Ekstraksi Topik Pada Aduan Mahasiswa Dengan Pendeketan Model Latent Dirichlet Allocation (LDA)

Gede Herdian Setiawan[1], I Made Budi Adnyana[2], I Gusti Rai Agung Sugiartha[3], Komang Budiarta[3]
Fakultas Informatika dan Komputer
Institut Teknologi dan Bisnis STIKOM Bali
Denpasar, Indonesia
e-mail: [1]herdian@stikom-bali.ac.id, [2]budi.adnyana@stikom-bali.ac.id, [3]sugiarta@stikom-bali.ac.id, [4]komang_budiarta@stikom-bali.ac.id

Abstrak

Aduan mahasiswa perlu dikelola dan dianalisa dengan baik dan dijadikan sebagai sumber informasi yang berharga bagi perguruan tinggi dalam mengembangkan sistem pendidikan yang lebih baik dan memenuhi harapan mahasiswa. Dalam mengelola aduan mahasiswa perguruan tinggi kerap mengalami kendala seperti tidak memiliki cukup sumber daya untuk mengumpulkan, menyimpan, dan menganalisis semua aduan mahasiswa dengan efektif. Pada pelitian ini menerapkan model Latent Dirichlet Allocation (LDA) yang bertujuan untuk ekstraksi topik aduan yang masuk berdasarkan kata-kata kunci yang terdapat dalam aduan mahasiswa. Dengan demikian, diharapkan dapat mempermudah proses pengelolaan aduan mahasiswa, sehingga masalah yang dihadapi oleh mahasiswa dapat diselesaikan dengan lebih cepat dan tepat. Data aduan mahasiswa akan dikumpulkan dari sistem pengelolaan aduan mahasiswa yang sudah ada. Selanjutnya, data tersebut akan diolah dengan menggunakan teknik model LDA untuk mengidentifikasi topik aduan yang masuk. Pengujian terhadap ekstraksi empat topik pada model LDA mendapat nilai perplexity sebesar -6.8026282, hal ini menujukan model LDA memiliki perplexity yang rendah dan menunjukkan kinerja yang baik. Uji koherensi dengan menggunakan metode u_mass perubahan nilai yang mengecil seiring dengan peningkatan jumlah topik menunjukkan adanya perubahan koherensi topik dalam model LDA, dapat disimpulkan pada dataset dan model LDA yang dibangun pada penelitian ini semakin tinggi jumlah topik maka performa pembentukan topik menjadi lebih buruk. Untuk mendapatkan jumlah topik yang ideal perlu dilakukan beberapa penyesuaian jumlah topik, dataset dan pengujian kohorensi dengan metode lain.

Kata kunci: *Aduan, Klasifikasi, Text Mining, LDA.*

Abstract

Student complaints need to be properly managed and analyzed and used as a valuable source of information for universities in developing a better education system and meeting student expectations. In managing student complaints, tertiary institutions often experience problems such as not having enough resources to collect, store, and analyze all student complaints effectively. This research applies the Latent Dirichlet Allocation (LDA) model which aims to classify incoming complaint topics based on the key words contained in student complaints. This is expected to facilitate the process of dealing with student complaints, so that problems faced by students can be resolved more quickly and accurately. Student complaint data will be collected from the existing student complaint management system. Furthermore, the data will be processed using the LDA model technique to identify the topic of the incoming complaint. Testing the extraction of the four topics in the LDA model obtained a perplexity value of -6.8026282, this indicates that the LDA model has a low perplexity and shows good performance. The coherence test using the u_mass method decreases in value as the number of topics increases, indicating a change in topic coherence in the LDA model. It can be concluded that in the dataset and LDA model built in this study, the higher the number of topics, the worse the topic formation performance. In order to get the ideal number of topics, it is necessary to make some adjustments to the number of topics, datasets and cohorrence testing with other methods.

Keywords*: Student complaints, Classification, Text Mining, LDA.*

1. Pendahuluan

Masukan atau aduan dari mahasiswa sangat penting bagi perguruan tinggi karena dapat memberikan informasi yang berharga terkait dengan kualitas dan keberhasilan sistem pendidikan yang ada. Aduan mahasiswa dapat memberikan informasi tentang hal-hal yang perlu diperbaiki dalam sistem pendidikan, seperti kurikulum yang kurang relevan, dosen yang kurang berkualitas, sarana dan prasarana yang tidak memadai, atau masalah administrasi.

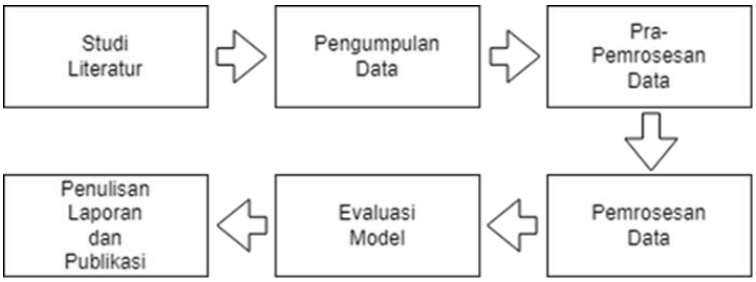
Aduan mahasiswa perlu dikelola dan dianalisa dengan baik dan dijadikan sebagai sumber informasi yang berharga bagi perguruan tinggi dalam mengembangkan sistem pendidikan yang lebih baik dan memenuhi harapan mahasiswa. Dalam mengelola aduan mahasiswa perguruan tinggi kerap mengalami kendala seperti tidak memiliki cukup sumber daya untuk mengumpulkan, menyimpan, dan menganalisis semua aduan mahasiswa dengan efektif [1], [2].

Berberapa teknik yang bisa digunakan untuk mengelola dan menganalisa aduan secara efektif seperti Analisis sentiment dan Analisis topik. dengan teknik analisis sentiment dapat menentukan apakah aduan tersebut berisi sentimen positif, negatif, atau netral. Dengan teknik ini, dapat memahami perasaan dan pandangan mahasiswa terhadap masalah tertentu [3]. Teknik analisis topik, dapat digunakan untuk menganalisis aduan mahasiswa dan mengidentifikasi topik atau masalah utama yang dihadapi mahasiswa. Dengan teknik ini, dapat memeberikan gambaran yang lebih jelas mengenai keluahan mahasiswa dan mengambil tindakan yang diperlukan untuk memperbaiki masalah dan memastikan kepuasan mahasiswa. Salah satu model yang dapat diterapkan untuk analisa topik adalah model *Latent Dirichlet Allocation* (LDA) merupakan salah satu model yang dapat dogunakan untuk mengelompokkan teks berdasarkan topik atau tema yang tersembunyi di dalamnya [4]–[6].

Penelitian terkait mengenai penerapan model LDA antara lain Manik Dkk menerapkan model LDA untuk ekstraksi topik dokumen, melalui pengujian penelitian ini menyimpulkan model LDA memiliki kinerja sangat baik dalam melakukan ektraksi berita digital berbahasa Indonesia [7]. Fajriyanto menerapkan model LDA untuk mengetahui topik berita yang sedang dominan di Twitter hasilnya mampu menampilkan topik yang sedang dominan [8]. Agus melalukan implementasi LDA untuk klasterisasi cerita bahasa bali menghasilkan akurasi hasil klasterisasi tertinggi mencapai 62% pada dokumen berbahasa bali [9].

Model LDA merupakan salah satu teknik *machine learning* yang dapat digunakan untuk melakukan klasifikasi topik pada dokumen. Pada elitian ini, penggunaan model LDA bertujuan untuk mengklasifikasikan topik aduan yang masuk berdasarkan kata-kata kunci yang terdapat dalam aduan mahasiswa. Dengan demikian, diharapkan dapat mempermudah proses pengelolaan aduan mahasiswa, sehingga masalah yang dihadapi oleh mahasiswa dapat diselesaikan dengan lebih cepat dan tepat. Data aduan mahasiswa akan dikumpulkan dari sistem pengelolaan aduan mahasiswa yang sudah ada. Selanjutnya, data tersebut akan diolah dengan menggunakan teknik model LDA untuk mengidentifikasi topik aduan yang masuk. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi yang signifikan dalam peningkatan sistem pengelolaan aduan mahasiswa yang ada, sehingga dapat meningkatkan kualitas pelayanan dan kepuasan mahasiswa terhadap sistem pendidikan yang ada.

2. Metode Penelitian



Gambar 1. Tahapan Penelitian

Gambar 1 merupakan diagram alir tahapan penelitian yang terdiri dari beberapa tahapan, antara lain : studi literatur, pengumpulan data, pra-pemrosesan data, pemrosesan data, evaluasi mode, penulisan laporan dan publikasi :

a. Studi Litaratur

   Membaca dan mempelajari penelitian-penelitian terkait dengan klasifikasi topik pada aduan mahasiswa dan metode LDA untuk pengolahan teks.

b. Pengumpulan Data

Data aduan mahasiswa dikumpulkan berdasarkan aduan mahasiswa melalui sistem Informasi aduan yang sudah ada, data aduan diambil pada salah satu Perguruan Tinggi Swasta (PTS) untuk selanjutnya dijadikan dataset.

c. Pra-pemrosesan data
   1. Membersihkan dan menganalisis teks, seperti menghapus tanda baca, kata-kata yang tidak relevan, dan mengubah semua huruf menjadi huruf kecil. Pemrosesan data
   2. Melakukan tokenisasi (memisahkan teks menjadi kata-kata) dan menghilangkan kata-kata yang umum *(stop words)*
   3. Melakukan stemming atau lemmatisasi untuk mengubah kata-kata menjadi bentuk dasar
d. Menerapkan *model Latent Dirichlet Allocation* (LDA) pada data yang telah dipra-pemroses sebelumnya
   1. Membangun kamus (*dictionary*) yang menghubungkan setiap kata dengan suatu indeks unik.
   2. Mengubah setiap dokumen menjadi vektor representasi *term frequency* (jumlah kemunculan kata dalam dokumen) menggunakan kamus yang telah dibangun. Vektor ini disebut *corpus*
   3. Menentukan jumlah topik yang diinginkan dalam model.
   4. Menginisialisasi model LDA dengan corpus dan kamus yang telah dibangun.
   5. Melakukan pelatihan (training) model LDA menggunakan algoritma pemodelan probabilistik untuk menemukan distribusi topik dalam dokumen dan distribusi kata dalam topik.
   6. Melakukan clustering terhadap distriubusi topik pada setiap dokumen
e. Evaluasi model
   Mengevaluasi dan memvalidasi model LDA, seperti dengan menggunakan metrik seperti *perplexity* atau *coherence score*.
f. Penulisan laporan dan publikasi
   Menulis laporan penelitian dan publikasi di jurnal ilmiah yang terindeks

3. Hasil dan Pembahasan

3.1. Penerapan Model LDA

Penelitian ini mengusulkan penerapan model LDA untuk kelasifikasi aduan mahasiswa. Sebelum penerapapan model LDA beberapa persiapkan dilakukan seperti : pengumpulan dataset, tahap pemrosesan awal / *preprocessing*, penerapan model LDA untuk pencarian topik, melalukan ekstraksi aduan berdasarkan topik dan dilajutkan dengan evaluasi dan analisa hasil.

Tahap preprocessing meliputi : tokenisasi yaitu membagi kalimat menjadi bagaian terkecil yang tediri dari kata. Sebagai contoh sebuah kalimat pada salah satu dokumen aduan : *"Ada peremajaan komputer untuk kelas-kelas yang membutuhkan"* setelah dilakukan tokenisasi menjadi : *['ada', 'peremajaan', 'komputer', 'untuk', 'kelas-kelas', 'yang', 'membutuhkan']*. Selanjutnya dilakukan proses stemming untuk melakukan penguraian bentuk kata menjadi kata dasar. Hasil stemming seperti berikut : *['ada', 'remaja', 'komputer', 'untuk', 'kelas', 'yang', 'butuh']*. Selanjutnya dilakukan proses *stopward removal* untuk mengilangkan atau eliminasi kata yang tidak penting. Hasil setelah proses *stopward* salah satu kalimat pada dokumen aduan adalah *['remaja', 'komputer', 'kelas', 'butuh']*.

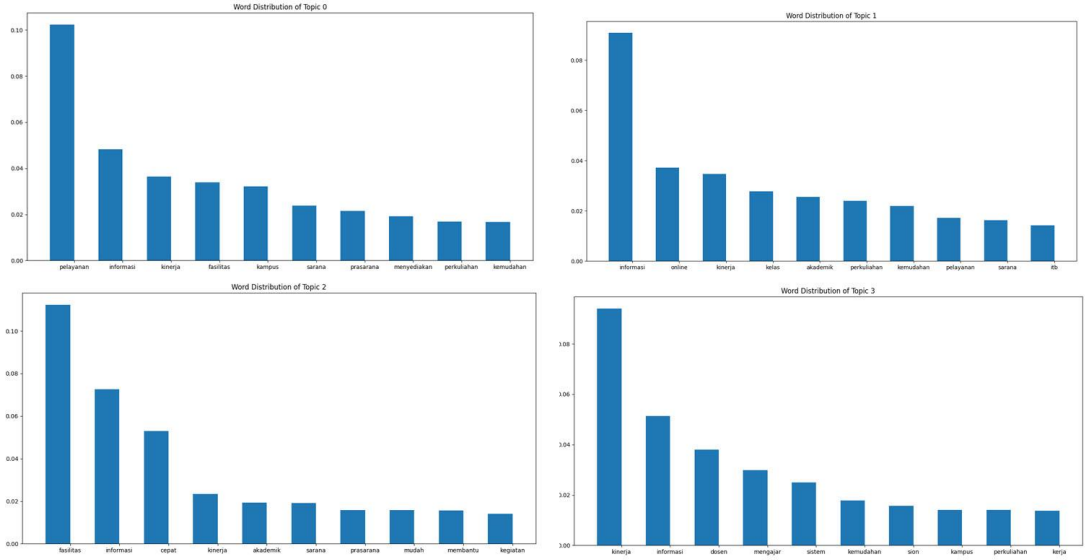
Setelah tahap *preprocessing* selesai dilakukan berikutnya melakukan pencarian topik-topik pada *dataset* aduan mahasiswa dengan menggunakan model LDA. Pada tahapannya LDA memerlukan hyper-parameter yaitu penentuan jumlah topik yang akan ditemukan oleh LDA (K). Pendekatan model LDA untuk menemukan topik aduan tidak memperhatikan urutan kata pada dokumen atau dianggap sebagai *bag-of-word* (BoW). Hasil dari proses LDA dengan jumlah topik (K) sebanyak 4 ditunjukkan pada tabel 1.

Tabel 1. Distribusi Topik dan Probabilitas

Distribusi topik (K) dan probabilitas			
1	**2**	**3**	**4**
Pelayanan (0.102)	Informasi (0.098)	Fasilitas (0.111)	Kinerja (0.093)
Informasi (0.048)	Online (0.037)	Informasi (0.072)	Informasi (0.051)

Kinerja (0.036)	Kinerja (0.034)	Cepat (0.052)	Dosen (0.038)
Fasilitas (0.033)	Kelas (0.027)	Kinerja (0.023)	Mengajar (0.029)
Kampus (0.032)	Akademik (0.025)	Akademik (0.019)	Sistem (0.024)
Sarana (0.023)	Perkuliahan (0.023)	Sarana (0.0191)	Kemudahan (0.017)
Prasarana (0.021)	Kemudahan (0.022)	Prasarana (0.015)	Sion (0.015)
Menyediakan (0.019)	Pelayanan (0.017)	Mudah (0.0157)	Kampus (0.014)
Perkuliahan (0.0169)	Sarana (0.016)	Membantu (0.015)	Perkuliahan (0.014)
Kemudahan (0.0167)	Prasarana (0.014)	Kegiatan (0.014)	Kerja (0.013)

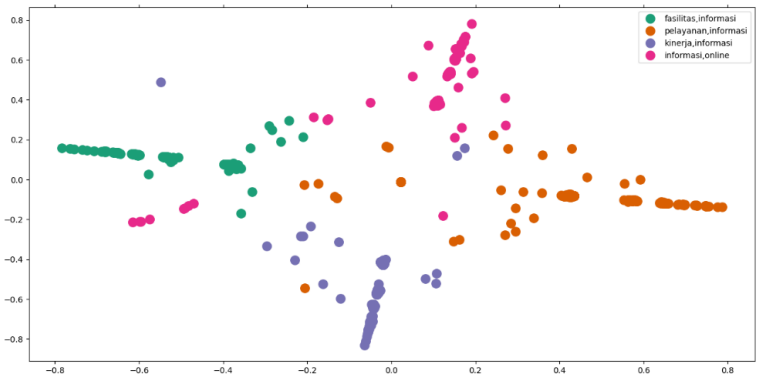
Tabel 1 menunjukkan luaran atau hasil dari model LDA berupa probabilitas pada setiap topik. Masing-masing topik memiliki kata pada setiap dokumen aduan dengan nilai rentang 0 sampai dengan 1 yang mengidikasikan nilai probabilitas kemunculan kata pada topik. Kemunculan kata pada setiap topik memiliki nilai probabilitas yang berbeda sebagai contoh kata kinerja pada topik 1 memiliki nilai probalitas yang berbeda pada kemunculan pada topik 2 sampai dengan 4. Nilai probabilitas kata pada setiap topik dapat menjelaskan topik pada dokumen aduan. Untuk dapat menggambarkan lebih jelas probabilitas kata pada setiap topik dapat ditampilkan dalam bentuk diagram seperti pada Gambar 2.



Gambar 2. Distribusi kata pada pada setiap topik

Gambar 2 menunjukan distribusi kata pada topik 1 sampai dengan topik 4, pelayanan menjadi isu aduan yang memiliki nilai probabilitas paling tinggi diikuti dengan informasi, kinerja, fasilitas dan lainnya pada topik 1. Informasi menjadi isu aduan yang memiliki nilai probabilitas paling tinggi diikuti *online*, kinerja, kelas dan lainnya pada topik 2.Fasilitas menjadi isu aduan yang memiliki nilai probabilitas paling tinggi diikuti informasi, cepat, kinerja dan lainnya. Kinerja menjadi isu aduan yang memiliki nilai probabilitas paling tinggi diikuti informasi, dosen, mengajar dan lainnya pada topik 4.Setelah mendapatkan distribusi topik pada setiap aduan selanjutnya dilakukan pengelompokan (*cluster*) topik aduan berdasarkan kedeketan topik. Peroses klustering menggunakan algoritma *KMeans* [10], [11].

Hasil dari proses cluster aduan ditujukan pada Gambar 3



Gambar 3 Hasil *cluster* empat topik

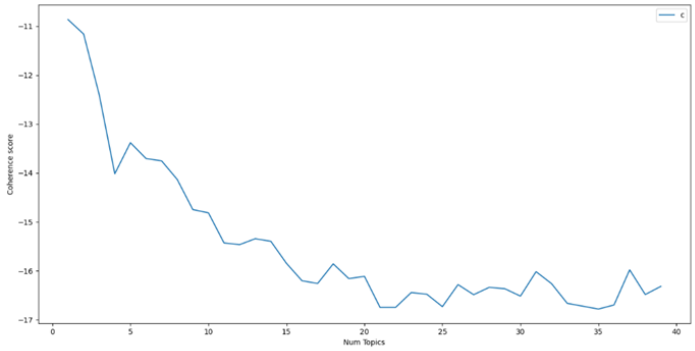
Gambar 3 menunjukan hasil *cluster* topik aduan. *Cluster* aduan diurutkan berdasarkan nilai probabilitasnya, terdapat dua wakil kata cluster yang memiliki nilai mayoritas di setiap *cluster*.

3.2. Analisis Hasil

Analisa hasil perapan model LDA dilakukan dengan menguji nilai *perplexity* dan *coherence* terhadap jumlah topik yang sudah ditentukan pada model. Uji *perplexity* pada model LDA digunakan untuk mengukur kualitas dan kinerja model. *Perplexity* adalah metrik evaluasi yang umum digunakan dalam model LDA untuk mengevaluasi sejauh mana model dapat memprediksi kata-kata yang muncul dalam korpus pelatihan. *Perplexity* menggambarkan sejauh mana model LDA dapat menyesuaikan distribusi probabilitas kata-kata dalam dokumen yang diberikan. Semakin rendah nilai *perplexity*, semakin baik model LDA dalam memprediksi kata-kata yang muncul dalam korpus pelatihan.

Setelah dilakukan pengujian terhadap empat topik pada model LDA mendapat nilai *perplexity* sebesar -6.8026282, hal ini menujukan model LDA memiliki perplexity yang rendah. Sebagai nilai perplexity, angka tersebut menunjukkan kinerja yang baik karena semakin rendah nilai *perplexity*, semakin baik kemampuan model dalam memprediksi kata-kata dalam korpus pelatihan. hasil *perplexity* yang sangat negatif menunjukkan bahwa model LDA dengan parameter yang digunakan mampu dengan baik dalam memodelkan data dan memiliki kemampuan yang baik dalam memprediksi kata-kata yang muncul dalam korpus. Model dengan *perplexity* rendah cenderung memiliki representasi yang lebih baik terhadap data yang diberikan.

Analisa uji *coherence* pada model LDA digunakan untuk mengukur kualitas topik yang dihasilkan oleh model tersebut. *Coherence* mengukur sejauh mana topik dalam model LDA memiliki hubungan yang bermakna antara kata-kata yang sering muncul Bersama [12]. Hasil uji *coherence* ditujukkan pada Gambar 7.



Gambar 7 Uji *coherence* dengan Metode *U_Mass*

Uji koherensi u_*mass* (*uMass Coherence*) mengukur kualitas koherensi topik dengan memperhatikan kemunculan kata-kata dalam dokumen secara berurutan. Evaluasi koherensi dilakukan dengan menghitung skor *log-likelihood* dari kemunculan kata-kata dalam konteks yang diberikan. Nilai koherensi u_*mass* pada Gambar 7 menunjukkan perubahan nilai yang mengecil seiring dengan peningkatan jumlah topik menunjukkan adanya perubahan koherensi topik dalam model LDA. Saat jumlah topik

meningkat, informasi dan kata-kata terdistribusi dengan lebih rinci pada topik yang lebih spesifik. Akibatnya, kata-kata dalam topik cenderung memiliki kemunculan yang lebih terbatas dan kurang tersebar di antara topik-topik lainnya. Hal ini dapat menyebabkan penurunan nilai koherensi u_*mass* karena kemunculan kata-kata dalam konteks yang diberikan menjadi lebih jarang dan kurang berurutan. Dapat disimpulkan pada dataset dan model LDA yang dibangun pada penelitian ini semakin tinggi jumlah topik maka performa pembentukan topik menjadi lebih buruk

4. Kesimpulan

Berdasarkan pembahasan dan analisa hasil penerapan model LDA untuk ekstraksi topik pada data aduan mahasiswa di salah satu perguruan tinggi swasta (PTS). Model LDA mampu menemukan topik pada setiap dokumen aduan yang telah dilakukan pra pemrosesan dan menghasilkan nilai probabilitas pada setiap topik. Masing-masing topik memiliki kata pada setiap dokumen aduan dengan nilai rentang 0 sampai dengan 1 yang mengidikasikan nilai probabilitas kemunculan kata pada topik. Kemunculan kata pada setiap topik memiliki nilai probabilitas yang berbeda. Pengujian terhadap ekstraksi empat topik pada model LDA mendapat nilai *perplexity* sebesar -6.8026282, hal ini menujukan model LDA memiliki *perplexity* yang rendah dan menunjukkan kinerja yang baik. Uji koherensi dengan menggunakan metode *u_mass* perubahan nilai yang mengecil seiring dengan peningkatan jumlah topik menunjukkan adanya perubahan koherensi topik dalam model LDA. Berdasarkan hal tersebut dapat disimpulkan pada dataset dan model LDA yang dibangun pada penelitian ini semakin tinggi jumlah topik maka performa pembentukan topik menjadi lebih buruk. Pada penelitian selanjutnya masih perlu dilakukan pengujian kohorensi dengan metode lain, sehingga dapat disimpulkan jumlah ekstraksi topik yang ideal diterapkan pada model LDA.

Daftar Pustaka

[1] I. Gede, T. Suryawan, I. Putu, And S. Handika, “Rancang Bangun Sistem Pengaduan Layanan Akademik Stmik Stikom Indonesia,” Online, 2018. [Online]. Available: Http://Jurnal.Stiki-Indonesia.Ac.Id/Index.Php/Sintechjournal

[2] M. F. Rahman *Et Al.*, “Peningkatan Kemampuan Manajemen Dalam Pengelolaan Elektronik Komplain Layanan Laboratorium Universitas Nurul Jadid,” Vol. 1, No. 1, Pp. 80–85, 2021.

[3] Q. Zhao, H. Zhang, And J. Shang, “Interpretable Sentiment Analysis Based On Sentiment Words’ Syntax Information,” In *2022 International Conference On Industrial Automation, Robotics And Control Engineering (Iarce)*, 2022, Pp. 80–85. Doi: 10.1109/Iarce57187.2022.00025.

[4] P. C. Kaur, T. Ghorpade, And V. Mane, “Extraction Of Unigram And Bigram Topic List By Using Latent Dirichlet Markov Allocation And Sentiment Classification,” In *2017 International Conference On Energy, Communication, Data Analytics And Soft Computing (Icecds)*, 2017, Pp. 2332–2339. Doi: 10.1109/Icecds.2017.8389869.

[5] E. Wahyudi And R. Kusumaningrum, “Aspect Based Sentiment Analysis In E-Commerce User Reviews Using Latent Dirichlet Allocation (Lda) And Sentiment Lexicon,” In *2019 3rd International Conference On Informatics And Computational Sciences (Icicos)*, 2019, Pp. 1–6. Doi: 10.1109/Icicos48119.2019.8982522.

[6] A. J. Nair, V. G, And A. Vinayak, “Comparative Study Of Twitter Sentiment On Covid - 19 Tweets,” In *2021 5th International Conference On Computing Methodologies And Communication (Iccmc)*, 2021, Pp. 1773–1778. Doi: 10.1109/Iccmc51019.2021.9418320.

[7] P. Manik, K. Suryawan, And N. Mandia, “Metode Latent Dirichlet Allocation Untuk Ekstraksi Topik Dokumen,” 2017.

[8] M. Fajriyanto, F. Matematika Dan Ilmu Pengetahuan Alam, And U. Negeri Yogyakarta, “Penerapan Metode Bayesian Dalam Model Latent Dirichlet Allocation Di Media Sosial Application Of Bayesian Methods In Latent Dirichlet Allocation Model In Social Media,” 2018. [Online]. Available: Https://T.Co/W03xq01cwb

[9] N. Agus And S. Er, “Implementasi Latent Dirichlet Allocation (Lda) Untuk Klasterisasi Cerita Berbahasa Bali,” Vol. 8, No. 1, Pp. 127–134, 2021, Doi: 10.25126/Jtiik.202183556.

[10] Y. W. Syaifudin And R. A. Irawan, “Implementasi Analisis Clustering Dan Sentimen Data Twitter Pada Opini Wisata Pantai Menggunakan Metode K-Means,” 2018.

[11] N. Widya Utami, I. Gede, And J. E. Putra, “Text Minig Clustering Untuk Pengelompokan Topik Dokumen Penelitian Menggunakan Algoritma K-Means Dengan Cosine Similarity,” 2022.

[12] U. Saidata Aesyi, “Analisis Tingkat Kebermanfaatan Mypertamina Menggunakan K-Means Clustering,” 2023.