PERBANDINGAN TF-IDF DAN BM25 UNTUK FEATURE EXTRACTION PADA DATA ULASAN BUKU

Sesillia Agustin[1*], Wahyu Syaifullah J.S.[1], Alfan Rizaldy Pratama[1]
[1]*Sains Data, UPN "Veteran" Jawa Timur, Surabaya, Indonesia*
**Penulis korespondensi: sesilliaagustinn@gmail.com*

ABSTRAK

Feature extraction atau ekstraksi fitur adalah salah satu langkah penting dalam *data mining* untuk *Natural Language Processing* (NLP), terutama dalam klasifikasi teks. Proses ini bertujuan mengubah dokumen menjadi suatu list fitur yang dapat diproses dengan lebih mudah dan akurat. Dalam penelitian ini, dilakukan studi komparatif antara dua metode *feature extraction* yaitu TF-IDF (*Term Frequency-Inverse Document Frequency*), dan BM25 pada data Ulasan Buku. Tujuan penelitian adalah mengevaluasi kemampuan kedua metode dalam mengekstraksi fitur yang paling relevan terhadap konten teks ulasan buku. Hasil analisis menunjukkan performa TF-IDF lebih baik dari BM25 dalam memberi nilai yang lebih tinggi pada kata-kata yang relevan terhadap data, dengan kata-kata yang mendapat nilai TF-IDF tertinggi seperti 'book', 'story', dan 'read', dibandingkan dengan BM25 yang mengekstraksi kata-kata umum seperti 'the', 'i', 'a', dan 'and'. Temuan penelitian ini dapat menjadi pendukung penelitian selanjutnya, terutama dalam bidang *text classification* dan *sentiment analysis*.

Kata kunci: TF-IDF, NLP, Text Preprocessing, BM25, Ulasan Buku

1 PENDAHULUAN

Feature Extraction (FE) telah menjadi suatu kebutuhan dalam berbagai proses yang melibatkan rekognisi dan *Machine Learning*. FE digunakan untuk mengekstraksi fitur paling meonjol dalam suatu *dataset*, baik dataset gambar, suara, maupun teks (Salau & Jain, 2019). Dalam pemrosesan teks atau *Natural Language Processing* (NLP), FE merupakan satu langkah penting dalam *Text Classification*, yang merupakan salah satu komponen penting dalam proses NLP seperti filter informasi email, analisis sentiment, mesin pencarian, dan lainnya (Dzisevic & Sesok, 2019). Inti dari FE pada NLP adalah mengekstraksi informasi yang merepresentasikan teks tersebut, hal ini adalah dasar dari banyak proses NLP. Proses FE dapat mengurangi dimensi secara signifikan dan menghapus fitur yang mengganggu untuk mendapatkan performa dan akurasi model yang lebih baik nantinya (Liang et al., 2017). Dua diantara banyak algoritma yang dapat digunakan untuk melakukan FE pada teks adalah TF-IDF yang merupakan gabungan dari TF (*Term Inverse*) dan IDF (*Inverse Document Frequency*)(Liu et al., 2018), dan BM25 yang telah banyak diketahui efektivitasnya dalam melakukan ekstraksi informasi(Mishchenko & Klymenko, 2023).

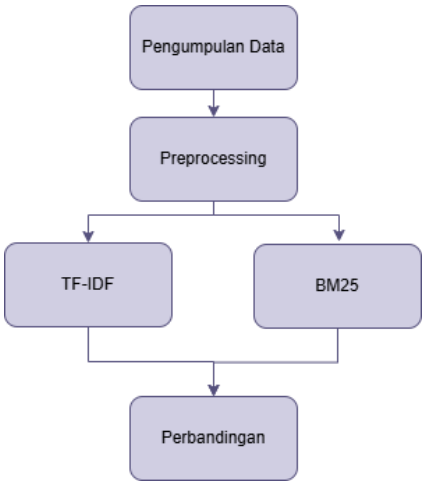
Sebelumnya Kadhim dkk telah melakukan penelitian membandingkan TF-IDF dan BM25 untuk dalam FE pada data twitter. Dalam penelitian ini, performa kedua metode dibandingkan dengan menggunakan *Term Weighting*, dan hasilnya menunjukkan bahwa TF-IDF bekerja sedikit lebih baik dengan nilai F1 89.77, dibandingkan dengan BM25 yang mendapatkan nilai F1 89.16 (Kadhim, 2019). Selanjutnya Qiu dan Yang juga melakukan FE dengan mengkombinasikan TF-IDF dan Word2Vec yang menunjukkan hasil yang mendukung bahwa penggunaan TF-IDF dalam NLP dapat memberikan dampak yang baik dalam mengekspresikan informasi dalam teks(Qiu &

Yang, 2021). Selain untuk mendukung proses NLP selanjutnya, proses FE juga dapat digunakan untuk mengukur relevansi kata-kata terhadap dokumen, hal ini penting dilakukan untuk memastikan fitur yang diambil merupakan fitur yang relevan dan merepresentasikan informasi dalam dokumen. Salah satu penelitian yang mengukur relevansi kata adalah penelitian Qaiser yang menggunakan TF-IDF untuk mengukur relevansi kata pada dokumen. Hasil penelitian ini menunjukkan bahwa TF-IDF dapat mengekstaksi kata yang relevan dengan baik (Shahzad Qaiser & Ramsha Ali, 2018). Penelitian mengenai relevansi kata menggunakan BM25 juga dilakuan oleh Peng dkk. Dalam penelitian ini algoritma berdasaarkan BM25 digunakan untuk membobot relevansi topik, dan hasilnya menunjukkan bahwa algoritma ini cukup efektif (Peng et al., 2021).

Data yang akan digunakan dalam penelitian ini adalah data mengenai ulasan buku. Pada masa modern ini, internet dianggap sebagai sumber utama bagi kustomer melalui berbagai *website*. *Website* ini membuat kustomer dapat memberikan opini yang pada berbagai hal, salah satunya buku (Almjawel et al., 2019a). Penelitian NLP yang biasanya diterapkan pada data ulasan buku adalah klasifikasi teks dan analisis sentimen, seperti data-data ulasan lain. Salah satu penelitian mengenai ulasan buku menggunakan NLP adalah penelitian Rahman dkk mengenai klasifikasi ulasan buku dalam Bahasa Bangla (Rahman et al., 2021). Penelitian lain adalah analisis sentiment dan visualisasi ulasan buku oleh Almjawel dkk yang menerapkan analisis sentiment pada data ulasan buku dari Amazon (Almjawel et al., 2019b). Semua penelitian yang telah disebutkan memerlukan proses FE dalam *preprocessing*-nya untuk mendapatkan hasil paling optimal dari model.

Penelitian ini bertujuan untuk membandingkan TF-IDF dan BM25 dalam *Feature Extraction*. Performa kedua algoritma akan dipelajari dan ditimbang melalui kemampuannya untuk mengekstraksi fitur yang relevan. Penelitian ini diharapkan dapat memberi pandangan terhadap penggunaan metode-metode FE, dan dapat digunakan oleh penelitian selanjutnya, terutama dalam bidang NLP dan text preprocessing.

2 METODE



Gambar 1. Diagram Alir Proses Penelitian

Gambar 1. menunjukkan alur penelitian ini, yang terdiri dari pengumpulan data, *preprocessing*, pembangunan model, dan pembandingan hasil. Data yang akan digunakan pada penelitian ini adalah data ‘Amazon Kindle Book Review’ yang berisi total 12.000 data ulasan buku

disertai beberapa parameter lain yang berguna dan relevan bagi ulasan buku yang diberikan. Variabel dari data yang akan digunakan dalam penelitian ini adalah variabel 'reviewText' yang berisikan ulasan buku dan variabel 'rating' yang menunjukkan nilai ulasan yang diberikan dari 1 hingga 5. Data ini melalui *preprocessing* dimana pembersihan teks dasar akan diterapkan seperti mengganti huruf kapital, menghapus link, dan membersihkan teks dari tanda baca. Perubahan kata tidak dilakukan dalam penelitian ini dengan tujuan untuk melihat performa model dalam meng-handle kata-kata umum. Setelah melalui preprocessing data akan diproses pada model, pemberlakuan model dilakukan secara terpisah antara TF-IDF dan BM25. Hasil dari pemrosesan sebelumnya akan dibandingkan di akhir penelitian.

2.1 TF-IDF

Term Frequency-Inverse Document Frequency atau yang disingkat sebagai TF-IDF mengukur pentingnya suatu kata bagi dokumen dimana kata tersebut berada, dan menyimpannya dalam suatu corpus (Shahzad Qaiser & Ramsha Ali, 2018). Metode ini kerap digunakan sebagai *weighting factor* dalam text mining, user modeling, dan information retrieval. Suatu survey yang dilakukan oleh Beel et al. pada tahun 2015 menunjukkan bahwa 83% dari text-based recommender system pada perpustakaan digital menggunakan tf-idf (Beel et al., 2016).

Bagian *Term-Frequency* (TF) adalah frekuensi relatif *term* t dalam dokumen d (Kadhim, 2019).

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{1}$$

Sedangkan bagian *Inverse Document-Frequency* (IDF) adalah

$$idf(t,D) = log\frac{N}{n_t} \tag{2}$$

Dengan D adalah suatu set dokumen dalam *corpus* atau *dataset*, N merupakan total seluruh dokumen dalam *corpus*, dan n_t adalah jumlah dokmen dimana *term* t muncul. Dengan begitu TF-IDF adalah sebagai berikut,

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D) \tag{3}$$

2.2 BM25

BM25 (*Best-Matching* 25) adalah salah satu fungsi peringkat (*ranking function*) yang umum digunakan di bidang information retrieval untuk menilai relevansi dokumen terhadap sebuah query pencarian. BM25 ini merupakan satu fungsi peringkat yang umum digunakan di bidang information retrieval untuk menilai relevansi dokumen terhadap sebuah query pencarian. Algoritma ini termasuk dalam keluarga model pencocokan leksikal (lexical matching) yang menghitung skor relevansi berdasarkan frekuensi istilah dalam dokumen dan kekhasan istilah di seluruh korpus (Kheng et al., 2025).

Secara matematis, fungsi BM25 menghitung skor relevansi antara dokumen D dan *query* Q sebagai jumlah kontribusi dari setiap istilah query. Misalkan Q berisi term q1, q2, …,qn, maka skor BM25 dihitung sebagai berikut (Zhang, 2021),

$$score\ (D,Q) = \sum_{i=1}^{n} IDF(qi) \cdot \frac{f(qi,D) \cdot (k1+1)}{f(qi.D) + k1(1-b+b \cdot \frac{|D|}{avgdl}}$$

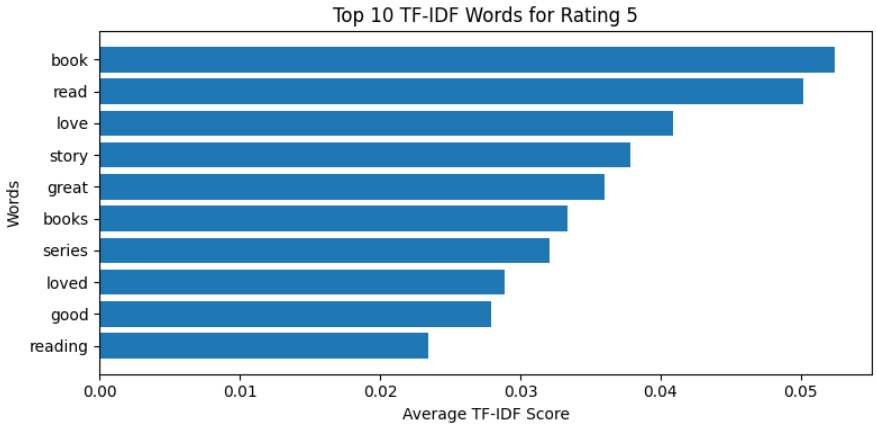
3 HASIL DAN PEMBAHASAN

Sebelum penerapan metode feature extraction, seluruh data ulasan buku melalui tahap preprocessing untuk mengurangi noise yang dapat memengaruhi hasil ekstraksi fitur. Tahapan preprocessing yang dilakukan meliputi case folding, penghapusan tanda baca, dan pembersihan karakter non-alfabet. Pada penelitian ini tidak dilakukan proses stemming maupun stopword removal dengan tujuan untuk mengamati kemampuan masing-masing metode dalam menangani noise berupa kata-kata umum secara alami.Setelah tahap preprocessing, data diproses secara terpisah menggunakan metode TF-IDF dan BM25. Hasil dari kedua metode kemudian dianalisis berdasarkan waktu komputasi, karakteristik kata yang dihasilkan, serta relevansi fitur terhadap topik dokumen ulasan buku.

Penerapan metode TF-IDF pada 12.000 dokumen ulasan buku menunjukkan waktu pemrosesan yang relatif cepat, yaitu kurang dari 20 detik. Hal ini menunjukkan bahwa TF-IDF memiliki efisiensi komputasi yang baik dan sesuai untuk diterapkan pada data teks berukuran menengah tanpa memerlukan sumber daya komputasi yang besar.

Nilai skor TF-IDF tertinggi yang diperoleh berada pada kisaran 0,06. Meskipun nilai tersebut relatif kecil secara absolut, skor TF-IDF bersifat relatif terhadap distribusi frekuensi kata dalam korpus. Oleh karena itu, evaluasi performa TF-IDF dilakukan dengan mengamati relevansi kata-kata yang memperoleh bobot tertinggi, bukan berdasarkan besar kecilnya skor semata.Hasil ekstraksi menunjukkan bahwa kata-kata dengan skor TF-IDF tertinggi didominasi oleh istilah yang secara langsung berkaitan dengan konteks ulasan buku, seperti “book”, “read”, “story”, dan “series”. Kata-kata tersebut merepresentasikan topik utama dokumen, yaitu pengalaman pembaca terhadap buku. Hal ini menunjukkan bahwa TF-IDF mampu menyaring noise berupa kata-kata umum dan mengekstraksi fitur yang relevan terhadap isi dokumen.

Temuan ini mendukung tujuan penelitian, yaitu mengevaluasi kemampuan metode feature extraction dalam mengekstraksi fitur yang relevan. TF-IDF terbukti mampu menghasilkan fitur yang merepresentasikan topik dokumen dengan baik tanpa terpengaruh secara signifikan oleh noise. Hasil ini juga sejalan dengan penelitian Shahzad Qaiser dan Ramsha Ali (2018) serta Kadhim (2019) yang menyatakan bahwa TF-IDF efektif dalam mengekstraksi kata-kata relevan pada data teks.



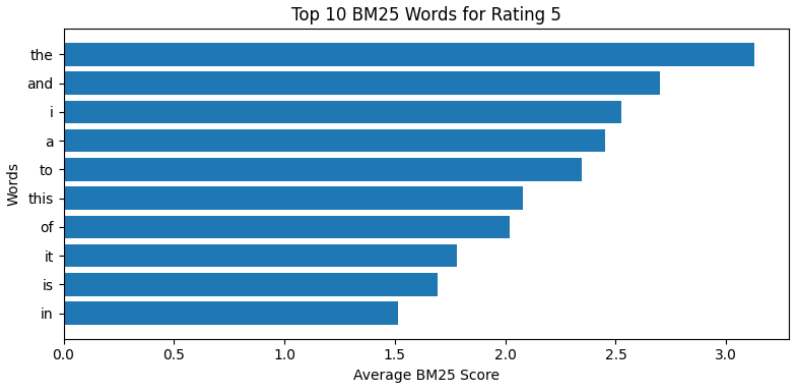
Gambar 2. Kata-kata dengan skor TF-IDF pada rating 5

Penerapan metode BM25 menunjukkan waktu komputasi yang jauh lebih lama dibandingkan TF-IDF. Proses pemrosesan terhadap 12.000 dokumen dapat memakan waktu hingga 20 menit dalam satu iterasi. Hal ini disebabkan oleh mekanisme BM25 yang mempertimbangkan frekuensi

istilah, panjang dokumen, serta normalisasi panjang dokumen dalam proses perhitungan skor relevansi.

Nilai skor tertinggi yang dihasilkan oleh BM25 mencapai 3,39. Namun, sebagaimana TF-IDF, skor BM25 tidak dapat dievaluasi secara absolut karena menggunakan skala pembobotan yang berbeda. Oleh karena itu, analisis difokuskan pada karakteristik kata yang memperoleh skor tertinggi. Hasil ekstraksi menunjukkan bahwa kata-kata dengan skor tertinggi pada BM25 didominasi oleh kata-kata umum seperti “the”, “and”, “I”, dan “a”. Kata-kata tersebut memiliki frekuensi kemunculan yang tinggi, namun tidak merepresentasikan topik ulasan buku secara spesifik. Hal ini menunjukkan bahwa BM25 kurang mampu menangani noise apabila tidak didukung oleh tahap preprocessing lanjutan seperti stopword removal.

Meskipun BM25 dikenal efektif dalam konteks information retrieval, seperti yang ditunjukkan oleh Peng et al. (2021) dan Mishchenko & Klymenko (2023), hasil penelitian ini menunjukkan bahwa dalam konteks feature extraction pada data ulasan buku, BM25 cenderung mengekstraksi kata-kata umum yang kurang relevan terhadap topik dokumen. Hal ini menjelaskan mengapa BM25 dinilai kurang optimal dalam penelitian ini

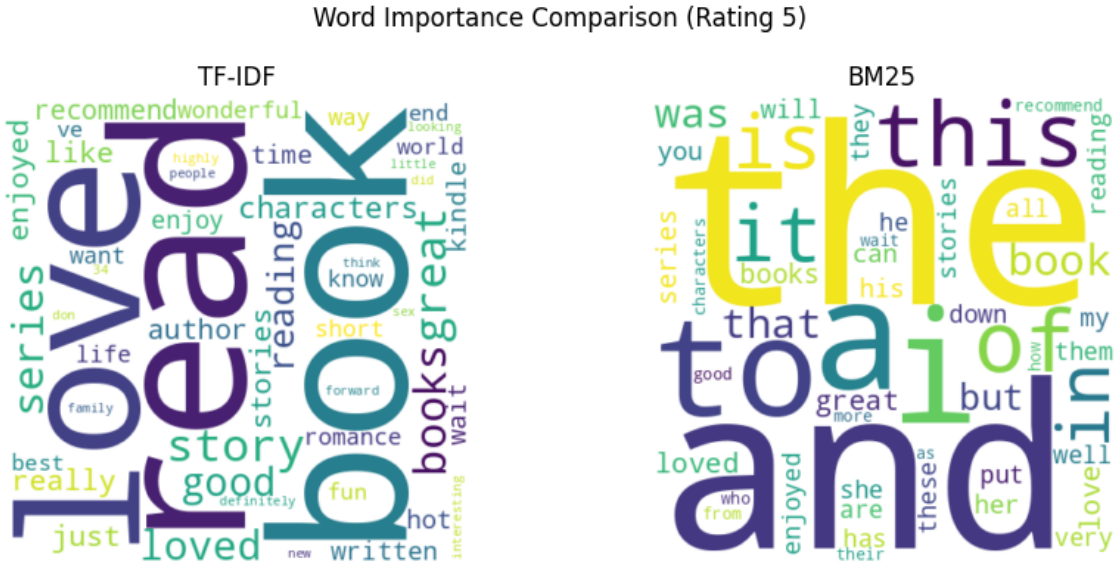


Gambar 3. Kata-kata dengan skor BM25 pada rating 5

Berdasarkan hasil penerapan kedua metode, TF-IDF dan BM25 menunjukkan perbedaan karakteristik yang cukup signifikan dalam proses feature extraction pada data ulasan buku. Perbedaan ini tidak hanya terlihat dari aspek waktu komputasi, tetapi juga dari jenis fitur yang dihasilkan serta tingkat relevansi fitur terhadap topik dokumen.

TF-IDF menunjukkan keunggulan dalam hal efisiensi komputasi dan kemampuan menyaring noise berupa kata-kata umum. Mekanisme pembobotan TF-IDF yang menggabungkan frekuensi kata dalam dokumen dan frekuensi kemunculan kata dalam seluruh korpus menyebabkan kata-kata yang sering muncul secara global, namun kurang informatif, memperoleh bobot yang lebih rendah. Hal ini membuat fitur yang dihasilkan oleh TF-IDF lebih fokus pada kata-kata yang memiliki keterkaitan langsung dengan isi dan topik ulasan buku, seperti pengalaman membaca dan elemen cerita. Karakteristik ini menjadikan TF-IDF efektif sebagai metode feature extraction dalam konteks text preprocessing, khususnya ketika tujuan penelitian adalah memperoleh representasi fitur yang ringkas dan relevan. Namun demikian, mekanisme penyaringan noise yang dimiliki TF-IDF juga berpotensi menjadi keterbatasan. Dalam beberapa konteks tertentu, kata-kata yang memiliki frekuensi tinggi dapat mengandung informasi penting, terutama jika digunakan dalam analisis lanjutan seperti sentiment analysis atau topic modeling. Oleh karena itu, penggunaan TF-IDF perlu disesuaikan dengan tujuan analisis serta karakteristik data yang digunakan.

Di sisi lain, BM25 memiliki pendekatan yang berbeda dalam menghitung relevansi kata. Metode ini mempertimbangkan panjang dokumen dan frekuensi istilah secara lebih eksplisit melalui parameter normalisasi panjang dokumen. Pendekatan ini membuat BM25 banyak digunakan dalam sistem information retrieval dan pencarian dokumen, sebagaimana ditunjukkan dalam penelitian Peng et al. (2021) dan Mishchenko & Klymenko (2023). Akan tetapi, hasil penelitian ini menunjukkan bahwa tanpa dukungan preprocessing lanjutan, BM25 cenderung memberikan bobot tinggi pada kata-kata umum yang tidak secara langsung merepresentasikan topik ulasan buku. Kurangnya sensitivitas BM25 terhadap noise pada penelitian ini mengindikasikan bahwa metode tersebut memerlukan tahapan preprocessing tambahan, seperti stopword removal atau pembobotan ulang parameter, agar dapat bekerja secara optimal dalam konteks feature extraction. Tanpa langkah tersebut, fitur yang dihasilkan oleh BM25 menjadi kurang selektif dan berpotensi menurunkan kualitas representasi teks.



Gambar 4. Wordcloud perbandingan kata-kata dengan skor tertinggi

Jika ditinjau dari tujuan penelitian, yaitu mengevaluasi kemampuan metode feature extraction dalam mengekstraksi fitur yang relevan, TF-IDF menunjukkan performa yang lebih sesuai pada data ulasan buku. Hal ini disebabkan oleh kemampuannya dalam menghasilkan fitur yang representatif, efisien, dan relatif tahan terhadap noise. Sementara itu, BM25 lebih cocok digunakan pada konteks pencarian atau pemeringkatan dokumen, dibandingkan sebagai metode utama feature extraction tanpa preprocessing tambahan.

Tabel 1. Perbandingan TF-IDF dan BM25

Faktor	TF-IDF	BM25
Kecepatan	✔	
Noise Handling	✔	
Keutuhan Dokumen		✔
Relevansi Kata	✔	

4 KESIMPULAN

Hasil penelitian menunjukkan kedua model memiliki performa yang cukup, namun tetap memiliki hasil yang berbeda. Dalam konteks Feature Extraction, TF-IDF memiliki performa yang lebih baik dikarenakan tidak terpengaruh oleh noise dan dapat mengambil kata atau fitur yang berhubungan dengan topik dalam dokumen. BM25 bekerj dengan lebih menyeluuh, namun tidak mampu bekerja dengan noise dan malah mengambil kata-kata yang lebih umum serta tidak berhubungan dengan topik dokumen. Hasil penelitian ini dapat digunakan sebagai pertimbangan dan sumber untuk melakukan Feature Extraction dan text preprocessing. Pada penelitian selanjutnya, dapat dikembangkan aspek preprocessing dan pengurangan noise pada dokumen data.

UCAPAN TERIMAKASIH

Ucapan terimakasih penulis sampaikan kepada seluruh pihak yang telah membantu proses penelitian ini.

DAFTAR PUSTAKA

Almjawel, A., Bayoumi, S., Alshehri, D., Alzahrani, S., & Alotaibi, M. (2019a). Sentiment Analysis and Visualization of Amazon Books' Reviews. *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)*, 1–6. https://doi.org/10.1109/CAIS.2019.8769589

Almjawel, A., Bayoumi, S., Alshehri, D., Alzahrani, S., & Alotaibi, M. (2019b). Sentiment Analysis and Visualization of Amazon Books' Reviews. *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)*, 1–6. https://doi.org/10.1109/CAIS.2019.8769589

Beel, J., Gipp, B., Langer, S., & Breitinger, C. (2016). Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries*, *17*(4), 305–338. https://doi.org/10.1007/s00799-015-0156-0

Dzisevic, R., & Sesok, D. (2019). Text Classification using Different Feature Extraction Approaches. *2019 Open Conference of Electrical, Electronic and Information Sciences (EStream)*, 1–4. https://doi.org/10.1109/eStream.2019.8732167

Kadhim, A. I. (2019). Term Weighting for Feature Extraction on Twitter: A Comparison Between BM25 and TF-IDF. *2019 International Conference on Advanced Science and Engineering (ICOASE)*, 124–128. https://doi.org/10.1109/ICOASE.2019.8723825

Kheng, T., Asri, J. S., Wahyu, S., & Yulhendri, Y. (2025). Penerapan Algoritma BM25 dalam Pencarian Lowongan Pekerjaan pada Website Job Portal. *Bulletin of Computer Science Research*, *5*(5), 1029–1038. https://doi.org/10.47065/bulletincsr.v5i5.760

Liang, H., Sun, X., Sun, Y., & Gao, Y. (2017). Text feature extraction based on deep learning: a review. *EURASIP Journal on Wireless Communications and Networking*, *2017*(1), 211. https://doi.org/10.1186/s13638-017-0993-1

Liu, Q., Wang, J., Zhang, D., Yang, Y., & Wang, N. (2018). Text Features Extraction based on TF-IDF Associating Semantic. *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, 2338–2343. https://doi.org/10.1109/CompComm.2018.8780663

Mishchenko, L., & Klymenko, I. (2023). Recognizing fake news based on natural language processing using the BM25 algorithm with fine-tuned parameters. *Eastern-European Journal of Enterprise Technologies*, *6*(2 (126)), 33–40. https://doi.org/10.15587/1729-4061.2023.293513

Peng, S., Xie, X., Zhai, J., Jia, Y., & Gong, Y. (2021). A Page-topic Relevance Algorithm Based on BM25 and Paragraph-Semantic Correlation. *Journal of Physics: Conference Series*, *1757*(1), 012115. https://doi.org/10.1088/1742-6596/1757/1/012115

Qiu, Y., & Yang, B. (2021). Research on Micro-blog Text Presentation Model Based on Word2vec and TF-IDF. *2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, 47–51. https://doi.org/10.1109/IPEC51340.2021.9421098

Rahman, Md. H., Islam, Md. S., Jowel, Md. M. U., Hasan, Md. M., & Latif, Ms. S. (2021). Classification of Book Review Sentiment in Bangla Language Using NLP, Machine Learning and LSTM. *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 1–5. https://doi.org/10.1109/ICCCNT51525.2021.9580116

Salau, A. O., & Jain, S. (2019). Feature Extraction: A Survey of the Types, Techniques, Applications. *2019 International Conference on Signal Processing and Communication (ICSC)*, 158–164. https://doi.org/10.1109/ICSC45622.2019.8938371

Shahzad Qaiser, & Ramsha Ali. (2018). Text mining: use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications* , *181*(1), 25–29.

Zhang, Z. (2021). An improved BM25 algorithm for clinical decision support in Precision Medicine based on co-word analysis and Cuckoo Search. *BMC Medical Informatics and Decision Making*, *21*(1), 81. https://doi.org/10.1186/s12911-021-01454-5