Berkala Ilmu Perpustakaan dan Informasi, Vol. 17, No. 2, Desember 2021 Hal. 168-180
https://doi.org/10.22146/bip.v17i1.2147
ISSN 1693-7740 (Print), ISSN 2477-0361 (Online)
Tersedia online di https://journal.ugm.ac.id/v3/BIP

Pemodelan topik pada dokumen paten terkait pupuk di Indonesia berbasis *Latent Dirichlet Allocation*

Aris Yaman[1], Bagus Sartono[2], Agus M. Soleh[3]
[1]Pascasarjana, Institut Pertanian Bogor
[2,3]Departemen Statistika dan Sains Data, Institut Pertanian Bogor
e-mail: [1]arisyaman@apps.ipb.ac.id, [2]bagusco@apps.ipb.ac.id, [3]agusms@apps.ipb.ac.id

Naskah diterima: 14 Juli 2021, direvisi: 2 September 2021, disetujui: 29 September 2021

ABSTRAK

Pendahuluan. Pupuk merupakan salah satu produksi terpenting dalam dunia pertanian. Peningkatan kapasitas teknologi terkait pupuk penting untuk dikembangkan. Analisis dokumen paten bisa menjadi salah satu cara dalam menganalisis perkembangan teknologi khususnya pupuk.
Metode penelitian. Data yang digunakan dalam penelitian ini adalah metadata khususnya bagian judul dan abstrak suatu dokumen paten berbahasa Indonesia yang terdaftar pada Dirjen Kekayaan Intelektual (DJKI-Kemenkumham). *Keyword* "pupuk" digunakan untuk memfilter hasil sesuai data yang diinginkan. Metadata Paten yang diproses pada periode 1945-2017.
Data analisis. Model LDA dapat memberikan interpretasi yang baik terkait pemodelan topik berbasis data teks. Pemodelan topik bertujuan untuk menemukan suatu kata atau kumpulan kata dari suatu dokumen yang merepresentasikan suatu topik.
Hasil dan Pembahasan. Tingkat koherensi bagian judul paten lebih baik daripada bagian abstrak paten. Pendekatan LDA dapat dengan baik memisahkan topik-topik teknologi paten pupuk sehingga tidak multi tafsir.
Kesimpulan dan Saran. Berdasarkan pendekatan ini pada dasarnya terdapat sembilan topik esensial dalam pengembangan teknologi pupuk. Implikasi teoritis yang dapat dilihat salah satunya tingkat koherensi judul pada dokumen paten jauh lebih baik dalam memberikan gambaran dalam pemodelan LDA dibanding koherensi pada bagian abstrak dokumen paten. Selain itu didapati bahwa memasukan parameter evaluasi tingkat keunikan kata, penting untuk dilakukan. Hal ini dapat memperkecil tingkat ambiguitas interpretasi topik.

Kata kunci: LDA; pemodelan topik; paten; ukuran koherensi topik

ABSTRACT

Introduction. *Fertilizer is one of the most important production in agriculture. It is crucial to increase the capacity of technology related to fertilizers. Analysis of patent documents can be one way to analyze technological developments, particularly fertilizers.*
Data Collection Methods. *The data used was metadata, particulatly the title and abstract of a patent document in Indonesia. With the keyword "fertilizer", patent metadata was harvested from 1945 to 2017.*
Data Analysis*. The Latent Dirichlet Allocation (LDA) model was used to provide a reasonable interpretation regarding topic modeling based on text data.*
Results and Discussion. *The results showed that the degree of the patent title is better than the abstract. The LDA approach can adequately separate the topics of fertilizer patent technology so that it does not have multiple interpretations.*
Conclusion. *Based on the findings, there are nine essential topics in the development of fertilizer technology. There is a phenomenon of the lack of technology collaboration between International Patent Classification (IPC) technology sections. In addition, it was found that the evaluation of the level of uniqueness of words is important to do, to minimize the level of ambiguity in the interpretation of the topic.*

Keywords: *LDA; topic modelling; paten; topic coherence*

A. PENDAHULUAN

Indonesia merupakan negara agraris yang telah lama dikenal dunia. Tetapi kenyataan ini tidak diimbangi dengan ketahanan teknologi pendukung khususnya bidang pertanian. Salah satu faktor produksi dalam usaha pertanian yang membutuhkan sumber daya ataupun biaya yang besar adalah pupuk. Alokasi Biaya 15-30% dari total cost untuk komoditi padi (FAO, 2016). Pemerintah dan pemangku kepentingan sudah selayakanya mendukung inovasi khususnya paten yang berkaitan dengan teknologi pupuk ini.

Dokumen-dokumen paten berisi berbagai informasi teknis yang berkaitan dengan hak kekayaan intelektual dan hasil penelitian (Hongshu *et al.*, 2017). Sebuah penelitian menunjukkan bahwa negara yang tidak melakukan analisis terhadap data paten berdampak terhadap tingginya biaya dan waktu yang panjang dalam proses pengembangan teknologi (Kim & Bae, 2017). Oleh karena itu, pemangku kebijakan seharusnya dapat memanfaatkan informasi yang terdapat pada penelitian ilmiah dan pengembangan industri yang telah dipatenkan sebelumnya untuk menentukan arah inovasi bangsa. Dalam hal ini, jika digunakan secara benar, informasi paten dapat menjadi salah satu faktor pendorong utama dalam perkembangan teknologi dan kemajuan sosial. Pemanfatan data paten dapat dijadikan sebagai landasan dalam menyusun *roadmap* inovasi. Hal ini tidak hanya memainkan peran penting dalam menemukan peluang teknologi baru, menghindari duplikasi investasi R&D, mencegah risiko pelanggaran paten, dan sebagainya, tetapi juga memiliki arti penting untuk memandu pengembangan industri secara keseluruhan (Yu & Zhang, 2019).

Studi terdahulu telah banyak berfokus pada pengklasifikasian paten, bukan berfokus pada representasi paten. Representasi dokumen sangat penting karena menentukan karakteristik paten tertentu. Selain itu, hal ini juga menjawab pertanyaan apa saja isi paten dan bagaimana isi paten tersebut dapat direpresentasikan secara efektif dalam bentuk yang terstruktur (Yun & Geum, 2020).

Penelitian terkait model representasi dokumen paten (pemodelan topik dalam paten) telah banyak dilakukan sebelumnya. Chen *et al*.(2016) memperkenalkan model Paten LDA dalam menganalisa akar topik dalam suatu dokumen paten. Penelitian ini mengklaim bahwa dengan teknik paten *Latent Dirichlet Allocation* (*Patent* LDA) dapat memberikan indikator nilai *perplexity* yang lebih baik dibandingkan dengan metode LDA konvensional. Hu et al., (2018) memperkenal metode *Hierarchical Feature Extraction Model* (HFEM) dan membandingkannya dengan tiga pemodelan lainnya yaitu *single neural network model* (CNN), *long–short-term memory* (LSTM), dan BiLSTM. Hasil studi memperlihatkan model yang diperkenalkan memiliki performa lebih baik dalam hal persisi dan recal ketika memodelkan topik dalam suatu dokumen paten. Kim *et al.*(2020) menerapkan model *Word2vec-based latent semantic analysis* (W2V-LSA) dalam memodelkan topik paten. Metode ini dapat menjadi alternatif dalam memodelkan paten teknologi *blockchain*, sehingga didapat tren dan riset kedepan yang dapat diteliti lebih lanjut dalam teknologi *blockchain*.

Pemaknaan dokumen paten dari peubah-peubah penjelas berupa teks sulit untuk dipahami pemaknaannya dalam pemodelan topik. Kendala ini dapat diatasi dengan melakukan transformasi pada sebaran kata peubah penjelas. *Latent Dirichlet Allocation* (LDA) dapat memberikan keluaran berupa daftar topik (kumpulan beberapa kata/frase yang menjadi intisari utama pembahasan dalam dokumen) yang diberi bobot untuk masing-masing dokumen. Pendekatan LDA dengan kata lain dapat melakukan transformasi peubah penjelas berupa sebaran kata menjadi peubah penjelas sebaran topik. Peubah penjelas berupa topik dapat memberikan pemaknaan pada peubah penjelas.

Berdasarkan paparan di atas, didapati bahwa masih terdapat gap kekosongan penelitian. Beberapa gap riset yang terjadi yaitu pertama belum adanya penelitian terdahulu yang melakukan pemodelan topik yang berbasis *corpus* dokumen paten dengan teks berbahasa

Indonesia. Kedua, masih belum ditemukannya penelitian terkait pemodelan topik yang dapat memberikan representasi akar kata dari suatu topik paten, terkhusus dokumen paten di Indonesia.

Penelitian ini berusaha melakukan pemodelan terhadap data teks pada dokumen paten, terkhusus bagian judul dan abstrak suatu dokumen paten. Pemodelan terhadap dokumen paten ini diharapkan nantinya didapat sekumpulan tema/topik tersembunyi yang terdapat dalam dokumen-dokumen paten di Indonesia, terkhusus dokumen bertemakan teknologi pupuk. Pendekatan yang dilakukan dalam pemdodelan untuk mendapatkan kumpulan topik tersembunyi dalam dokumen paten dengan menerapkan pemodelan berbasis LDA. Analisa melalui dokumen paten diharapkan didapati tren teknologi dan topik yang representatif terkait pendaftaran paten bidang teknologi pupuk ini.

B. TINJAUAN PUSTAKA

Paten

Negara memberikan hak ekslusif atas penemuan teknologi di bidang tertentu kepada inventor selama kurun waktu tertentu dalam bentuk paten (Presiden Republik Indonesia, 2016). Menurut amanat UU No. 13 tahun 2016, pengelolaan paten baik proses pendaftaran, publikasi dan pemberian, dilakukan oleh Direktorat Jenderal Kekayaan Intelektual - Kemenkumham (DJKI). Sebagai bentuk transparansi pengelolaan kekayaan intelektual, DJKI memberikan laporan/publikasi status paten di laman https://pdki-indonesia.dgip.go.id/. Pada situs tersebut ditampilkan metadata terkait paten-paten yang di daftarkan ke DJKI. Metadata informasi paten inilah yang akan dianalisis lanjut pada penelitian ini.

International Patent Classification (IPC)

International Patent Classification (IPC) merupakan suatu sistem klasifikasi paten secara internasional yang dikembangkan oleh *World Intellectual Property Organization* (WIPO). Sistem klasifikasi IPC merupakan cara yang paling tepat untuk menggali informasi paten karena pengkategorian dan pengindeksannya yang konsisten. Untuk mengklasifikasikan semua fitur teknologi yang relevan, seluruh paten diklasifikasikan setidaknya oleh satu kode IPC. Ketika terdapat lebih dari satu aspek teknologi yang berbeda dalam satu paten tunggal, maka paten tersebut akan memiliki lebih dari satu kode IPC. Sebagai hasilnya, dapat dengan mudah untuk mengidentifikasi berapa banyak teknologi dan bidang teknologi mana yang saling terkait dalam suatu paten. Aplikasi paten di masing-masing bidang menunjukkan akumulasi dari pengetahuan dan kemajuan dalam lintasan teknologi. Oleh karena itu, dengan menganalisis data kode IPC yang diekstraksi dari dokumen paten memungkinkan untuk memahami perkembangan teknologi dan membuat prediksi tren pengembangan teknologi (Hongshu *et al*., 2017).

Kode IPC adalah suatu hirarki yang menetapkan keberadaan paten dalam suatu kategori. Terdapat 8 bagian/seksi, 128 kelas, 648 subkelas, sekitar 7.200 kelompok utama, dan sekitar 72.000 subkelompok. Adapun 8 Seksi IPC terdiri dari: (A) Kebutuhan Manusia (*Human Necessities*); (B) Peogoperasian, Transportasi (*Perfoming Operation, Transportation*); (C) Kimia, Metalurgi (Chemistry, Metallurgy); (D) Tekstil, Kertas (Textiles, Paper); (E) Konstruksi (*Fixed Constructions*); (F) Teknik Permesinan, Penerangan, Pemanasan, Senjata, Peledakan (*Mechanical Engineering, Lighting, Heating, Weapons, Blasting*); (G) Fisika (*Physiscs*); (H) Kelistrikan (*Electricity*) (WIPO, 2018). Penelitian ini menggunakan versi IPC terbaru (2018). Sebagai contoh, definisi dari A61 adalah "Ilmu Medis dan Kedokteran Hewan; Ilmu Kesehatan" dibagi dalam 16 subkelas, diantaranya A61B -"Diagnosa, Operasi, dan Identifikasi"; A61D-"Instrumen, Pengaplikasian, Alat, dan Metode untuk Kedokteran Hewan", A61K-"Persiapan untuk tujuan medis, gigi, atau toilet".

Latent Dirichlet Allocation (LDA)

Hal yang mendasari munculnya ide terkait pemodelan topik adalah bahwa sebuah topik merupakan turunan dari kata-kata tertentu yang

menyusun topik tersebut, dan beberapa topik tersusun dalam satu dokumen, dimana tiap topik memiliki peluang kemunculan masing-masing dalam dokumen tersebut. Sehingga pemodelan topik bertujuan untuk menemukan suatu kata atau kumpulan kata dari suatu dokumen yang merepresentasikan suatu topik Blei *et al*.(2012) berdasarkan hal ini mendefinisikan pemodelan topik sebagai suatu rangkaian algoritma yang sedemikan sehingga dapat menemukan tema utama dari suatu teks dokumen yang besar dan tidak terstruktur.

Latent Dirichlet Allocation (LDA) termasuk salah satu metode pemodelan topik yang paling banyak digunakan dibandingkan metode pemodelan topik lainnya (Vayansky & Kumar, 2020) (Vayansky & Kumar, 2020). Proses peringkasan, klasterisasi, menghubungkan ataupun pemrosesan data yang sangat besar dapat dilakukan oleh metode LDA. Hal ini dikarenakan algoritma LDA memberikan output daftar topik yang diberi bobot untuk masing-masing dokumen (Campbell et al., 2015). Pendekatan ini menyusun data menjadi tiga tingkatan : kata, topik, dan dokumen. LDA mengasumsikan dokumen dihasilkan dari campuran acak topik tersembunyi, yang dilihat sebagai distribusi probabilitas atas kata-kata.

Pemahaman mendasar yang perlu disepakati dalam proses LDA yaitu terkait definisi dan parameter dalam LDA. Secara formal LDA mendefinisikan istilah-istilah sebagai berikut :

- Kata adalah bagian terkecil dari kosakata, unit terkecil dari kata yang bersifat diskrit, yang didefinisikan sebagai bagian dari kosakata yang diindeks oleh {1, ...,...,..., *V*}.
- Topik merupakan kumpulan kata atau istilah yang menjadi inti utama pembahasan dalam suatu dokumen. Banyaknya topik dilambangkan dengan *k*.
- Dokumen adalah serangkaian dari N buah kata terurut yang dilambangkan dengan $w = (w_1, w_2, ..., w_N)$, di mana w_N adalah kata ke-n dalam urutan tersebut.
- Korpus adalah kumpulan dari M dokumen yang dilambangkan dengan D = $\{w_1, w_2, ..., w_M\}$

Dalam menentukan topik dalam setiap dokumen, sample pertama dari sebuah vector-k θ merepresentasikan campuran dari *k* buah topik dari sebaran prior dirichlet $p(\theta|\alpha)$. Variabel *k* tidak hanya akan menentukan dimensi dari distribusi ini dan juga dimensi dari variabel topik *z*, tetapi juga mewakili jumlah total topik yang akan dikembalikan dalam model. Selain itu, matriks β dengan dimensi $k \times v$ memparameterisasi probabilitas kata sedemikian rupa sehingga $\beta_{ij} = p\ (w^j=1 \mid z^i=1)$ dimana $i = 0, 1, ..., K$ dan $j = 0, 1, ... ,,V$. Untuk nilai $\theta_i \geq 0$ dan $\sum_{i=1}^{k} \theta_i = 1$, peubah acak dirichlet θ berdimensi-k terletak pada (k-1) – simpleks, maka fkp dari simpleks ini akan menyeber menurut :

$$P(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma\left(\alpha_i\right)} \theta_1^{\alpha_1 - 1} \ldots \ldots \theta_k^{\alpha_k - 1}$$

D. M. Blei et al., (2003) memvisualkan penggunaan model LDA sebagai model peluang. Representasi LDA terbagi dalam tiga tingkatan, yaitu tingkat corpus, dokumen dan kata. Peubah α dan β merupakan peubah pada level corpus (kumpulan dari *M* dokumen). Parameter α menyatakan distribusi topik dalam kumpulan dokumen. Sementara β parameter penentu sebaran kata dalam topik. Variabel θ_m melambangkan peubah pada tingkatan dokumen *M* (kumpulan dari *N* kata), sampel diambil satu kali per dokumen. Peubah θ merupakan peubah yang menjelaskan distribusi topik dalam dokumen tertentu. Peubah Z_n dan W_n merupak peubah pada tingkatan kata. Variabel *Z* menggambarkan topik dari kata tertentu pada sebuah dokumen dan variabel *W* merepresentasikan kata yang berkaitan dengan topik tertentu yang terdapat dalam dokumen (D. M. Blei *et al*., 2003).

Pendugaan Parameter : *Gibbs Sampler* LDA

Gibbs sampler pertama kali diperkenalkan ketika diaplikasikan dalam menganalisa sebaran kisi gibbs. Gibbs sampler banyak diterapkan untuk sebaran multivariate. Untuk menurunkan Gibbs sampler pada LDA, kita terapkan metode pencarian peubah tersembunyi. Variabel tersembunyi dalam model kita adalah $Z_{m,n}$, topik yang muncul dengan kata-kata pada korpus $W_{m,n}$.

Tujuan utama dari inferensi ini adalah mencari fkp $p(\vec{z}|\vec{w})$, dalam hal ini dapat didekati dengan fungsi kepekatan peluang bersyarat :

$$p(\vec{z}|\vec{w}) = \frac{p(\vec{z},\vec{w})}{p(\vec{w})} = \frac{\prod_{i=1}^{W} p(z_i, w_i)}{\prod_{i=1}^{W} \sum_{k=1}^{K} p(z_i = k, w_i)}$$

Setelah dilakukan serangkaian proses dengan prinsip *marcov chain monte carlo* (MCMC), didapat persamaan untuk menentukkan serangkaian topik seperti tampak pada persamaan 3 dan 4. Dimana $\vec{n}_m$ menyatakan vector dari banyaknya topik yang terobservasi pada dokumen *m*, sedang $\vec{n}_k$, banyaknya term yang terobservasi pada topik *k*. Fungsi pada persamaan 2 menyebar menurut sebaran *dirichlet.* Fungsi ini mempunyai informasi bahwa nilai harapan sebaran diri chlet, $\langle \mathrm{Dir}\,(\vec{a})\rangle = \frac{a_i}{\sum_i a_i}$ maka :

$$\varphi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^{V} n_k^{(t)} + \beta_t}$$

$$\vartheta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_m^{(k)} + \alpha_k}$$

Visualisasi dan Intrepretasi Pemodelan Topik LDA

Beberapa hal mendasar yang perlu dielaborasi setelah didapatkan serangakaian topik dan kata dalam topik tersebut yaitu pemaknaan dari setiap topik yang terbentuk, kelaziman setiap topik yang terbentuk, hubungan antar topik yang terbentuk satu sama lain serta hubungan antar dokumen dalam corpus yang terbentuk (Mabey, 2015). Beberapa penelitian sebelumnya telah memberikan beberapa alat ukur untuk menjawab hal-hal ini. Salah dua diantaranya ukuran *saliency* dan *relevance*.

Saliency

Ukuran *saliency* diklaim dapat memberikan urutan kata teratas yang bersifat koheren yang mendukung suatu topik. Didapati urutan kata dengan ukuran saliency ini dapat diperoleh lebih cepat dan koheren dibanding metode urutan kata berdasarkan frekuensi kemunculan maupun urutan kata berdasarkan ukuran distinctiveness (Chuang *et al*., 2012). Secara formulasi matematis ukuran *saliency* ini merupakan pengembangan dari pengurutan kata berdasarkan ukuran *distinctiveness*. Persamaan ukuran saliency dirumuskan sebagai berikut:

$$saliency(w) = P(w) \times distinctiveness(w)$$

$$distinctiveness(w) = \sum_{T} P(T|w) \ln \left(\frac{P(T|w)}{P(T)}\right)$$

Relevance

Metode untuk menentukan pilihan kata/istilah apa saja yang akan diberikan kepada pengguna/narsumber sehingga memudahkan dalam interpretasi topik (Sievert & Shirley, 2014). Alat ukur ini dapat membantu kebimbangan ketika terjadi fenomena bahwa seringkali kata yang muncul dalam rangkaian teratas merupakan kata-kata atau istilah yang berisifat umum, sehingga kata/istilah yang bersifat umum tidak lagi berada dalam rangking teratas. Hal ini akan berdampak mempermudah dalam upaya intrepretasi topik.

$$r(w, k \mid \lambda) = \lambda \log(\phi_{kw}) + (1 - \lambda) \log\left(\frac{\phi_{kw}}{p_w}\right)$$

dimana *r*, menyatakan nilai relevansi, ϕ_{kw} peluang munculnya kata ke '*w*' dalam topik ke '*k*'. p_w, peluang munculnya kata ke –*w* dalam keseluruhan corpus. λ , merupakan parameter bobot berkisar antara 0 dan 1.

C. METODE PENELITIAN

Data

Metadata khususnya bagian judul dan abstrak suatu dokumen paten berbahasa Indonesia yang terdaftar pada Dirjen Kekayaan Intelektual (DJKI-Kemenkumham) digunakan sebauh peubah yang akan dielaborasi lebih lanjut. Dengan keyword "pupuk", metadata Paten yang diproses pada periode 1945-2017. Data dapat diunduh pada alamat : https://pdki-indonesia.dgip.go.id/. Didapat sebanyak 762 dokumen paten dengan filter seperti tersebut.

Prosedur Analisis Data

Beberapa tahapan analisis data pada penelitian ini dapat disajikan sebagai berikut:

1. Preprosesing data teks

   Pada dasarnya sebuah dokumen berisi kumpulan data-data yang tidak terstruktur, sehingga diperlukan suatu preprosesing data. Data yang berbentuk teks ini supaya menjadi terstruktur agar dapat dilakukan analisa lanjutan. Pada prinsipnya proses preprosessing teks dibagi menjadi beberapa tahap, yaitu :

   a. Case folding

   Prose ini bertujuan merubah semua huruf capital menjadi non kapital. Hanya huruf 'a' sampai 'z' yang diterima. Karakter selain huruf dihilangkan. Terdapat beberapa cara yang dapat digunakan dalam tahap case folding, tergantung konteks dan kebutuhan analisa nantinya. Beberapa cara tersebut diantaranya, mengubah teks menjadi huruf kecil semua, menghapus karakter angka (tergantung kebutuhan), menghapus tanda baca dan menghapus white space (karakter kosong).

   b. Tokenizing

   Tahapan ini melakukan proses memotong-motong kata dalam suatu kalimat atau paragraf. Potongan-potongan kata ini selanjutnya disebut sebagai token, yang akan dianalisa lanjut

   c. Filtering (Stopword Removal)

   Tahap mengambil kata-kata penting dari hasil token dengan menggunakan algoritma *stoplist* (membuang kata non informatif) atau *wordlist* (memasukan kata penting). Dalam tahapan ini bisa jadi kata-kata penting tersebut lebih dari satu kata. Dan untuk melihat kata atau kumpulan kata mana yang dianggap penting menerapkan prosedur *N-gram.*

   d. Stemming

   Stemming adalah proses membuang imbuhan pada suatu kata (awalan dan akhiran kata), namun bentuk dasar tersebut tidak berarti sama dengan akar kata (*root word*). Stemming merupakan proses yang unik, berbeda bahasa maka kemungkinan besar proses *stemming* yang dilakukan juga berbeda.

   Terdapat beberapa algoritma yang telah dikembangkan untuk melakukan preposesing teks berbahasa Indonesia khususnya *stemming*. Algoritma tersebut diantaranya Nazief and Adriani Algorithm, Algoritma Arifin dan Setiani, Algoritma Vega dan Ahmad, Yusuf dan Sembak. Adriani et al., (2005) menyimpulkan bahwa Algoritma Nazief dan Adriani lebih baik dibanding tiga algoritma lainnya. Oleh karena itu, tahapan preprosesing dalam penelitian ini menggunakan algoritma Nazief dan Adriani (Asian et al., 2005).

2. Pemodelan Topik : *Latent Dirichlet Allocation* (LDA)

   Sebelum dilakukan pemodelan topik dengan LDA, terlebih dahulu dilakukan pendugaan parameter-parameter pada distribusi LDA. Proses pendugaan parameter distribusi LDA dilakukan dengan teknik *Gibbs Sampler*, pada set data training.

3. Kebaikan Pemodelan Topik : Ukuran Koherensi Topik

   Sekumpulan fakta yang koheren dapat diinterpretasikan dalam konteks yang mencakup semua atau sebagian besar fakta. Menurut Röder et al. (2015), terdapat empat dimensi pengukuran tingkat koherensi yang bersifat saling bebas sehingga dapat digabungkan. Koherensi sekumpulan kata mengukur keterikatan dan kesesuaian kata-kata tunggal atau subset dari kata-kata itu. Jadi, dimensi pertama adalah jenis segmentasi yang digunakan untuk membagi kumpulan kata menjadi bagian-bagian yang lebih kecil. Potongan-potongan ini dibandingkan satu sama lain, misalnya, segmentasi menjadi pasangan kata. Himpunan berbagai jenis segmentasi adalah *S*.

Dimensi kedua adalah ukuran konfirmasi yang menilai kesepakatan pasangan tertentu, misalnya, NPMI dari dua kata. Himpunan ukuran konfirmasi adalah *M*. Himpunan metode untuk memperkirakan probabilitas kata adalah *P*, yang membentuk dimensi ketiga dari ruang konfigurasi. Terakhir, metode bagaimana menggabungkan nilai skalar yang dihitung oleh

ukuran konfirmasi membentuk dimensi keempat. Himpunan fungsi agregasi adalah Σ. Secara sederhana ukuran koherensi merupakan produk silang dari C = S ×M ×P ×∑ (Röder et al., 2015).

4. Penghitungan Tingkat Keunikan Kata dalam Topik

Pada tahap 3, akan kita dapatkan nilai koherensi pada berbagai skema jumlah topik. Seringkali terjadi penambahan skema jumlah topik tidak menyebabkan keefektifan dalam intrepretasi. Kejadian yang sering terjadi ketika semakin banyak skema jumlah topik, maka semakin sering beberapa istilah/kata muncul pada beberapa topik. Hal ini menimbulkan ambiguitas dalam interpretasi. Fakta yang mendasari hal ini adalah ketika suatu kata muncul diberbagai topik maka dapat dikatakan penambahan jumlah topik sudah tidak efektif. Dengan kata lain semakin besar tingkat keunikan suatu kata/frase (*term*) dalam suatu skema jumlah topik maka semakin baik skema tersebut. Penelitian ini memperkenalkan parameter sederhana tingkat keunikan kata/istilah sebagai parameter evaluasi lanjutan setelah penentuan tingkat koherensi. Formulasi sederahana tingkat keunikan dapat dilihat seperti pada persamaan 8.

$$\%\ keunikan\ kata = \frac{jumlah\ kata\ unik}{total\ jumlah\ kata}\ x\ 100\%$$

Secara mendetail diagram alir analisis data yang dilakukan pada penelitian ini dapat dijelaskan seperti berikut. Tahap pertama dilakukan *crawling* metadata dokumen paten dengan filter berupa kata kunci "pupuk", status paten "terdaftar" dan "diberi", periode paten dari tahun 1945 - 2017. Metadata yang dianalisis lebih lanjut bagian judul dan abstrak dokumen paten. Tahap ini didapat sebanyak 762 dokumen dengan filter tersebut.

Tahapan kedua dilakukan preprosesing data teks terkhusus bagian judul dan abstrak metadata dokumen paten. Pada tahapan ini akan didapatakn judul paten terfilter (teks judul yang telah mengalami proses *case folding, tokenizing, filtering,* dan *stemming*). Secara paralel juga dikenai proses yang sama pada bagian abstrak paten sehingga didapat serangkaian abstrak terfilter. Pada tahapan ini akan dihasilkan matriks berupa banyaknya kemunculan kata atau istilah yang muncul pada tiap dokumen paten, yang kemudian output ini dinamai dengan istilah *bag of word* (BoW).

Tahapan berikutnya, ditahap ketiga ini output BoW yang didapat pada tahapan sebelumnya dikenai proses pemodelan berbasis LDA. Sebelum dilakukan pemodelan LDA, dilakukan pendugaan parameter dengan teknik *Gibbs Sampler*. Setalah didapat dugaan parameter LDA, dilakukan pemodelan LDA sehingga didapat kumpulan kata untuk setiapa topik yang diwakili. Pemodelan LDA ini dilakukan secara parallel baik untuk *corpus* judul paten maupun *corpus* abstrak paten.

Tahapan selanjutnya, dilakukan perbandingan kumpulan kata pada berbagai skema jumlah topik antara pemodelan LDA dengan data berupa *corpus* bagian judul dan pemodelan LDA dengan data berupa *corpus* bagian abstrak paten. Pada tahap ini, nilai koherensi dijadikan dasar dalam memperbandingkan. Semakin tinggi nilai koherensi semakin baik pemodelan LDA yang dilakukan. Tahapan ini akan menentukan bagian *corpus* mana yang terbaik dalam pemodelan LDA. Apakah bagian judul ataukah bagian abstrak dokumen paten?

Tahapan kelima akan dilakukan evaluasi untuk menentukkan skema banyaknya jumlah topik yang optimal dalam pemodelan LDA. Pada bagian pertama ditahapan ini akan dilakukan identifikasi jumlah topik yang menghasilkan nilai optimal pada parameter evaluasi koherensi. Sebagai kesempurnaan evaluasi, ditambahkan parameter tingkat keunikan kata sebagai evaluasi lanjutan. Pada tahap akhir ini akan dihasilkan kumpulan topik beserta kata-kata yang mewakilinya.

D. HASIL DAN PEMBAHASAN

Statistika Deskriptif Paten Pupuk di Indonesia

Gambar 1. memperlihatkan sebaran negara-negara pemilik paten di Indonesia. Terlihat bahwa perbandingan sebaran negara pemilik paten terkait pupuk dengan sebaran negara pemilik paten secara umum memiliki perbedaan sebaran kepemilikan. Apabila kita melihat dari sisi paten secara umum, aktor utama pemegang paten di Indonesia bukanlah negara kita sendiri melainkan Jepang dan Amerika. Temuan ini mendukung laporan WEF bahwa salah satu penyebab rendahnya daya saing global Indonesia adalah rendahnya tingkat inovasi bangsa (salah satunya paten). Fenomena yang berbeda terjadi untuk paten terkait pupuk. Paten-paten (teknologi) terkait pupuk lebih dari 64% masih dikuasai bangsa Indonesia. Hal ini menjadi indikasi bahwa bangsa Indonesia masih memiliki keunggulan dari sisi teknologi pertanian (terkhusus teknologi pupuk) dibanding teknologi pupuk dari luar/asing.

Perkembangan spesifik teknologi paten pupuk di Indonesia dapat dilihat dalam Gambar 2. Secara deskriptif tampak bahwa dari tahun 1992-2017 terjadi ketimpangan perkembangan teknologi (seksi IPC). Dapat kita analisis, paten-paten pupuk hanya berkembang di seksi teknologi (seksi IPC) "A" dan "C". Seksi IPC "A" diartikan sebagai teknologi-teknologi yang secara garis besar berkaitan dengan kebutuhan manusia (*Human Necessities*). Sedangkan seksi IPC "C", diartikan sebagai teknologi-teknologi yang secara umum berkaitan dengan bidang kimia dan metalurgi. Seksi IPC "A" dan "C" mendominasi dibandingkan seksi IPC lainnya. Minimnya jumlah seksi IPC lainnya mengindikasikan kemungkinan kecil terjadinya radikal invensi terkait teknologi pupuk terhadap paten yang didaftarkan.

Penelitian ini berusaha memberikan representasi (pemodelan topik) dari serangkaian kata dan paragraf dalam peubah penjelas. Sebelum melangkah ke tahap pemodelan topik, dilakukan gambaran secara deksriptif dari kumpulan kata, baik yang terdapat dalam *corpus* bagian judul kumpulan dokumen paten maupun bagian abstrak dokumen paten. Penggambaran deskriptif sebaran kata dalam penelitian ini menggunakan teknik *wordcloud*. Gambar 3. sebelah kiri menampilkan deskriptif sebaran kumpulan kata yang terdapat pada bagian judul *corpus* dokumen paten. Gambar 3. Sebelah kanan, memperlihatkan sebaran kata yang terdapat dalam bagian abstrak suatu *corpus* paten.

Wordcloud memberikan gambaran terkait frekuensi kemunculan suatu kata dalam suatu *corpus*. Semakin sering suatu kata muncul, maka digambarkan dengan ukuran huruf yang semakin besar untuk kata tersebut. Berdasarkan hal ini, pendeskripsian Gambar 3. bagian kiri dapat dimaknai bahwa sebagian besar dokumen (terkhusus bagian judul paten) membahas mengenai keunggulan atau temuan teknologi pupuk yang berkaitan dengan komposisi, metode pembuatan, proses, pupuk organik hayati, dan atau formulasinya.

Gambar 3. sebelah kanan memberikan gambaran deskriptif terkait sebaran frekuensi kemunculan kata di bagian abstrak suatu *corpus* paten. Gambar tersebut memberikan informasi bahwa sebagian besar dokumen paten (bagian abstrak) lebih banyak memberikan klaim temuan teknologi pupuk berkisar mengenai dampak/hasil terhadap tanaman, proses campur pembuatan pupuk, pupuk organik (kompos dan lain-lain), keunggulan kandungan, keunggulan bentuk pupuk (cair, padat atau bentuk lainnya). Penggambaran dengan teknik *wordcloud* ini dilanjutkan secara inferensi dengan menggunaan metode pemodelan topik pendekatan LDA. Diharapkan dengan pendekatan LDA ini didapat topik-topik yang mungkin terdapat dalam tiap dokumen paten yang dianalisis.

Pemodelan Topik (Pendekatan LDA)

Gambar 4. menunjukkan perbandingan berbagai nilai koherensi pada berbagai skema banyaknya topik. Bagian sebelah kiri pada Gambar 4. menunjukkan nilai-nilai tingkat koherensi yang didapat pada *corpus* judul paten, yang disandingkan dengan skema jumlah topik. Sedangkan, bagian kiri pada gambar tersebut memeprlihatkan hubunngan antara skema jumlah topik terhadap nilai koherensi yang

didapat berdasarkan data teks pada bagian *corpus* abstrak dokumen paten. Apabila kita perbandingkan antara kedua corpus tersebut, terlihat bahwa pada seluruh skema jumlah topik, nilai koherensi corpus pada bagian Judul dokumen paten lebih tinggi (lebih baik) dibandingkan dengan corpus pada bagian abstrak paten. Atas dasar hal ini dapat kita pahami bersama bahwa untuk analisa lanjutan (multi-label klasifikasi) akan lebih baik dilakukan pada corpus bagian judul suatu dokumen paten. Setelah didapati corpus yang akan dianalisa lanjutan adalah corpus pada bagian judul dokumen paten. Langkah selanjut akan ditentukan banyaknya topik yang akan dibangun dalam pemodelan LDA. Dapat dilihat bahwa pada dasarnya model LDA akan semakin baik ketika nilai koherensi topik yang dihasilkan juga tinggi.

Terlihat secara grafis pada Gambar 4. sebelah kiri, topik dengan jumlah 9 dan 18 memberikan perbedaan yang signifikan nilai koherensinya dibandingkan jumlah topik lainnya. Hal ini mendasari untuk dilakukan perbandingan antara jumlah topik sebanyak 9 ataukah 18 yang layak dijadikan dasar pemodelan LDA pada dokumen paten, terkhusus bagian judul.

Penentuan jumlah topik efektif dapat dilakukan dengan melihat tingkat keunikan kata untuk setiap skema jumlah topik. Fakta yang mendasari hal ini adalah ketika suatu kata muncul diberbagai topik maka dapat dikatakan penambahan jumlah topik sudah tidak efektif. Dengan kata lain semakin besar tingkat keunikan suatu term dalam suatu skema jumlah topik maka semakin baik skema tersebut. Sebagai gambaran, misalkan kata "organic" muncul pada topik 1, topik 4, topik 7, dan topik 8 pada skema jumlah topik sebanyak 18. Sementara itu pada skema jumlah topik sejumlah Sembilan, kata "organic" hanya muncul pada topik 1. Maka, secara sederhana dapat kita katakan topik 1 pada skema jumlah topik Sembilan bertemakan pupuk organik. Akan tetapi kita tidak dapat mengklaim tema ini untuk skema jumlah topik sebanyak 18, karena kata ini dimiliki lebih dari satu topik. Sehingga penentuan persentase keunikan kata penting untuk dilakukan dan dipertimbangkan sebagai parameter evaluasi lanjutan setelah evaluasi tingkat koherensi.

Tabel 1, memperlihatkan tingkat keunikan kata untuk skema sembilan topik dan 18 topik. Terlihat skema sembilan topik lebih baik daripada skema 18 topik. Oleh karena itu untuk analisa lanjutan akan diterapkan pada skema sembilan topik. Berbeda dengan penelitian Yun & Geum (2020) pada tahap evaluasi topik, evaluasi jumlah topik hanya didasarkan pada skema jumlah topik yang akan menghasilkan akurasi terbaik pada tahapan berikutnya yaitu tahapan klasifikasi. Dampak yang terjadi pada penelitian tersebut aspek interpretabilitas dikesampingkan. Padahal klaim penelitian tersebut menghasilkan peubah penjelas yang mudah ditafsirkan. Sehingga penambahan tahapan evaluasi dengan mempertimbangkan tingkat keunikan kata dalam topik penting adanya untuk dilakukan. Terkhusus pemodelan topik pada dokumen paten, penelitian yang dilakukan Momeni & Rost (2016), dalam menginterpretasikan pemodelan topik yang dihasilkan dengan melibatkan banyak ekspert terkait untuk memberikan informasi dan gambaran topik yang dihasilkan. Hal ini pada dasarnya baik dilakukan, akan tetapi di era big data saat ini proses ini cukup memakan waktu dan kurang efisiens. Sehingga penambahan parameter keunikan kata dapat lebih mempercepat dan mengefisienkan waktu dalam interpretasinya.

Tabel 2. memberikan informasi terkait sepuluh kata yang mewakili masing-masing topik. Hal ini disajikan untuk memudahkan dalam intrepretasi (Sievert & Shirley, 2014). Sebagai bentuk gambaran analisa, topik satu berkaitan dengan kata "organik", "proses", "cair", "buat", "hayati", "fosfat", "kelapa sawit", "produksi", "asam", dan "Fosfor". Kesepuluh kata terurut berdasarkan nilai *saliency* tertinggi ke terendah ini apabila kita elaborasi lebih lanjut menggambarkan topik mengenai pupuk organik cair. Berpedoman dengan cara analisa yang sama, sembilan topik lainnya dapat dielaborasi terkait pemaknaannya, sebagaimana terdapat dalam Tabel 2.

E. KESIMPULAN

Tren Perkembangan teknologi pupuk berkembang ke arah teknologi yang berkaitan dengan kebutuhan manusia, proses kimiawi dan metalurgi. Minimnya pengembangan teknologi di area seksi teknologi selain A dan C, berdampak pada munculnya inovasi radikal menjadi kecil. Pendekatan LDA dengan baik dapat memberikan informasi terkait topik yang relevan dalam pengembangan teknologi dalam bidang pupuk. Berdasarkan pendekatan ini pada dasarnya terdapat sembilan topik esensial dalam pengemabngan teknologi pupuk. Implikasi teoritis yang dapat dilihat salah satunya tingkat koherensi judul pada dokumen paten jauh lebih baik dalam memberikan gambaran dalam pemodelan LDA dibanding koherensi pada bagian abstrak dokumen paten. Selain itu didapati bahwa memasukan parameter evaluasi tingkat keunikan kata penting untuk dilakukan. Hal ini untuk memperkecil tingkat ambiguitas interpretasi topik.

UCAPAN TERIMAKASIH

Ucapan terima kasih disampaikan kepada Program Beasiswa Saintek Kementerian Riset dan Teknologi / Badan Riset dan Inovasi Nasional atas segala fasilitas dan kesempatan yang diberikan.

DAFTAR PUSTAKA

Adriani, M., Asian, J., Nazief, B., Williams, H. E., & Tahaghoghi, S. M. M. (2005). Stemming Indonesian : A Confix-Stripping Approach. *Conferences in Research and Practice in Information Technology Series*, *38*(4), 307–314. https://doi.org/10.1145/1316457.1316459

Asian, J., Williams, H. E., & Tahaghoghi, S. M. M. (2005). Stemming Indonesian. *Conferences in Research and Practice in Information Technology Series*, *38*(January), 307–314. https://doi.org/10.1145/1316457.1316459

Blei, D., Carin, L., & Dunson, D. (2012). Probabilistic topic models. *Communications of the Acm*, *27*(6), 55–65. https://doi.org/10.1109/MSP.2010.938079

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *The Art and Science of Analyzing Software Data*, *3*, 139–159. https://doi.org/10.1016/B978-0-12-411519-4.00006-9

Campbell, J. C., Hindle, A., & Stroulia, E. (2015). Latent Dirichlet Allocation: Extracting topics from software engineering data. *The Art and Science of Analyzing Software Data*, 139–159. https://doi.org/10.1016/B978-0-12-411519-4.00006-9

Chuang, J., Manning, C. D., & Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models. *Proceedings of the Workshop on Advanced Visual Interfaces AVI*, 74–77. https://doi.org/10.1145/2254556.2254572

FAO. (2016). *Agricultural Cost of Production Statistics :Guidelines for Data Collection, Compilation and Dissemination* (FAO (ed.)). Food and Agriculture Organization of the United Nations.

Hongshu, C., Guangquan, Z., Donghua, Z., & Jie, L. (2017). Topic-based technological forecasting based on patent data: A case study of Australian patents from 2000 to 2014. *Technological Forecasting and Social Change*, *119*, 39–52. https://doi.org/10.1016/j.techfore.2017.03.009

Hu, J., Li, S., Hu, J., & Yang, G. (2018). A hierarchical feature extraction model for multi-label mechanical patent classification. *Sustainability (Switzerland)*, *10*(1), 219. https://doi.org/10.3390/su10010219

Kim, G., & Bae, J. (2017). A novel approach to forecast promising technology through patent analysis. *Technological Forecasting and Social Change*, *117*, 228–237. https://doi.org/10.1016/j.techfore.2016.11.023

Liang, C., Weijiao, S., Guancan, Y., Jing, Z., & Xiaoping, L. (2016). A topic model integrating patent classification information for patent analysis. *Wuhan Daxue Xuebao (Xinxi Kexue Ban)/Geomatics and Information Science*

of Wuhan University, *41*(October), 123–126.

Mabey, B. (2015). Visualizing topic models. In Dato (Ed.), *Data Science Summit and Dato Conference 2015*. Dato, Inc.

Momeni, A., & Rost, K. (2016). Identification and monitoring of possible disruptive technologies by patent-development paths and topic modeling. *Technological Forecasting and Social Change*, *104*, 16–29. https://doi.org/10.1016/j.techfore.2015.12.003

Presiden Republik Indonesia. (2016). *Undang-Undang No 13 Tahun 2016:Paten* (Issue 1).

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, 399–408. https://doi.org/10.1145/2684822.2685324

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. *Workshop on Interactive Language Learning, Visualization, and Interfaces*, 63–70. https://doi.org/10.3115/v1/w14-3110

Suhyeon, K., Haecheong, P., & Junghye, L. (2020). Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis. *Expert Systems with Applications*, *152*. https://doi.org/10.1016/j.eswa.2020.113401

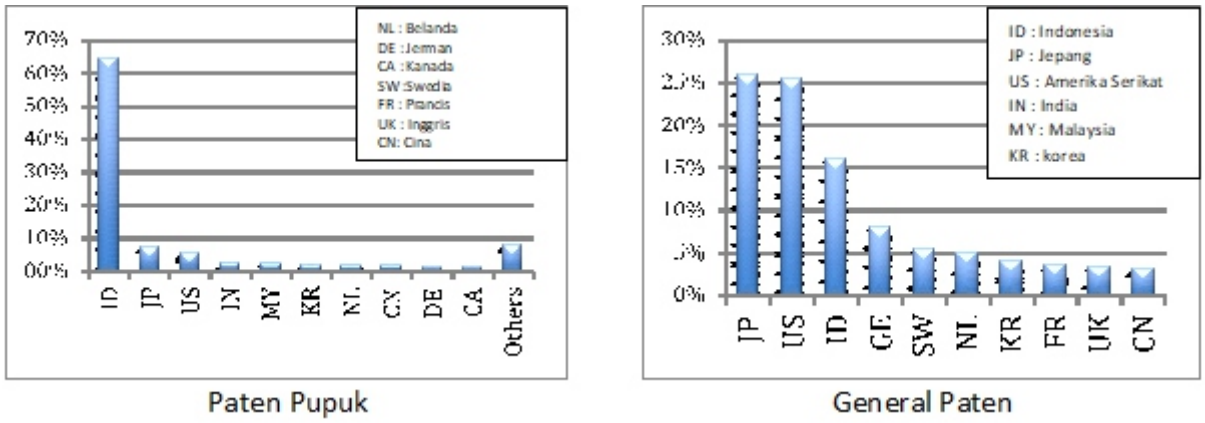
Vayansky, I., & Kumar, S. A. P. (2020). A review of topic modeling methods. *Information Systems*, *94*. https://doi.org/10.1016/j.is.2020.101582

WIPO. (2018). Guide to the International Patent Classification. *WIPO (World Intellectual Property Organization)*. https://www.wipo.int/export/sites/www/classifications/ipc/en/guide/guide_ipc.pdf.

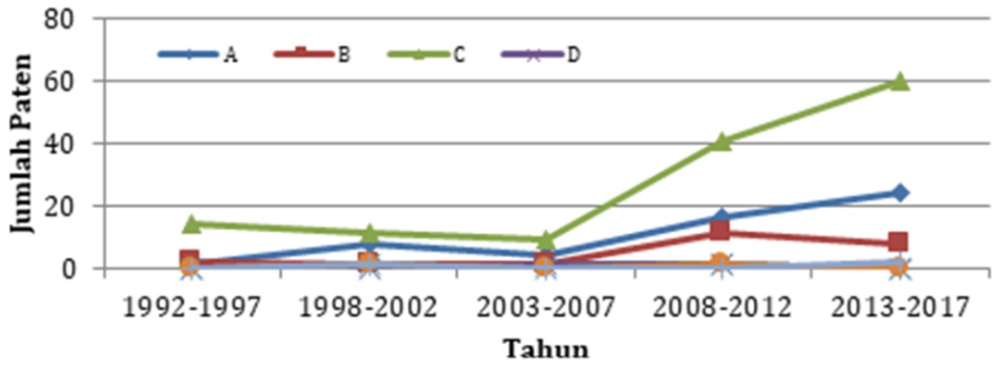
Yu, X., & Zhang, B. (2019). Obtaining advantages from technology revolution: A patent roadmap for competition analysis and strategy planning. *Technological Forecasting and Social Change*, *145*(April), 273–283. https://doi.org/10.1016/j.techfore.2017.10.008

Yun, J., & 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

DAFTAR GAMBAR



Gambar 1. Sebaran Negara Pemegang Paten di Indonesia



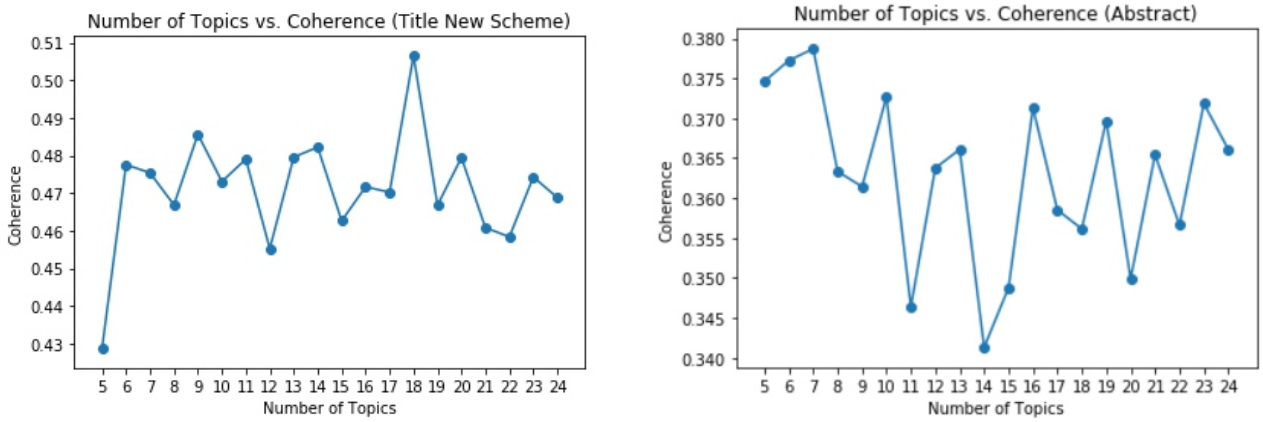
Gambar 2. Tren Perkembangan Paten untuk setiap seksi IPC
Sumber : Data Primer diolah, tahun 2020

Corpus bagian Judul Paten

Corpus bagian Abstrak Paten

Gambar 3. Sebaran Kata dalam *Corpus*
Sumber : Data Primer diolah, tahun 2020

DAFTAR GAMBAR



Gambar 4. Hubungan Jumlah Topik terhadap Nilai Koherensi
Sumber : Data Primer diolah, tahun 2020

DAFTAR TABEL

Tabel 1. Persentase keunikan Istilah/Kata

Skema Topik	∑ kata per topik	∑ kata	∑ Kata Unik	% unik
9 Topik	10	90	70	78%
18 Topik	10	180	133	74%

Tabel 2. Distribusi Topik terhadap Keterwakilan Kata

	term 1	*term 2*	*term 3*	*term 4*	*term 5*	*term 6*	*term 7*	*term 8*	*term 9*	*term 10*	Intrepretasi
Topik 1	organik	proses	cair	buat	hayati	fosfat	kelapa sawit	produksi	asam	Fosfor	Pupuk Cair
Topik 2	metode	bahan	tanah	produksi	hayati	tingkat	mikoriza	terak	baku	Baja	Metode Pembuatan Benah Tanah
Topik 3	padat	formulasi	metode	organik	air	fermentasi	urin	limbah	kombinasi	Biopestisida	Formulasi Pupuk
Topik 4	metode	sistem	laut	limbah	lumpur	ubah	rumput	utama	salur	Pasir	Pupuk dari Limbah
Topik 5	semprot	produk	tumbuh	pupuk pestisida	alat	cair	lengkap	elektrik	baterai	Bentuk	Pupuk dan pestisida alami
Topik 6	kandung	butir	kaya	sulfur	pabrik	metoda	bubuk	komposisi	gulma	Sawit	Pupuk granul
Topik 7	komposisi	urea	budidaya	lap	polimer	kotor	nitrogen	majemuk	arang	Zeolite	Kompoisisi Pupuk Urea
Topik 8	bio	kendali	lepas lambat	granul	blok	bas	buat	lepas	proses	Salut	Pupuk slow release
Topik 9	alat	mesin	benih	transplanter	aplikasi	tahan	senyawa	granular	enzim	Tipe	Mesin Pembuat Pupuk

Sumber : Data Primer diolah, tahun 2020