PENGELOMPOKAN JURNAL ILMIAH BERDASARKAN JUDUL MENGGUNAKAN LDA

[1]Yosefina Oktaviani Santoso, [2]R.Setiawan Aji Nugroho
[1,2]Program Studi Teknik Informatika Fakultas Ilmu Komputer,
Universitas Katolik Soegijapranata
[1]15K10014@student.unika.ac.id, [2]nugroho@unika.ac.id

ABSTRACT

Scientific journals develop very rapidly along with the development of science. Reporting from labs.semanticscholar.org/corpus, the number of scientific journals has reached over 39 million. The large number of scientific journals makes it challenging to grouping scientific journals. Grouping become more difficult because each scientific journal can have more than one topic. Therefore, special methods are needed to group the scientific journals.One of the well-known topic modeling methods is Latent Dirichlet Allocation (LDA). This research is an implementation of the LDA algorithm to do topic modeling in scientific journals. The topic modeling in this study uses the title as a corpus. Various titles are processed into bag of words in the pre-processing process so that they can be used to distribute. The results of the distribution stage are used for sampling with the Gibbs Sampling method. Through the sampling process, testing can also be done to determine the optimal parameters. The testing in this study used perplexity to find the most optimal number of iterations and topics. The result from this research are that LDA Algorithm successfully performs topic modeling in scientific journals by generating a list of keywords for each topic and grouping documents on each topic. The optimal parameters based on the results of perplexity comparison are 3 topics and 500 iterations.

Keywords: Topic Modeling, LDA, Perplexity, Scientific Journal

PENDAHULUAN

Saat ini, jumlah jurnal ilmiah meningkat pesat seiring dengan perkembangan ilmu pengetahuan. Jurnal ilmiah adalah artikel yang berisi pengetahuan dan informasi tentang penemuan yang sesuai dengan aturan dan dipublikasikan dengan baik di media elektronik atau media konvensional. Salah satu negara yang berkontribusi pada jumlah jurnal ilmiah adalah Indonesia. Dikutip dari news.okezone.com yang ditulis oleh Susi Fatimah, menyatakan bahwa jurnal ilmiah internasional di Indonesia mencapai 5.125 per tanggal 6 April 2018. Jumlah ini tidak sebanding jika dibandingkan dengan jumlah jurnal ilmiah dari berbagai belahan dunia. Di web labs.semanticscholar.org/corpus menyatakan bahwa jurnal ilmiah yang ada mencapai lebih dari 39 juta.

Jumlah jurnal yang tumbuh secara eksponensial menciptakan tantangannya sendiri untuk menganalisis topik jurnal ilmiah yang berkembang pesat. Kesulitan juga meningkat karena kurangnya kepastian topik dalam jurnal. Dalam setiap jurnal ilmiah, ada peluang untuk memiliki lebih dari satu topik. Berdasarkan semua challeges, pemodelan topik diperlukan untuk secara otomatis menangkap topik koleksi jurnal.

Pemodelan Topik dapat membantu memahami data dalam jumlah besar dan teks tidak terstruktur. Ada banyak metode yang ada untuk melakukan pemodelan topik. Latent Dirichlet Allocation (LDA) adalah salah satunya. Latent Dirichlet Allocation (LDA) adalah model

probabilistik generatif yang menggunakan kata-kata distribusi acak dan dokumen [1]. Banyak studi pemodelan topik telah dilakukan menggunakan algoritma LDA. Sebagai contoh, Putra & Kusumawardani [2], menganalisis topik media sosial di Surabaya menggunakan Latent Dirichlet Allocation menyatakan bahwa metode LDA mampu menemukan pola tertentu dalam dokumen dan menghasilkan beberapa jenis topik yang berbeda. Ada juga Anupriya & Karpagavalli [3], yang membandingkan dua metode LDA ke jurnal kelompok berdasarkan pada abstrak. Kedua metode tersebut adalah Collapsed Variational Bayes dan Gibbs Sampling. Hasil penelitian menyatakan bahwa kinerja metode sampling gibs lebih efektif.

Berdasarkan kesulitan yang ada dan hasil dari penelitian sebelumnya, penelitian ini menggunakan Algoritma Latent Dirichlet Allocation (LDA) dengan gibbs sampling sebagai metode sampling untuk mengelompokkan jurnal ilmiah berdasarkan judul.Dalam penelitian ini, algoritma LDA berpusat pada distribusi probabilitas korpus dan membagi jurnal ilmiah menjadi beberapa topik. Dalam satu topik, berisi jurnal ilmiah yang serupa dan berkelanjutan. Sedangkan antara topik memiliki tingkat kemiripan yang sangat jauh. Pada akhirnya, sistem ini menghasilkan daftar daftar dokumen untuk setiap topik dan daftar kata-kata yang paling menggambarkan topik. Korpus penelitian ini diperoleh dari judul jurnal ilmiah melalui labs.semanticscholar.org/corpus.

Adapun tujuan dalam penelitian ini meliputi :

(1) menemukan daftar kata kunci setiap topik.

(2) menemukan daftar dokumen pada setiap topik.

(3) mengetahui jumlah topik dan pengulangan untuk menghasilkan peluang optimal.

LANDASAN TEORI

Topic Modeling

Pemodelan topik adalah teknik yang digunakan untuk menganalisis distribusi kata sehingga membentuk kelompok kata dan daftar dokumen pada topik tertentu [4]. Pemodelan topik adalah metode yang digunakan untuk melakukan pengelompokan. Dalam klasifikasinya, pemodelan topik dimasukkan dalam pembelajaran tanpa pengawasan yang berarti tidak memiliki identitas yang pasti, dengan kata lain pada topik pemodelan memungkinkan suatu objek memiliki lebih dari satu identitas. Identitas pada objek disebut sebagai topik. Pemodelan topik telah terbukti menjadi alat yang efektif dan efisien untuk menggantikan manusia dalam memproses sejumlah besar data [5]. Kemampuan pemodelan topik untuk mengelompokkan dan mengatur data berukuran besar membuat pemodelan topik menjadi salah satu cara paling populer untuk memproses data teks. Selain keuntungan dari topik pemodelan adalah kemampuannya untuk menemukan pola tersembunyi dalam dokumen [6]. Kemampuan ini berguna untuk menemukan kelompok kata dalam dokumen sehingga daftar kata kunci dapat dibentuk untuk setiap topik dan daftar dokumen untuk setiap topik. Model Topik

Latent Dirichlet Allocation

Ada berbagai macam metode dalam pemodelan topik. Salah satu metode yang cukup terkenal adalah Latent Dirichlet Allocation. LDA adalah model topik yang paling sederhana [7]. LDA adalah model probabilistik generatif untuk menemukan topik tersembunyi dalam dokumen besar dan salah satu metode pemodelan topik yang paling sederhana [1]. LDA terkait erat dengan dokumen, corpus dan kata-kata. Objek utama dalam algoritma LDA adalah kata-kata yang disimpan dalam kantong kata-kata. Pengumpulan kata-kata juga disebut corpus. Setiap corpus yang ada saling berhubungan dalam sebuah dokumen.

Langkah pertama dalam LDA adalah mendistribusikan nilai ke topik kata dan topik dokumen. "variabel utama yang menarik dalam model adalah distribusi topik-kata" dan distribusi topik θ untuk setiap dokumen" [8]. Dalam proses LDA, beberapa tahapan pengelompokan dilakukan. Pertama, ada pengelompokan kata dengan melakukan distribusi topik di setiap kata. Kumpulan kata-kata yang sudah memiliki topik akan bergabung untuk membentuk korpus pada setiap topik. Kumpulan korpus terkait akan membentuk dokumen. Inilah yang menyebabkan dokumen memiliki lebih dari satu topik. Oleh karena itu, perlu kalkulasi probabilitas untuk menemukan peluang topik tertinggi dalam sebuah dokumen.

Algoritma LDA membutuhkan beberapa parameter seperti alpha yang digunakan dalam distribusi dokumen, betta digunakan dalam distribusi kata, jumlah topik, dan iterasi untuk perhitungan sampling dan *perplexity*. LDA memiliki berbagai jenis metode untuk melakukan pemodelan topik, dalam penelitian ini menggunakan metode Gibbs Sampling. Gibbs Sampling adalah metode terkenal dan banyak digunakan dalam pemodelan penelitian topik [9]. Penggunaan Gibbs sampling dalam penelitian ini adalah karena metode ini mampu menganalisis kata-kata dalam corpus sangat dalam [10]. Alasan lain untuk menggunakan Gibbs Sampling adalah karena sebuah penelitian yang berjudul "Online Inference of Topics with Latent Dirichlet Allocation" yang dilakukan oleh Canini, et al. [11], membandingkan tiga metode LDA. Hasil penelitian menyatakan bahwa Gibbs sampling memiliki kinerja terbaik dan diikuti oleh filter partikel, sedangkan o-LDA berada di posisi terendah. Gibs Samplingg membutuhkan parameter dan mendistribusikan kata dan dokumen ke dalam topik word matrix dan dokumen topik. Pada langkah sampling, proses akan berjalan berulang kali sesuai dengan jumlah iterasi.

Perplexity

Proses terakhir adalah perhitungan kebingungan. Perplexity adalah cara untuk mengevaluasi topik pemodelan yang sering digunakan. Perplexity menggunakan hasil perhitungan model probabilitas untuk melakukan perhitungan lebih lanjut. Menurut Blei, et al. [1], itu mengungkapkan bahwa nilai yang lebih kecil dari kompleksitas yang dihasilkan menunjukkan kinerja LDA yang lebih baik. Perplexity memiliki kemampuan untuk menemukan jumlah topik yang optimal tanpa melakukan pelatihan data. Ini dilakukan dengan perulangan untuk menemukan jumlah topik dengan nilai kebingungan terkecil

METODOLOGI PENELITIAN

Studi Literatur

Langkah pertama dalam penelitian ini adalah studi pustaka. Dalam pelaksanaannya dilakukan dengan mencari jurnal ilmiah, buku , dan berbaga artikel . Dalam langkah ini, material atau referensi yang dicari terkait dengan algoritma LDA dan pemodelan topik. Langkah ini merupakan langkah utama yang sangat penting, karena melalui langkah ini, proses penelitian memiliki landasan implementasi yang jelas.S

Mengumpulkan data

Data atau korpus diperoleh dari labs.semanticscholar.org/corpus, yang merupakan penyedia data jurnal yang dipublikasikan. Namun, dalam penelitian ini hanya menggunakan setengah dari data yang telah disediakan. Data yang digunakan adalah judul jurnal ilmiah yang berbahasa Inggris dan memiliki berbagai topik. Data yang digunakan adalah dalam bentuk daftar judul jurnal ilmiah dan disimpan dalam file CSV, sehingga semua proses yang menggunakan korpus mengarah pada panggilan untuk CSV. Pengambilan Corpus dilakukan pada 17 September 2018.

Pra Pengolahan

Langkah pertama dari penelitian ini adalah pra-pemrosesan. Dalam pra-pemrosesan, proses yang dilakukan menghasilkan Bag of Words. Menurut Sriurai [12], menyatakan bahwa kantong kata adalah hasil dari mengumpulkan kata-kata atau istilah tertentu dan tidak mengetahui arti sebenarnya dari setiap kata. Dalam Bags of Word sendiri pengulangan kata-kata tidak diijinkan.

Kemudian, menurut Schofield, et al. [13], mengatakan bahwa dalam kegiatan pra-pemrosesan, ada beberapa proses seperti cleanning, stopwords dan stemming. Bahkan dalam jurnalnya juga menyatakan bahwa melakukan stopword, dapat meningkatkan kompatibilitas dan kualitas model.

Tokenizing

Langkah pertama dalam pra-pemrosesan adalah tokenizing. Tokenizing adalah proses memotong kalimat atau teks menjadi kata demi kata yang sering disebut token. Setiap kata disimpan menjadi satu yang disebut Bag of Words.

Cleanning

Pada tahap cleanning, digunakan untuk menghapus semua simbol seperti tanda baca atau angka. Melalui proses ini, data yang diperoleh hanya terdiri dari alfabetis. Surat-surat sisanya adalah hasil dari mengubah huruf-huruf dari surat-surat yang ada menjadi huruf kecil.

Stopwords

Tahap ini digunakan untuk menyingkirkan kata sambung atau kata-kata yang sering digunakan tetapi tidak memiliki arti. Meskipun ada beberapa kata atau token yang terbuang dalam tahap ini, itu tidak akan mempengaruhi makna teks atau kalimat secara keseluruhan. Dalam python, salah satu cara melakukan stopwords menggunakan library NLTK (Natural Language ToolKit). Impor stopwords di NLTK untuk mendapatkan daftar kata-kata yang tidak berguna. Penelitian saat ini menggunakan stopword NLTK dengan bahasa Inggris.

Stemming

Tahap stemming adalah langkah untuk mengubah setiap kata menjadi bentuk dasar atau menjadi kata standar. Dalam proses perubahan menjadi kata standar, setiap awalan atau akhiran dihilangkan. Ada berbagai cara untuk melakukan stemming. Salah satu algoritma yang cukup populer untuk stemming dalam bahasa Inggris adalah stemmer porter. Dalam python, stemmer porter disediakan pada NLTK. Algoritma Porter Stemmer sangat bergantung pada konsonan, vokal dan kombinasi vokal-konsonan yang disebut VC. Menghapus akhiran dilakukan jika jumlah vc lebih dari 0.

The Bag of Words yang merupakan hasil dari proses Pra-Pengolahan kemudian akan digunakan untuk menerapkan algoritma Algoritma Dirichlet Laten.

Inisialisasi

Fase Inisialisasi adalah langkah untuk menetapkan nilai ke parameter yang digunakan. Dalam Alokasi Alokasi Dirichlet algoritma, itu membutuhkan pembentukan nilai pada parameter. Beberapa parameter yang perlu diatur adalah alpha, betta, jumlah topik dan jumlah iterasi. Parameter alfa digunakan untuk mendistribusikan topik pada setiap dokumen, sementara parameter beta digunakan untuk mendistribusikan topik dalam setiap kata. Jumlah topik dan iterasi dapat dicari melalui kebingungan.

Pemodelan topik menggunakan LDA

Pada tahap ini algoritma Alokasi Diri Terpadu Laten digunakan untuk melakukan pemodelan topik. Penggunaan LDA itu sendiri karena LDA mampu melakukan pemodelan topik pada data yang sangat jarang. Algoritma ini dapat menghasilkan distribusi kata dan dokumen dengan mencari probabilitas. Metode LDA yang digunakan dalam penelitian ini adalah metode sampling gibbs. Dalam metode Gibbs Sampling, ada 2 rumus. Formula pertama disebut phi. Phi adalah hasil dari distribusi matriks kata-topik. Kemudian, formula kedua adalah theta. Theta adalah hasil dari distribusi topik dokumen. Perkalian kedua rumus ini akan menghasilkan peluang. Ada berbagai cara untuk membantu menghitung peluang. Di Python, perpustakaan membantu menghitung klien adalah numpy. Numpy adalah pustaka untuk mengubah array menjadi matriks. Penggunaan Numpy dapat membantu mempersingkat waktu perhitungan karena perkalian dan pembagian antar matriks lebih mudah daripada perkalian dan pembagian antar array. Selain itu, pada tahap ini proses perhitungan kebingungan juga dilakukan untuk pemeriksaan. Semakin kecil nilainya menunjukkan pemodelan yang lebih baik dari topik.

Derajat Kebingungan

Perplexity adalah proses memanfaatkan peluang distribusi untuk menemukan hasil yang optimal tanpa data pelatihan. Nilai perplexity yang lebih kecil menunjukkan informasi yang lebih tepat dari perhitungan probabilitas. Perplexity dapat digunakan untuk menguji kesesuaian dokumen untuk identitas topik. Perplexity adalah metode umum yang digunakan untuk melihat keakuratan probabilitas model dalam memprediksi sampel dan memproses selama iterasi. Dalam studi ini, perplexity digunakan untuk menemukan jumlah topik dan jumlah iterattions yang optimal.

HASIL DAN PEMBAHASAN

Langkah pertama dalam penelitian ini adalah pra-pemrosesan yang termasuk dalam kategori text-mining. Dalam pelaksanaannya, terdapat 4 tahapan yang terdiri dari *tokennizing*, *cleanning*, *stopword*, dan *stemming.*

Tabel 1. Hasil pra-pemrosesan

Sebelum	Sesudah	Tahapan
Effect of occlusion of the aorta on the coronary and pulmonary circulation	Effect, of, occlusion, aorta, on, coronary, pulmonary, circulation.	Tokenizing
Effect, of, occlusion, aorta, on, coronary, pulmonary, circulation.	effect, of, occlusion, aorta, on, coronary, pulmonary, circulation	Cleanning + lowercase
effect, of, occlusion, aorta, on, coronary, pulmonary, circulation	effect, occlusion, aorta, coronary, pulmonary, circulation	Stopwords
effect, occlusion, aorta, coronary, pulmonary, circulation	effect, occlus, aorta, coronari, pulmonari, circul	Stemming

Setelah melalui tahap pra-pemrosesan, dilanjutkan kepada tahap analisa data. Hal ini untuk memastikan penggunaan algoritma sesuai dengan sifat data. Proses analisa dilakukan dengan membuat tabel tf-idf dan mencari *desity-nya.* Diketahui nilai density dari hasil perhitungan tabel TF-IDF sebesar 0.197346369779%.

Tabel 2. Tabel Term Frequency

No	larg	Renal	angiomyolipoma	digit	subtract
1	1	1	1	1	1
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	0	0	0	0	0

Setelah memastikan persebaran data sesuai dengan penggunaan algoritma, tahap selanjutnya merupakan penyebaran tahap pertama topik dengan kata maupun topik dengan dokumen. Persebaran topik pada setiap kata dipengaruhi oleh parameter betta, sedangkan persebaran topik pada setiap dokumen dipengaruhi oleh parameter alpha.

Tabel 3. Distribusi alpha pada Dokumen-Topik

Dokumen	Topik 0	Topik 1	Topik 3
1	0.5	0.5	0.5
2	0.5	0.5	0.5

Dokumen	Topik 0	Topik 1	Topik 3
3	0.5	0.5	0.5
4	0.5	0.5	0.5
5	0.5	0.5	0.5

Tabel 3 merupakan cotoh hasil pendistribusian awal alpha di setiap topik pada setiap dokumen, sedangkan tabel 4 merupakan contoh hasil pendistribusian awal betta di setiap topik pada setiap kata. Parameter alpha dan betta yang dideklarasi adalah 0.5.

Tabel 4. Distribusi beta pada Word-Topik

Word	Topik 0	Topik 1	Topik 3
1	0.5	0.5	0.5
2	0.5	0.5	0.5
3	0.5	0.5	0.5
4	0.5	0.5	0.5
5	0.5	0.5	0.5

Kemudian, dilakukan perulangan dalam persebaran topik. Perulangan dilakukan sebanyak iteration yang telah ditetapkan. Melalui perulangan ini persebaran topik pada setiap kata dan dokumen menjadi berubah dan terbentuk perbedaan angka antar topik. Topik yang memiliki nilai tertinggi dianggap sebagai topik utama.

Tabel 5. Distribusi ulang alpha pada Dokumen-Topik

Dokumen	Topik 0	Topik 1	Topik 3
1	2.5	**4.5**	3.5
2	0.5	2.5	**3.5**
3	0.5	0.5	**9.5**
...	...	...	...

Melalui Tabel 5 dapat diliat bahwa setiap dokumen memiliki peluang pada tiap topiik, namun setiap topik memiliki peluang yag berbeda. Perbedaan peluang antar topik mengakibatkan setiap dokumen memiliki topik yang unggul. Topik yang memiliki peluang terbesar pada setiap dokumen dianggap sebagai topik dari dokumen tersebut. Hal yang sama juga terjadi pada setiap kata. Persebaran topik pada setiap kata dapat dilihat di tabel 6.

Tabel 6. Distribusi ulang betta pada Word-Topik

Kata	Topik 0	Topik 1	Topik 3
1	0.5	1.5	**3.5**

Kata	Topik 0	Topik 1	Topik 3
2	3.5	**8.5**	2.5
3	0.5	**1.5**	**0.5**
...	...	...	...

Hasil persebaran parameter pada setiap topik digunakan sebagai bahan untuk melakukan proses perhitungan peluang meggunakan metode Gibs Sampling. Metode Gibs Sampling mencari peluang topik tertingi pada setiap kata. Berdasarkan hasil perhitungan peluang, setiap kata dikelompokan dan terbentuk daftar kata.

```
 topic:  0    2951  words
patient  :  0.010418815762327213
studi  :  0.008974623478440273
clinic  :  0.005467299360429131
treatment  :  0.005467299360429131
diseas  :  0.004848359810191871
case  :  0.004848359810191871
effect  :  0.004642046626779452
diagnosi  :  0.003816793893129771
health  :  0.0034041675263049307
manag  :  0.003197854342892511
associ  :  0.0027852279760676706
children  :  0.0027852279760676706
arteri  :  0.0027852279760676706
report  :  0.0027852279760676706
use  :  0.0023726016092428305
review  :  0.0023726016092428305
syndrom  :  0.0023726016092428305
renal  :  0.0023726016092428305
cancer  :  0.0023726016092428305
care  :  0.0023726016092428305
```

Gambar 1. Daftar Kata Kunci Topik ke-0

Pada saat terjadi perulangan, diadakan pula proses pencarian nilai perplexity minimum. Data dengan hasil perplexity minimum dapat dianggap sebagai hasil akhir yang paling optimal. Pada penelitian ini, dilakukan beberapa kali uji coba dengan beberapa kandidat jumlah iterasi. Setiap kandidat iterasi dicari nilai minimumnya dan rata-ratanya, setelah itu dilakukan perbandingan.

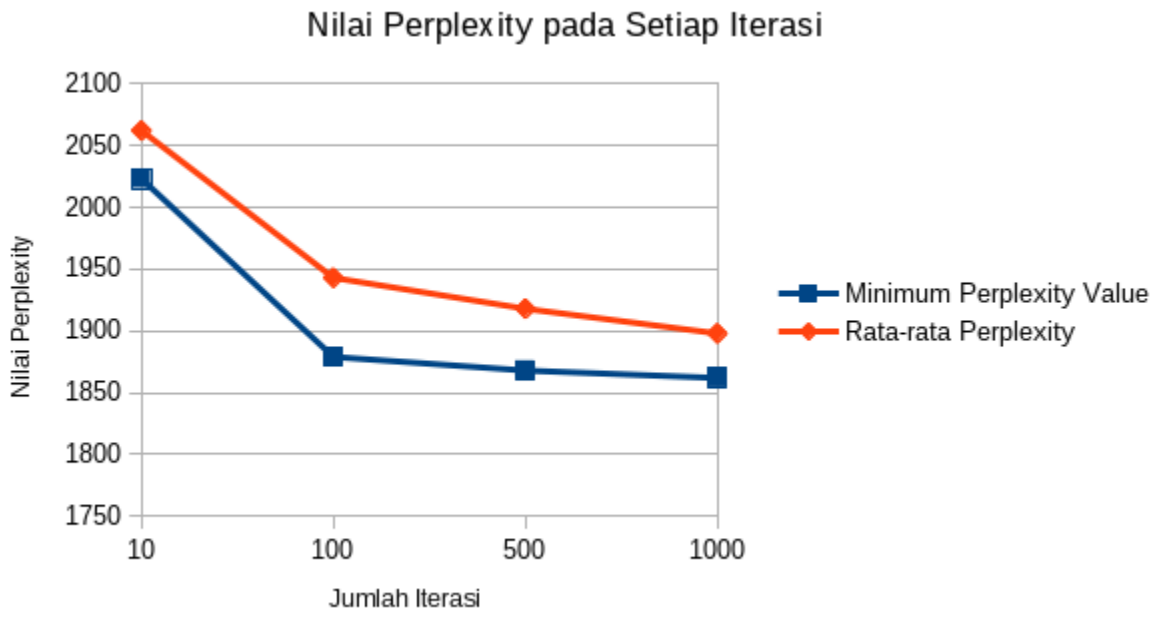
Tabel 7. Hasil Perplexity Setiap Kandidat Iterasi

	Iteration			
	10	***100***	***500***	***1000***
Minimum Perplexity Value	2022.9909411348	1878.65194473	1868.26020952	1862.03909192
Average Perplexity	2061.4654304118	1942.51672212	1918.39376598	1898.24403684

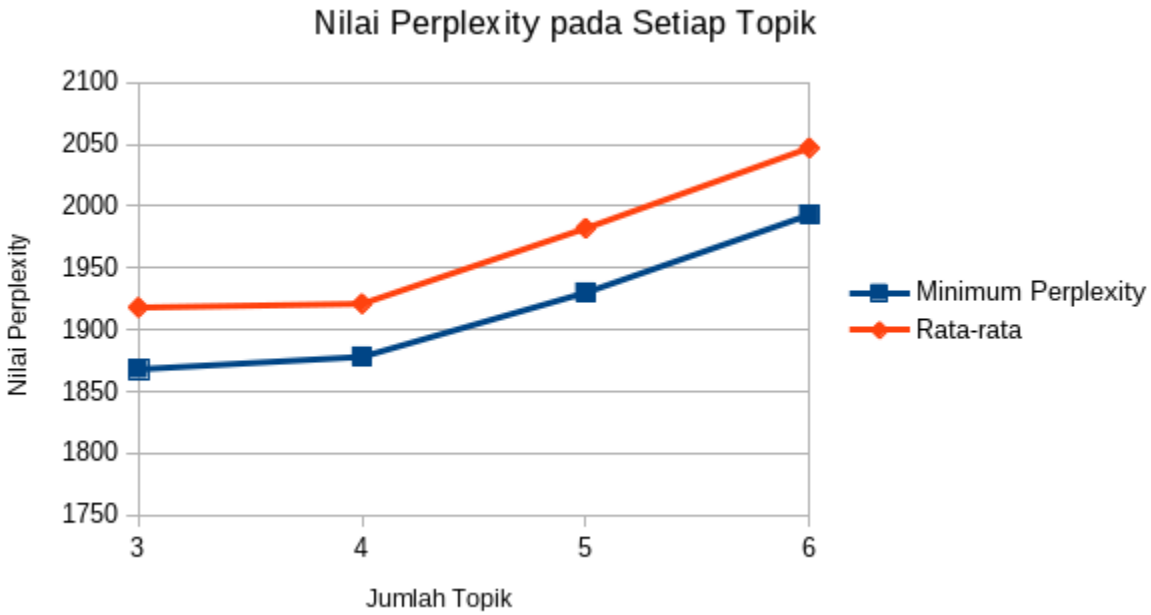
Selain itu, dilakukan pula beberapa kali pengujian menggunakan beberapa kandidat topik. Setiap kandidat topik dicari nilai minimum dan rata ratanya. Nilai perplexity dari setiap kandidat topik dibandingkan dan digunakan sebagai acuan untuk mencari jumlah topik teroptimum.

Tabel 8. Hasil Perplexity Setiap Kandidat Topik

	Topik			
	3	***4***	***5***	***6***
Minimum Perplexity Value	1868.26020952	1878.40587740	1930.21580322	1993.01004882
Average Perplexity	1918.39376598	1921.22378592	1981.74674769	2046.77632076



Gambar 2. Grafik Nilai Perplexity setiap iterasi



Gambar 3. Grafik Nilai Perplexity setiap topik

KESIMPULAN

Melalui penelitian ini, dapat disimpulkan bahwa Alokasi Algoritma Dirichlet Laten dengan metode Gibbs Sampling terbukti mampu melakukan pemodelan topik jurnal ilmiah berdasarkan judul.Penggunaan algoritma LDA adalah karena kebutuhan untuk algoritma khusus untuk menangani jenis data yang terlalu jarang. Data jurnal dinyatakan jarang setelah melewati perhitungan densitas yang mencapai angka 0,197346369779%. Hasil perhitungan densitas diperoleh setelah melalui proses pra-pemrosesan yang mengubah 1047 dokumen menjadi 4016 kata.Hasil kata yang dihasilkan melalui proses pra-pemrosesan diproses ulang untuk didistribusikan ke setiap topik menggunakan alpha dan betta 0,5 sebagai parameter hiper. Ada dua jenis distribusi yang dihasilkan, yaitu distribusi kata dan dokumen topik. Jumlah topik yang digunakan adalah 3 topik. Menentukan jumlah topik dilakukan pada tahap pengujian dengan membandingkan kebingungan.Dalam tahap pengambilan sampel, dilakukan dengan menggunakan metode sampling gibbs. Proses pengambilan sampel dilakukan berulang kali sebanyak iterasi. Iterasi adalah parameter yang ditentukan. Proses penentuan jumlah iterasi juga menggunakan kebingungan dengan membandingkan kebingungan masing-masing kandidat. Berdasarkan hasil perbandingan, iterasi 500 dianggap paling optimal dalam penelitian ini. Saran untuk penelitian lebih lanjut adalah menggunakan sejumlah besar data dokumen sehingga distribusi data semakin jelas.

DAFTAR PUSTAKA

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *J. Mach. Learn. Res.*, vol. 3, no. null, pp. 993–1022, Mar. 2003.

[2] K. B. Putra and R. P. Kusumawardani, “Analisis Topik Informasi Publik Media Sosial di Surabaya Menggunakan Pemodelan Latent Dirichlet Allocation (LDA),” 2017. Accessed: May 15, 2021. [Online]. Available: http://ejurnal.its.ac.id/index.php/teknik/article/view/23205
[3] P. Anupriya and S. Karpagavalli, “LDA based topic modeling of journal abstracts,” in *2015 International Conference on Advanced Computing and Communication Systems*, Jan. 2015, pp. 1–5. doi: 10.1109/ICACCS.2015.7324058.
[4] C. Jacobi, W. van Atteveldt, and K. Welbers, “Quantitative analysis of large amounts of journalistic texts using topic modelling,” *Digital Journalism*, vol. 4, no. 1, pp. 89–106, Jan. 2016, doi: 10.1080/21670811.2015.1093271.
[5] X. Quan, C. Kit, Y. Ge, and S. J. Pan, “Short and sparse text topic modeling via self-aggregation,” in *Proceedings of the 24th International Conference on Artificial Intelligence*, Buenos Aires, Argentina, Jul. 2015, pp. 2270–2276.
[6] R. Alghamdi and K. Alfalqi, “A Survey of Topic Modeling in Text Mining,” *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 1, 2015, doi: 10.14569/IJACSA.2015.060121.
[7] D. M. Blei, “Probabilistic topic models,” *Commun. ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012, doi: 10.1145/2133806.2133826.
[8] T. L. Griffiths and M. Steyvers, “Finding scientific topics,” *PNAS*, vol. 101, no. suppl 1, pp. 5228–5235, Apr. 2004, doi: 10.1073/pnas.0307752101.
[9] P. M. Prihatini, I. K. Suryawan, and I. N. Mandia, “METODE LATENT DIRICHLET ALLOCATION UNTUK EKSTRAKSI TOPIK DOKUMEN,” *Logic : Jurnal Rancang Bangun dan Teknologi*, vol. 17, no. 3, pp. 153–157, 2017, doi: 10.31940/logic.v17i3.604.
[10] R. Y. K. Lau, Y. Xia, and Y. Ye, “A Probabilistic Generative Model for Mining Cybercriminal Networks from Online Social Media,” *IEEE Computational Intelligence Magazine*, vol. 9, no. 1, pp. 31–43, Feb. 2014, doi: 10.1109/MCI.2013.2291689.
[11] K. Canini, L. Shi, and T. Griffiths, “Online Inference of Topics with Latent Dirichlet Allocation,” in *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, Apr. 2009, vol. 5, pp. 65–72. [Online]. Available: http://proceedings.mlr.press/v5/canini09a.html
[12] W. Sriurai, “Improving Text Categorization By Using A Topic Model,” *Advanced Computing : an International Journal*, vol. 2, Dec. 2011, doi: 10.5121/acij.2011.2603.
[13] A. Schofield, M. Magnusson, L. Thompson, and D. Mimno, “Pre-Processing for Latent Dirichlet Allocation,” 2017.