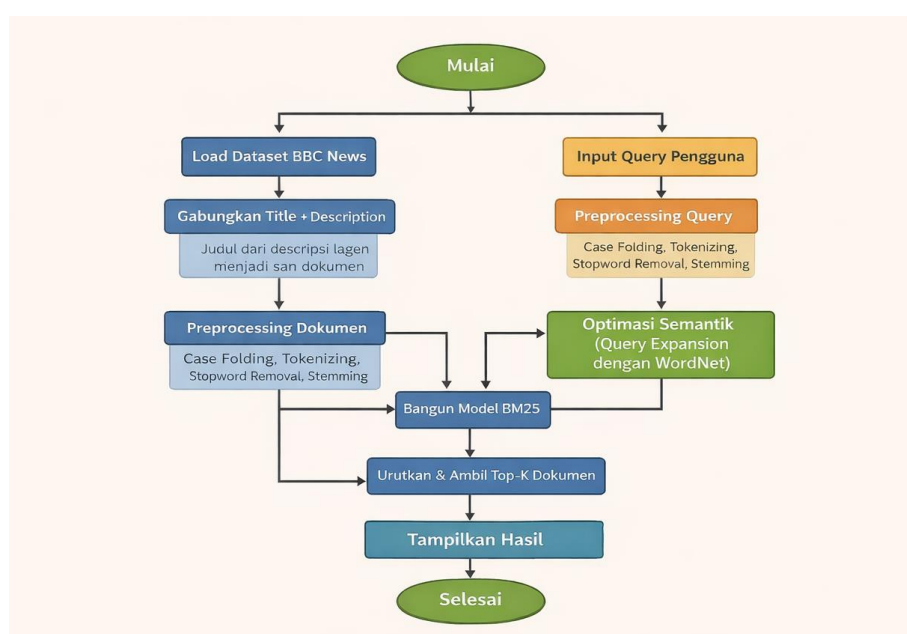
2. METODE



Penelitian ini menggunakan pendekatan temu balik informasi (*Information Retrieval*) dengan mengombinasikan metode pemeringkatan leksikal dan optimasi semantik pada *query*. Sistem yang diusulkan menerapkan algoritma BM25 sebagai metode utama pemeringkatan dokumen, kemudian dioptimasi menggunakan *semantic query expansion* berbasis WordNet pada tahap pemrosesan *query*. Pendekatan semantik dalam perluasan *query* terbukti secara empiris dapat meningkatkan relevansi hasil pencarian dengan mempertimbangkan relasi makna antar istilah yang lebih efektif dibanding pendekatan berbasis leksikal tradisional seperti BM25 saja [7]. Hal ini bertujuan untuk meningkatkan relevansi hasil pencarian dengan mengatasi permasalahan *vocabulary mismatch* antara *query* pengguna dan dokumen berita, tanpa menambah kompleksitas sistem secara signifikan.

2.1. Desain Metode Sistem

Desain metode sistem disusun dalam bentuk alur kerja terstruktur yang menggambarkan integrasi antara pemrosesan teks, pemeringkatan dokumen berbasis BM25 dan optimasi semantik *query*. Alur sistem divisualisasikan dalam bentuk flowchart, sebagaimana ditunjukkan pada Gambar 1.



Gambar 1. Desain metode sistem

Flowchart sistem menunjukkan pencarian informasi dalam sistem ini diawali dengan pemuatan dataset BBC News yang digunakan sebagai koleksi dokumen. Untuk membentuk representasi dokumen yang lebih kaya konteks, atribut judul (*title*) dan deskripsi (*description*) digabungkan menjadi satu kesatuan teks. Seluruh dokumen selanjutnya melalui tahapan *preprocessing* teks yang meliputi penyeragaman huruf, pemotongan kata, penghapusan kata umum, dan stemming. Dokumen hasil *preprocessing* kemudian diindeks dan dibobot menggunakan algoritma BM25 sebagai metode utama pemeringkatan dokumen [8].

Pada sisi pengguna, *query* pencarian yang dimasukkan juga diproses melalui tahapan *preprocessing* yang sama dengan dokumen untuk menjaga konsistensi representasi term [9]. *Query* yang telah diproses selanjutnya dievaluasi dalam dua skenario, yaitu pencarian menggunakan BM25 murni tanpa optimasi semantik dan pencarian menggunakan BM25 yang dioptimasi melalui teknik *semantic query expansion* berbasis WordNet. Pada skenario kedua, *query* diperluas dengan menambahkan istilah-istilah sinonim yang memiliki keterkaitan makna untuk mengurangi permasalahan perbedaan kosakata.

Query hasil pemrosesan pada masing-masing skenario kemudian digunakan untuk menghitung skor relevansi setiap dokumen dalam koleksi menggunakan algoritma BM25. Dokumen diurutkan berdasarkan skor relevansi yang diperoleh, dan sejumlah dokumen dengan skor tertinggi ditampilkan sebagai hasil pencarian. Hasil dari kedua skenario tersebut selanjutnya dibandingkan untuk menganalisis pengaruh optimasi semantik terhadap relevansi dokumen yang dihasilkan.

2.2. Dataset dan Representasi Dokumen

Dataset yang digunakan dalam penelitian ini adalah BBC News Dataset berbahasa Inggris yang bersifat multitema, mencakup kategori seperti *business*, *entertainment*, *politics*, *sport*, dan *technology*. Dataset ini dipilih karena merepresentasikan karakteristik berita daring yang memiliki variasi topik dan kosakata yang tinggi.

Untuk membentuk representasi dokumen yang lebih informatif, atribut judul (*title*) dan deskripsi (*description*) digabungkan menjadi satu teks dokumen [10]. Penggabungan ini bertujuan untuk memperkaya konteks dokumen sehingga proses pemeringkatan dapat menangkap informasi yang lebih representatif.

2.3. Preprocessing Teks

Preprocessing teks dilakukan untuk menyeragamkan bentuk kata dan mengurangi *noise* pada data teks sebelum digunakan dalam proses pemeringkatan [11]. Tahapan *preprocessing* ini diterapkan baik pada dokumen maupun *query* pengguna agar representasi *term* menjadi konsisten [12]. Tahapan *preprocessing* yang diterapkan meliputi:

- a. *Case folding*, untuk mengubah seluruh teks menjadi huruf kecil,
- b. *Tokenizing*, untuk memecah teks menjadi unit kata,
- c. *Stopword removal*, untuk menghapus kata-kata umum yang tidak memiliki makna diskriminatif,
- d. *Stemming*, untuk mengubah kata ke bentuk dasarnya menggunakan *Porter Stemmer* [13].

2.4. Pemeringkatan Dokumen menggunakan BM25

BM25 (*Best Match 25*) digunakan sebagai metode utama pemeringkatan dokumen. BM25 menghitung skor relevansi dokumen terhadap *query* berdasarkan frekuensi kemunculan *term* dalam dokumen (*term frequency*), frekuensi *term* dalam koleksi dokumen (*inverse document frequency*), serta panjang dokumen [14].

Dalam penelitian ini, BM25 digunakan sebagai *baseline system* untuk menghasilkan hasil pencarian tanpa optimasi semantik [15]. Hasil ini kemudian dibandingkan dengan hasil pencarian setelah diterapkan *semantic query expansion* untuk mengevaluasi dampak optimasi semantik terhadap relevansi dokumen yang dihasilkan.

2.5. Optimasi Semantik Kata dengan *Semantic Query Expansion*

Optimasi semantik dilakukan melalui teknik *semantic query expansion* berbasis WordNet. Pada tahap ini, *query* pengguna yang telah melalui *preprocessing* diperluas dengan menambahkan kata-kata sinonim yang memiliki keterkaitan makna [16].

Proses ekspansi dilakukan secara terkontrol, di mana: kata asli pada *query* tetap dipertahankan sebagai *term* utama, sinonim yang diambil dari WordNet dibatasi jumlahnya untuk setiap kata, dan ekspansi hanya bertujuan memperkaya representasi makna tanpa menggeser konteks *query* secara berlebihan. Pendekatan ini dirancang untuk mengurangi risiko *query drift* dan menjaga relevansi tematik hasil pencarian [17].

2.6. Proses Pencarian Dokumen

Setelah *query* dioptimasi secara semantik, sistem menghitung skor relevansi antara *query* dan seluruh dokumen menggunakan algoritma BM25. Dokumen kemudian diurutkan berdasarkan skor relevansi tertinggi dan sejumlah dokumen teratas ditampilkan sebagai hasil pencarian.

Untuk mendukung analisis hasil, sistem juga menampilkan: perbandingan hasil pencarian tanpa semantik dan dengan semantik, daftar sinonim hasil ekspansi *query*, dan visualisasi *term* yang cocok antara

query, sinonim, dan dokumen. Hasil pencarian dianalisis secara kualitatif dan kuantitatif untuk melihat perubahan relevansi dokumen yang diperoleh setelah diterapkan optimasi semantik kata.

3. HASIL DAN PEMBAHASAN

Pengujian menggunakan dataset BBC News yang bersifat multitema menunjukkan bahwa sistem dengan *semantic query expansion* cenderung menghasilkan jumlah dokumen relevan yang lebih banyak dibandingkan pencarian tanpa semantik.

3.1. Load Dataset Dan Preprocessing Dokumen

Pada tahap ini, kolom *title* dan *description* digabungkan menjadi satu teks dokumen. Penggabungan ini bertujuan untuk memperkaya konteks dokumen sehingga informasi yang terkandung di dalamnya lebih lengkap sebelum dilakukan proses pencarian. Dari proses ini, diperoleh total 42.115 dokumen yang digunakan sebagai koleksi data.

Selanjutnya, dilakukan *preprocessing* teks yang meliputi *case folding*, tokenisasi, penghapusan *stopword*, serta *stemming* menggunakan algoritma *Porter Stemmer*. Tahapan *preprocessing* ini bertujuan untuk mengurangi *noise* pada data teks dan menyamakan bentuk kata, sehingga proses pencocokan *term* pada tahap pemerinkatan dapat berjalan lebih efektif.

3.2. Semantic Query Expansion

Semantic query expansion diterapkan untuk mengatasi permasalahan perbedaan kosakata antara *query* pengguna dan dokumen berita. Proses ini dilakukan dengan memanfaatkan WordNet sebagai sumber pengetahuan semantik untuk memperoleh sinonim dari setiap kata pada *query*.

Dalam implementasinya, sistem melakukan perluasan *query* dengan menambahkan kata-kata yang memiliki makna serupa dengan kata asli *query*. Selain itu, diterapkan pembatasan jumlah sinonim untuk menghindari perluasan makna yang terlalu luas. Kata asli *query* tetap dipertahankan sebagai prioritas utama dalam pencarian.

Pendekatan ini memungkinkan sistem untuk tidak hanya mencari dokumen yang mengandung kata yang sama persis dengan *query*, tetapi juga dokumen yang menggunakan istilah berbeda namun memiliki makna yang berkaitan secara semantik.

3.3. Perbandingan Hasil Pencarian

Perbandingan hasil pencarian dilakukan antara sistem tanpa *semantic query expansion* dan sistem dengan *semantic query expansion* menggunakan *query* uji, seperti kata "*sport*". Pada pencarian tanpa semantik, sistem hanya mengambil dokumen yang mengandung kata "*sport*" secara eksplisit.

```

--- Masukkan query pencarian: sport

--- HASIL TANPA SEMANTIK ---
{'doc_id': 7729, 'title': 'In pictures: Sporting memories of the Queen', 'score': np.float64(7.4833), 'link': 'https://www.bbc.co.uk/sport/42430154?at_medium=RSS&at_campaign=KARANGA'},
{'doc_id': 3408, 'title': 'We Know Our Place: Women's sport campaign launched by BBC Sport', 'score': np.float64(7.075), 'link': 'https://www.bbc.co.uk/sport/61567371?at_medium=RSS&at_campaign=KARANGA'},
{'doc_id': 41189, 'title': 'What impact could Trump's return have on sport?', 'score': np.float64(6.9511), 'link': 'https://www.bbc.com/sport/articles/c170h0y0z0v0'},
{'doc_id': 11874, 'title': 'Lionel Messi named BBC Sports Personality's World Sport Star of the Year', 'score': np.float64(6.9246), 'link': 'https://www.bbc.co.uk/sport/sports-person'},
{'doc_id': 12070, 'title': 'Quiz: How many memorable sporting moments from 2022 do you recall?', 'score': np.float64(6.7588), 'link': 'https://www.bbc.co.uk/sport/64086596?at_medium=RSS&at_campaign=KARANGA'},
{'doc_id': 24064, 'title': 'Does it matter if other countries develop your budding sports stars?', 'score': np.float64(6.5804), 'link': 'https://www.bbc.co.uk/news/uk-wales-67193671'},
{'doc_id': 8029, 'title': 'Roger Federer retires: Swiss legend a tennis great who reached sporting perfection', 'score': np.float64(6.5769), 'link': 'https://www.bbc.co.uk/sport/ten'},
{'doc_id': 2791, 'title': 'The "Jockstrap" that revolutionised women's sports', 'score': np.float64(6.4981), 'link': 'https://www.bbc.co.uk/news/sport-us-canada-61187881?at_medium=RSS&at_campaign=KARANGA'},
{'doc_id': 26680, 'title': 'Sports Personality of the Year: Relive Muhammad Ali's classic humour', 'score': np.float64(6.4045), 'link': 'https://www.bbc.co.uk/sport/av/sports-person'},
{'doc_id': 7712, 'title': 'Sport pays tribute to Queen Elizabeth II', 'score': np.float64(6.3249), 'link': 'https://www.bbc.co.uk/sport/62843668?at_medium=RSS&at_campaign=KARANGA'}

--- HASIL DENGAN SEMANTIK ---
{'doc_id': 14982, 'title': 'Stranger Things prequel play combines darkness and fun, creators say', 'score': np.float64(15.5532), 'link': 'https://www.bbc.co.uk/news/entertainment-arts-63615618'},
{'doc_id': 40203, 'title': 'AI Pacino says "it's fun" to be a new dad at 84', 'score': np.float64(12.8825), 'link': 'https://www.bbc.com/news/articles/cdemk709z1o'},
{'doc_id': 19877, 'title': 'Teenage Mutant Ninja Turtles reviews: Mutant Mayhem impresses critics', 'score': np.float64(12.778), 'link': 'https://www.bbc.co.uk/news/entertainment-arts-63615618'},
{'doc_id': 20958, 'title': 'I worked on the Teenage Mutant Ninja Turtles movie from thousands of miles away', 'score': np.float64(12.5326), 'link': 'https://www.bbc.co.uk/news/uk-scotland-62522340?at_medium=RSS&at_campaign=KARANGA'},
{'doc_id': 6706, 'title': '"I got told off for playing football"', 'score': np.float64(11.0978), 'link': 'https://www.bbc.co.uk/news/uk-england-leicestershire-62522340?at_medium=RSS&at_campaign=KARANGA'},
{'doc_id': 3119, 'title': 'More people playing golf in Scotland than before pandemic', 'score': np.float64(10.8421), 'link': 'https://www.bbc.co.uk/news/uk-scotland-north-east-orkney-shetland-62522340?at_medium=RSS&at_campaign=KARANGA'},
{'doc_id': 34355, 'title': 'Women better at sport during period, study suggests', 'score': np.float64(10.1552), 'link': 'https://www.bbc.com/news/articles/c1446z0wv0'},
{'doc_id': 36531, 'title': 'The police station listed on Rightmove', 'score': np.float64(10.1336), 'link': 'https://www.bbc.com/news/articles/c151d0y0z0v0'},
{'doc_id': 16718, 'title': 'Sport pays coronation tribute to King Charles III', 'score': np.float64(9.7384), 'link': 'https://www.bbc.co.uk/sport/65510193?at_medium=RSS&at_campaign=KARANGA'},
{'doc_id': 10338, 'title': 'Strictly Come Dancing: Tony Adams withdraws with an injury', 'score': np.float64(9.5816), 'link': 'https://www.bbc.co.uk/news/entertainment-arts-63615618'}

```

Gambar 2. Perbandingan hasil pencarian

Sebaliknya, setelah diterapkan *semantic query expansion*, sistem mampu mengambil dokumen tambahan yang tidak mengandung kata "*sport*" secara langsung, tetapi memiliki keterkaitan makna, seperti



dokumen yang mengandung kata *play* atau *fun*. Hal ini menunjukkan bahwa pendekatan semantik berhasil memperluas cakupan hasil pencarian.

Namun, pada dataset multitema, perluasan ini juga menyebabkan beberapa dokumen dari domain lain, seperti hiburan, ikut terambil. Kondisi ini menunjukkan bahwa meskipun *semantic query expansion* meningkatkan *recall*, relevansi tematik tidak selalu terjaga sepenuhnya.

3.4. Evaluasi Hasil Pencarian

Evaluasi hasil pencarian dilakukan secara visual dengan menampilkan *query* asli, daftar sinonim hasil ekspansi semantik, serta dokumen-dokumen yang dihasilkan beserta skor BM25. Evaluasi ini bertujuan untuk menganalisis pengaruh *semantic query expansion* terhadap hasil pencarian secara kualitatif.

```

QUERY ASLI:
['sport']

SINONIM (HASIL EKSPANSI SEMANTIK):
['disport', 'feature', 'sportsman', 'run_around', 'lark', 'sportsman', 'frolic', 'variation', 'gambo', 'play', 'boast', 'athletics', 'cavort', 'rollick', 'summercater']

--- HASIL PENCARIAN (VISUAL EVALUASI) ---

1. Stranger Things prequel play combines darkness and fun, creators say
   Score: 15.5532
   Kata yang cocok:
   - fun [5]
   - play [5]

2. Al Pacino says 'it's fun' to be a new dad at 84
   Score: 12.8825
   Kata yang cocok:
   - fun [5]
   - play [5]

3. Teenage Mutant Ninja Turtles reviews: Mutant Rayhem impresses critics
   Score: 12.778
   Kata yang cocok:
   - mutant [5]

4. I worked on the Teenage Mutant Ninja Turtles movie from thousands of miles away
   Score: 12.5326
   Kata yang cocok:
   - play [5]
   - mutant [5]

5. 'I got told off for playing football'
   Score: 11.8978
   Kata yang cocok:
   - sport [0]
  
```

Gambar 3. Visualisasi term

Hasil evaluasi menunjukkan bahwa beberapa dokumen yang tidak muncul pada pencarian tanpa semantik dapat terambil setelah *query* diperluas. Hal ini membuktikan bahwa *semantic query expansion* mampu meningkatkan kemampuan sistem dalam menemukan dokumen yang relevan secara konteks.

Di sisi lain, evaluasi juga memperlihatkan adanya dokumen yang kurang relevan secara topik akibat perluasan makna yang terlalu umum. Oleh karena itu, hasil penelitian ini menegaskan bahwa penerapan *semantic query expansion* perlu dikombinasikan dengan mekanisme pengendalian konteks agar keseimbangan antara *precision* dan *recall* dapat tercapai secara optimal.

4. KESIMPULAN

Penelitian ini bertujuan untuk mengatasi keterbatasan sistem temu balik informasi berbasis leksikal dalam menghadapi perbedaan kosakata antara *query* pengguna dan dokumen berita, sebagaimana diuraikan pada bagian pendahuluan. Berdasarkan hasil dan pembahasan, penerapan *semantic query expansion* berbasis WordNet yang dikombinasikan dengan metode pemeringkatan BM25 terbukti mampu menjawab tujuan tersebut dengan memperluas cakupan hasil pencarian secara semantik pada dataset BBC News yang bersifat multitema.

Hasil pengujian menunjukkan bahwa sistem dengan *semantic query expansion* mampu mengambil dokumen yang relevan secara makna meskipun tidak mengandung kata yang sama persis dengan *query*, sehingga meningkatkan nilai *recall*. Dari sisi implikasi praktis, pendekatan ini bermanfaat bagi pengguna sistem pencarian dokumen karena membantu menemukan informasi yang relevan meskipun pengguna tidak menggunakan istilah yang sama dengan dokumen, serta mengurangi risiko kehilangan informasi akibat keterbatasan pemilihan kata kunci.

Namun demikian, hasil penelitian juga menunjukkan adanya penurunan relevansi tematik pada sebagian hasil pencarian akibat perluasan makna yang terlalu umum, yang menandakan adanya *trade-off* antara *precision* dan *recall*. Oleh karena itu, penelitian selanjutnya disarankan untuk mengintegrasikan mekanisme

pengendalian konteks, seperti pembobotan sinonim atau pembatasan berdasarkan domain, guna memperoleh hasil pencarian yang lebih seimbang dan optimal.

REFERENSI

- [1] I Wayan Yudik Pradnyana, I Wayan Ady Juliantara, I Nyoman Ariana Guna, and I Made Agus Widiana Putra, "Analisis Strategi Sistem Informasi dengan Aplikasi e-Purchasing Menggunakan Metode Ward and Peppard (Studi Kasus Toko ASHA)," *Jurnal Sutasoma*, vol. 4, no. 1, pp. 51–61, Dec. 2025, doi: 10.58878/sutasoma.v4i1.419.
- [2] P. Natalia Hutapea, P. Pandu Adikara, and Y. Arum Sari, "Pemerolehan Informasi Pada Alkitab Berbahasa Indonesia Menggunakan Metode BM25 Dan Query Expansion Word2vec," 2021. [Online]. Available: <http://j-ptiik.ub.ac.id>
- [3] L. Ashari and A. Suyanto, "A New Approach in Query Expansion Methods for Improving Information Retrieval," 2021.
- [4] A. Suparto, M. J. Clement, A. Rahman, and H. Irsyad, "Implementation of Preprocessing and Synonym Expansion for Indonesian News Retrieval System," 2025.
- [5] R. Ridlo Baihaqi, "Temu Kembali Informasi pada Berita Olahraga Berbahasa Indonesia dengan Seleksi Fitur Term Frequency dan Metode BM25," 2020. [Online]. Available: <http://j-ptiik.ub.ac.id>
- [6] M. Pan, W. Xiong, S. Zhou, M. Gao, and J. Chen, "LLM-Based Query Expansion with Gaussian Kernel Semantic Enhancement for Dense Retrieval," *Electronics (Switzerland)*, vol. 14, no. 9, May 2025, doi: 10.3390/electronics14091744.
- [7] D. A. Suryaningrum, R. Syaifudin, and H. R. P. Putra, "INTEGRASI WORD EMBEDDINGS DAN INVERSE BOOK FREQUENCY DALAM PEMBOBOTAN TERM UNTUK PENINGKATAN PENCARIAN DOKUMEN," *JIPi (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 9, no. 4, pp. 2529–2537, Dec. 2024, doi: 10.29100/jipi.v9i4.7557.
- [8] I. Darmawan, M. Nazir Arifin, N. Ramadhani, and N. Puspa Dewi, "ALGORITMA OKAPI BM25 DALAM SISTEM PENCARIAN INFORMASI BERBASIS TEKS," *Jurnal Insand Comtech*, vol. 10, no. 1, 2025.
- [9] D. Rodiah, K. Januar Miraswan, J. Kurniati, D. Irawan, and V. Terra Ardani, "Representasi Pengetahuan Ontologi untuk Klasifikasi Topik Penelitian pada Bidang Ilmu Informatika," *JSI : Jurnal Sistem Informasi (E-Journal)*, vol. 17, no. 1, 2025.
- [10] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. S. Weld, "SPECTER: Document-level Representation Learning using Citation-informed Transformers," May 2020, [Online]. Available: <http://arxiv.org/abs/2004.07180>
- [11] D. Rifaldi, Abdul Fadlil, and Herman, "Teknik Preprocessing Pada Text Mining Menggunakan Data Tweet 'Mental Health,'" *Decode: Jurnal Pendidikan Teknologi Informasi*, vol. 3, no. 2, pp. 161–171, Apr. 2023, doi: 10.51454/decode.v3i2.131.
- [12] M. Siino, I. Tinnirello, and M. La Cascia, "Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers," *Inf. Syst.*, vol. 121, Mar. 2024, doi: 10.1016/j.is.2023.102342.
- [13] I. Risma Huriyah, A. Ismania Sita Widianingrum, T. Informatika, and U. Muhammadiyah Riau, "Optimasi Augmentasi Data Berbasis Synonym Replacement pada Klasifikasi Teks Berita Menggunakan Neural Network Optimization of Data Augmentation Based on Synonym Replacement in News Text Classification Using Neural Network," vol. 14, no. 1, pp. 2715–7849, 2025, doi: 10.34010/komputa.v14i1.
- [14] A. Allahim, A. Cherif, and A. Imine, "Semantic approaches for query expansion: taxonomy, challenges, and future research directions," *PeerJ Comput. Sci.*, vol. 11, 2025, doi: 10.7717/peerj-cs.2664.
- [15] M. Syarif *et al.*, "Analisis Perbandingan Algoritma BM25 dan TF-IDF untuk Temu Kembali Metadata Jurnal Indonesia pada Temujurnal.com," *Jurnal ICT: Information Communication & Technology*, vol. 25, pp. 173–180, 2025, doi: 10.36054/jict-ikmi.v25i2.339.
- [16] Muhammad Rafli Aditya H., Muhammad Ilham, Dewi Fatmarani Surianto, and Abdul Muis Mappalotteng, "Evaluasi Pengukuran Semantik Sinonim KBBI Menggunakan Pendekatan Word Embedding," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 14, no. 2, pp. 112–120, May 2025, doi: 10.22146/jnteti.v14i2.17117.
- [17] M. Dwi Laxmi and M. Ali Fauzi, "Query Expansion Pada Sistem Temu Kembali Informasi Berbahasa Indonesia Dengan Metode Pembobotan TF-IDF Dan Algoritme Cosine Similarity Berbasis Wordnet," 2022. [Online]. Available: <http://j-ptiik.ub.ac.id>

