

Temu Balik Dokumen Berita Berbasis Kemiripan Semantik Menggunakan FastText

Ary Dian Pratama¹, Yardhan Zaendy A², Nur Alfani Wahid³, Rifqi Dwi Aryanto⁴, Aditya Nugroho⁵,
Muhammad Adiib Fauzi Hidayat⁶

^{1,2,3,4,5}Jurusan Informatika, Fakultas Ilmu Komputer, Universitas Amikom Purwokerto, Purwokerto, Indonesia

⁶Jurusan Sistem Informasi, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika, Purwokerto, Indonesia

surel:¹arydianprtma@gmail.com,²yardanzaendy@gmail.com,³wahidzeronine09@gmail.com,⁴rizqicahyandika24@gmail.com,

⁵nugrohaditya010@gmail.com,⁶adiibfauzi4@gmail.com

Info Artikel

Sejarah artikel:

Diterima 01-25-2026

Revisi 28-06-2026

Diterima 17-08-2026

Kata kunci:

Temu Balik Informasi
Dokumen Berita
Kemiripan Semantik
FastText

ABSTRAK

Perkembangan informasi digital, khususnya dokumen berita daring, menuntut adanya sistem temu balik informasi yang mampu menyajikan dokumen relevan secara cepat dan akurat. Sistem temu balik informasi tradisional umumnya masih mengandalkan pendekatan berbasis kata kunci, yang memiliki keterbatasan dalam memahami hubungan makna antar kata sehingga berpotensi menghasilkan hasil pencarian yang kurang optimal.

Penelitian ini mengusulkan sebuah sistem temu balik dokumen berita berbasis kemiripan semantik dengan memanfaatkan FastText sebagai metode representasi teks. Dalam sistem yang diusulkan, dokumen berita dan query pengguna direpresentasikan dalam bentuk vektor embedding untuk menangkap makna semantik, kemudian tingkat kemiripan antara query dan dokumen dihitung untuk menentukan relevansi. Pendekatan ini diharapkan mampu mengatasi keterbatasan metode berbasis kata kunci dan menghasilkan dokumen berita yang lebih relevan secara makna, khususnya pada pencarian berita berbahasa Indonesia.

Penulis yang sesuai:

Nur Alfani Wahid

Program Studi Informatika Fakultas Ilmu Komputer Universitas Amikom Purwokerto

Email: wahidzeronine09@gmail.com

1. PENDAHULUAN

Perkembangan teknologi informasi telah mendorong pertumbuhan data digital dalam jumlah yang sangat besar, khususnya dokumen teks seperti artikel dan berita daring. Informasi yang tersedia secara masif tersebut menuntut adanya sistem yang mampu membantu pengguna menemukan dokumen yang relevan secara cepat dan akurat sesuai dengan kebutuhan informasi. Oleh karena itu, sistem temu balik informasi (*information retrieval*) memegang peranan penting dalam pengelolaan dan pencarian dokumen teks, terutama pada domain berita yang bersifat dinamis dan terus diperbarui [1].

Pada praktiknya, sistem temu balik informasi tradisional masih banyak menggunakan pendekatan berbasis kata kunci (*lexical-based retrieval*), seperti model Boolean, Extended Boolean, dan Vector Space Model. Beberapa penelitian menunjukkan bahwa metode-metode tersebut menilai relevansi dokumen berdasarkan kecocokan kata secara literal antara query pengguna dan dokumen [2][3]. Meskipun relatif sederhana dan mudah diimplementasikan, pendekatan berbasis kata kunci memiliki keterbatasan dalam memahami hubungan makna antar kata, sehingga hasil pencarian seringkali belum optimal.

Permasalahan tersebut semakin terlihat pada pencarian dokumen berita. Sebuah topik berita dapat diekspresikan dengan beragam istilah yang berbeda namun memiliki makna yang sama. Penelitian yang membandingkan efektivitas temu kembali informasi pada portal berita menunjukkan bahwa perbedaan penggunaan istilah antar dokumen dapat mempengaruhi kualitas hasil pencarian ketika sistem hanya mengandalkan pencocokan kata secara langsung [4]. Akibatnya, dokumen yang relevan secara semantik berpotensi tidak terambil oleh sistem.

Seiring dengan berkembangnya penelitian di bidang temu balik informasi, pendekatan berbasis kemiripan semantik (*semantic-based retrieval*) mulai banyak dikembangkan untuk mengatasi keterbatasan metode berbasis kata kunci. Pendekatan ini merepresentasikan kata atau dokumen ke dalam bentuk vektor numerik yang mencerminkan makna semantik, sehingga memungkinkan sistem untuk mengukur kedekatan makna antara query dan dokumen. Beberapa penelitian terbaru menunjukkan bahwa pemanfaatan representasi semantik mampu meningkatkan relevansi hasil pencarian dibandingkan metode konvensional [1], [5].

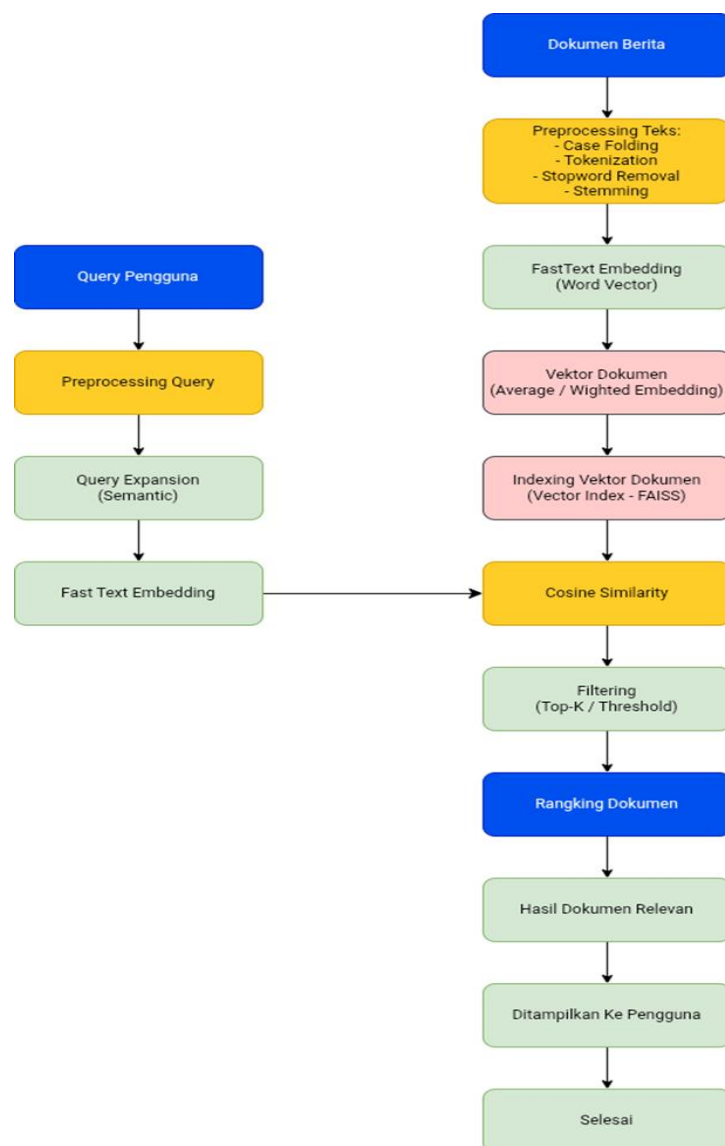
Salah satu metode representasi kata yang banyak digunakan dalam pendekatan semantik adalah FastText. FastText mampu merepresentasikan kata berdasarkan informasi sub-kata (*subword*), sehingga efektif dalam menangani variasi morfologi dan kata tidak baku. Hal ini menjadi keunggulan tersendiri ketika diterapkan pada bahasa Indonesia yang memiliki banyak imbuhan dan variasi bentuk kata [5], [2]. Dengan karakteristik tersebut, FastText dinilai sesuai untuk diterapkan pada sistem temu balik dokumen berita berbahasa Indonesia.

Berdasarkan latar belakang tersebut, penelitian ini mengusulkan sebuah sistem temu balik dokumen berita berbasis kemiripan semantik dengan memanfaatkan FastText sebagai metode representasi teks. Dalam sistem yang diusulkan, dokumen berita dan query pengguna direpresentasikan dalam bentuk vektor embedding untuk selanjutnya dihitung tingkat kemiripannya guna menentukan relevansi dokumen. Pendekatan ini diharapkan mampu mengatasi keterbatasan metode temu balik informasi berbasis kata kunci dan menghasilkan dokumen yang lebih relevan secara makna.

Adapun tujuan dari penelitian ini adalah untuk merancang dan mendeskripsikan sistem temu balik dokumen berita berbasis kemiripan semantik menggunakan FastText. Kontribusi utama penelitian ini terletak pada penerapan pendekatan semantic retrieval pada domain berita berbahasa Indonesia sebagai alternatif solusi terhadap keterbatasan metode temu balik informasi konvensional.

2. METODE

Penelitian ini merupakan penelitian kuantitatif eksperimental yang bertujuan untuk membangun dan mengevaluasi sistem temu balik informasi (Information Retrieval) berbasis semantik menggunakan representasi word embedding FastText dan pengukuran kemiripan cosine similarity. Pendekatan yang digunakan berfokus pada pengujian efektivitas model dalam menemukan dokumen berita yang relevan terhadap query pengguna berdasarkan kemiripan makna, bukan hanya kesamaan kata secara leksikal [6] [7]. *Dataset* yang digunakan diambil dari situs [Kaggle.com](https://www.kaggle.com/datasets/indonesiannews/indonesia-news-dataset-2024) "*Indonesia News Dataset 2024*". *Dataset* tersebut berisi kumpulan artikel berita berbahasa Indonesia dari berbagai kategori seperti politik, ekonomi, teknologi, kesehatan, dan olahraga. Dalam penelitian ini, bagian isi berita digunakan sebagai dokumen utama dalam sistem temu balik informasi [8].



Gambar 1. Flowchart Sistem Temu Balik Informasi

2.1 Preprocessing

Preprocessing dilakukan untuk membersihkan dan menyamakan teks sehingga dapat direpresentasikan secara efektif dalam bentuk vektor numerik. Tahapan *preprocessing* meliputi *case folding* untuk mengubah seluruh karakter menjadi huruf kecil, *tokenization* untuk memecah teks menjadi unit-unit kata, *stopword-removal* untuk menghapus kata-kata umum yang tidak memiliki kontribusi besar terhadap makna [9]. Dan *stemming* untuk mengubah kata ke bentuk dasarnya. Tahap *preprocessing* ini dilakukan dengan tujuan agar kualitas embedding terjamin.

2.2 FastText Word Embedding

FastText merupakan pengembangan dari Word2Vec yang memanfaatkan subword information berupa karakter n-gram. Model ini tidak hanya mempelajari kata secara utuh, tetapi juga bagian kata, sehingga mampu menangani kata yang tidak pernah muncul sebelumnya (out-of-vocabulary). Keunggulan FastText menjadikannya cocok untuk bahasa yang memiliki morfologi kompleks seperti Bahasa Indonesia. Model FastText pralatih Bahasa Indonesia digunakan untuk memetakan setiap kata hasil pra-pemrosesan menjadi vektor berdimensi tetap [10].

2.3 Pengindeksan Vektor Dokumen

Seluruh vektor dokumen disimpan ke dalam struktur indeks menggunakan FAISS (Facebook AI Similarity Search). Pengindeksan ini bertujuan untuk mempercepat proses pencarian vektor terdekat ketika query pengguna dimasukkan ke sistem.

2.4 Pemrosesan Query Pengguna

Query yang dimasukkan pengguna melalui proses yang sama dengan dokumen, yaitu *preprocessing*, *query expansion*. Konversi ke vektor menggunakan FastText. Hasilnya adalah satu vektor query yang siap dibandingkan dengan vektor dokumen.

2.5 Cosine Similarity

Kemiripan antara vektor query dan vektor dokumen dihitung menggunakan cosine similarity, yang mengukur sudut antara dua vektor dalam ruang berdimensi tinggi.

Rumus cosine similarity:

$$\text{Cosine}(A, B) = \frac{A \cdot B}{||A|| \times ||B||}$$

Nilai cosine similarity berada pada rentang -1 hingga 1, di mana nilai yang lebih mendekati 1 menunjukkan kemiripan yang lebih tinggi. Metode ini banyak digunakan dalam pengukuran kemiripan semantik berbasis embedding.

2.6 Filtering dan Perangkingan

Dokumen diseleksi berdasarkan dari *Top-K* (hanya K dokumen dengan skor tertinggi yang diambil) dan *Threshold* (dokumen dengan nilai *similarity* di atas ambang batas tertentu), selanjutnya dokumen diurutkan dari nilai *similarity* tertinggi ke terendah untuk menghasilkan daftar hasil akhir (scaling up top k).

2.7 Output

Sistem menampilkan daftar dokumen berita yang paling relevan terhadap query pengguna, beserta skor kemiripannya. Hasil ini menjadi dasar evaluasi kualitas sistem temu balik informasi yang dibangun.

3. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil implementasi sistem temu balik dokumen berita berbasis kemiripan semantik menggunakan FastText serta pembahasan terhadap kinerja sistem yang dibangun. Pengujian dilakukan dengan memberikan sejumlah query berita berbahasa Indonesia dan mengamati relevansi dokumen yang dihasilkan oleh sistem.

3.1 Hasil Implementasi Sistem

Sistem berhasil mengimplementasikan seluruh tahapan yang dirancang, mulai dari preprocessing teks, pembentukan word embedding menggunakan FastText, pengindeksan vektor dokumen dengan FAISS, hingga proses pencarian berbasis cosine similarity. Antarmuka sistem memungkinkan pengguna memasukkan query berupa kalimat atau frasa, kemudian sistem menampilkan daftar dokumen berita yang diurutkan berdasarkan nilai kemiripan tertinggi.

Setiap hasil pencarian menampilkan judul berita, potongan isi berita, serta nilai similarity sebagai indikator tingkat relevansi. Keberadaan skor kemiripan ini membantu pengguna memahami tingkat kedekatan makna antara query dan dokumen yang ditampilkan.

Hasil pengujian menunjukkan bahwa sistem mampu mengembalikan dokumen yang relevan meskipun kata-kata pada query tidak selalu sama persis dengan kata-kata yang terdapat dalam dokumen. Hal ini mengindikasikan bahwa representasi semantik menggunakan FastText mampu menangkap hubungan makna antar kata.



3.2. Analisis Relevansi Hasil Pencarian

Pada pengujian dengan beberapa query bertema berita, sistem menghasilkan daftar dokumen yang secara umum sesuai dengan topik yang dimaksud pengguna. Sebagai contoh, ketika pengguna memasukkan query mengenai “kenaikan harga bahan pokok”, sistem menampilkan dokumen berita yang membahas inflasi, harga kebutuhan pokok, serta kondisi pasar, meskipun tidak semua dokumen menggunakan frasa yang sama persis dengan query.

Fenomena tersebut menunjukkan bahwa sistem tidak hanya mengandalkan pencocokan kata secara literal, tetapi mempertimbangkan kedekatan makna antara query dan dokumen. Dengan demikian, pendekatan berbasis kemiripan semantik lebih fleksibel dibandingkan metode berbasis kata kunci murni.

3.2. Pengaruh Penggunaan FastText

FastText memberikan kontribusi penting dalam meningkatkan kualitas representasi teks karena memanfaatkan informasi subword. Dengan karakteristik ini, sistem tetap dapat membentuk vektor untuk kata yang jarang muncul atau memiliki variasi imbuhan. Hal ini sangat relevan untuk bahasa Indonesia yang memiliki banyak bentuk turunan kata.

Hasil pengujian menunjukkan bahwa dokumen dengan variasi bentuk kata tetap dapat terhubung secara semantik dengan query. Sebagai contoh, kata “pendidikan”, “mendidik”, dan “pendidik” cenderung memiliki kedekatan vektor yang relatif tinggi, sehingga dokumen yang mengandung variasi kata tersebut tetap dapat terambil.

3.2.2. Efektivitas Cosine Similarity dan Pengindeksan FAISS

Penggunaan cosine similarity memungkinkan sistem mengukur tingkat kemiripan antara query dan dokumen secara numerik dalam ruang vektor berdimensi tinggi. Nilai similarity yang dihasilkan digunakan sebagai dasar perankingan dokumen, sehingga dokumen dengan tingkat kemiripan tertinggi ditampilkan pada urutan teratas. Selain itu, penerapan FAISS sebagai indeks vektor mempercepat proses pencarian, terutama ketika jumlah dokumen cukup besar. Sistem mampu melakukan pencarian vektor terdekat secara efisien tanpa harus membandingkan query dengan seluruh dokumen secara langsung.

3.2.2. Efektivitas Cosine Similarity dan Pengindeksan FAISS

Meskipun sistem menunjukkan kinerja yang baik dalam menghasilkan dokumen relevan secara semantik, terdapat beberapa keterbatasan. Pertama, kualitas hasil pencarian sangat bergantung pada kualitas model FastText yang digunakan. Kedua, sistem belum mempertimbangkan konteks kalimat secara mendalam seperti pada model transformer berbasis BERT. Ketiga, evaluasi kuantitatif menggunakan metrik seperti precision, recall, dan F1-score belum dilakukan secara menyeluruh. Keterbatasan tersebut membuka peluang pengembangan lanjutan, seperti penggunaan model embedding kontekstual, optimasi query expansion, serta penerapan evaluasi berbasis ground truth.

4. KESIMPULAN

Penelitian ini bertujuan untuk merancang sistem temu balik dokumen berita yang mampu menghasilkan dokumen relevan berdasarkan kemiripan makna, sebagaimana dipaparkan pada bagian pendahuluan, dan hasil yang diperoleh menunjukkan bahwa tujuan tersebut telah tercapai. Sistem yang dibangun mampu merepresentasikan dokumen dan query dalam bentuk vektor semantik menggunakan FastText serta melakukan perankingan dokumen berdasarkan nilai kemiripan, sehingga dokumen yang dihasilkan tetap relevan meskipun tidak selalu memiliki kecocokan kata secara langsung dengan query. Hal ini sejalan dengan hasil dan pembahasan yang menunjukkan bahwa pendekatan berbasis kemiripan semantik lebih fleksibel dibandingkan metode berbasis kata kunci murni dalam menangani variasi istilah pada berita berbahasa Indonesia. Selain itu, penerapan pengindeksan vektor menggunakan FAISS turut mendukung efisiensi proses pencarian. Ke depan, sistem ini masih berpotensi dikembangkan dengan memanfaatkan model embedding kontekstual, optimalisasi mekanisme query expansion, serta evaluasi kinerja menggunakan metrik standar agar kualitas temu balik informasi dapat semakin ditingkatkan.

REFERENSI

- [1] Song, R., Tian, J., Zhu, P., & Chen, B. (2025). Improving News Retrieval with a Learnable Alignment Module for Multimodal Text–Image Matching. *Electronics*, 14(3098). doi: 10.3390/electronics14153098
- [2] Husni, H., Muntasa, A., Putro, S. S., & Osman, Z. (2023). Cross-Language Tourism News Retrieval System Using Google Translate API on SEBI Search Engine. *ELINVO (Electronics, Informatics, and Vocational Education)*, 8(1), 113–120. doi: 10.21831/elinvo.v8i1.55851
- [3] Suhardi, S., Walim, W., Priyandaru, H., Prabowo, W., & Priatmojo, H. (2021). Implementasi Information Retrieval System untuk Klasifikasi Berita Offline di Indonesia Menggunakan Metode Extended Boolean. *CERMIN: Jurnal Penelitian*, 5(1), 124–132. doi: 10.36841/cermin_unars.v5i1.951
- [4] 4. Perbandingan Efektivitas Temu Kembali Informasi pada Portal Berita Online di Yogyakarta (Harian Jogja dan Tribun Jogja). *JIPKA*, 1(1), 59–69. doi: 10.26418/jipka.v1i1.50295
- [5] Fauzi, Ahmad Ginabila, Ginabila. (2019) ‘Information Retrieval System Pada File Pencarian Dokumen Tesis Berbasis Text Menggunakan Metode Vector Space Model. *Jurnal Pilar Nusa Mandiri*’. doi: 10.33480/pilar.v15i1.61
- [6] P. Delgado, C. Vargasb, R. Ackermanc, dan L. Salmerón, “Jangan membuang buku cetak Anda: Sebuah meta-analisis tentang pengaruh media membaca terhadap pemahaman membaca,”*Mendidik. Res. Putaran.*, jilid. 25, hlm. 23–38, 2018, doi: 10.1016/j.edurev.2018.09.003.
- [7] F. Reichert, D. Lange, dan L. Chow, “Keyakinan pendidikan penting untuk pengajaran di kelas: Analisis komparatif keyakinan guru tentang tujuan pendidikan kewarganegaraan,”*Mengajar. Mengajar. Mendidik.*, jilid. 98, hal. 1–13, 2020, doi: 10.1016/j.tate.2020.103248.
- [8] J. Roick dan T. Ringeisen, “Prestasi matematika siswa di pendidikan tinggi: menguji peran pembelajaran mandiri dan efikasi diri,”*Mempelajari. Individu. Berbeda.*, jilid. 65, hlm.148–158, 2018.
- [9] G. Ocak dan A. Yamaç, “Pemeriksaan hubungan antara strategi pembelajaran mandiri siswa kelas lima, keyakinan motivasi, sikap, dan prestasi,”*Mendidik. Sains. Praktek Teori.*, jilid. 13, tidak. 1, hal.380–387, 2013.
- [10] S. Li dan J. Zheng, “Hubungan antara efikasi diri dan pembelajaran yang diatur sendiri dalam lingkungan komputasi satu-ke-satu: Peran yang dimediasi dari nilai-nilai tugas,”*Pendidikan Asia-Pasifik. Res.*, jilid. 27, tidak. 6, hal.455–463, 2018, doi: 10.1007/s40299-018-0405-2.
- [11] BJ Zimmerman dan AR Moylan, “Pengaturan diri: di mana metakognisi dan motivasi berpotongan,” diDJ Hacker, J. Dunlosky, dan AC Graesser, Eds.,*Buku Pegangan Metakognisi dalam Pendidikan*, 2009, hlm.299–315.
- [12] PR Pintrich, DAF Smith, T. Duncan, dan W. Mckeachie,*Panduan penggunaan kuesioner strategi motivasi untuk belajar (MSLQ)*. Ann Arbor, Michigan, 1991.
- [13] M.Pressley dan CB McCormick,*Psikologi pendidikan tingkat lanjut untuk pendidik, peneliti, dan pembuat kebijakan*. New York, AS: HarperCollins College Publishers, 1995.
- [14] A.bandura,*Seri Prentice-Hall dalam teori pembelajaran sosial. Landasan sosial pemikiran dan tindakan: Sebuah teori kognitif sosial*. Prentice-Hall, Inc., 1985.
- [15] AL Dent dan AC Koenka, “Hubungan antara pembelajaran mandiri dan prestasi akademik pada masa kanak-kanak dan remaja: sebuah meta-analisis,”*Mendidik. Psikologi. Putaran.*, jilid. 28, tidak. 3, hal.425–474, 2015, doi: 10.1007/s10648-015-9320-8.
- [16] TJ Cleary dan A. Kitsantas, “Motivasi dan pengaturan diri belajar berpengaruh terhadap prestasi matematika sekolah menengah,”*Psikologi Sekolah. Putaran.*, jilid. 46, tidak. 1, hal.88–107, 2017.
- [17] PR Pintrich, “Bab 14 - Peran Orientasi Tujuan dalam Pembelajaran Mandiri,” didalam M. Boekaerts, P. Pintrich, M. Zeidner, Eds.,*Buku Pegangan Pengaturan Mandiri*, San Diego, California: Academic Press, 2000, hlm.451–502.
- [18] H. Vonkova dan J. Hrabak, “Perbandingan pengetahuan TIK dan penilaian diri keterampilan di kalangan siswa sekolah menengah atas: Penggunaan metode sketsa penahan,”*Hitung. Mendidik.*, jilid. 85, hlm. 191–202, 2015, doi: 10.1016/j.compedu.2015.03.003.
- [19] F. Baier, A.-T. Decker, T. Voss, T. Kleickmann, U. Klusmann, dan M. Kunter, “Apa yang dimaksud dengan guru yang baik? Pentingnya relatif kemampuan kognitif, kepribadian, pengetahuan, keyakinan, dan motivasi guru matematika terhadap kualitas pengajaran,”*Sdr. J.Mendidik. Psikologi.*, jilid. 89, tidak. 4, hal.767–786, 2019, doi: 10.1111/bjep.12256.
- [20] AM Flanagan, DC Cormier, dan O. Bulut, “Prestasi mungkin berakar pada ekspektasi guru: mengkaji perbedaan pengaruh etnis, masa mengajar, dan perilaku di kelas,”*sosial. Psikologi. Mendidik.*, jilid. 23, hal. 1429–1448, 2020, doi: 10.1007/s11218-020-09590-y.

