PAWIYATAN XXVIII (3) (2021) 1 - 9
Pawiyatan
Universitas IVET
http://e-journal.ikip-veteran.ac.id/index.php/pawiyatan

Pemanfaatan Text-Mining untuk Ekstraksi Informasi Artikel Ilmiah Bidang PAUD

Dewi Nugrahastuti Wirahno[1)], Widya Damayanti[2)]
Program Studi Pendidikan Guru-Pendidikan Anak Usia Dini
Fakultas Keguruan dan Ilmu Pendidikan Universitas IVET di Semarang.
E-mail: dwnugrahastuti@gmail.com

Diterima: September 2021, Di publikasikan: Oktober 2021

ABSTRAK

Pentingnya isu pendidikan anak usia dini menuntut penelitan yang berkualitas dan melampaui topik-topik dasar. Artikel ini menginvestigasi topik-topik penelitian bidang PAUD yang telah diteliti pada artikel jurnal yang terbit tahun 2018-2020. Dua ratus enam puluh satu (261) abstrak didapatkan dari pencarian melalui Directory of Open Access Journals (DOAJ). Penelitian ini bertujuan mendapatkan data yang bermakna dari abstrak dan kata kunci artikel-artikel ilmiah bidang PAUD dan memvisualisasikannya. Teknik text mining digunakan untuk mengidentifikasi topik-topik dari abstrak artikel-artikel jurnal tersebut. Melalui topic modelling yang menggunakan latent semantic analysis diperoleh sepuluh kelompok topik, dan melalui word cloud divisualisasikan kata kunci-kata kunci sesuai dengan frekuensi penggunaan. Hasil penelitian menunjukkan sepuluh kluster atau kelompok kata kunci yang merepresentasi topik atau tema penelitian dari abstrak artikel yang diteliti. Dalam word cloud yang dihasilkan dari abstrak dan dari kata kunci pengarang terdapat enam kata yang paling sering digunakan: anak, usia, pembelajaran, pendidikan, metode, dan guru.

Kata Kunci: Text-Mining, Ekstraksi Informasi, Artikel Ilmiah Bidang PAUD

PENDAHULUAN

Pendidikan anak usia dini merupakan isu pendidikan penting yang menjadi agenda pembangunan terdepan di banyak negara (South-East Asian Ministers of Education Organization, 2016). Pendidikan pada periode ini memiliki implikasi keberhasilan pada semua jenjang pendidikan. Penelitian dari bidang-bidang lain menunjukkan bahwa PAUD yang berkualitas berpengaruh secara signifikan dalam menentukan keberhasilan anak di masa depan (Alitajer et al., 2016 dalam Khodabandelou et al., 2018). Dalam agenda pendidikan 2030, UNESCO mendorong pengembangan dan peningkatan program-program PAUD di tingkat nasional, regional dan internasional, serta mendukung program-program PAUD yang secara khusus fokus pada pembelajaran yang berkualitas, dan mengedepankan perkembangan anak yang holistik (UNESCO International Institute for Capacity-Building in Africa, 2013). Seiring dengan pentingnya isu PAUD ini, muncul kebutuhan untuk memastikan pertumbuhan kajian bidang PAUD. Oleh karenanya, penelitian bidang PAUD seharusnya bergerak melampaui topik-topik dasar seperti karakteristik pendidik, pengalaman siswa, dan pengembangan kurikulum (Sheridan et al., 2009).

Untuk mengetahui topik-topik penelitian bidang PAUD yang telah dilakukan, Khodabandelou et al. (2018) melakukan kajian bibliometrik atas 6,730 rekod artikel Web of Science periode 2000-2016 untuk mengamati tren penelitian bidang PAUD. Hasil kajian menunjukkan bahwa topik-topik penelitian yang banyak dibahas adalah 1) penelitian berbasis kurikulum dan pendidikan; 2) kesehatan, keamanan/keselamatan, nutrisi; 3) terkait dengan kegiatan fisik; serta 4) kajian gender dan keluarga. Di Turki, sebuah penelitian analisis konten atas 136 laporan penelitian bidang PAUD menunjukkan bahwa topik penelitian yang diminati peneliti adalah 1) pendidikan lingkungan; 2) ketrampilan proses ilmiah; serta 3) metode, sikap, dan perilaku mengajar (İlhan & Erden, 2019). Sementara di Indonesia, penelitian terkait pengelompokan artikel hasil penelitian telah dilakukan oleh Sabna (2020) yang menerapkan text mining untuk mengelompokkan penelitian dosen di Program Studi Ilmu Kesehatan Masyarakat, STIKES Hang Tuah, Pekanbaru. Dalam penelitian tersebut ditemukan dua (2) topik yang paling banyak diteliti, yaitu Covid dan HIV. Beberapa penelitian selanjutnya berusaha mengelompokkan dokumen-dokumen ilmiah dengan beragam metode. Mushlihudin & Zahrotun (2017) merancang pengelompokan penelitian dosen Universitas Ahmad Dahlan, Yogyakarta dengan menggunakan metode shared nearest neighbor dengan euclidean similarity. Nurmalasari & Syahrial (2014) memanfatkan sequential pattern untuk mengelompokkan dokumen; sedangkan Priandini et al. (2017) menggunakan pendekatan Fuzzy C-means dan K-Nearest Neighbors.

Meskipun telah dilakukan dalam bidang lain, penelitian dengan menggunakan teknik text mining belum pernah dilakukan terkait PAUD. Khodabandelou et al. (2018) dan İlhan & Erden (2019) menginvestigasi topik penelitian atau tema-tema yang diangkat dalam bidang PAUD meskipun tidak menggunakan metode text mining. Sementara penelitian-penelitian lainnya menggunakan text mining, namun tidak dalam bidang PAUD. Berangkat dari hal ini, artikel ini akan melihat gambaran umum topik-topik yang telah diteliti dalam

literatur terkait PAUD di Indonesia dengan menggunakan teknik text mining. Penelitian ini merupakan preliminary research atau penelitian awal untuk mengetahui struktur pengetahuan bidang PAUD dengan menginvestigasi topik atau tema penelitian yang telah dibahas. Dengan melihat manfaat text mining dalam penelitian-penelitian sebelumnya, penelitian ini bertujuan mendapatkan data yang bermakna dari abstrak dan kata kunci artikel-artikel ilmiah bidang PAUD dan memvisualisasikannya.

Text mining adalah teknik yang biasa digunakan untuk menggali data tersembunyi dari data yang berbentuk teks. Dengan kata lain, text mining merupakan strategi pencarian informasi nontradisional yang digunakan untuk memperoleh informasi dari kumpulan besar dokumen teks (N & S, 2016). Selain itu, text mining memberi solusi pada masalah yang muncul dalam memproses, mengorganisasi, dan menganalisa teks tak-terstruktur atau unstructured text dalam jumlah besar. Dalam menganalisa sebagian atau keseluruhan unstructured text, text mining mengasosiasikan satu bagian teks dengan yang lainnya berdasarkan aturan-aturan tertentu (Adiwijaya, 2006). Text mining merupakan bagian dari data mining (Akilan, 2015), yaitu proses menemukan pola-pola bermakna dari sekelompok teks yang tak terstruktur. Teknik ini banyak digunakan dalam bidang ilmu komputer, ilmu informasi, matematika, dan manajemen untuk mendulang wawasan dari data besar, atau big data (Humphreys & Wang, 2018).

Salah satu teknik yang banyak diterapkan dalam text mining adalah Latent Semantic Analysis (LSA), yang merupakan metode untuk menganalisis hubungan interrelasi semantik antara kalimat dan kata dalam sekelompok dokumen (Süzek, 2017). Hal ini masuk dalam cakupan topic modelling, yang merupakan teknik penting dalam text mining (Adil et al., 2019). Dengan menggunakan pendekatan matematis, LSA mencari nilai kemiripan antara dua buah segmen teks tanpa memperhatikan susunan kata. Prinsipnya, konsep yang muncul dalam suatu teks dapat diketahui dengan memperhatikan kata yang muncul atau diksi yang digunakan, karena tiap kata kunci memiliki makna yang terkait dengan dokumen dan dianggap merepresentasikan ide (Nakov, 2000 dalam Jadhira et al., 2018).

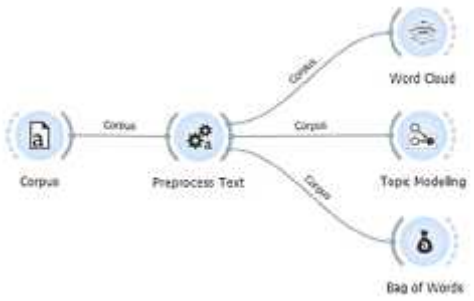
Word cloud disebut juga sebagai tag cloud didefinisikan sebagai representasi visual teks tertulis yang terstruktur menurut frekuensinya (Jayashankar & Sridaran, 2017). Cara ini merupakan teknik yang paling sering digunakan untuk menampilkan data teks secara grafis, yang mempermudah proses analisis data teks seperti esai makalah, jawaban tertulis, dan sejenisnya. Word cloud sering difungsikan sebagai tahap awal dalam analisis mendalam atas suatu materi tertulis atau teks (Sinclair et al., 2008).

METODE PENELITIAN

Data yang digunakan dalam penelitian ini berasal dari abstrak dan kata kunci pada artikel-artikel ilmiah yang diterbitkan oleh jurnal-jurnal akses terbuka di Indonesia periode 2018-2020. Abstrak dan kata kunci diperoleh melalui halaman pencarian pada Directory of Open Access Journals https://www.doaj.org. Pencarian dilakukan dengan menggunakan kata kunci “pendidikan anak usia dini,” “pendidikan usia dini,” dan “PAUD.” Hasil pencarian dibatasi hanya menampilkan rekod artikel pada periode tahun 2018-2020. Dari pencarian ini dihasilkan 339

rekod artikel dan 1035 kata kunci. Data ini kemudian dibersihkan dengan cara menyingkirkan duplikasi, dan hanya menggunakan abstrak dan kata kunci berbahasa Indonesia, yang menghasilkan 261 rekod. Pembersihan duplikasi tidak dilakukan pada data kata kunci karena frekuensi penggunaan kata kunci menjadi faktor penentu untuk menghasilkan word cloud.

Untuk penelitian ini, perangkat lunak open-source Orange (https://orangedatamining.com/) digunakan untuk tahap preprocess, topic modelling, dan visualisasi data. Tahap pertama adalah preprocess, di mana teks abstrak dan kata kunci diubah menjadi huruf kecil, serta menghilangkan tanda baca. Data yang telah dibersihkan kemudian disimpan dalam bag of word atau database dalam Orange. Data tersimpan adalah abstrak dan tahun terbit. Tahap selanjutnya adalah menggambarkan word cloud untuk melihat kata kunci yang sering digunakan. Secara ringkas, Gambar 1 berikut menunjukkan alur kerja dalam Orange yang digunakan dalam penelitian ini.



Gambar 1. Alur kerja Orange untuk penelitian ini

Penggunaan software Orange untuk melihat kalimat apa yang sering muncul dalam sebuah dokumen dan menentukan berapa banyak yang akan ditampilkan dalam hal ini melihat abstrak dan kata kunci.

HASIL DAN PEMBAHASAN

HASIL

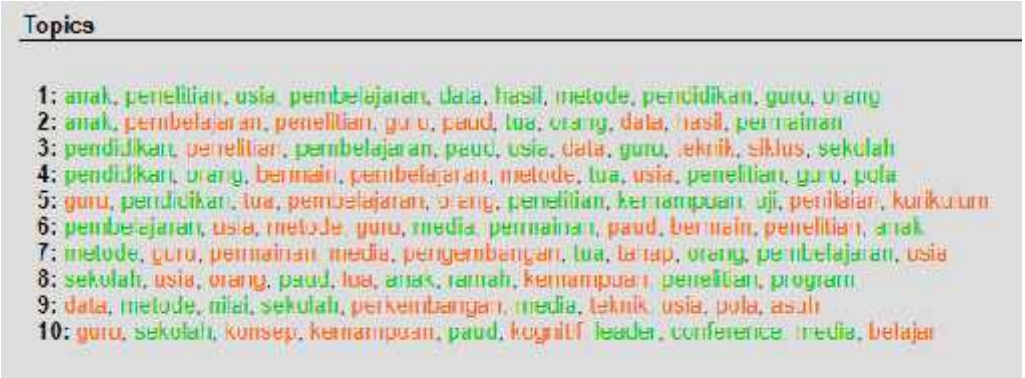
Tabel 1 menunjukkan jumlah abstrak dan kata kunci per tahun sebelum dan sesudah dibersihkan melalui tahap preprocess.

Tabel 1. Jumlah abstrak dan kata kunci

Abstrak			Kata kunci	
Sebelum Preprocess	Sesudah Preprocess	Tahun	Sebelum Preprocess	Sesudah Preprocess
183	157	2018	91	61
121	84	2019	354	265
35	20	2020	590	479
339	261	Jumlah	1035	805

Data abstrak yang telah melewati tahap preprocessing kemudian digunakan untuk tahap topic modelling, atau pemodelan topik.

Pada tahap topic modelling ini, latent semantic analysis (LSA) dipilih untuk mengidentifikasi topik-topik laten dalam data abstrak yang telah dimasukkan ke dalam Orange. Hasilnya nampak dalam Gambar 2.



Gambar 2. Topik-topik yang dihasilkan dalam tahap *topic modelling*

Tahap selanjutnya adalah word coud untuk memvisualisasikan kata-kata yang sering muncul atau digunakan dalam abstrak yang diteliti. Gambar 3 merupakan visualisasi word cloud yang diperoleh dari abstrak.

Gambar 3. *Word cloud* dari abstrak

Word cloud juga dapat digunakan untuk melihat visualisasi frekuensi kata kunci yang disediakan oleh para penulis artikel, seperti ditunjukkan dalam Gambar 4.

Gambar 4. *Word cloud* dari kata kunci pengarang

Tabel 2. Sepuluh (10) kata yang paling banyak digunakan

Abstrak			Kata Kunci Pengarang		
No.	Frekuensi	Kata	No.	Frekuensi	Kata
1	1185	anak	1	162	anak
2	1011	penelitian	2	133	usia
3	451	usia	3	50	pendidikan
4	387	pembelajaran	4	44	pembelajaran
5	357	hasil	5	37	paud
6	349	data	6	21	kemampuan
7	289	pendidikan	7	20	guru
8	276	metode	8	19	permainan
9	275	guru	9	19	covid
10	210	orang	10	18	metode

Terdapat enam kata yang paling banyak muncul baik dalam abstrak maupun kata kunci pengarang, yaitu: anak, usia, pembelajaran, pendidikan, metode, dan guru.

PEMBAHASAN

Topik-topik yang dihasilkan dalam tahap topic modelling menampilkan sepuluh kluster atau kelompok kata kunci yang terasosiasi atau terhubung dengan masing-masing dokumen/abstrak. Metode LSA diterapkan untuk mendapatkan bobot topik yang positif dan negatif. Dengan merujuk pada Gambar 2, masing-masing bobot topik ini memiliki arti yang berbeda; kata-kata yang berwarna hijau mewakili bobot positif, yang berarti mereka lebih mewakili suatu topik, sementara yang berwarna merah kurang mewakili.

Kelompok kata yang ditunjukkan pada Gambar 2 tersebut merupakan kata kunci dokumen atau abstrak artikel yang sedang diteliti; masing-masing kelompok dapat membentuk topik. Dalam satu kelompok kata kunci dapat ditemukan lebih dari satu konteks atau tema. Contohnya kelompok kata kunci berwarna hijau pada nomor 1: anak, penelitian, usia, pembelajaran, data, hasil, metode, pendidikan, guru, dan orang dapat merujuk pada topik penelitian tentang metode pembelajaran, pendidikan guru, atau metode penelitian tentang anak.

Dalam word cloud dari abstrak dapat ditemukan beberapa kata yang tidak tampil dalam word cloud dari kata kunci pengarang meskipun bukan merupakan kata yang berukuran besar, seperti kemuhammadiyahan, kooperatif, daring, dan lain-lain. Hal ini terjadi karena penulis tidak mencantumkan kata kunci penelitiannya dalam abstrak, sehingga kata kunci yang seharusnya mencerminkan isi artikel justru tidak didapati dalam abstrak. Dari gambar word cloud tersebut dapat dipahami bahwa semakin besar ukuran suatu kata, semakin sering frekuensi penggunaannya.

Dari 10 (sepuluh) kata yang paling banyak muncul dalam artikel yang diteliti dari tahun 2018-2020, baik dalam abstrak maupun kata kunci pengarang, kata anak, usia, pembelajaran, pendidikan, metode, dan guru yang mempunyai frekuensi tertinggi dibanding dengan kata lain yang tidak ditemukan secara bersama dalam abstrak dan kata kunci pengarang.

PENUTUP

Kesimpulan

Penelitian ini menyajikan analisis teks atas artikel-artikel penelitian bidang PAUD berbahasa Indonesia yang diterbitkan pada tahun 2018-2020 dengan menggunakan teknik text mining, yaitu topic modelling dan word cloud. Hasil penelitian menunjukkan sepuluh kluster atau kelompok kata kunci yang merepresentasi topik atau tema penelitian dari abstrak artikel yang diteliti. Dalam word cloud yang dihasilkan dari abstrak dan dari kata kunci pengarang terdapat enam kata yang paling sering digunakan: anak, usia, pembelajaran, pendidikan, metode, dan guru. Kata kunci-kata kunci yang teridentifikasi dapat digunakan untuk pencarian atau pengelompokkan artikel penelitian PAUD yang relevan. Sebagai penutup, teknik text mining dapat menolong peneliti menemukan pengetahuan atau informasi tersembunyi dari banyaknya artikel hasil penelitian.

Saran

Untuk mengetahui kluster topik penelitian yang telah dilakukan di bidang PAUD, penelitian selanjutnya dapat memanfaatkan teknik lainnya, seperti hierarchical clustering. Teknik ini akan mengelompokkan artikel-artikel penelitian dalam suatu struktur hierarki berdasarkan kemiripan abstrak artikel. Metode lain yang dapat dimanfaatkan adalah menerapkan analisis isi atau content analysis untuk melihat cakupan penelitian. Terakhir, menarik jika ada perbandingan hasil kluster topik penelitian bidang PAUD menggunakan kedua metode tersebut.

DAFTAR PUSTAKA

Adil, S. H., Ebrahim, M., Ali, S. S. A., & Raza, K. (2019). Identifying Trends in Data Science Articles using Text Mining. 2019 4th International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST), 1–7. https://doi.org/10.1109/ICEEST48626.2019.8981702

Adiwijaya, I. (2006). Text Mining dan Knowledge Discovery. Kolokium Bersama Komunitas Datamining Indonesia & Soft-Computing Indonesia, 1–9.

Akilan, A. (2015). Text mining: Challenges and future directions. 2015 2nd International Conference on Electronics and Communication Systems (ICECS), 1679–1684. https://doi.org/10.1109/ECS.2015.7124872

Humphreys, A., & Wang, R. J.-H. (2018). Automated Text Analysis for Consumer Research. Journal of Consumer Research, 44(6), 1274–1306. https://doi.org/10.1093/jcr/ucx104

İlhan, S. D., & Erden, F. T. (2019). A Content analysis of graduate theses concerning early childhood education in Turkey. Turkish Journal of Education, 8(2), 86–108. https://doi.org/10.19128/turje.489226

Jadhira, A. A., Bijaksana, M. A., & Wahyudi, B. A. (2018). Deteksi Kemiripan Bagian-bagian Terjemah Al-Qur'an dengan Menggunakan Metode. e-Proceeding of Engineering, 5(3), 9.

Jayashankar, S., & Sridaran, R. (2017). Superlative model using word cloud for short answers evaluation in eLearning | SpringerLink. Education and

Information Technologies, 22, 2283–2402. https://doi.org/10.1007/s10639-016-9547-0

Khodabandelou, R., Mehran, G., & Nimehchisalem, V. (2018). A Bibliometric Analysis of 21st Century Research Trends in Early Childhood Education. Revista Publicando, 5(15), 137–163. https://papers.ssrn.com/abstract=3234130

Mushlihudin, & Zahrotun, L. (2017). Perancangan Text Mining Pengelompokkan Penelitian Dosen Menggunakan Metode Shared Nearest Neighbor Dengan. Prosiding SNATIF Ke -4, 849–855.

N, Y., & S, M. (2016). A Review on Text Mining in Data Mining. International Journal on Soft Computing, 7(2/3), 01–08. https://doi.org/10.5121/ijsc.2016.7301

Nurmalasari, D., & Syahrial, Z. (2014). Pemanfaatan Sequential Pattern Sebagai Representasi Data Untuk Pengelompokan Dokumen Penelitian. shorturl.at/eiyFT

Priandini, N., Zaman, B., & Purwanti, E. (2017). Categorizing document by fuzzy C-Means and K-nearest neighbors approach. 020012. https://doi.org/10.1063/1.4994415

Sabna, E. (2020). Penerapan Text Mining Untuk Pengelompokan Penelitian Dosen. Jurnal Ilmu Komputer, 9(2), 161–164. https://doi.org/10.33060/JIK/2020/Vol9.Iss2.183

Sheridan, S. M., Edwards, C. P., Marvin, C. A., & Knoche, L. L. (2009). Professional Development in Early Childhood Programs: Process Issues and Research Needs. Early Education and Development, 20(3), 377–401. https://doi.org/10.1080/10409280802582795

Sinclair, J., Cardew-hall, M., Sinclair, J., & Cardew-hall, M. (2008). The folksonomy tag cloud: When is it useful? Journal of Information Science, 34(15). https://doi.org/10.1177/0165551506078083

South-East Asian Ministers of Education Organization. (2016). Southeast Asian guidelines for early childhood teacher development and management. UNESCO Office Bangkok/SEAMEO. https://unesdoc.unesco.org/ark:/48223/pf0000244370

Süzek, T. O. (2017). Using latent semantic analysis for automated keyword extraction from large document corpora. Turkish Journal of Electrical Engineering & Computer Sciences, 25, 1784–1794. https://doi.org/10.3906/elk-1511-203

UNESCO International Institute for Capacity-Building in Africa. (2013). Indigenous early childhood care and education (IECCE) curriculum framework for Africa: A focus on context and contents. IICBA. https://unesdoc.unesco.org/ark:/48223/pf0000229201