

## **PENGELOMPOKAN TOPIK CUITAN PENGGUNA TWITTER TERHADAP KULIAH KERJA NYATA (KKN) DI INDONESIA MENGUNAKAN *LATENT DIRICHLET ALLOCATION***

<sup>1</sup>Anis Fitri Nur Masruriyah

<sup>2</sup>Cici Emilia Sukmawati

<sup>3</sup>Hilda Yulia Novita

Teknik Informatika, Fakultas Ilmu Komputer, Universitas Buana Perjuangan Karawang  
anis.masruriyah@ubpkarawang.ac.id<sup>1</sup>, cici.emilia@ubpkarawang.ac.id<sup>2</sup>,  
hildayulia@ubpkarawang.ac.id<sup>3</sup>

### **ABSTRAK**

Kuliah Kerja Nyata (KKN) menjadi kegiatan rutin yang dilakukan setiap tahun oleh universitas di Indonesia. Peserta KKN terdiri dari mahasiswa aktif yang telah menempuh kurang lebih 100 Satuan Kredit Semester (SKS). Saat ini, tinjauan literatur eksplorasi manual harus menjadi sesuatu dari masa lalu, karena teknologi dan pengembangan metode *Machine Learning* telah matang. Kurva pembelajaran untuk menggunakan metode pembelajaran mesin menurun dengan cepat, memungkinkan kemungkinan baru bagi semua peneliti. Kerangka kerja disajikan tentang bagaimana menggunakan pemodelan topik pada koleksi besar makalah untuk tinjauan literatur eksplorasi dan bagaimana itu dapat digunakan untuk tinjauan literatur lengkap. Tujuan dari naskah ilmiah untuk membangun pemodelan topik terkait KKN bagi para peneliti dengan menyajikan kerangka langkah-demi-langkah pada kasus dan berbagi template kode. Kerangka kerja ini terdiri dari tiga langkah; pra-pemrosesan, pemodelan topik, dan pasca-pemrosesan, di mana model topik *Latent Dirichlet Allocation* digunakan. Kerangka kerja ini memungkinkan sejumlah besar makalah untuk ditinjau dengan cara yang transparan, andal, lebih cepat, dan dapat direproduksi.

Kata kunci: Kuliah Kerja Nyata, Latent Dirichlet Allocation, Mahasiswa, Pemodelan Topik

### **ABSTRACT**

*Kuliah Kerja Nyata* (KKN) is routine activities carried out annually by universities in Indonesia. KKN participants consist of active students who have taken approximately 100 Semester Credit Units (SKS). Nowadays, the manual exploratory literature review should be a thing of the past, as the technology and development of Machine Learning methods have matured. The learning curve for using machine learning methods is decreasing rapidly, allowing new possibilities for all researchers. A framework was presented on how to use topic modelling on a large collection of papers for an exploratory literature review and how it can be used for a full literature review. The purpose of the scientific paper is to build modelling of KKN-related topics for researchers by presenting a step-by-step framework for cases and sharing code templates. This framework consists of three steps; pre-processing, topic modelling, and post-processing, where the Latent Dirichlet Allocation topic model is used. This framework allows a large number of papers to be reviewed in a transparent, reliable, faster and reproducible manner.

Keywords: *Kuliah Kerja Nyata*, Latent Dirichlet Allocation, Student, Topic Modeling  
Karawang, 24 Maret 2022

## PENDAHULUAN

KKN adalah salah satu model pengabdian kepada masyarakat yang dilaksanakan Perguruan Tinggi (Syardiansah 2019). Kegiatan ini dapat menjadi wahana penerapan ilmu pengetahuan dan hasil-hasil penelitian dalam upaya memberikan sumbangan dalam pemecahaan masalah yang ada di masyarakat. Adanya KKN menunjukkan bahwa Perguruan Tinggi memiliki kepedulian dan berperan nyata kepada masyarakat, baik masyarakat sekitar kampus atau masyarakat luas. Kegiatan KKN yang telah terlaksana umumnya didokumentasikan dan diunggah pada media sosial dalam berbagai *platform*. Kemudian, Twitter menjadi salah satu sosial media yang digunakan untuk menceritakan kegiatan KKN. Twitter adalah salah satu media sosial dan microblog daring yang memungkinkan penggunaanya untuk mengirim dan membaca pesan berbasis teks hingga 280 karakter, yang dikenal dengan sebutan cuitan (tweet) (Ratnawati 2018). Mahasiswa yang menggunakan Twitter banyak yang membuat utas tentang kegiatan KKN, inovasi yang dilakukan, serta menceritakan hal-hal yang terjadi selama KKN. Utas di Twitter sering mendapatkan perhatian khusus dari masyarakat (Arief dan Imanuel 2019; Felicia dan Loisa 2019; Liang *et al.* 2019; Gruzd dan Mai 2020) sehingga diperlukan pengelompokan topik untuk memahami utas yang ada di Twitter memiliki manfaat, berita bohong atau tidak.

Penelitian sebelumnya (Curiskis *et al.* 2020) telah melakukan pemodelan topik pada Twitter dan Reddit untuk memahami volume dari konten yang dibuat pengguna. Penelitian tersebut mengevaluasi beberapa teknik untuk pengelompokan dokumen dan pemodelan topik pada tiga kumpulan data dari Twitter dan Reddit. Selanjutnya, membandingkan empat representasi fitur berbeda yang diturunkan dari matriks *Term-Frequency Inverse-Document-Frequency* (TF-IDF) dan penyisipan kata yang dikombinasikan dengan empat metode pengelompokan, dan menyertakan model topik *Latent Dirichlet Allocation* (LDA) untuk perbandingan. Hasil penelitian tersebut menunjukkan bahwa teknik pengelompokan yang diterapkan pada representasi fitur penyematan saraf memberikan kinerja terbaik di atas semua kumpulan data menggunakan ukuran evaluasi ekstrinsik yang sesuai.

Selanjutnya, Negara *et al.* (2019) juga menggunakan cuitan pada Twitter untuk mendapatkan informasi terkait olahraga. Pada penelitian tersebut LDA digunakan untuk menghasilkan pemodelan topik, kesamaan topik masing-masing, dan visualisasi pengelompokan topik dari data tweet. Penelitian tersebut mendapatkan model pengelompokan

topik yang relevan dengan olahraga yaitu Ekonomi, Militer, Olahraga, Teknologi dalam bahasa Indonesia. Metode LDA yang digunakan dalam pengolahan data tweet berhasil dilakukan dan bekerja secara optimal, pada setiap ekstraksi topik, pemodelan topik, pembangkitan kata indeks yang berada pada setiap cluster topik dan visualisasi komputer pada topik tersebut. pengindeksan kata dalam topik Olahraga dengan 1260 tweet dengan akurasi 98% lebih baik daripada metode LSI dalam Pemodelan Topik.

Kemudian, terdapat penelitian yang mengevaluasi efek samping dari vaksin COVID-19 berdasarkan cuitan di Twitter (Habibabadi dan Haghghi 2019). Penelitian tersebut bertujuan untuk mengevaluasi model topik untuk mengidentifikasi posting pengguna di Twitter yang kemungkinan besar mencuitkan efek samping setelah melakukan vaksin. Teknik yang digunakan difokuskan pada identifikasi desain model dan jumlah topik yang paling baik mengungkapkan dokumen yang berisi sinyal keamanan vaksin, untuk membantu pengurangan dimensi dan pelabelan data teks selanjutnya. Studi ini membandingkan model Gensim LDA, MALLET, dan jLDADMM DMM untuk menentukan model yang paling efektif untuk mendeteksi sinyal keamanan vaksin, dibantu oleh proses evaluasi yang menggunakan teknik F-Scoring yang disesuaikan pada subset dokumen yang berlabel.

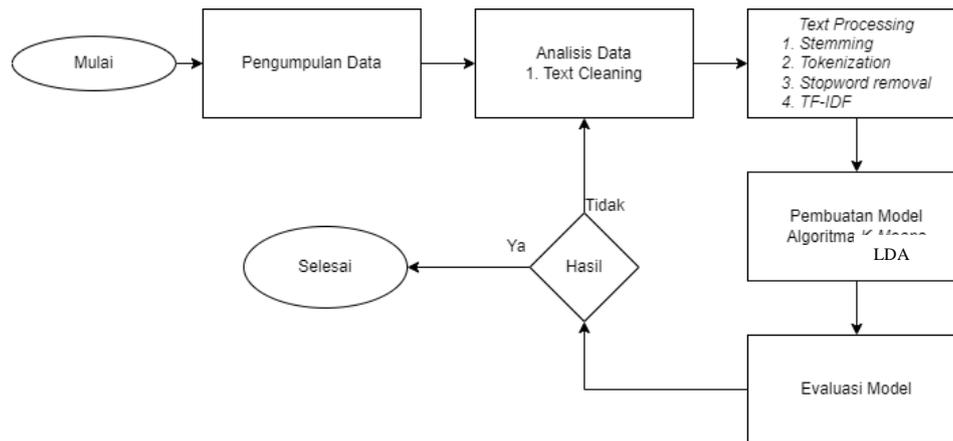
Berdasarkan masalah yang telah dipaparkan dan bukti penelitian yang relevan, maka penelitian ini bertujuan untuk membuat model pengelompokan topik terkait kegiatan KKN. Selanjutnya algoritma LDA digunakan untuk menghasilkan pemodelan topik, kesamaan topik masing-masing, dan visualisasi pengelompokan.

## **METODE PENELITIAN**

### **Waktu dan Tempat Penelitian**

Penelitian ini dilakukan sejak bulan Juli 2022 hingga Agustus 2022 di Laboratorium Penelitian Universitas Buana Perjuangan Karawang.

### **Prosedur Penelitian**



Gambar 1 Prosedur Penelitian

Berdasarkan Gambar 1 penelitian dimulai dengan mengumpulkan data cuitan dari Twitter, kemudian dilakukan pembersihan dokumen dan pra-pemrosesan data. Pada pra-pemrosesan data *stemming* berfungsi untuk mengubah semua kata menjadi kata dasar, kemudian *tokenization* dilakukan untuk memecah kalimat menjadi kata-kata. Selanjutnya, kata-kata umum yang tidak bermakna dan tanda baca dihilangkan dengan *stopword removal*, dan terakhir TF-IDF digunakan untuk mengetahui frekuensi kata muncul dalam keseluruhan cuitan Twitter. Setelah data siap tahapan selanjutnya membuat model dengan LDA dan dilakukan evaluasi. *Pseudocode* untuk cara kerja LDA ditunjukkan pada Gambar 2.

```

Input:  $S$  (instance set),  $K$  (number of clusters),  $T_0$  (initial temperature),
 $T_f$  (final temperature),  $c$  (temperature reducing constant)
Output: clusters
1: Randomly select  $p_0$  which is a  $K$ -partition of  $S$ . Compute the squared
error value  $E(p_0)$ .
2: while  $T_0 > T_f$  do
3:   Select a neighbor  $p_1$  of the last partition  $p_0$ .
4:   if  $E(p_1) < E(p_0)$  then
5:      $p_0 \leftarrow p_1$  with a probability that depends on  $T_0$ 
6:   else
7:      $p_0 \leftarrow p_1$ 
8:   end if
9:    $T_0 \leftarrow c * T_0$ 
10: end while
    
```

Gambar 2 Pseudocode LDA

## HASIL PENELITIAN DAN PEMBAHASAN

Hasil pemodelan topik ditunjukkan pada Tabel 1 dengan 5 topik yang saling berkaitan. Berdasarkan hasil pengumpulan data, ditemukan korelasi antara film dengan judul “KKN Desa Penari” yang diproduksi berdasarkan utas yang ada pada Twitter. Film tersebut memiliki *genre* horror sehingga bersamaan dengan hal tersebut terjadi keterkaitan topik dengan film “Pengabdian Setan” yang dalam waktu bersamaan rilis.

Tabel 1 Pemodelan Topik

No.	Kata Kunci Topik
1.	KKN, Desa, Penari, Nonton, Film, Horor
2.	KKN, Desa, Penari, Film, Pengabdi, Setan
3.	KKN, Desa, Penari, Kayak
4.	KKN, Desa, Penari, Pengabdi, Setan, Penonton
5.	KKN, Desa, Penari, Setan, Pengabdi, Film

Selanjutnya, ditemukan kumpulan kata yang saling bertaut dengan kata kunci “Kuliah Kerja Nyata” atau “KKN” yang ditunjukkan pada Gambar 3 dalam bentuk awan kata (*Word Cloud*). Nampak kata yang dominan adalah “KKN” dan diikuti kata “Pengabdi”, “Penari” dan “Nonton”. Hal ini menandakan bahwa kegiatan KKN menjadi cukup populer di Twitter berkaitan dengan film horror yang menyertainya.



Gambar 3 Hasil Awan Kata (*WordCloud*)

## KESIMPULAN DAN IMPLIKASI

Berdasarkan hasil pengelompokan topik, LDA terbukti mampu mengelompokkan 5 topik yang berkaitan dengan KKN. Namun, penelitian ini belum mampu membedakan kalimat sarkas

dalam cuitan, sehingga hasil pengelompokan masih terlalu umum. Maka untuk penelitian selanjutnya, disarankan melakukan pengembangan terkait pendeteksi kelimat sarkas.

#### DAFTAR PUSTAKA

- Arief R, Imanuel K. 2019. Analisis Sentimen Topik Viral Desa Penari Pada Media Sosial Twitter Dengan Metode Lexicon Based. *J Ilm Matrik*. 21(3):242–250. doi:10.33557/jurnalmatrik.v21i3.727.
- Curiskis SA, Drake B, Osborn TR, Kennedy PJ. 2020. An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. *Inf Process Manag*. 57(2):102034. doi:10.1016/j.ipm.2019.04.002.
- Felicia, Loisa R. 2019. Peran Buzzer Politik dalam Aktivitas Kampanye di Media Sosial Twitter. *Koneksi*. 2(2):352. doi:10.24912/kn.v2i2.3906.
- Gruzd A, Mai P. 2020. Going viral: How a single tweet spawned a COVID-19 conspiracy theory on Twitter. *Big Data Soc*. 7(2). doi:10.1177/2053951720938405.
- Habibabadi SK, Haghighi PD. 2019. Topic Modelling for Identification of Vaccine Reactions in Twitter. *ACM Int Conf Proceeding Ser.*, siap terbit.
- Liang H, Fung ICH, Tse ZTH, Yin J, Chan CH, Pechta LE, Smith BJ, Marquez-Lameda RD, Meltzer MI, Lubell KM, *et al*. 2019. How did Ebola information spread on twitter: Broadcasting or viral spreading? *BMC Public Health*. 19(1):1–11. doi:10.1186/s12889-019-6747-8.
- Negara ES, Triadi D, Andryani R. 2019. Topic Modelling Twitter Data with Latent Dirichlet Allocation Method. *ICECOS 2019 - 3rd Int Conf Electr Eng Comput Sci Proceeding.*, siap terbit.
- Ratnawati F. 2018. Implementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter. *INOVTEK Polbeng - Seri Inform*. 3(1):50. doi:10.35314/isi.v3i1.335.
- Syardiansah S. 2019. Peranan Kuliah Kerja Nyata Sebagai Bagian Dari Pengembangan Kompetensi Mahasiswa. *JIM UPB (Jurnal Ilm Manaj Univ Puter Batam)*. 7(1):57–68. doi:10.33884/jimupb.v7i1.915.