Pemodelan Topik Menggunakan Metode *Latent Dirichlet Allocation* dan *Gibbs Sampling*

Rizki Ramadandi [1], Novi Yusliani [2], Osvari Arsalan [3], Rizki Kurniati [4], Rahmat Fadli Isnanto[5]

[1,2,3,4,5]*Fakultas Ilmu Komputer, Universitas Sriwijaya, Palembang, Indonesia*

rramadandi16@gmail.com, novi_yusliani@unsri.ac.id, osvari.arsalan@unsri.ac.id, [4]rizki.kurniati@unsri.ac.id, rahmatfadliisnanto@ilkom.unsri.ac.id

***Abstrak*— Pemodelan topik adalah suatu alat yangdigunakan untuk menemukan topik laten pada sekelompok dokumen. Pada penelitian ini dilakukan pemodelan topik dengan menggunakan metode *Latent Dirichlet Allocation* dan *Gibbs Sampling*. Enam artikel berita Bahasa Indonesia telah dikumpulkan dari portal berita detiknews dengan menggunakan metode *Web Scrapper*. Artikel berita dibagi menjadi dua kategori utama yaitu, narkoba dan COVID-19. Analisis model LDA dilakukan dengan menggunakan metode koherensi topik pengukuran skor UCI dengan hasil penelitian menyebutkan diperoleh lima buah topik optimal pada kedua konfigurasi pengujian.**

Kata kunci*— Pemodelan Topik, *Latent Dirichlet Allocation*, *Gibbs Sampling*, Koherensi Topik,UCI, *Web Scrapper

I. PENDAHULUAN

Dengan semakin banyaknya teks tidak terstruktur di internet, diperlukan sebuah alat yang dapat membantu untuk menemukan topik tersembunyi yang ada pada sekumpulan teks. Untuk menyelesaikan masalah tersebut, maka para peneliti telah mengembangkan suatu teknik yang dinamakan pemodelan topik [20]. Pemodelan topik merupakan bidang penelitian Pemrosesan Bahasa Alami (PBA) yang digunakan dalam berbagai macam bidang, seperti *medical sciences*, *software engineering, geography, political science,* dan lain-lain. Dalam pemodelan topik, teks yang digunakan dapat berupa *e-mail*, jurnal, artikel berita, dan segala bentuk teks yang tidak terstruktur lainnya.

Dalam pemodelan topik dikenal beberapa metode yang secara umum digunakan oleh para peneliti, seperti *Latent Semantic Analytic* (LSA) dan *Latent Dirichlet Allocation* (LDA). LSA merupakan metode yang bertujuan membentuk vektor representasi teks [20]. Sedangkan metode LDA merupakan algoritma untuk mendeteksi topik melalui pemodelan probabilistik dalam sekumpulan data [5]. LDA juga dapat digunakan untuk meringkas, melakukan pengelompokan, menghubungkan maupun memproses data [5].

Latent Dirichlet Allocation (LDA) merupakan salah satu metode populer dikalangan peneliti pemrosesan bahasa alami di bidang pemodelan topik. Meskipun begitu, LDA memiliki distribusi yang kompleks, sehingga menyebabkan perhitungan estimasi dari distribusi posterior untuk model LDA menjadi sangat sulit dilakukan secara manual [16]. Untuk menyelesaikan permasalahan distribusi kompleks tersebut, maka digunakan algoritma *Gibbs Sampling* [19].

II. METODE

A. Pengumpulan Data

Web Scraping adalah proses pengambilan sebuah dokumen semi-terstruktur dari internet, umumnya berupa halaman-halaman *web* dalam bahasa markup sepertiHTML atau XHTML, dan menganalisis dokumen tersebut untuk diambil data tertentu dari halaman tersebut untuk digunakan bagi kepentingan lain [27]. *Web scraping* memiliki sejumlah langkah, sebagai berikut.

1) *Create Scraping Template*: Pembuat program mempelajari dokumen HTML dari *website* yang akan diambil informasinya untuk tag HTML yang berisi informasi yang diperlukan.
2) *Explore Site Navigation*: Pembuat program mempelajari teknik navigasi pada website yang akan diambil informasinya untuk ditirukan pada aplikasi *web scrapper* yang akan dibuat.
3) *Automate Navigation and Extraction*: Berdasarkan informasi yang didapat pada langkat 1 dan 2 di atas, aplikasi *web scrapper* dibuatuntuk mengotomatisasi pengambilan informasi dari website yang ditentukan.
4) *Extracted Data and Package History*: Informasi yang didapat dari langkah 3 kemudian disimpan dalam tabel atau basis data sesuai kebutuhan.

Data yang digunakan berupa data berbentuk teks yang diperoleh dari portal berita detiknews. Pengumpulan data yang digunakan adalah metode *web scraping*. Sebanyak dua puluh artikel berita yang telah dikumpulkan dengan

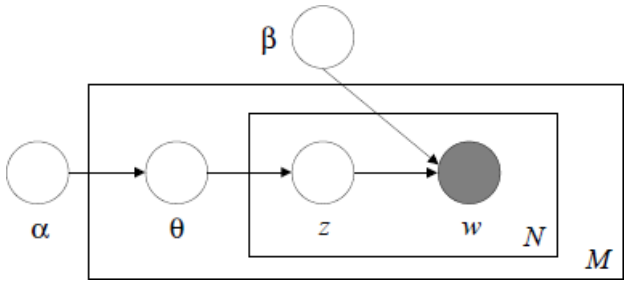
pembagian dua kategori utama yaitu, COVID-19 dan Narkoba.

B. *Text preprocessing*

Text preprocessing merupakan tahap pada Pemrosesan Bahasa Alami (PBA) untuk menyeleksi data yang akan diproses pada setiap dokumen, dimana proses tersebut disesuaikan dengan kebutuhan penelitian pada bidang penelitian terkait. Pada penelitian ini digunakan beberapa *text preprocessing* sebagai berikut.

1) *Case Folding*: Tahap merubah teks pada dokumen menjadi huruf kecil (*lowercase*) dan menghapus karakter selain huruf.
2) *Tokenizing*: Tahap pemotongan kalimat (s*tring*) menjadi kumpulan satuan kata (*terms*).
3) *Stemming*: Tahap merubah kata menjadi kata dasar. Contoh : kata "memakan" menjadi kata "makan".
4) *Stopword Removal*: Tahap menghapus kata-kata yang dianggap tidak terlalu penting atau tidak memiliki makna pada suatu teks.

C. *Pembentukan Model LDA*



Gambar. 1 Representasi Model LDA

Model LDA dapat direpresentasikan sebagai model grafis probabilistik seperti pada Gambar. 1 yang menunjukkan bahwa terdapat tiga tingkatan representasi LDA. Parameter α dan β merupakan parameter distribusi topik yang berada pada tingkatan korpus, yaitu kumpulan dari M dokumen. Parameter α digunakan dalam menentukan distribusi topik dalam dokumen, semakin besar nilai alpha dalam suatu dokumen, menandakan campuran topik yang dibahas dalam dokumen semakin banyak. Parameter β digunakan untuk menentukan distribusi kata dalam topik. Semakin tinggi nilai beta, maka semakin banyak kata-kata yang ada di dalam topik, sedangkan semakin kecil nilai beta, maka semakin sedikit kata-kata yang ada di dalam topik sehingga topik tersebut mengandung kata-kata yang lebih spesifik. Variabel θ adalah variabel yang berada di tingkat dokumen (M). Variabel θ merepresentasikan distribusi topik untuk dokumen tertentu. Semakin tinggi nilai θ, maka semakin banyak topik yang ada di dalam dokumen, sedangkan semakin kecil nilai θ, maka dapat dikatakan dokumen tersebut semakin spesifik pada topik tertentu. Variabel Z dan W adalah variabel tingkat kata (N). Variabel Z merepresentasikan topik dari kata tertentu pada sebuah dokumen. Sedangkan variabel W merepresentasikan kata yang berkaitan dengan topik tertentu yang terdapat dalam dokumen [5].

Gibbs sampling adalah salah satu algoritma keluarga dari Markov Chain Monte Carlo (MCMC) [23]. Gibbs sampling mengambil sampel untuk membangkitkan nilai sampel berikutnya secara acak. Sebagai contoh, untuk melakukan sampling terhadap nilai x pada distribusi p(x) = p(x1, , xm), dimana tidak ada solusi utuh untuk p(x) namun terdapat hasil yang representatif, maka hasil tersebut dapat ditemukan dengan menggunakan Gibbs Sampling dengan langkah-langkah sebagai berikut.

1) Buat variabel x_i, secara acak.
2) Untuk $t = 1, \dots, T$:
   a. $x_1^{t+1} \sim p(x_1 | x_2^{(t)}, x_3^{(t)}, \dots, x_m^{(t)})$
   b. $x_2^{t+1} \sim p(x_2 | x_1^{(t+1)}, x_3^{(t)}, \dots, x_m^{(t)})$
   c. $x_m^{t+1} \sim p(x_m | x_1^{(t+1)}, x_2^{(t+1)}, \dots, x_{m-1}^{(t+1)})$

Prosedur ini diulang hingga nilai sampel mendekatinilai distribusi sebenarnya. Dalam kasus LDA, nilai yang ingin diuji adalah porsi dokumen topik (θ), distribusi topik-kata φ (z), dan penugasan topik untuk tiap kata z.

Setelah tahap *text preprocessing*, data kemudian diproses dengan menggunakan metode *Latent Dirichlet Allocation* (LDA) dengan bantuan *Gibbs Sampling* untuk membantu penyebaran topiknya.

D. *Evaluasi Model LDA*

Ada beberapa metode evaluasi model LDA, salah satunya adalah koherensi topik. Pengukuran skorkoherensi topik terhadap sebuah topik dilakukan dengan mengukur derajat kemiripan semantik antar kata yang memiliki skor atau probabilitas tertinggi pada suatu topik. Dalam pengukuran koherensi topik, terdapat beberapa cara pengukuran, seperti UCI [7], dan uMass [6]. Untuk mengevaluasi model LDA yang telah dibentuk, kami menggunakan metode pengukuran skor koherensi topik UCI.

Metode evaluasi skor UCI dilakukan dengan menghitung pasangan kata menggunakan korpus eksternal. Ada tiga korpus eksternal yang disarankan yaitu, *Wikipedia*, *Google 2-grams*, dan *Medline* [7]. Untuk nilai UCI, nilai yang lebih baik adalah nilai yang lebih besar [28].

III. HASIL DAN PEMBAHASAN

A. *Konfigurasi Percobaan*

Proses pengujian dilakukan secara berulang sebanyak K. Pada pengujian ini akan digunakan parameter K = 5, maka pemodelan topik dilakukan dari K = 1, K = 2, K = 3, K = 4, dan K = 5. Hasil pemodelan topik tersebut kemudian dievaluasi menggunakan metode koherensi topik skor UCI untuk mengetahui kualitas topik yang

dihasilkan. Skor UCI dihitung dengan menggunakan data dari luar, dalam hal ini telah dipersiapkan korpus dokumen eksternal yang telah dikumpulkan dari banyak artikel Wikipedia Bahasa Indonesia sebanyak sepuluh artikel. Setelah didapat skor UCInya, kemudian dicari nilai K optimalnya berdasarkan skor rata-rata UCI.

B. Konfigurasi 1

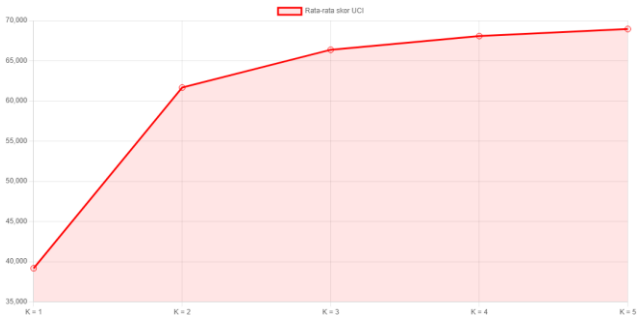
Pengujian pertama ini menggunakan konfigurasi yang ditunjukkan pada Tabel 1, dan hasil pengujian ditunjukkan pada Tabel 2.

TABEL I
KONFIGURASI PARAMETER 1

Parameter	Nilai
K	5
Jumlah iterasi *Gibbs Sampling*	100
Alpha	0,1
Beta	0,1

TABEL II
DATA HASIL PENGUJIAN KONFIGURASI 1

K	Topik	Daftar Kata	UCI
1	1	kota kabupaten pasien laku polisi covid nur ain rumah tangkap	39.171,635
2	1	laku polisi tangkap narkoba aman kali oknum bandung jakarta hasil	61.668,207
	2	kota kabupaten pasien covid nur ain trenggalek wisma atlet arifin	
3	1	nur ain wisma atlet nikah oknum rsdc mayor virtual pasien	66.376,412
	2	polisi laku tangkap narkoba aman kali bandung jakarta hasil tembakau	
	3	kota kabupaten pasien covid trenggalek jatim malang orang probolinggo pasuruan	
4	1	nur pasien ain wisma atlet arifin nikah sakit covid rsdc	68.085,821
	2	laku jakarta polisi hasil tembakau aman iptu santi rumah pabrik	
	3	polisi narkoba tangkap oknum bandung laku zaky sabu polres periksa	
	4	kota kabupaten covid jatim malang pasien orang probolinggo pasuruan mojokerto	
5	1	pasien sakit arifin ambah covid trenggalek rumah darurat manfaat corona	68.955,743
	2	laku polisi tembakau aman iptu santi jakarta rumah pabrik gorila	
	3	polisi narkoba tangkap oknum bandung zaky sabu laku periksa had	
	4	kota kabupaten covid jatim malang pasien probolinggo pasuruan mojokerto total	
	5	ain nur wisma atlet nikah rsdc mayor virtual bahagia langsung	



Gambar. 2 Grafik skor rata-rata UCI

C. Analisis Konfigurasi 1

Berdasarkan hasil pengujian dengan konfigurasi parameter 1 yang ditunjukkan pada Tabel 2, nilai K optimal adalah K = 5 berdasarkan skor rata-rata UCI sebesar 68.955,743. Daftar kata probabilitas tertinggi masing-masing topik dapat dilihat pada Tabel 3.

TABEL III
DAFTAR KATA PROBABILITAS TERTINGGI KONFIGURASI 1

Topik	Daftar Kata	UCI
1	pasien sakit arifin ambah covid trenggalek rumah darurat manfaat corona	69.163,045
2	laku polisi tembakau aman iptu santi jakarta rumah pabrik gorila	68.233,513
3	polisi narkoba tangkap oknum bandung zaky sabu laku periksa had	67.972,725
4	kota kabupaten covid jatim malang pasien probolinggo pasuruan mojokerto total	70.240,621
5	ain nur wisma atlet nikah rsdc mayor virtual bahagia langsung	69.168,811

Gambar. 3 Grafik skor UCI topik

Berdasarkan Tabel 3 dan Gambar. 3., topik 3 merupakan topik kualitas terendah dengan skor UCI sebesar 67.972,725. Sedangkan topik 4 merupakan topik kualitas tertinggi dengan skor UCI sebesar 70.240,621. Distribusi topik untuk setiap dokumen ditunjukkan pada Gambar. 4.

Gambar. 4 Distribusi Topik Dokumen (Konfigurasi 1)

D. Konfigurasi 2

Pengujian kedua ini menggunakan konfigurasi yang ditunjukkan pada Tabel 4 dan hasil pengujian ditunjukkan pada Tabel 5.

TABEL IV
KONFIGURASI PARAMETER 2

Parameter	Nilai
K	5
Jumlah iterasi *Gibbs Sampling*	1
Alpha	0,1
Beta	0,1

TABEL V
DATA HASIL PENGUJIAN KONFIGURASI 2

K	Topik	Daftar Kata	UCI
1	1	kota kabupaten pasien laku polisi covid nur ain rumah tangkap	39.171,635
2	1	kabupaten polisi pasien kota tangkap orang positif covid total trenggalek	60.698,429
	2	kota laku kabupaten wisma atlet arifin jatim rumah sakit ambah	
3	1	nur ain orang wisma laku tangkap oknum pasien kabupaten rsdc	65.856,413
	2	kabupaten covid kota malang jakarta sakit pasien jatim rumah blitar	
	3	kota polisi total positif kali narkoba hasil zaky sabu pasien	
4	1	rumah kota total nur tembakau iptu kali blitar laku aman	67.417,935
	2	pasien laku jatim tangkap probolinggo langsung nikah rsdc zaky jumat	
	3	polisi covid kota atlet ain virtual ambah madiun polres santi	
	4	kabupaten narkoba wisma kota pasuruan positif sembuh orang mojokerto trenggelek	
5	1	covid laku polisi aman langsung ain tembakau surabaya total awat	68.630,539
	2	pasien kabupaten madiun sakit polisi tangkap tuban skip manfaat orang	
	3	kota probolinggo hasil orang virtual proses oknum sidoarjo tinggal jakarta	
	4	kabupaten kota nur malang	

		pasuruan mojokerto wisma rumah capai laku	
	5	atlet positif jatim narkoba trenggalek iptu kota persen bandung jember	

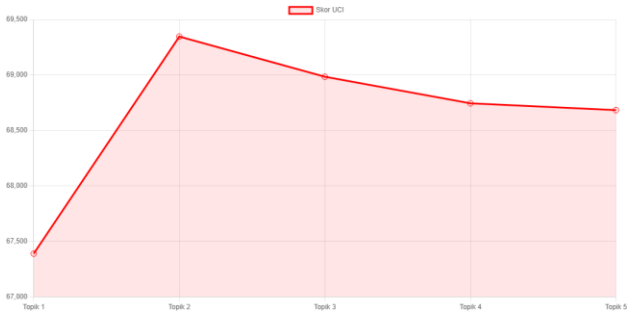
Gambar. 5 Grafik skor rata-rata UCI

E. Analisis Konfigurasi 2

Berdasarkan hasil pengujian dengan konfigurasi parameter 2 yang ditunjukkan pada Tabel 5., nilai K optimal berdasarkan skor rata-rata total adalah K = 5 dengan skor rata-rata UCI sebesar 5.277,580. Daftar kata probabilitas tertinggi setiap topik dapat dilihat pada Tabel 6.

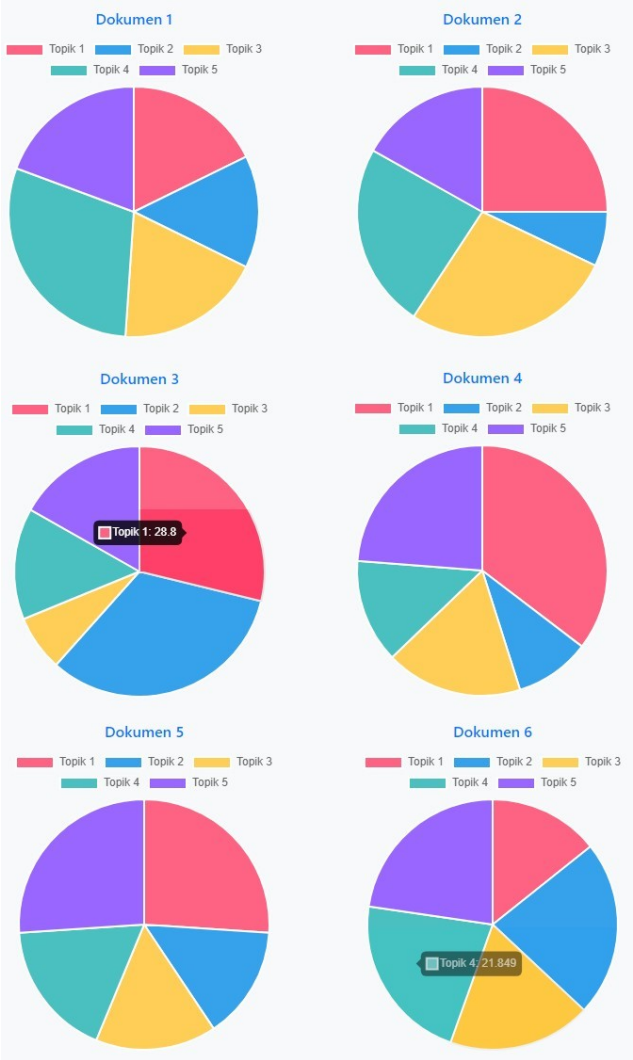
TABEL VI
DAFTAR KATA PROBABILITAS TERTINGGI KONFIGURASI 2

Topik	Daftar Kata	UCI
1	covid laku polisi aman langsung ain tembakau surabaya total awat	67.390,540
2	pasien kabupaten madiun sakit polisi tangkap tuban skip manfaat orang	69.347,197
3	kota probolinggo hasil orang virtual proses oknum sidoarjo tinggal jakarta	68.985,115
4	kabupaten kota nur malang pasuruan mojokerto wisma rumah capai laku	68.745,347
5	atlet positif jatim narkoba trenggalek iptu kota persen bandung jember	68.684,497



Gambar. 6 Grafik skor UCI topik

Berdasarkan Tabel 6 dan Gambar. 6, topik 2 merupakan topik kualitas tertinggi dengan skor UCI sebesar 69.347,197. Sedangkan topik 1 merupakan topik kualitas terendah dengan skor UCI sebesar 67.390,540. Distribusi topik untuk setiap dokumen ditunjukkan pada Gambar. 7.



Gambar. 7 Distribusi Topik Dokumen (Konfigurasi 2)

IV. KESIMPULAN

Berdasarkan hasil pengujian terhadap kedua konfigurasi parameter, didapatkan nilai K optimal untuk masing-masing konfigurasi adalah K = 5. Pada konfigurasi pertama, distribusi kata masing-masing topik dapat tersebar dengan baik. Sedangkan pada konfigurasi kedua, topik-topik yang dihasilkan merupakan topik dengan distribusi kata yang terlalu tersebar karena faktor konfigurasi parameternya yaitu jumlah iterasi Gibbs Sampling yang mengakibatkan kualitas topik yang kurang baik.

Skor UCI pada kedua konfigurasi tidak jauh berbeda. Hal ini dikarenakan perhitungan skor UCI yang bergantung pada faktor eksternal yang sama yaitu korpus Wikipedia. Pada konfigurasi pertama, distribusi topik setiap dokumen benar-benar tersebar dengan baik. Sedangkan pada konfigurasi kedua, distribusi topik setiap dokumen tidak tersebar dengan baik yang mengakibatkan kurang jelasnya topik yang dibahas pada masing-masing dokumen.

REFERENSI

[1] A. F. Octaviansyah, D. Darwis, A. Surahman. 2019. "SISTEM PENCARIAN LOKASI BENGKEL MOBIL RESMI MENGGUNAKAN TEKNIK PENGOLAHAN SUARA DAN PEMROSESAN BAHASA ALAMI." *Jurnal TEKNOINFO, Vol. 13, No. 2* 71-77.

[2] A. Piepenbrink, A. S. Gaur. 2017. "Topic models as a novel approach to identify themes in content analysis: The example of organization research methods."

[3] B. W. Arianto, G. Anuraga. 2020. "Pemodelan Topik Pengguna Twitter Mengenai Aplikasi "Ruangguru"." *Jurnal ILMU DASAR, Vol. 21 No. 2* 149-154.

[4] D. Koren˘ci´c, S. Ristov,J. Snajder. 2018. "Document-based Topic Coherence Measures for News Media Text."

[5] D. M. Blei, A. Y. Ng, M. I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research 3* 993-1022.

[6] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, A. McCallum. 2011. "Optimizing Semantic Coherence in Topic Models." *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* 262-272.

[7] D. Newman, J.H. Lau, K. Grieser, T. Baldwin. 2010. "Automatic Evaluation of Topic Coherence." *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL* 100-108.

[8] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, L. Zhao. 2018. "Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey."

[9] H. Jiang, R. Zhou,L. Zhang,H. Wang,Y. Zhang. 2018. "Sentence Level Topic Models for Associated Topics Extraction."

[10] H. M. Wallach, I. Murray, R. Salakhutdinov, D. Mimno. 2009. "Evaluation Methods for Topic Models."

[11] J. Qiang, P. Chen,T. Wang,X. Wu. 2016. "Topic Modeling over Short Texts by Incorporating Word Embeddings."

[12] J. Su, J. Xu,X. Qiu,X.Huang. 2018. "Incorporating Discriminator in Sentence Generation: A Gibbs Sampling Method." *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)* 5496-5503.

[13] K. B. Putra, R. P. Kusumawardani. 2017. "ANALISIS TOPIK INFORMASI PUBLIK MEDIA SOSIAL DI SURABAYA MENGGUNAKAN PEMODELAN LATENT DIRICHLET ALLOCATION (LDA)." *Jurnal Teknik ITS Vol 6 No. 2.*

[14] K. Beck, M. Hendrickson, M. Fowler. 2001. *Planning Extreme Programming.* Boston: Pearson Education, Inc. Right and Contracts Department.

[15] K. Stevens, P. Kegelmeyer, D. Andrzejewski, D. Buttler. 2012. "Exploring Topic Coherence over many models and many topics." *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural* 952-961.

[16] Kengken, R. I. 2014. "PEMODELAN TOPIK UNTUK MEDIA SOSIAL MENGGUNAKAN LATENT DIRICHLET ALLOCATION."

[17] L. A. Wirasakti, R. Permadi, A. D. Hartanto, Hartatik. 2020. "Pembuatan Kata Kunci Otomatis Dalam Artikel Dengan Pemodelan Topik." *JURNAL MEDIA INFORMATIKA BUDIDARMA Volume 4, Nomor 1* 27-31.

[18] M. Cendana, S. D. H. Permana. 2019. "PRA-PEMROSESAN TEKS PADA GRUP WHATSAPP UNTUK PEMODELAN TOPIK." *Jurnal Mantik Penusa Vol. 3 No.3* 107-116.

[19] Qiu, X. 2014. "Topic words analysis based on LDA model."

[20] R. Alghamdi, K. Alfalqi. 2015. "A Survey of Topic Modeling in Text Mining." *(IJACSA) International Journal of Advanced Computer Science and Applications, Vol 6 No.1.*

[21] R. Ding, R. Nallapati,B. Xiang. 2018. "Coherence-Aware Neural Topic Modeling."

[22] R. Y. K. Lau, Y. Xia, Y. Ye. 2014. "A Probabilistic Generative Model for Mining Cybercriminal Networks from Online Social Media." *IEEE Computational intelligence magazine.*

[23] S. German, D. German. 1984. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. PAMI-6, NO. 6.*

[24] S. J. Lee, E. Brennan, L. A. Gibson, A. S. L. Tan, A. Kybert-Momjian, J. Liu, R. Hornik. 2016. "Predictive Validity of an Empirical Approach for Selecting Promising Message Topics: A Randomized-Controlled Study." *Journal of Communication, Volume 66, Issue 3* 433-453.

[25] Supriyatna, A. 2018. "Metode Extreme Programming Pada Pembangunan Web Aplikasi Seleksi Peserta Pelatihan Kerja." *Jurnal Teknik Informatika Vol 11 No.1.*

[26] T. Williams, J. Betak. 2018. "A Comparison of LSA and LDA for the Analysis of Railroad Accident Text." *Procedia Computer Science 130* 98-102.

[27] Turland, M. 2010. *Php-Architect's Guide to Web Scraping.*

[28] Y.U. Al-khairi, Y. Wibisono, B.L. Putro. 2019. "DETEKSI TOPIK FASHION PADA TWITTER DENGAN LATENT DIRICHLET ALLOCATION."

[29] Zulhanif, Sudartianto, B. Tantular, I. G. N. M. Jaya. 2017. "APLIKASI LATENT DIRICHLET ALLOCATION (LDA) PADA CLUSTERING DATA TEKS." *Jurnal "LOG!K@" , Jilid 7, No. 1* 46-51.