# Topic Modelling Latent Dirichlet Allocation untuk Klasifikasi Komentar Kuliah Pada Twitter X

Bambang Subeno[1]

bangsubeno@telkomuniversity.ac.id

[1] Telkom University

***Abstract***

*In recent years, social media data analysis has become crucial to understand public opinion, feedback, and engagement in various domains, including education. This study uses the Latent Dirichlet Allocation (LDA) method to perform topic modeling on college-related comments uploaded on Twitter X. The goal is to classify these comments into meaningful topics to make it easier for educators and institutions to identify areas of interest, concern, and overall sentiment. From the dataset of Twitter comments X, as many as 1200 comments related to college, four topics have been produced, labeled topic-0: "Tuition Fees", topic-1: "College Scholarships", topic-2: "Universities in Bandung", and topic-3: "Student Activities". The highest probability of the word from the four topics is the word "college" at 0.046. Based on these results, the most searched and discussed in social media are related to financing, scholarships, universities and student activities. For this reason, with these results, it is hoped that Educational Institutions need to pay attention to promotional content media related to the four topics properly.*

***Keywords***: clasification, LDA, topic modeling

**Abstrak**

Dalam beberapa tahun terakhir, analisis data media sosial menjadi krusial untuk memahami opini publik, umpan balik, dan keterlibatan dalam berbagai domain, termasuk pendidikan. Penelitian ini menggunakan metode Latent Dirichlet Allocation (LDA) untuk melakukan pemodelan topik pada komentar terkait kuliah yang diunggah di twitter X. Tujuannya adalah untuk mengklasifikasikan komentar-komentar ini ke dalam topik-topik yang bermakna untuk memudahkan pendidik dan lembaga dalam mengidentifikasi bidang minat, perhatian, dan sentimen secara keseluruhan. Dari dataset kumpulan komentar twitter X sebanyak 1200 komentar yang berkaitan tentang kuliah telah menghasilkan empat topic yang diberi label topic-0: “Biaya Kuliah”, topic-1: “Beasiswa Kuliah”, topic-2: “Universitas di Bandung”, dan topic-3: “Kegiatan Mahasiswa”. Probabilitas kata yang paling tinggi dari ke-empat topic adalah kata “kuliah” sebesar 0.046. Berdasarkan hasil tersebut bahwa dalam media social yang paling banyak dicari dan diperbicangkan adalah terkait pembiayaan, beasiswa, universitas dan kegiatan mahasiswanya. Untuk itu dengan hasil ini diharapkan para Lembaga Pendidikan perlu memperhatikan media konten promosi terkait empat topic tersebut dengan baik.

***Kata kunci***: klasifikasi, LDA, *topic modeling*

## Pendahuluan

Media sosial telah menjadi sumber utama komunikasi dan informasi bagi banyak orang di era komputer dan internet saat ini. Platform seperti Twitter (X) memungkinkan orang berinteraksi, berpikir, dan berbicara tentang berbagai hal, seperti pendidikan, kuliah, ekonomi, budaya. Munculnya platform media sosial seperti Twitter yang berganti nama menjadi X telah merevolusi cara individu berkomunikasi dan berbagi informasi. Dalam konteks pendidikan, mahasiswa sering menggunakan platform ini untuk membahas kuliah, berbagi wawasan, dan mengungkapkan pendapat mereka. Menganalisis komentar-komentar ini dapat memberikan umpan balik yang berharga bagi para pendidik dan lembaga.
Seiring dengan volume data yang dihasilkan dari diskusi tersebut semakin tumbuh menjadi data yang besar, maka pendidik, lembaga akan mengalami kesulitan intisari dari pembicaraan melalui komentar tersebut. Oleh karena itu, dibutuhkan analisis untuk melakukan klasifikasi komentar. Salah satu pendekatan yang efektif untuk menganalisis data teks adalah Topic Modeling (Vulić, De Smet, Tang, & Moens, 2015) (Hua, Nambiar, & Kemp, 2020)

Topic Modeling LDA merupakan teknik statistik yang digunakan untuk menemukan topik abstrak dalam kumpulan dokumen (Blei & David, 2012). LDA merupaka metode yang paling banyak digunakan untuk pemodelan topik karena kesederhanaan dan efektivitasnya dalam mengungkap topik laten. LDA telah diterapkan secara luas dalam berbagai domain, mulai dari menganalisis artikel berita (Alghamdi & Alfalqi, 2015) hingga publikasi ilmiah (Griffiths & Steyvers, 2004), dan yang terbaru, konten media sosial (Kaur & Singh, 2019). Dalam konteks media sosial, para peneliti telah memanfaatkan LDA untuk mengekstrak topik dari platform seperti Twitter, mengidentifikasi pola dalam diskusi dan sentimen pengguna. Seperti, memanfaatkan LDA untuk menganalisis komentar Twitter selama acara berlangsung, mengidentifikasi tema utama dan sentimen public (Zhao, et al., 2011). menerapkan LDA pada data Twitter untuk mengidentifikasi topik yang terkait dengan berbagai komunitas pengguna dan minat mereka (Hong & Davison, 2010).

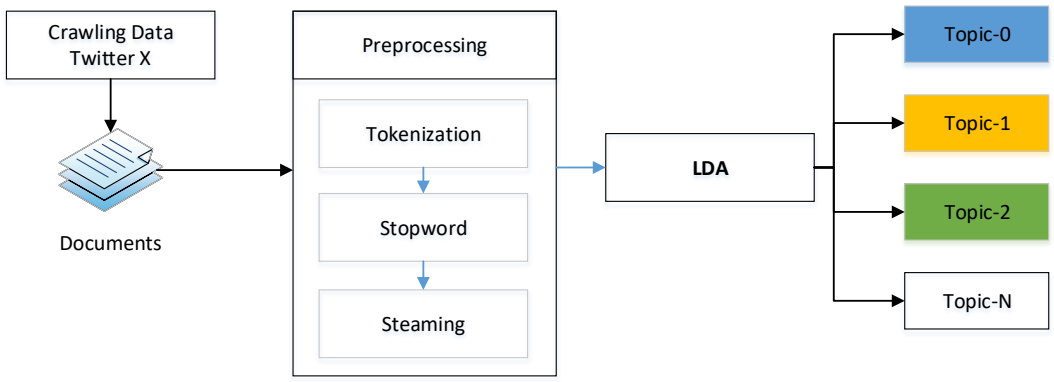
Klasifikasi data teks menggunakan pendekatan pemodelan topik dapat dilakukan, seperti Penggunaan LDA untuk meningkatkan kinerja klasifikasi dokumen dengan menggabungkan distribusi topik sebagai fitur untuk pengklasifikasi (Meng, Huang, Li, & Gu, 2012)**,** dan menggabungkan LDA dengan support vector machines (SVM) untuk mengklasifikasikan email phishing memiliki akurasi lebih tinggi dibandingkan dengan metode tradisional (Bursztein, Benko, Pietraszek, & Grier, 2014). Dalam konteks pendidikan, pemodelan topik telah diterapkan untuk menganalisis umpan balik siswa, evaluasi kursus, dan makalah akademis (Sun, Cheng, Sun, & Li, 2017) menggunakan LDA untuk menganalisis evaluasi kursus, mengidentifikasi tema-tema utama yang terkait dengan kualitas pengajaran dan konten kursus. LDA untuk meneliti makalah akademis, mengungkap tren dan bidang penelitian yang muncul dalam Pendidikan (Nelson, Burk, Knudsen, & McCall, 2018).

Berdasarkan latar belakang tersebut, Penelitian ini berfokus pada pemanfaatan metode topic modelling Latent Dirichlet Allocation (LDA) untuk

mengklasifikasikan dan menganalisis komentar kuliah di pada Twitter X dengan tujuan mengetahui topik yang tersembunyi dari berbagai komentar yang ada.

## Metode Penelitian

Pada tahap ini membahas metode yang digunakan untuk melakukan klasifikasi data teks komentar twitter X. Tahapan proses yang dilakukan yaitu scarping data, preprocessing, penentuan jumlah topik K, proses LDA. Untuk lebih jelasnya dapat dilihat pada Gambar 1.




Gambar 1. Proses Tahapan Klasifikasi topik

Berdasarkan gambar 1 proses awal yang dilakukan adalah melakukan crawling data yang berasal dari twitter X (https://x.com), kemudian menghasilkan koleksi data dokumen, dilakukan preprocessing, setelah selesai preprocessing dokumen komentar diolah menggunakan pemodelan topic LDA.

## Dataset

Data yang digunakan adalah data hasil crawling twitter X, untuk query search yang digunakan menggunakan kata “minat kuliah”. Berdasarkan hasil crawling data dihasilkan kumpulan data 1200 komentar. Satu komentar direpresentasikan sebagai satu dokumen.

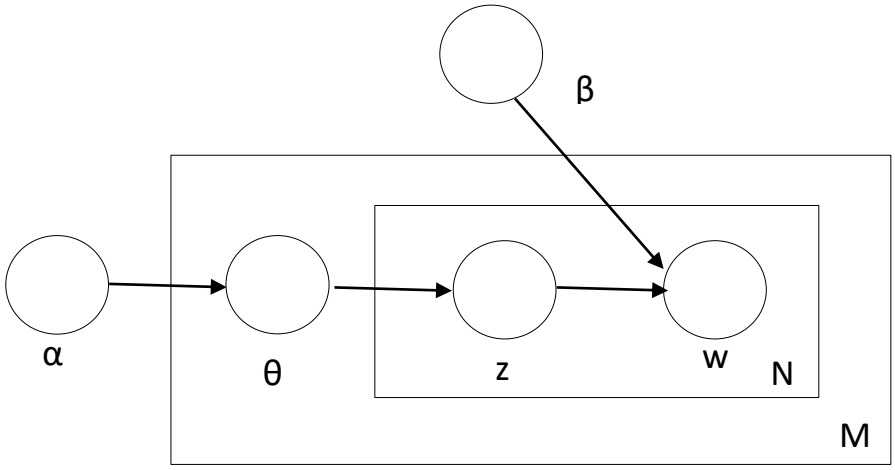
## Preprocessing

Pada tahap ini dilakukan pemrosesan data komentar, komentar akan dibentuk menjadi token-token kata, kemudian kata-kata yang dianggap tidak penting akan dihapus dan kata-kata yang dianggap penting kemudian dibentuk menjadi kata dasar.

## Model LDA

Model LDA yang digunakan dapat dilihat pada gambar 2. Pada gambar 2 digambarkan representasi LDA untuk menemukan topik tersembunyi didalam suatu dokumen, dalam satu kali prosedur inference dapat dihasilan topic dokumen dan

topic dari berbagai dokumen. Setiap topik terdiri dari kumpulan kosakata yang memiliki keterkaitan kata yang sama (Blei, Ng, & Jordan, 2003) (Blei & David, 2012) (Li, et al., 2020).




Gambar 2. Representasi LDA Model **(LI, ET AL., 2020)**

dengan :

D : Dokumen korpus
K : Jumlah topik
k : Topik
N : Vocabulary
$\alpha$ : Nilai *Hyperparameter* untuk proporsi topik untuk dokumen ke-d
$\beta$ : Nilai *Hyperparameter* untuk probabilatas kata-topik dari kata w di dalam vocabulary
$w_{d,i}$ : kata w pada token (d,i) yaitu token kata ke-i pada dokumen korpus ke-d
$z_{d,i}$ : topik dari tiap kata yang ada dalam dokumen

## Hasil dan Pembahasan

Pada proses ekseperiman, digunakan google colab untuk melakukan proses sesuai tahapan pada gambar 1. Spesifikasi komputer yang digunakan Core(TM) i5-13420H memory 8 GB. Pada tahap crawling data dihasiljan 1200 komentar yang dijadikan sebagai dataset untuk dilakukan proses ekseperimen. Untuk tahap preprocessing menggunakan library sastrawi, karena dataset yang digunaakan dalam bentuk Bahasa Indonesia. Setelah dilakukan preprocessing dan proses LDA maka dihasilkan empat topik dengan representasi kata dan probabilitasnya dapat dilihat pada table 1.

Table 1 Hasil probabilitas topik terhadap kata

| Topic-0 | Topic-1 | Topic-2 | Topic-3 |
|---|---|---|---|
| kuliah (0.034) | kuliah (0.046) | kuliah (0.041) | bandung (0.039) |
| bandung (0.033) | bandung (0.042) | bandung (0.041) | kuliah (0.039) |
| mau (0.012) | ga (0.010) | dana (0.012) | actually (0.028) |
| ga (0.010) | dulu (0.007) | jadi (0.008) | you (0.017) |
| ken (0.008) | anak (0.007) | mau (0.007) | kalo (0.007) |
| tp (0.008) | sama (0.006) | pas (0.007) | look (0.006) |
| sama (0.007) | pas (0.006) | sama (0.005) | from (0.006) |
| buat (0.006) | kerja (0.006) | pernah (0.005) | sama (0.006) |
| bgt (0.005) | buat (0.006) | biaya (0.005) | status (0.006) |
| dulu (0.005) | gak (0.005) | kampus (0.005) | are (0.006) |
| selalu (0.005) | aja (0.005) | ga (0.004) | seem (0.006) |
| gw (0.004) | beli (0.005) | aja (0.004) | ga (0.006) |
| waktu (0.004) | jadi (0.005) | banget (0.004) | mau (0.005) |
| enak (0.004) | kakak (0.005) | kerja (0.004) | aja (0.005) |
| jg (0.004) | kalo (0.004) | kota (0.004) | gw (0.004) |
| tpi (0.004) | mau (0.004) | anak (0.004) | bgt (0.004) |
| lulus (0.004) | jakarta (0.004) | nya (0.004) | temen (0.004) |
| pindah (0.003) | nyata (0.004) | buat (0.003) | terus (0.004) |
| keluarga (0.003) | banget (0.004) | daftar (0.003) | dah (0.004) |
| gak (0.003) | padahal (0.004) | rumah (0.003) | apa (0.004) |

Pada table 1 ditampilkan 20 probalitas kata tertinggi, terlihat bahwa untuk topic-0 probalitas kata “kuliah” adalah 0.034, untuk topic-1 probalitas kata “kuliah” adalah 0.046, untuk topic-2 probalitas kata “kuliah” adalah 0.041 dan untuk topic-3 probalitias kata “bandung” adalah 0.039. Kumpulan kata yang terdapat pada topic menggambarkan topic yang berbeda dengan topic yang lainnya, Gambaran setiap topic digambarkan dalam bentuk word cloud topic sesuai gambar 3. Untuk memperjelas topic apa, maka dilakukan pelabelan setiap topic berdasarkan kumpulan kata yang terkandung didalamnya. Hasil pelabelan topic dapat dilihat pada table 2.

Table 2 Hasil Pelabelan Setiap Topic

| Topic | Nama Pelabelan Topic |
|---|---|
| Topic-0 | "Biaya Kuliah” |
| Topic-1 | “Beasiswa Kuliah” |
| Topic-2 | “Universitas di Bandung” |
| Topic-3 | “Kegiatan Mahasiswa” |


Gambar 3. Word Cloud Topic

## Kesimpulan

Penggunaan pemodelan topik dengan LDA untuk menganalisis komentar kuliah di media social twitter X dapat diimplementasikan dan menghasilkan klasifikasi topic yang relevan, dengan mengetahui klasifikasi topic berdasarkan table 2 para pendidik dan institusi dapat memperoleh wawasan berharga tentang umpan balik mahasiswa, sehingga memungkinkan keputusan berdasarkan data untuk meningkatkan layanan Pendidikan yang lebih baik. Berdasarkan table 1 dihasilkan empat topic dengan masing-masing probalibilat kata tertinggi yaitu topic-0 probabilitas kata “kuliah” 0.034, topic-1 probabilitas kata “kuliah” 0.046, topic-2 probabilitas kata “kuliah” 0.041, topic-3 probabilitas kata “bandung” 0.039. Kumpulan kata dari probabilitas tertinggi sampai terkecil digambarkan dalam bentuk word cloud untuk lebih memhami kata yang sering digunakan dalam komentar. Dari dataset komentar dihasilkan topic dengan label “Biaya Kuliah”, “Beasiswa Kuliah”, “Universitas di Bandung”, dan “Kegiatan Mahasiswa”. Penulis menyadari bahwa penelitian ini masih perlu dikembangkan kedepannya untuk dapat membandingkan dengan model topic modeling terbaru seperti BERTopic.

## Daftar Pustaka

Alghamdi, R., & Alfalqi. (2015). A survey of topic modeling in text mining. *International Journal of Advanced Computer Science and Applications*.

Blei, & David, M. (2012). Probabilistic Topic Models. *Communication of The ACM, 55*(4), 77-84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*, 993-1022.

Bursztein, E., Benko, T., Pietraszek, T., & Grier. (2014). Using Latent Dirichlet Allocation to Filter Spam. *In LEET*.

Griffiths, L., & Steyvers. (2004). Finding scientific topics. Proceedings of the National Academy of Sciences. *Proceedings of the National Academy of Sciences*.

Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in Twitter. *Proceedings of the First Workshop on Social Media Analytics*.

Hua, X., Nambiar, A., & Kemp, R. (2020). Topic modeling for healthcare tweets: An exploratory study. *International Journal of Environmental Research and Public Health*.

Kaur, H., & Singh. (2019). A survey of topic modeling techniques and trends. *Journal of Information Science*.

Li, B., Wang, Y., Li, X., Chen, Q., Bao, J., & Zheng, T. (2020). Characteristics and Evolution of Citation Distance Based on LDA Method. *Advances in Intelligent Systems and Interactive Applications (IISA2019)* (pp. 303-311). Springer.

Meng, X., Huang, S., Li, B., & Gu, J. (2012). Enhancing Text Classification Performance Using Latent Dirichlet Allocation and Support Vector Machines. *Journal of Computer Science and Technology*.

Nelson, L. K., Burk, D., Knudsen, M., & McCall, L. (2018). The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. *Sociological Methods & Research*.

Sun, Z., Cheng, X., Sun, D., & Li, H. (2017). A novel method for analyzing course evaluations using topic modeling. *IEEE Transactions on Education*.

Vulić, I., De Smet, W., Tang, J., & Moens, M. F. (2015). Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Information Processing & Management*.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., & Li, X. (2011). Comparing Twitter and traditional media using topic models. *Advances in Information Retrieval*.