

# Analisis Komen *Youtube* dan *Topic Modelling* Pernikahan Kaesang dan Erina

1 <sup>st</sup> Ade Cahyaning Palupi	2 <sup>st</sup> Kristine Angelina Simanjuntak	3 <sup>st</sup> Angelia Regina Dwi Kartika	4 <sup>st</sup> Yolla Putri Ervanisari
105220001	105220009	105220039	105220049
<i>Ilmu Komputer</i> <i>Universitas Pertamina</i>	<i>Ilmu Komputer</i> <i>Universitas Pertamina</i>	<i>Ilmu Komputer</i> <i>Universitas Pertamina</i>	<i>Ilmu Komputer</i> <i>Universitas Pertamina</i>
Jakarta, Indonesia	Jakarta, Indonesia	Jakarta, Indonesia	Jakarta, Indonesia

## I. LATAR BELAKANG

*Big Data*. Menurut MGI dan McKinsey's *Business Technology Office*, banyaknya jumlah data yang ada di dunia ini telah semakin meningkat, sehingga menganalisis kumpulan data yang besar atau yang biasanya disebut sebagai *big data* menjadi kunci utama dalam persaingan yang menopang gelombang baru pertumbuhan inovasi, produktivitas, dan *surplus* konsumen. Para pemimpin pada setiap bidang juga harus bersaing dengan implikasi dari *big data* itu sendiri, tidak hanya beberapa hal yang berorientasi pada data. Dengan meningkatnya *detail* informasi dan volume yang diperoleh dari organisasi, dengan munculnya media sosial, *Internet of Things (IoT)*, dan multimedia yang akan mendorong pertumbuhan data secara eksponensial di masa mendatang (Manyika, 2011).

*Big data* telah mengubah cara bekerja sebuah teknologi dan praktisi dalam membangun sebuah sistem *data analytics*. *Youtube* sendiri merupakan sebuah *platform search engine* terbesar kedua yang menyediakan berbagai fitur yang memungkinkan pengguna untuk melihat, mengomentari, menautkan, dan *memposting* sebuah video. *Youtube* juga telah memiliki lebih dari satu miliar pengguna aktif yang dapat melihat saluran (*channel*) yang direkomendasikan berdasarkan seberapa seringnya pengguna tersebut menonton video tersebut.

Pada *youtube* sendiri mengimplementasikan *big data* untuk mendapatkan banyak perhatian untuk mendapatkan manfaat dan peluang yang belum pernah terjadi sebelumnya. Hasil analisis *big data* dapat meningkatkan perubahan bisnis dan pengambilan keputusan dengan menerapkan teknik analitik yang tinggi pada *big data*, dan mengungkapkan informasi dan pengetahuan berharga yang belum pernah *dieksplor*. Seperti contoh, *trending* pada *youtube* sendiri didasarkan pada *tren* video berdasarkan jumlah klik yang diperoleh dari video tersebut setiap harinya, serta banyaknya ataupun sebagai *feedback* komen yang diberikan oleh masyarakat (Andry, 2021).

Dalam *paper* ini kami menganalisis komen video *youtube* dengan menggunakan teknik *topic modelling* dalam membantu memvisualisasikan hasil data yang telah dianalisis. *Topic modelling* merupakan semacam algoritma yang memberikan wawasan tentang tema *semantic* atau topik dalam korpus dokumen. *Topic modelling* sendiri didasari dari asumsi setiap dokumen yang memiliki tema atau topik (D. M. Blei, A. Y. Ng, and M. I. Jordan, 2003). *Topic modelling* digunakan karena untuk mengekstraksi kata-kata dari dokumen lebih rumit dan membutuhkan waktu lebih banyak daripada mengekstraksi dari topik dalam dokumen. Misalnya, ada 1000 dokumen dan 500 kata dalam setiap dokumen. Jadi untuk memproses ini membutuhkan  $500 \times 1000 = 500000$  utas. Jadi ketika sebuah dokumen dibagi menjadi topik tertentu maka, jika ada 5 topik yang ada di dalamnya, pemrosesannya hanya  $5 \times 500$  kata = 2500 utas. *Topic modelling* akan membuatnya menjadi lebih sederhana daripada memproses seluruh dokumen dan *topic modelling* dapat memecahkan masalah atau memvisualisasikan hal-hal dengan lebih baik. Selain dokumen, *topic modelling* juga dapat memberikan topik terhadap komen yang ada pada sebuah video dalam *youtube*.

*Latent Dirichlet Allocation (LDA)* merupakan salah satu model topik probabilistik untuk menemukan struktur topik yang tersembunyi dalam beberapa komen *youtube* yang ada. Komen tersebut dilihat sebagai campuran probabilistik dari topik-topik yang tersembunyi atau topik yang memiliki distribusi probabilitas atas kata-kata dan setiap komen dimodelkan berdasarkan model berbasis kata. Akan tetapi, *LDA* tidak memperhitungkan semantik yang lebih dalam dari sebuah komen *youtube*. Melainkan cukup baik dalam mempelajari topik-topik yang tersembunyi (Ekinci, 2019).

*Paper* ini bertujuan untuk memberikan informasi bagaimana memahami dan mengimplementasikan *big data* dalam komen video *pada youtube*. Data yang telah divisualisasikan dapat memberikan prediksi yang lebih baik tentang pendapat masyarakat luas dalam sebuah video “Pernikahan Kaesang dan Erika”, sehingga hal tersebut dapat digunakan untuk mempengaruhi pengambilan keputusan dalam sebuah video.

## II. METODE

*Paper* ini mengekstrak beberapa hasil komen masyarakat pada video Pernikahan Kaesang dan Erika yang dibagi menjadi topik-topik dengan menggunakan *LDA*. Selain menggunakan *LDA*, kami juga menggunakan pendekatan kualitatif untuk menjadikan komen-komen dalam video *youtube* menjadi *label-label* sesuai dengan kategori yang sama pada setiap beberapa komen yang ada.

### A. Data Collection

Pengumpulan data-data yang digunakan pada *paper* ini menggunakan *keywords* “Pernikahan Kaesang” dari beberapa video *youtube*. Dari video-video tersebut, komen yang ada pada setiap video digunakan sebagai data yang akan dianalisis atau dimanipulasi untuk divisualisasikan agar mendapat *insight* atau informasi yang tersembunyi yang ada pada komen-komen video *youtube* tersebut.

### B. Data Preprocessing

Pada tahap *preprocessing* terdapat langkah-langkah yang dilakukan untuk menghasilkan data agar dapat siap diolah. Langkah pertama yang dilakukan adalah *data cleaning* agar data yang diperoleh dapat digunakan menjadi akurasi, karena data tersebut diseleksi dan dibuang atau istilah lainnya dibersihkan agar tidak terdapat data yang bermasalah. Langkah kedua adalah *data integration* yang mengambil beberapa sumber referensi yang disatukan dan digabungkan menjadi data yang lebih besar dalam data yang memiliki format yang serupa. Langkah ketiga adalah *data transformation* dilakukan untuk menjadikan beberapa sumber data yang telah terkumpul menjadi serupa, dengan mengubah format, struktur, atau nilai data agar diperoleh *dataset* yang dapat dilakukan untuk proses *data mining*. Langkah terakhir yang dilakukan adalah *data reduction* yang dilakukan untuk mengurangi data yang tidak perlukan, akan tetapi hal tersebut tidak akan mengubah hasil analisis dari data tersebut.

Teknik *preprocessing* yang digunakan dalam pengolahan data yang dilakukan antara lain *Data Cleaning*, *Tokenize*, *Stopwords*, dan *Lemmatizing*. Pada teknik ini menggunakan library khusus yaitu NLTK (*Natural Language Toolkit*) yang merupakan library *Python* untuk pemodelan teks. Teknik yang pertama adalah *Data Cleaning*. Pada teknik ini dilakukan pembersihan data untuk menghasilkan data yang lebih akurat dengan cara menghilangkan *Missing Value*, *Noisy data*, dan *Inconsistent data*, serta penerapan yang dilakukan dengan mengubah ‘\n’ menjadi ‘*string*’ kosong, menghilangkan tanda baca seperti *!,-./;<=>?@[]^\_`{|}~`"#%&'()\*+` kemudian seluruh kalimat diubah menjadi *Lowercase*. Selanjutnya teknik *preprocessing* yang kedua adalah *Tokenize* yang merupakan memisahkan kalimat menjadi kata-kata seperti contohnya “Selamat menempuh hidup baru untuk mas Kaesang” menjadi [selamat, menempuh, hidup, baru, untuk, mas, kaesang].*

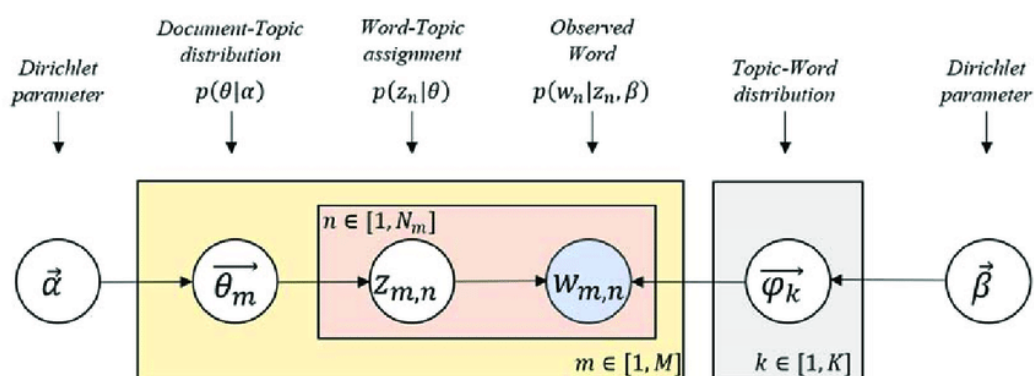
Teknik *Preprocessing* ketiga adalah *Stopwords* yang digunakan untuk menghilangkan kata yang sering muncul dan tidak memiliki makna atau bisa disebut

sebagai *noise* dalam teks seperti halnya kata hubung “yang”, “dan”, “di”, “dari”, dll. Teknik yang terakhir merupakan *Lemmatizing* digunakan untuk mereduksi variasi kata dalam kalimat aktif atau pasif untuk mendapatkan kata dasarnya atau asal kata tersebut. Keempat teknik *Preprocessing* dapat dilihat dalam tabel dibawah ini.

Teknik <i>Preprocessing</i>	Contoh Data
Data Asli	Selamat Raden kaisang sama mbak Erina.. semuga cepat dapat momongan yang Sholeh dan Sholeh
<i>Data Cleaning</i>	selamat raden kaisang sama mbak erina semuga cepat dapat momongan yang sholeh dan sholeh
<i>Tokenize</i>	['selamat', 'raden', 'kaisang', 'sama', 'mbak', 'erina', 'semuga', 'cepat', 'dapat', 'momongan', 'yang', 'sholeh', 'dan', 'sholeh']
<i>Stopwords</i>	['selamat', 'raden', 'kaisang', 'mbak', 'erina', 'semuga', 'cepat', 'momongan', 'sholeh', 'sholeh']
<i>Lemmatizing</i>	selamat raden kaisang mbak erina semuga cepat momong sholeh sholeh

### C. Topic Modelling

*Topic modelling* merupakan suatu metode pengelompokan data berdasarkan suatu topik tertentu. *Topic modelling* biasa disebut sebagai *unsupervised learning* yang berarti tidak membutuhkan data *label* atau dapat dikatakan *topic modelling* ini bekerja seperti *clustering* yaitu mengelompokkan data berdasarkan kemiripannya, tetapi data yang dikelompokkan pada *topic modelling* lebih spesifik seperti mencari pola topik secara abstrak pada kumpulan data.



**Gambar 1. Graphical Model Of Latent Dirichlet Allocation LDA**

*Latent Dirichlet Allocation (LDA)* merupakan sebuah metode untuk mendeteksi topik-topik yang ada pada kolom data beserta proporsi kemunculan topik yang telah ditentukan. Ide dasar yang digunakan pada *LDA* terdiri dari beberapa topik data. Proses pada *LDA* bersifat generatif melalui *imaginary random process* pada model yang mengasumsikan bahwa data berasal dari topik tertentu, dimana setiap topik terdiri dari distribusi kata-kata.

Tujuan dari menggunakan *topic modelling* dengan metode *LDA* yaitu untuk menentukan topik secara otomatis. *Topic modelling* dengan metode *LDA* ini dapat menentukan distribusi topiknya yang mendeskripsikan kumpulan kata pada data yang diambil. Digunakannya kumpulan data tersebut untuk *inference* struktur topik tersembunyi, karena data yang diteliti memiliki struktur tersembunyi.

## III. HASIL DAN PEMBAHASAN

Pada penelitian ini, *LDA* dengan banyak topik 2 memiliki nilai *perplexity* sebesar -5,87614328307074. *LDA* dengan banyak topik 10 memiliki nilai *perplexity*

paling rendah. Nilai *perplexity* sendiri adalah cara untuk menentukan jumlah topik sebagai sebuah *metric* untuk menguji ketepatan informasi dari dokumen terhadap topik yang dihasilkan. Perhitungan *perplexity* dilakukan dengan menentukan kemungkinan dari *log teks* dokumen yang tidak terlihat, semakin rendah nilai *perplexity* semakin baik model yang dihasilkan. Pada penelitian ini, nilai *perplexity* yang dihasilkan bernilai negatif. Pada dasarnya, nilai *perplexity* akan semakin rendah pada negatif yang semakin kecil. Memiliki nilai *perplexity* negatif tampaknya disebabkan oleh probabilitas sangat kecil yang dikonversi ke skala log secara otomatis oleh *Gensim*, tetapi meskipun menginginkan *perplexity* yang lebih rendah, nilai batas bawah menunjukkan penurunan, sehingga nilai *perplexity* yang lebih rendah memburuk dengan lebih banyak topik. Nilai *perplexity* untuk setiap jumlah topik dapat dilihat pada Gambar 2.

2	:	-5.876143283070743
3	:	-5.95229687597474
4	:	-6.064294650174504
5	:	-6.164646621444311
6	:	-6.2089923663131
7	:	-6.308743392746424
8	:	-6.359197375955493
9	:	-6.416016079491676
10	:	-6.488776860122811

Gambar 2. Nilai *Perplexity* Setiap Topik

Tabel 1. Tema Distribusi Marginal Setiap Topik

Topik-0		Topik-1	
Tema	Marginal distribution	Tema	Marginal distribution
Mas_kaesang	0.024	Yg	0.036
Ya	0.020	Jokowi	0.022
Keluarga	0.020	Mas_kaesang	0.021
Yg	0.020	Bahagia	0.020
Moga	0.020	Kaesang	0.016
Banget	0.017	Nikah	0.015
Anak	0.015	Moga	0.014
Selamat	0.014	Keluarga	0.013
Mas	0.013	Selamat	0.012

Tabel 1 menunjukkan tema dan distribusi marginal dari setiap topik untuk menentukan *label* setiap topiknya. Setelah itu, kami menggunakan pendekatan kualitatif untuk melabeli masing-masing topik. Dilakukan analisis kualitatif yang menghasilkan 2 label yaitu “Doa” dan “Keluarga”. Berikut ini akan kami jelaskan setiap *label*.

#### A. Doa

Topik-0 diberi *label* “Doa” dengan alasan memiliki 3 tema berisi “moga”, “anak”, dan “selamat” yang masing-masing memiliki distribusi marginal sebesar

0.020, 0.015, dan 0.014. Ini berarti bahwa mayoritas orang berkomentar berisi doa terhadap pernikahan Kaesang dan Erina.

#### B. *Keluarga*

Topik-1 diberi *label* “Keluarga” dengan alasan memiliki distribusi marginal tema “Jokowi” sebesar 0.022, kemudian diikuti oleh tema “mas\_kaesang” sebesar 0.021 dan “Kaesang” sebesar 0.016, lalu tema “keluarga” memiliki distribusi marginal sebesar 0.013. Pernikahan Kaesang sendiri sangat disoroti karena merupakan keluarga dari Presiden Indonesia yaitu Pak Jokowi. Tidak sedikit juga masyarakat yang berkomentar terkait keluarganya.

### IV. KESIMPULAN

Penelitian ini mencoba melakukan telaah literatur dari komentar *youtube* pada keyword tertentu. Penelitian ini mengekstraksi informasi menggunakan algoritma pemodelan topik dari 23 *video* terkait keyword “pernikahan kaesang”. Kemudian peneliti menerapkan analisis kualitatif untuk menentukan *label* topik dari distribusi marginal. Topik tersebut diberi label: “Doa” dan “Keluarga”. Metodologi yang kami terapkan dalam penelitian ini dapat diterapkan pada penelitian serupa namun pada konteks yang berbeda.

### REFERENSI

- Andry, J.F., Tannady, H., Limawal, I.I., Rembulan, G.D., & Marta, R.F. (2021). “BIG DATA ANALYSIS ON YOUTUBE WITH TABLEAU”.
- D. M. Blei, A. Y. Ng, and M. I. Jordan. (2003). “Latent dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, no. Jan, pp. 993– 1022.
- Ekinci, E., & İlhan Omurca, S. (2019). Concept-LDA: Incorporating BabelFy into LDA for aspect extraction. *Journal of Information Science*, 46, 406 - 418.
- Manyika, J. (2011). Big data: The next frontier for innovation, competition, and productivity.