JTT JURNAL TEKNOLOGI TERAPAN
POLITEKNIK NEGERI INDRAMAYU

Volume 11, Nomor 2, Juli 2025

DOI 10.31884/jtt.v11i2.793
p-ISSN 2477-3506
e-ISSN 2549-1938

INFORMATION RETRIEVAL BERBASIS LATENT DIRICHLET ALLOCATION PADA DATA KEKAYAAN INTELEKTUAL

Hashri Hayati [1*], Muhammad Riza Alifi[1]

[1] Politeknik Negeri Bandung, Bandung, Indonesia 40559
* Correspondence: hashri.hayati@polban.ac.id

Abstrak

Perubahan menuju ekonomi berbasis pengetahuan menekankan pentingnya pengelolaan kekayaan intelektual. Sayangnya, metode pencarian konvensional berbasis kata kunci sering gagal menangkap hubungan semantik antar konsep dalam dokumen, khususnya dokumen kompleks seperti paten dan hak cipta. Penelitian ini mengusulkan pendekatan berbasis topic modeling menggunakan metode Latent Dirichlet Allocation (LDA) untuk meningkatkan relevansi dan akurasi pencarian informasi dalam data kekayaan intelektual. Penelitian mengembangkan 76 model berdasarkan empat skenario: dengan dan tanpa penerjemahan bahasa, serta dengan dan tanpa tokenisasi n-gram, menggunakan jumlah topik antara 1 hingga 19. Empat model terbaik dari tiap skenario menghasilkan coherence score 0,4411–0,4581. Evaluasi menggunakan Mean Average Precision (MAP) terhadap 10 dokumen teratas menunjukkan bahwa model tanpa translasi dan dengan tokenisasi unigram (10 topik) memberikan hasil terbaik dengan MAP rata-rata 78%. Hasil penelitian menunjukkan bahwa penyamaan bahasa dan penggunaan n-gram tidak secara signifikan mempengaruhi coherence score. Namun, model tanpa penggunaan tokenisasi n-gram (kombinasi bigram dan tigram) menghasilkan pencarian yang relatif lebih relevan secara semantik berdasarkan nilai MAP. Penerjemahan otomatis pada penelitian ini menghasilkan nilai MAP yang lebih kecil dibandingkan model tanpa penerjemahan.

Kata Kunci: LDA; *Topic modeling; Intelectual property; Information retrieval*

Abstract

The shift toward a knowledge-based economy underscores the importance of intellectual property (IP) management. Unfortunately, conventional keyword-based search methods often fail to capture the semantic relationships between concepts in documents—particularly complex ones like patents and copyrights. This study proposes a topic modeling approach using the Latent Dirichlet Allocation (LDA) method to improve the relevance and accuracy of information retrieval in IP data. The research developed 76 models based on four scenarios: with and without language translation, and with and without n-gram tokenization, using topic numbers ranging from 1 to 19. The best four models from each scenario yielded coherence scores between 0.4411 and 0.4581. Evaluation using Mean Average Precision (MAP) on the top 10 documents showed that the model without translation and with unigram tokenization (10 topics) achieved the best results with an average MAP of 78%. The findings indicate that language translation and n-gram tokenization do not significantly impact the coherence score. However, models without n-gram tokenization (bigram and trigram combinations) yielded relatively more semantically relevant search results based on MAP values. Automatic translation in this study resulted in lower MAP scores compared to models without translation.

Keywords: LDA; *Topic modeling; Intelectual property; Information retrieval*

Received: 23 February 2025
Revised: 21 May 2025
Accepted: 25 May 2025
Published: 02 July 2025

DOI: 10.31884/jtt.v11i2.793

1. PENDAHULUAN

Paradigma ekonomi mulai beralih dari ekonomi berbasis aset menuju ekonomi berbasis pengetahuan. Hal ini menyebabkan manajemen kekayaan intelektual menjadi semakin penting (Jeong, Park and Yoon, 2019). Pergeseran paradigma ekonomi ini juga berdampak pada pola luaran penelitian yang semula didominasi oleh publikasi ilmiah, kini mulai diimbangi dengan publikasi kekayaan intelektual.

Kekayaan intelektual di Indonesia terdiri dari Hak Cipta, Merek, Desain Industri, Paten, Desain Tata Letak Sirkuit Terpadu, Rahasia Dagang, Indikasi Geografis, dan Kekayaan Intelektual Komunal. Setiap jenis kekayaan intelektual ini memiliki karakteristik unik dalam dokumen yang dihasilkan. Di Indonesia, khususnya perguruan tinggi, kekayaan intelektual yang terdaftar didominasi oleh Hak Cipta dan Paten.

Manajemen kekayaan intelektual terdiri dari proses yang terkait dengan pengenalan, publikasi, dan eksploitasi terhadap kreasi pemikiran manusia (Hanbury et al., 2014). Akumulasi data kekayaan intelektual menjadi semakin kompleks karena kuantitas yang meningkat, membuat munculnya kebutuhan metode yang lebih baik untuk mendapatkan dan mengakses informasi ini (Jochim, 2014). Ketika data kekayaan intelektual yang banyak ini terkoneksi, ekosistem baru untuk inovasi yang lebih terbuka dapat terbangun (Modic et al., 2019).

Information retrieval (IR) sudah dikenal sebagai bidang di komputasi yang dapat digunakan untuk menilai aspek kebaruan dari proses kekayaan intelektual dengan menyediakan metode untuk mencari, membandingkan, dan menganalisis dokumen paten (Hanbury et al., 2014). Penelitian Aristodemou mengkategorikan hasil analisis kekayaan intelektual dengan menggunakan metode kecerdasan buatan dengan pendekatan *machine learning* dan *deep learning*, terdiri dari *knowledge management, technology management, economic value*, dan *extraction and effective management of information* (Aristodemou and Tietze, 2018).

Dokumen kekayaan intelektual memiliki kekhususan, sehingga *query* berbasis kata kunci biasa terbukti tidak efektif untuk pengambilan data (Khode and Jambhorkar, 2022). Metode pencarian tradisional berbasis kata kunci sering kali gagal dalam menangkap hubungan semantik dan relevansi tematik dokumen. Analisis semantik merupakan proses menghubungkan struktur sintaksis yang mengandalkan *domain knowledge* dan menciptakan hubungan antar konsep di domain yang spesifik (Aristodemou and Tietze, 2018). Analisis semantik penting dalam pencarian dokumen kekayaan intelektual, karena ragam domain dalam dokumen memungkinkan terjadinya penyimpangan pemaknaan topik jika hanya mengandalkan penyamaan kata kunci pada pencarian. Penyimpangan topik merupakan masalah penting dalam pengambilan informasi paten karena orang cenderung menggunakan ekspresi berbeda untuk menggambarkan konsep serupa yang menyebabkan rendahnya presisi dan perolehan kembali pada saat yang bersamaan (Al-Shboul and Myaeng, 2014).

Penelitian (Khode and Jambhorkar, 2017) mengkaji teknik dan pendekatan pengambilan informasi paten. Berdasarkan penelitian tersebut, beberapa teknik yang telah digunakan di penelitian-penelitian sebelumnya diantaranya adalah *query formulation, query expansion, summarization, relevance feedback*, dan lain-lain. Sementara dari sisi pemodelan, beberapa publikasi telah menggunakan *Vector Space Model* (VSM), *semantic based processing, Latent Semantic Analysis* (LSA), *language model, weighting techniques, probabilistic model*, dan lain-lain. Terdapat beberapa pendekatan lain yang juga digunakan, yaitu metodologi Bibliometric, *data mining, text mining, database management tools*, dan analisis sitasi (Khode and Jambhorkar, 2017). Dengan kemajuan teknik *deep learning*, beberapa teknologi mutahir dari model NLP digunakan untuk mengekstraksi informasi berharga dari dokumen paten (Chen et al., 2022).

Salah satu model NLP yang digunakan adalah *topic modeling*. *Topic Modeling* merupakan komponen yang meningkat popularitasnya di riset kekayaan intelektual (Lehmann, 2023). Penelitian Lehmann pada 2023 menggunakan *topic modeling* untuk sekumpulan data merek di kantor kekayaan intelektual Kanada. Penelitian lain menggunakan Latent Dirichlet Allocation (LDA) pada data abstrak untuk mengekstrak topik inovasi dengan melihat hubungan antar subjek, dan menganalisis peraturan yang berlaku di Korea terkait inovasi terbuka (Cho, Shin and Kang, 2018). Penelitian lain mengkombinasikan LDA dan *Support Vector Machine* (SVM) untuk klasifikasi paten otomatis (Yun and Geum, 2020). Pada penelitian (Yun and Geum, 2020), LDA digunakan untuk representasi teks dan hasilnya digunakan sebagai input untuk SVM. Sementara penelitian (Jochim, 2014) melakukan analisis sitasi pada literatur ilmiah dengan mengekstrak fitur melalui analisis sentimen, *Named Entity Recognition* (NER), dan keterhubungannya.

Penelitian ini mengusulkan pendekatan *topic modeling* yang memanfaatkan Latent Dirichlet Allocation (LDA) untuk pengambilan informasi dari data kekayaan intelektual. Penelitian-penelitian lain kebanyakan menggunakan data paten saja, atau data merek, sementara pada penelitian ini dokumen yang digunakan adalah data dari dokumen hak cipta dan paten yang bersumber dari Direktorat Jenderal Kekayaan Intelektual Kemenkumham RI. Dengan mengungkap topik laten dalam dokumen, model ini akan dimanfaatkan untuk mengambil dokumen serupa berdasarkan kesamaan topik. Metodologi yang diusulkan bertujuan untuk meningkatkan akurasi dan relevansi pengambilan dokumen kekayaan intelektual dibandingkan dengan metode berbasis kata kunci tradisional.

Berdasarkan masalah yang telah dipaparkan, penelitian ini akan fokus pada bagaimana menggunakan teknik *Information Retrieval* (IR) dan pendekatan *machine learning* untuk menganalisis dan mengekstraksi informasi dari data kekayaan intelektual, khususnya dalam konteks ekstraksi topik, dengan tujuan memfasilitasi proses pencarian dan manajemen kekayaan intelektual yang lebih efektif.

Penelitian ini akan melakukan eksplorasi penggunaan algoritma *topic modeling* untuk meningkatkan relevansi pencarian pada data kekayaan intelektual dengan mempertimbangkan semantik pada isi dokumen.

2. METODE

Eksperimen pada penelitian ini dijelaskan pada Gambar 1. *Dataset* yang digunakan pada penelitian ini diperoleh dari Pangkalan Data Kekayaan Intelektual, Kementerian Hukum dan HAM, Republik Indonesia. Pengumpulan data dilakukan dengan Teknik *scrapping*, yaitu prosedur untuk ekstraksi otomatis dari data pada *website* menggunakan *software* (Khder, 2021). Jumlah dan atribut data yang digunakan dalam penelitian ini dijelaskan pada Tabel 1.

Tabel 1. Jumlah Data Kekayaan Intelektual.

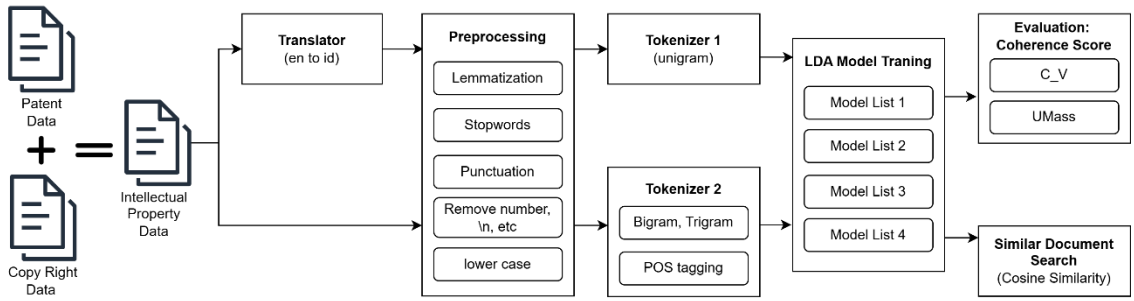
No	Jenis KI	Atribut yang Digunakan	Jumlah Data
1	Paten/Paten Sederhana	• Judul • Abstrak	420
2	Hak Cipta	• Judul • Uraian Ciptaan	60
TOTAL			480

Sebelum dimasukan menjadi data latih untuk membentuk model LDA, data yang terkumpul harus melalui tahap *pre-processing*. Setelah diamati, terdapat beberapa data yang menggunakan bahasa inggris. Oleh karena penelitian ini juga merancang tahap penerjemahan bahasa sebagai tahap yang dilakukan sebelum *pre-processing*. Agar

mengetahui dampak dari proses penerjemahan ini, dilakukan 2 skenario berbeda, yaitu skenario dengan data berbahasa inggris secara terpisah diterjemahkan terlebih dahulu dan skenario tanpa penerjemahan.

Beberapa *pre-processing* yang dilakukan pada penelitian ini adalah *lemmatization* (mengembalikan kata ke bentuk dasarnya), penghapusan *stopwords* (kata-kata yang sangat umum dan biasanya diabaikan dalam pemrosesan bahasa oleh komputer), penghapusan *punctuation* (tanda baca), mengubah seluruh huruf menjadi *lower case*, serta beberapa penghapusan kata yang dianggap tidak merepresentasikan substansi kalimat seperti spasi, baris baru, angka, dan lain-lain.

Setelah data melewati tahap *pre-processing*, data tersebut ditokenisasi menjadi *list* kata. Tokenisasi ini dilakukan untuk membentuk *dictionary* dan *corpus* dari model LDA yang akan dibuat. Penelitian ini membuat 2 skenario berbeda untuk tokenisasi, yaitu tokenisasi per kata dengan cara *split string* (unigram) dan tokenisasi yang mempertimbangkan pemotongan n buah kata (n-gram). Penelitian ini menggunakan n-gram dengan pemotongan 2 dan 3 kata, atau tokenisasi bigram dan trigram secara sekaligus dalam satu skenario *tokenizer* kedua. Pada jenis tokenizer yang kedua, token atau daftar frase juga diproses dengan *Part of Speech* (POS) *Tagging* untuk memilah kembali kata yang relevan, yaitu dengan membuang jenis kata tertentu seperti kata hubung.



Gambar 1. Desain Eksperimen.

Data yang sudah ditokenisasi dijadikan kamus dan *corpus*. Pada penelitian ini, *corpus* dibuat dengan metode *Bag of Word*. Sebagaimana dijelaskan sebelumnya, eksperimen ini menggunakan 2 jenis skenario penerjemahan dan 2 skenario tokenisasi, sehingga dihasilkan 4 eksperimen pembangunan model LDA seperti pada Tabel 2. LDA merupakan *unsupervised machine learning* sehingga pengelompokan topik dilakukan tanpa perlu menyediakan label, melainkan dengan *hyperparameter* jumlah topik yang menjadi dasar jumlah *cluster* dari model yang dihasilkan. Penelitian ini melakukan percobaan jumlah topik 1 sampai 19 dan jumlah iterasi 19.

Tabel 2. Variasi Skenario Eksperimen.

Skenario	Terjemahan ke Bahasa Indonesia	n-gram *tokenization* (Kombinasi bigram dan trigram)	Hasil Model LDA
1	Tidak	Tidak	Model *List* 1
2	Ya	Tidak	Model *List* 2
3	Tidak	Ya	Model *List* 3
4	Ya	Ya	Model *List* 4

Dari empat eksperimen tersebut, dihasilkan 76 Model LDA yang dievaluasi menggunakan *coherence score*. Tahap berikutnya dari penelitian ini adalah

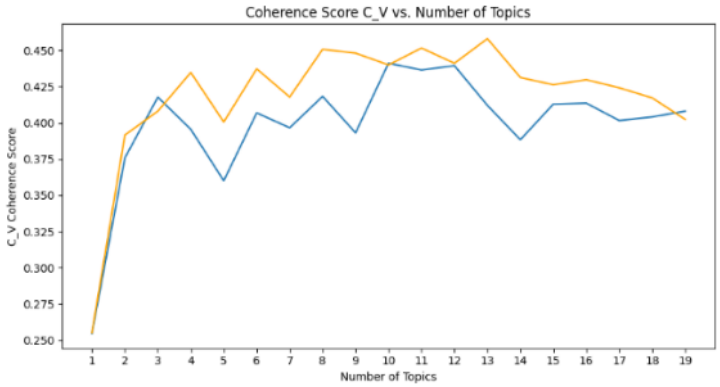
mengimplementasikan model LDA untuk melakukan *similar document searching* dengan menghitung nilai *cossine similarity*. Hasil pencarian berdasarkan nilai *cosine similarity* kemudian digunakan untuk menghitung *Mean Average Precision* (MAP) dari relevansi dokumen yang dihasilkan.

3. HASIL DAN PEMBAHASAN

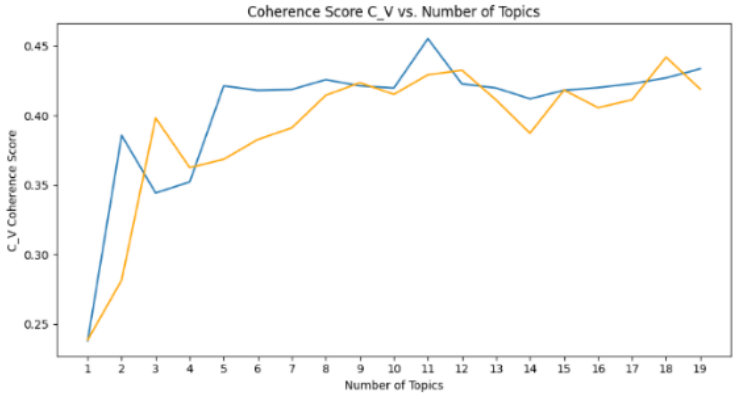
Coherence score merupakan salah satu metrik evaluasi dari *topic modeling* yang mengukur interpretabilitas dan kualitas topik yang dihasilkan model (Belford and Greene, 2019). Terdapat beberapa metode yang bisa digunakan untuk mengukur nilai *coherence* (Röder, Both and Hinneburg, 2015). Penelitian ini menggunakan *tools* Gensim pada Python (Řehřek, Sojka and others, 2011) dengan default metrik evaluasi yang digunakan adalah Coherence C_V. Nilai Coherence C_V diperoleh dengan merata-ratakan *cosine similarity* (s_{cos}) untuk seluruh index topik (k) dan index kata dalam topik (n) menggunakan formula (1).

$$c_v = \frac{\sum_{k=1}^{K} \sum_{n=1}^{N} s_{cos}\left(\vec{w}_{n,k}, \vec{w}_k^*\right)}{N \times K} \qquad (1)$$

Gambar 2 dan Gambar 3 menunjukkan hasil evaluasi *coherence score* dari model yang dihasilkan pada penelitian ini. Grafik pada Gambar 2 menunjukan *coherence score* dari model *list* 1 dan 2 dengan percobaan jumlah topik 1 sampai 19, sementara Gambar 3 menunjukan nilai tersebut untuk model *list* 3 dan 4. Tabel 3 menunjukan model terbaik dan jumlah topik berdasarkan nilai *coherence* tertinggi untuk setiap Model *List* sesuai skenario pada Tabel 2.



Gambar 2. Coherence Score C_V Model *List* 1 (biru) dan 2 (oranye).



Gambar 3. Coherence Score C_V Model List 3 (biru) dan 4 (oranye).

Tabel 3. Coherence Score Model Terbaik.

Model LDA	*Coherence Score*	Jumlah Topik
Model_List_1 [9]	0.4411	10
Model_List_2 [12]	0.4581	13
Model_List_3 [10]	0.4552	11
Model_List_4 [17]	0.4419	18

Dari hasil penelitian ini, diperoleh beberapa temuan sebagai berikut:

1. Pengaruh Translasi terhadap nilai *coherence*
   Model 2 menunjukkan bahwa translasi ke bahasa Indonesia tidak secara signifikan meningkatkan nilai *coherence* dibandingkan model lain. Hal ini mengindikasikan bahwa penerjemahan tidak memberikan dampak besar terhadap pemahaman semantik topik dalam pemodelan. Namun hal ini juga bisa jadi disebabkan proporsi data dalam bahasa asing yang perlu penerjemahan tidak berjumlah signifikan.
2. Peran *Tokenizer* N-Gram
   Meskipun teknik n-gram (bigram dan trigram) sering digunakan untuk menangkap konteks lebih luas dalam teks, hasil penelitian ini menunjukkan bahwa penggunaannya dalam model tidak memiliki pengaruh signifikan terhadap nilai *coherence*. Hal ini bisa berarti bahwa pendekatan lain dalam pemrosesan teks mungkin lebih efektif dalam meningkatkan kualitas pemodelan topik.
3. Variasi Skor Antar Model
   Perbedaan nilai *coherence* antar model terbaik memiliki selisih kurang dari 2% menunjukkan bahwa semua pendekatan yang digunakan memiliki hasil yang relatif serupa. Hal ini dapat mengindikasikan bahwa pemilihan metode tidak memberikan perbedaan yang cukup besar untuk meningkatkan kualitas topik yang dihasilkan.
4. Implikasi untuk Pemodelan Topik
   Berdasarkan temuan ini, dapat disimpulkan bahwa penyamaan bahasa dan penggunaan *tokenizer* n-gram bukanlah faktor utama dalam meningkatkan nilai *coherence*. Peneliti atau praktisi dapat mempertimbangkan faktor lain, seperti pemilihan algoritma pemodelan topik atau parameter pemrosesan teks lainnya, untuk mencapai hasil yang lebih optimal.

Pemodelan topik pada penelitian ini memiliki tujuan akhir untuk mencari dokumen yang relevan secara semantik, oleh karena itu setelah mendapatkan model terbaik, penelitian dilanjutkan dengan menyediakan fitur pencarian yang memanfaatkan model LDA. Fitur pencarian telah dikembangkan dengan menggunakan metode *cosine similarity*.

Salah satu metrik pengujian relevansi dokumen hasil pencarian adalah MAP, karena tugas penting dari sistem pencarian adalah memaksimalkan nilai presisi dan *recall* (Shukla, Das and Kumar, 2021). Presisi adalah proporsi dari dokumen relevan yang berhasil ditemukan dalam hasil pencarian (Kishida, 2005). Pada kasus pencarian informasi, urutan (*rank*) kemunculan dokumen seringkali perlu dipertimbangkan, sehingga penelitian ini menggunakan metrik *Mean Average Precision* @10 (MAP@10) yang menghitung nilai relevansi dokumen pada 10 dokumen dengan nilai *cosine similarity* tertinggi. Penelitian ini menggunakan tiga data uji (Q) yang dimasukan ke empat model terbaik sesuai Tabel 3 untuk menghasilkan 10 dokumen dengan nilai *cosine similarity* terbesar. Setiap dokumen yang dihasilkan dimasukan dalam kategori relevan (R) atau tidak relevan (N) sehingga diperoleh Tabel 4 Relevansi Dokumen.

Tabel 4. Relevansi Dokumen.

	Q1				Q2				Q3			
	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
1	**R**	**R**	**R**	N	**R**	**R**	**R**	N	**R**	N	**R**	N
2	N	N	**R**	N	**R**	N	**R**	**R**	**R**	**R**	**R**	**R**
3	**R**	N	**R**	**R**	**R**	N	N	N	**R**	**R**	**R**	**R**
4	**R**	N	N	N	**R**	**R**	N	**R**	**R**	**R**	**R**	**R**
5	**R**	N	**R**	**R**	**R**	N	**R**	**R**	**R**	**R**	N	**R**
6	N	**R**	**R**	**R**	**R**	N	**R**	N	**R**	N	**R**	**R**
7	**R**	**R**	**R**	**R**	**R**	N	N	N	**R**	**R**	N	**R**
8	N	**R**	**R**	N	**R**	N	**R**	N	**R**	N	**R**	N
9	**R**	N	**R**	N	N	N	**R**	N	**R**	**R**	**R**	N
10	**R**	N	N	N	N	N	**R**	N	**R**	N	**R**	N
n(R)	**7**	**4**	**8**	**4**	**8**	**2**	**7**	**3**	**10**	**6**	**8**	**6**

Pada setiap *query* di masing-masing model, dilakukan perhitungan *Average Precision* @10 (AP@10) sesuai formula (2). Nilai akhir MAP@10 diperoleh dengan merata-ratakan nilai AP@10 untuk 3 *query* yang diuji sesuai formula (3). Tabel 5 menunjukan nilai AP@10 untuk setiap *query* dan MAP@10 untuk setiap model.

$$AP@K = \frac{1}{N}\sum_{k=1}^{K} precision(k) \times rel(k) \qquad (2)$$

$AP@K = Average\ Precision\ pada\ Top\ K\ recommendation$
$N = jumlah\ dokumen\ relevan$
$precision\ (k) = nilai\ precision\ pada\ posisi\ k$
$rel(k) = bernilai\ 1\ jika\ dokumen\ pada\ posisi\ k\ relevan\ atau\ 0\ jika\ tidak\ relevan$

$$MAP@K = \frac{1}{Q}\sum_{q=1}^{Q} AP@K_q \qquad (2)$$

$MAP@K = Mean\ Average\ Precision\ pada\ Top\ K\ recommendation$
$Q = jumlah\ query\ yang\ diujikan$

Tabel 5. MAP@10 Relevansi Dokumen.

	$AP@10_1$	$AP@10_2$	$AP@10_3$	$MAP@10$
M1	53%	80%	100%	78%
M2	23%	15%	41%	26%
M3	73%	53%	72%	66%
M4	18%	16%	44%	26%

Berdasarkan Tabel 5, Model 1 dan 3 menghasilkan pencarian dokumen yang jauh lebih relevan daripada model 2 dan 4. Dari eksperimen yang dilakukan, berikut beberapa temuan yang dihasilkan:

1. Pengaruh tokenisasi pada relevansi hasil pencarian
   Secara umum, MAP dari model 1 yang menggunakan tokenisasi sederhana (unigram) memiliki nilai tertinggi, namun hal ini tidak secara konsisten ditunjukan pada perbandingan M2 dan M4 dimana keduanya menggunakan teknik tokenisasi berbeda dan menghasilkan nilai MAP yang sama. Pada penelitian ini dapat disimpulkan penggunaan kombinasi n-gram yang diharapkan dapat

meningkatkan relevansi pencarian tidak selalu bisa terjadi. Hal ini dapat disebabkan karakteristik dokumen yang relatif pendek atau penggunaan istilah teknis yang cukup kuat meski hanya dengan satu kata (contoh: “regresi”, “klasifikasi”).

2. Pengaruh translasi otomatis pada relevansi hasil pencarian
   Berdasarkan eksperimen yang dilakukan, translasi otomatis menurunkan nilai presisi secara signifikan. Pada penelitian ini translasi dilakukan pada dokumen yang terklasifikasi menggunakan Bahasa Inggris untuk kemudian diterjemahkan ke Bahasa Indonesia. Pada dokumen kekayaan intelektual, banyak istilah teknis yang sudah baku menjadi kabur maknanya saat diterjemahkan dalam Bahasa Indonesia.

4. PENUTUP

Kesimpulan

Dari 76 model LDA yang dibangun, diperoleh 4 model terbaik dengan *coherence score* tertinggi yaitu 0,4411 – 0,4581 dan jumlah topik 10 – 18. Relevansi dokumen diuji dengan menghitung MAP pada 10 urutan *cosine similarity* tertinggi (MAP@10) untuk 4 model tersebut dan didapat nilai tertinggi dari model 1 dengan rata-rata 78%, yaitu model tanpa translasi bahasa dan menggunakan tokenisasi unigram dengan jumlah topik 10. Hasil penelitian menunjukkan bahwa penyamaan bahasa dan penggunaan *tokenizer* n-gram (kombinasi bigram dan trigram) tidak secara signifikan mempengaruhi *coherence score*. Berdasarkan nilai MAP, model tanpa penggunaan *tokenizer* n-gram menghasilkan pencarian yang relatif lebih relevan secara semantik. Penggunaan penerjemahan otomatis pada penelitian ini menghasilkan nilai rata-rata presisi yang jauh lebih rendah daripada model tanpa proses penerjemahan dokumen.

Saran

Perlu dilakukan penelitian lebih lanjut mengenai dampak penerjemahan bahasa dengan jumlah data yang lebih representatif. Pengujian relevansi juga dapat dibandingkan dengan metode *information retrieval* lainnya.

Daftar Pustaka

Al-Shboul, B. and Myaeng, S.H., 2014. Wikipedia-based query phrase expansion in patent class search. *Information Retrieval*, 17(5–6), pp.430–451. https://doi.org/10.1007/s10791-013-9233-4.

Aristodemou, L. and Tietze, F., 2018. *The state-of-the-art on Intellectual Property Analytics (IPA): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data*. *World Patent Information*, https://doi.org/10.1016/j.wpi.2018.07.002.

Belford, M. and Greene, D., 2019. Comparison of Embedding Techniques for Topic Modeling Coherence Measures. In: *LDK (Posters)*. pp.1–5.

Chen, L., Xu, S., Zhu, L., Zhang, J., Yang, G. and Xu, H., 2022. A deep learning based method benefiting from characteristics of patents for semantic relation classification. *Journal of Informetrics*, 16(3), p.101312.

Cho, S.-B., Shin, S. and Kang, D.-S., 2018. A study on the research trends on open innovation using topic modeling. *Informatization policy*, 25(3), pp.52–74.

Hanbury, A., Lupu, M., Kando, N., Diallo, B. and Adams, S., 2014. *Guest editorial: Special issue on information retrieval in the intellectual property domain*. *Information Retrieval*, https://doi.org/10.1007/s10791-014-9245-8.

Jeong, Y., Park, I. and Yoon, B., 2019. Identifying emerging Research and Business Development (R&BD) areas based on topic modeling and visualization with intellectual property right data. *Technological Forecasting and Social Change*, 146, pp.655–672. https://doi.org/10.1016/j.techfore.2018.05.010.

Jochim, C., 2014. *Natural Language Processing and Information Retrieval Methods for Intellectual Property Analysis*.

Khder, M., 2021. Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. *International Journal of Advances in Soft Computing and its Applications*, 13(3), pp.145–168. https://doi.org/10.15849/IJASCA.211128.11.

Khode, A. and Jambhorkar, S., 2017. A Literature Review on Patent Information Retrieval Techniques. *Indian Journal of Science and Technology*, [online] 10(36), pp.1–13. https://doi.org/10.17485/ijst/2017/v10i37/116435.

Khode, A. and Jambhorkar, S., 2022. Augmenting keyword-based patent prior art search using weighted classification code hierarchies. *International Journal of Business Intelligence and Data Mining*, 21(4), pp.397–418.

Kishida, K., 2005. *Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments*. National Institute of Informatics Tokyo, Japan.

Lehmann, A., 2023. *Topic Modeling for Intellectual Property Research: Comparing Methods Through Simulation and Application*.

Modic, D., Hafner, A., Damij, N. and Cehovin Zajc, L., 2019. Innovations in intellectual property rights management: Their potential benefits and limitations. *European Journal of Management and Business Economics*, 28(2), pp.189–203. https://doi.org/10.1108/EJMBE-12-2018-0139.

Řehřek, R., Sojka, P. and others, 2011. Gensim—statistical semantics in python. *Retrieved from genism. org*.

Röder, M., Both, A. and Hinneburg, A., 2015. Exploring the Space of Topic Coherence Measures. In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. New York, NY, USA: ACM. pp.399–408. https://doi.org/10.1145/2684822.2685324.

Shukla, A.K., Das, S. and Kumar, P., 2021. WordNet Based Hybrid Model for Query Expansion. In: *2021 IEEE International Conference on Technology, Research, and Innovation for Betterment of Society (TRIBES)*. IEEE. pp.1–6. https://doi.org/10.1109/TRIBES52498.2021.9751671.

Yun, J. and Geum, Y., 2020. Automated classification of patents: A topic modeling approach. *Computers and Industrial Engineering*, 147. https://doi.org/10.1016/j.cie.2020.106636.