TOPIC ANALYSIS VIDEO DEBAT JELANG PEMILU PRESIDEN DAN WAKIL PRESIDEN TAHUN 2024 DENGAN LATENT DIRICHLET ALLOCATION

Ivana Valentina[1)], Aziz Mu'min[2)], Devion Tanrico[3)], Oscar Karnalim[4)]

[1,2,3,4]Magister Ilmu Komputer
[1,2,3,4]Fakultas Teknologi Informasi
[1,2,3,4]Universitas Kristen Maranatha
E-mail : 22790006@maranatha.ac.id [1)], 2279010@maranatha.ac.id [2)], 2279008@maranatha.ac.id [3)], oscar.karnalim@it.maranatha.edu [4)]

Abstract

Several debates were held among presidential and vice presidential candidates to convey their ideas for the 2024 presidential general election (PEMILU). This research analyzes the topics discussed in the debates using Latent Dirichlet Allocation (LDA), K-means Clustering, and word tagging methods for each candidate pair. The K-Means Clustering method yielded more diverse and evenly distributed topics for each candidate pair, while LDA produced fewer topics but was more effective in identifying topics for candidate A. The K-Means Clustering method yielded more diverse and evenly distributed topics for each candidate pair, while LDA produced fewer topics but was more effective in identifying topics for candidate. These are somewhat consistent with previos works. A. In dataset 1 using the LDA model, candidate pairs A have a probability of 60%, B have a probability of 25%, and C have a probability of 0%. In dataset 2 using the K-Means model, candidate pairs A have a probability of 37.04%, B have a probability of 25%, and C have a probability of 17.24%. In dataset 2 using the LDA model, candidate pairs A have a probability of 100%, B have a probability of 40%, and C have a probability of 0%. In dataset 2 using the K-Means model, candidate pairs A have a probability of 35.71%, B have a probability of 14.29%, and C have a probability of 28.57%.

Keywords *- Topic Modeling, Latent Dirichlet Allocation (LDA), K-Means*

Intisari

Beberapa debat calon presiden dan wakil presiden kerap dilaksanakan menjelang pemilihan umum presiden (PEMILU) sebagai sarana menyampaikan gagasan-gagasan. Untuk dapat menemukan topik dari gagasan-gagasan yang ada dalam debat bakal calon presiden dan wakil presiden yang ditayangkan dalam video debat, penelitian ini melakukan analisis topik terhadap video debat calon presiden dan wakil presiden dengan menggunakan metode Latent Dirichlet Allocation (LDA), K-means Clustering,* dan *tagging* kata*, untuk masing-masing pasangan calon: A (Anies-Muhaimin), B (Prabowo-Gibran)), dan C (Ganjar-Mahfud). Hasilnya dengan menggunakan K-Means Clustering topik yang didapatkan lebih banyak, beragam, dan lebih merata topiknya untuk setiap paslon. Sedangkan topik LDA mendapatkan topik lebih sedikit dan lebih berhasil mendapatkan topik untuk paslon A. Berdasarkan standar yang telah dibuat, efektivitas pencarian topik berbeda-beda keberhasilannya disetiap dataset. Hal ini cukup konsisten dengan penelitian-penelitian terdahulu. Pada dataset 1 dengan model LDA didapati gagasan pasangan calon untuk paslon A 60%, B 25%, dan C 0%. Pada dataset 2 dengan model K-Means didapati gagasan pasangan calon untuk paslon A 37.04%, B 25%, dan C 17.24%. Pada dataset 2 dengan model LDA didapati gagasan pasangan calon untuk paslon A 100%, B 40%, dan C 0%. Pada dataset 2 dengan model K-Means didapati gagasan pasangan calon untuk paslon A 35.71%, B 14.29%, dan C 28.57%.

Kata Kunci *- Topic Modeling, Latent Dirichlet Allocation (LDA), K-Means*

1. PENDAHULUAN

Mendekati Pemilihan Umum Presiden memang banyak diadakan debat bahkan sebelum pendaftaran calon presiden. Debat-debat ini dibuat sebagai sarana bakal calon presiden (bacapres) dan bakal calon presiden (bacawapres) untuk memperkenalkan gagasan-gagasan mereka kepada masyarakat. Gagasan-gagasan tersebut yang sering disampaikan berkali-kali dalam kesempatan debat manapun sehingga masyarakat dapat familier dengan gagasan yang ditawarkan. Selain gagasan sering juga digunakan untuk menjelaskan strategi bacapres untuk menang, mulai dari koalisi hingga memilih bacawapres.

DOI : https://doi.org/10.21063/jtif.2024.V12.1.29-35

Penelitian terkait pencarian topik sudah banyak dilakukan. Zulhanif *et. al.* [1] melakukan penelitian untuk *clustering* data *text* dari 1500 *tweet* di *X (Twitter)* dengan #Bandung. *Preprocessing* dilakukan dalam data tersebut. Metode LDA dipakai dalam menemukan *terms* yang paling sering muncul. Hasilnya didapatkan 24 topik. *Fuadi et. al.* [2] melakukan penelitian *automasi* penentuan tren topik skripsi untuk menghasilkan sebuah aplikasi yang dapat mengatasi kendala mahasiswa yaitu kesulitan dalam menentukan topik judul skripsi sehingga lama dalam proses pembuatan proposal skripsi. Penelitian ini menggunakan metode algoritma *K-means clustering*. Aplikasi yang dirancang dapat berjalan dengan baik dengan tingkat akurasi sebesar 84% dari 70 data uji. Setiawan *et. al.* [3] melakukan penelitian yang bertujuan untuk mengatasi kendala dalam mengelola aduan mahasiswa di perguruan tinggi dengan menerapkan model *Latent Dirichlet Allocation* (LDA). Hasil penelitian menunjukkan bahwa model LDA memiliki kinerja yang baik.

Penelitian ini juga hendak menemukan topik. Topik yang ingin dicari adalah kata-kata gagasan yang sering dikeluarkan oleh pasangan calon dalam menyongsong Pemilihan Umum Presiden Tahun 2024. Gagasan yang dimaksud tidak hanya dikatakan langsung oleh pasangan calon tetapi oleh perwakilan/juru bicara dari pasangan calon.

Dalam mencapai tujuan penelitian, penelitian ini akan menganalisis kemunculan kata yang sering dipergunakan dalam debat pasangan calon. Kata-kata yang sering muncul ini diasumsikan sebagai gagasan dari pasangan calon. Data yang akan dipergunakan adalah data percakapan debat pasangan calon.

Dalam mengetahui spesifik gagasan masing-masing pasangan calon, penelitian ini akan menganalisis kata yang sering muncul dalam debat pasangan calon dengan *tagging* kata [4] yang sering muncul tersebut diucapkan oleh calon itu sendiri dan perwakilan/juru bicara dari pasangan calon masing-masing.

Pada penelitian ini tidak ada unsur memihak ke pasangan calon manapun. Penelitian murni untuk menemukan topik yang sering disebutkan setiap pasangan calon di setiap kesempatan debat menjelang Pemilu 2024 untuk kepentingan ilmu pengetahuan.

Penelitian terkait pencarian topik sering dilakukan dalam berbagai kesempatan. Setijohatmo *et. al.* [5] melakukan penelitian untuk menemukan topik yang tersembunyi dalam laporan tugas akhir. Menggunakan metode *Latent Dirichlet Allocation* (LDA). Hasilnya probabilitas sebuah kata bergantung kepada banyaknya jumlah topik dan dokumen. LDA dapat mengelompokkan dokumen dengan topik tertentu namun tidak berlabel.

LDA juga dipakai Tondang *et. al.* [6] melakukan analisis pemodelan topik ulasan pada aplikasi BNI, BCA dan BRI menganalisis hubungan antara teknologi dan *mobile banking* melalui pengelompokan ulasan yang diterima oleh aplikasi. Kesimpulan dari penelitian ini adalah bahwa *topic modelling* dengan LDA adalah metode yang berguna untuk mengevaluasi dan memahami ulasan aplikasi.

Matira *et. al.* [7] melakukan penelitian dengan metode LDA, yang bertujuan untuk melakukan ekstraksi topik dari data berita yang terus bertambah di portal berita *online* Detikcom. Dari penelitian ini, ditemukan bahwa jumlah topik yang terbentuk sebanyak 3. Masing-masing topik memiliki karakteristik tertentu yang mencerminkan pola-pola dalam data berita Detikcom. *Coherence score*, yang merupakan metrik untuk mengukur seberapa baik topik-topik tersebut terbentuk, diberikan nilai sebesar 0,7586. Dengan demikian, penelitian ini berhasil mengidentifikasi dan mengelompokkan topik-topik utama yang muncul dalam data berita Detikcom

Rusdhi dan Sari [8] menyebutkan bahwa algoritma LDA adalah untuk menemukan representasi yang paling optimal dari *document topic matrix* dan *topic word matrix* untuk menemukan distribusi dokumen-topik dan distribusi topik-kata yang paling optimal. LDA merepresentasikan kumpulan dokumen sebagai campuran topik-topik dan topik direpresentasikan sebagai campuran kata yang tersembunyi dan belum belum diketahui. LDA adalah algoritma untuk *topic modeling*.

Pada *topic modeling* kumpulan dokumen disebut *corpus* yang selanjutnya direpresentasikan sebagai *Document Term Matrix* (DTM), DTM berorder M x N. M adalah jumlah dokumen, N adalah total kata unik pada semua dokumen. LDA mendekomposisi DTM

DOI : https://doi.org/10.21063/jtif.2024.V12.1.29-35

menjadi dua matrik, *document topic matrix* berorde M x K dan *topic word matrix* K x N. K adalah jumlah topik dalam *corpus* yang nilainya dimasukkan oleh pengguna.

Hudin *et. al.* [9] melakukan pengelompokan berdasarkan tema, objek, dan metode yang tertulis dalam laporan skripsi. Dalam pengelompokan yang dilakukan, diekstraksi dokumen skripsi dengan *text mining* lalu dikelompokkan menggunakan *K-Means Clustering*.

Siringoringo *et. al.* [10] melakukan pemodelan topik berita dengan membandingkan dua model yaitu dengan menggunakan *Latent Dirichlet Allocation* (LDA) dan *K-Means*. Hasil penelitian menunjukkan bahwa baik LDA maupun KMC mampu melakukan pemodelan topik berita dengan baik.

Pada penelitian ini juga digunakan LDA dan *K-Means Clustering* sebagai metode pencarian topik untuk gagasan yang ada dalam data debat. Tahapan penelitian lebih jelasnya adalah sebagai berikut:

2. METODOLOGI

Bagian ini berisi penjelasan mengenai berbagai tahap metodologi dan pendekatan yang digunakan dalam setiap tahap. Penelitian dilakukan dalam beberapa tahap, mulai dari tahap *Data Acquisition, Preprocessing, Model designing, Model Execution*, dan *Result Analysis*. Berikut penjelasan dari tahapan-tahapan dalam penelitian ini :

1. Data Acquisition

Pengumpulan data dilakukan dengan pengambilan transkrip debat dari berbagai video debat di *YouTube* dari *channel* yang berbeda-beda. Transkrip dipisahkan dalam baris-baris kalimat yang diucapkan oleh narasumber. Kemudian baris-baris kalimat dipisahkan berdasarkan narasumber yang menjadi wakil dari tiga capares-cawapres yaitu Anies Rasyid Baswedan-Muhaimin Iskandar (label A), Prabowo Subianto-Gibran Rakabuming Raka (label B), dan Ganjar Pranowo-Mahfud MD (label C).

2. Preprocessing Data

Selanjutnya kalimat-kalimat yang ada di dalam baris-baris data dilakukan *preprocessing* yaitu:

a. Case folding untuk menyeragamkan seluruh kata menjadi huruf kecil. Hal ini dilakukan agar sebuah kata, walaupun terdapat huruf besar dan ada variasi huruf kecil, tetap dianggap kata yang sama dan masuk hitungan kemunculan kata.
b. *Remove Punctuation* untuk menghilangkan tanda baca yang tidak akan dihitung sebagai topik.
c. Tokenizing untuk memecah-mecah kata per kata. Hal ini digunakan untuk proses *translate* dan *remove stopwords*.
d. *Translate* untuk mengubah kata yang tidak dalam Bahasa Indonesia, sehingga walaupun diucapkan dalam Bahasa Inggris, kemunculan kata dapat dihitung bersamaan dengan kata Bahasa Indonesianya.
e. *Remove Stopwords* yaitu melakukan penghilangan untuk kata-kata yang tidak memiliki makna sehingga tidak mengganggu perhitungan kata sebagai topik yang sering muncul. Hal ini dilakukan agar topik yang didapatkan benar-benar memiliki makna.
f. *Stemming* untuk membuang imbuhan menjadi kata dasar, sehingga perhitungan kemunculan sebuah kata dapat lebih kolektif dengan dibuangnya imbuhan.
g. *Bigram* dan *Trigram* untuk menghasilkan padanan dua kata dan tiga kata, sehingga didapatkan hitungan kemunculan kata dengan padanan dua atau tiga kata yang bisa jadi memiliki arti berbeda.

3. Model Designing

Dalam menemukan topik dari data debat yang ada maka akan dilakukan pencarian topik menggunakan *Latent Dirichlet Allocation (LDA)*. Sebelum pencarian topik dengan LDA dipersiapkan dulu *Dictionary* berupa *id* dari kata juga frekuensi kata itu keluar, dan juga *Corpus* berupa list dokumen.

Model LDA yang dibuat untuk masing-masing *dataset* yang mewakili pasangan calon dibuat dengan *input Dictionary* dan

DOI : https://doi.org/10.21063/jtif.2024.V12.1.29-35

Corpus yang sudah dibentuk sebelumnya, juga *input* banyak topik yang akan dicari berdasarkan *coherence score* tertinggi.

Dictionary di Model LDA memerlukan *parameter* berupa *Maximum Document Frequency* dan *Minimum Document Frequency*. Dua parameter ini akan dieksperimenkan nilainya untuk dilihat nilai yang membuat model lebih maksimal. Nilai yang akan dicoba untuk *Minimum Document Frequency* 15 dan 20; *Maximum Document Frequency* 0.3 dan 0.8.

Dalam menggunakan *K-Means Clustering* perlu dibuat dari *text* data menjadi *matrix* menggunakan TF-IDF, Lalu dilakukan *setting* dengan 3 *cluster* dan *random state* 42.

4. ***Model Execution***

Hasil dari eksekusi LDA adalah sejumlah topik dengan *term frequency* setiap kata yang ada di dalam topik. Eksekusi yang sudah dilakukan selanjutnya divisualisasi dengan menggunakan *library pyLDAvis.* Pada visualisasi *pyLDAvis* terdapat *Intertopic Distance Map* dari topik-topik yang ada dan *bar chart* yang menunjukkan *term frequency* dari kata-kata yang ada dalam topik.

Hasil eksekusi menggunakan *K-Means Clustering* berupa *array* berisi topik-topik yang dibagi kedalam 3 *clustering*. Kemudian divisualisasi dengan *matplotlib* untuk dapat melihat *scatter plot* persebaran topik.

5. ***Result Analysis***

Berdasarkan hasil pencarian topik dan visualisasi topik yang didapat dilakukan analisis terhadap hasil tersebut.

3. HASIL DAN PEMBAHASAN

Pertama-tama dilakukan *preprocessing* sehingga dari total baris yang ada di *raw data* masing-masing *dataset* bertambah karena adanya *preprocessing* juga pembuatan *bigram* dan *trigram*. Banyaknya total baris akhir terlihat di Tabel 1.

Tabel 1. Pembagian data setiap pasangan calon

Dataset	**Total Baris**
Dataset A	*1076*
Dataset B	*1934*
Dataset C	*1412*

3.1. Metode LDA Eksperimen 1

Seperti dijelaskan dalam metodologi, metode LDA perlu dibuat *Dictionary* dan *Corpus,* dan dibutuhkan jumlah topik yang akan diambil. Dalam pembentukan *Dictionary* sendiri terdapat parameter *Minimum Document Frequency* dan *Maximum Document Frequency* untuk menentukan minimal kata keluar dalam data dan maksimum kata keluar dalam data. Dua parameter ini akan dieksperimen dengan angka *Minimum Document Frequency* 15 dan 20; *Maximum Document Frequency* 0.3 dan 0.8. Hasilnya dapat terlihat di Tabel 2.

Tabel 2. Pembagian data setiap pasangan calon

Dataset	***Min Doc. Freq.***	***Max Doc. Freq.***	**Jumlah Padanan Kata**
A	*15*	*0.3*	*12*
B	*15*	*0.3*	*27*
C	*15*	*0.3*	*16*
A	*15*	*0.8*	*12*
B	*15*	*0.8*	*27*
C	*15*	*0.8*	*16*
A	*20*	*0.3*	*5*
B	*20*	*0.3*	*23*
C	*20*	*0.3*	*8*
A	*20*	*0.8*	*5*
B	*20*	*0.8*	*25*
C	*20*	*0.8*	*8*

Berdasarkan padanan kata dan jumlah padanan kata yang keluar, maka topik yang terfokus didapatkan ketika *Minimum Document Frequency* lebih besar, sedangkan *Maximum Document Frequency* tidak memiliki pengaruh signifikan dalam topik yang dihasilkan.

Penggunaan *Minimum Document Frequency* terbaik di range 15-20 kali keluar, sehingga kata/topik yang keluar benar-benar kata yang memang sering disebutkan dan menjadi topik. Jika kurang dari 15 akan terlalu banyak kata kurang bermakna keluar sebagai topik. Sedangkan jika lebih dari 20 tidak ada kata yang keluar.

DOI : https://doi.org/10.21063/jtif.2024.V12.1.29-35

Sedangkan *Maximum Document Frequency* lebih baik memiliki nilai lebih tinggi, pada penelitian ini di angka 0.8 atau minimal di angka 0.5 agar hanya kata-kata bermakna yang keluar sebagai topik. Jika kurang dari 0.5 akan banyak kata yang berulang namun tidak memiliki makna, sedangkan jika diatas 0.8 akan banyak kata yang berulang namun bermakna tidak keluar sebagai topik.

Tabel 3 berisi kemunculan kata-kata yang merupakan topik debat yang sering diucapkan pasangan A, pasangan B, dan pasangan C.

Tabel 3. Topik masing-masing pasangan

Pasangan	***Topik***
A	*pks, pkb, indonesia, ubah, proses*
B	*prabowo, jokowi, perintah, bicara. partai politik, politik, partai, pilih, ketua, dukung, rakyat, nama, masuk, pimpin, ubah, anis, koalisi, putus, jalan, indonesia, presiden, orang, menang, demokrat, lihat, proses*
C	*ganjar, bicara, politik, jokowi, orang, nama, pilih, prabowo*

Jika dilihat topik yang muncul dari seluruh *dataset* beberapa topik yang sering muncul adalah terkait nama calon, partai pengusung/koalisi, dan nama Presiden Jokowi. Sedangkan terkait gagasan hanya dapat dilihat di *dataset A* yang mewakili pasangan calon Anies Baswedan dan Muhaimin Iskandar. Hal ini dapat terjadi karena kebanyakan data yang berasal dari video debat banyak membahas terkait koalisi dan video terkait gagasan lebih sedikit dari video terkait koalisi. Tentu nama calon akan menjadi kata atau padanan kata yang sering keluar. Selain itu terkait Pemilihan Umum Presiden tentunya akan banyak membahas terkait pemerintahan saat ini yang dipimpin oleh Presiden Jokowi.

3.2. Model LDA Eksperimen 2

Berdasarkan eksperimen pertama maka dilakukan eksperimen lanjutan. Jika sebelumnya video debat yang menjadi *data source* membahas debat terkait Pemilihan Umum Presiden 2024 secara umum, baik koalisi maupun penyampaian gagasan. Eksperimen berikutnya dilakukan *filtering* video sebagai *source data*.

Seperti eksperimen sebelumnya dilakukan *preprocessing* terhadap data, dengan metode-metode *preprocessing* yang sama. Banyak total baris untuk masing-masing *dataset* yang mewakili pasangan calon berkurang, terlihat pada Tabel 4.

Pada *topic modeling* menggunakan algoritma LDA akan diset *Minimum Document Frequency* 15 dan *Maximum Document Frequency* 0.5. Total kata yang didapat dari masing-masing *dataset* dapat dilihat dari Tabel 4.

Tabel 4. Pembagian data dan total padanan kata setiap pasangan calon

Dataset	**Total Baris**	**Total Padanan Kata**
Dataset A	*770*	*4*
Dataset B	*644*	*5*
Dataset C	*776*	*3*

Eksperimen dilanjutkan dengan visualisasi dengan *library pyLDAvis*. Masing-masing topik untuk pasangan A, pasangan B, dan pasangan C dapat dilihat dalam Tabel 5.

Tabel 5. Topik masing-masing pasangan eksperimen ke-2

Pasangan	***Topik***
A	*ubah, bangun, proses, indonesia*
B	*rakyat, jokowi, presiden, prabowo, indonesia*
C	*orang, ganjar, jokowi*

Berdasarkan hasil eksperimen kedua, didapatkan hasil lebih spesifik untuk topik yang dibahas. Memang kata Indonesia dan jokowi akan menjadi kata yang sering keluar karena video debat sering menyangkutkan dengan pemerintah sekarang. Sama seperti eksperimen sebelumnya, nama calon akan keluar menjadi topik utama pada masing-masing dataset.

3.3. Metode *K-Means Clustering Eksperimen 1*

Menggunakan data seluruh video yang jumlahnya tertera di Tabel 2, dilakukan *clustering* dengan *K-Means*. Jumlah *cluster*

DOI : https://doi.org/10.21063/jtif.2024.V12.1.29-35

adalah 3 dengan *random state* 42. Didapatkan topik untuk masing-masing pasangan A, pasangan B, dan pasangan C seperti pada Tabel 6.

Tabel 6. Topik masing-masing pasangan dengan *K-Means Clustering*

Pasangan	***Topik***
A	*undang, teman, sehat, proses, revisi, demokrat, pontianak, koalisi, pdip, tolak, ubah, indonesia, proses, bangun, partai, rakyat, anis, presiden, jakarta, pkb, nasdem, pks, bicara, muhaimin, prioritas, basis, nu*
B	*teman, komitmen, pks, solid, etika, koalisi, hormat, jalan, sayang, prabowo, nama, jokowi, politik, orang, presiden, indonesia, koalisi, rakyat, pimpin, partai, politik, putus, demokrat, ketua, mekanisme, golkar, orang, prabowo, majelis*
C	*jokowi, ganjar, prabowo, lanjut, bicara, bilang, rakyat, partai, mahfud, ubah, orang, muda, bersih, generasi, projo, pilih, head, angka, dukung, kalah, politik, proses, partai, survei, nu, apa, fakta, jalan, opini*

3.4. Metode *K-Means Clustering Eksperimen 2*

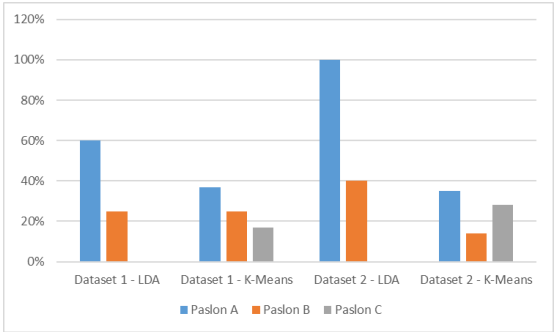
Menggunakan data video yang sudah diseleksi, dilakukan *clustering* dengan *K-Means*. Jumlah *cluster* adalah 3 dengan *random state* 42. Didapatkan topik untuk masing-masing pasangan A, pasangan B, dan pasangan C seperti pada Tabel 7.

Menggunakan metode *K-Means Clustering*, baik dengan Eksperimen 1 ataupun dengan Eksperimen 2 didapatkan topik yang tidak jauh berbeda. Kebanyakan pasangan berbicara tentang ubah, Presiden Jokowi, dan tentu terdapat nama calon beserta wakilnya.

Tabel 7. Topik masing-masing pasangan dengan *K-Means Clustering*

Pasangan	***Topik***
A	*bangun, proses, manusia, kritik, indonesia, bangsa, tingkat, mental, urut, lihat, teman, nama, partai, tugas, butuh, jakarta, muncul, pupuk, ikut, pks, ubah, undang, indonesia, jokowi, orang, presiden, rakyat, jalan, jakarta, prabowo*
B	*orang, negara, personal, hadap, kali, tentara, bicara, kadang, prajurit, hidup, politik, ajar, akrab, erektoral, logika, keliling, ubah, kadang, hidup, ketemu, prabowo, jokowi, presiden, pimpin, indonesia, perintah, lanjut, 2024, rakyat, dukung*
C	*jokowi, mental, revolusi, kerja, bangun, citra, presiden, ganjar, simbol, orang, belok, undang, bagus, juang, luan, cari, diam, ngarang, tarung, bilang, mahfud, nama, rakyat, anak, partai, 212, anis, ubah*

Gambar 1 menunjukkan hasil pengelompokkan dengan menggunakan standar yang telah kita tentukan diantaranya kata bangun, bersih, indonesia, jakarta, juang, kerja, komitmen, lanjut, mekanisme, mental, pimpin, prioritas, proses, pupuk, rakyat, revisi, revolusi, solid, tolak, tugas, ubah dan undang. Pada *dataset* 1 dengan model LDA didapati gagasan pasangan calon untuk paslon A 60%, B 25%, dan C 0%. Pada *dataset* 1 dengan model *K-Means* didapati gagasan pasangan calon untuk paslon A 37.04%, B 25%, dan C 17.24%. Pada *dataset* 2 dengan model LDA didapati gagasan pasangan calon untuk paslon A 100%, B 40%, dan C 0%. Pada *dataset* 2 dengan model *K-Means* didapati gagasan pasangan calon untuk paslon A 35.71%, B 14.29%, dan C 28.57%.



Gambar 1. Hasil Pengelompokkan dengan Standar yang Diterapkan.

DOI : https://doi.org/10.21063/jtif.2024.V12.1.29-35

4. KESIMPULAN DAN SARAN

Berdasarkan penelitian yang ada sudah bisa didapatkan topik dari video-video debat yang menjadi data untuk penelitian ini. Namun untuk spesifik kepada gagasan pasangan calon Presiden dan Wakil Presiden belum sepenuhnya dapat mencapai tujuan tersebut. Untuk parameter sendiri, pada penelitian pencarian topik dari video debat ini didapatkan topik yang terfokus didapatkan ketika *Minimum Document Frequency* lebih besar, sedangkan *Maximum Document Frequency* tidak memiliki pengaruh signifikan dalam topik yang dihasilkan.

Menggunakan *K-Means Clustering* topik yang didapatkan lebih banyak dan beragam juga lebih merata untuk setiap paslon. Sedangkan topik LDA mendapatkan topik lebih sedikit dan lebih berhasil mendapatkan topik untuk paslon A. Baik dengan metode LDA ataupun *K-Means Clustering* didapatkan garis merah topik yang sama. Sudah cukup menggambarkan hal-hal yang sering dibicarakan pasangan calon baik secara langsung ataupun melalui perwakilannya

Penelitian selanjutnya dapat dikembangkan dengan memilih dataset yang lebih relevan untuk menjadi data untuk penelitian. Sehingga dataset yang dihasilkan juga lebih relevan dengan tujuan penelitian dilakukan. Dataset yang relevan akan membantu mencapai tujuan untuk menemukan gagasan pasangan calon Presiden dan Wakil Presiden dapat tercapai.

DAFTAR PUSTAKA

[1] Zulhanif, Sudartianto, B. Tantular dan I. G. N. M. Jaya, "Aplikasi Latent Dirichlet Allocation (LDA) Pada Clustering Data Teks," *Jurnal Logika,* vol. 7, no. 1, pp. 46-51, 2017.

[2] W. Fuadi, A. Razi og D. Fariadi, «Automasi Penentuan Tren Topik Skripsi Menggunakan Algoritma K-Means Clustering,» *Serambi Engineering,* vol. 7, no. 2, pp. 3072-3077, 2022.

[3] G. H. Setiawan, I. M. B. Adnyana, I. G. R. A. Sugiartha dan K. Budiarta, "Ekstraksi Topik Pada Aduan Mahasiswa Dengan Pendeketan Model Latent Dirichlet Allocation (LDA)," dalam *Seminar Nasional Penelitian dan Pengabdian Kepada Masyarakat CORISINDO*, Bali, 2023.

[4] H. V. Halteren, J. Zavrel, and W. Daelemans, W. "Improving accuracy in word class tagging through the combination of machine learning systems ," *Computational linguistics*, vol. 27, no. 2, pp 199-229, 2011.

[5] U. T. Setijohatmo, S. Rachmat, T. Susilawati dan Y. Rahman, "Analisis Metoda Latent Dirichlet Allocation untuk Klasifikasi Dokumen Laporan Tugas Akhir Berdasarkan Pemodelan Topik," dalam *Prosiding 11th Industrial Research Workshop and National Seminar (IRWNS)*, Bandung, 2020.

[6] B. A. Tondang, M. R. Fadhil, M. N. Perdana, A. Fauzi dan U. S. Janitra, "Analisis Pemodelan Topik Ulasan Aplikasi BNI, BCA, dan BRI Menggunakan Latent Dirichlet Allocation," *INFOTECH: Jurnal Informatika & Teknologi,* vol. 4, no. 1, pp. 114-127, 2023.

[7] Y. Matira, Junaidi dan I. Setiawan, "Pemodelan Topik pada Judul Berita Online Detikcom Menggunakan Latent Dirichlet Allocation," *ESTIMASI: Journal of Statistics and Its Application,* vol. 4, no. 1, pp. 53-63, 2023.

[8] V. F. Rusdhi dan I. Sari, "Identifikasi Topik Artikel Berita Menggunakan Topic Modelling dengan Latent Dirichlet Allocation," *Jurnal Ilmiah Informatika Komputer,* vol. 27, no. 2, pp. 169-176, 2022.

[9] M. S. Hudin, M. A. Fauzi dan S. Adinugroho, "Implementasi Metode Text Mining dan K-Means Clustering untuk Pengelompokan Dokumen Skripsi (Studi Kasus: Universitas Brawijaya)," *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer,* vol. 2, no. 11, pp. 5518-5524, 2018.

[10] R. Siringoringo, J. Jamaluddin dan R. Perangin-Angin, "Pemodelan Topik Berita Menggunakan Latent Dirichlet Allocation dan K-Means Clustering," *Jurnal Informatika Kaputama (JIK),* vol. 4, no. 2, pp. 216-222, 2020.

DOI : https://doi.org/10.21063/jtif.2024.V12.1.29-35