PENERAPAN METODE BAYESIAN DALAM MODEL LATENT DIRICHLET ALLOCATION DI MEDIA SOSIAL

APPLICATION OF BAYESIAN METHODS IN LATENT DIRICHLET ALLOCATION MODEL IN SOCIAL MEDIA

Oleh: Muh. Fajriyanto, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Yogyakarta
muhfajriyanto@gmail.com

Abstrak

Penelitian ini bertujuan untuk mengetahui berita apa yang dominan dibahas di masyarakat pada periode waktu tertentu. Model *Latent Dirichlet Allocation* (LDA), sebuah model probabilitas dimana akan menghasilkan beberapa macam topik yang berbeda. Diawali dengan mengambil data data *tweet* dari *twitter.* Data yang semula vektor diubah menjadi *corpus* lalu dilakukan pre-processing pada data agar bisa dibentuk model. Selanjutnya pembentukan model pada data dan estimasi parameter yang digunakan adalah metode Bayesian dengan estimasi *Gibbs Sampling.* Setelah mendapatkan model dari data training maka model tersebut digunakan pada data testing untuk mendapatkan berita yang dominan dibahas di masyarakat. Hasil penelitian menunjukkan nilai *loglikehood* paling tinggi -1759487 dengan 10 topik dan topik yang dominan dibahas di masyarakat yang diterbitkan @kompascom pada tanggal 11 Mei 2018 sampai 25 Mei 2018 adalah Menyebarkan gambar atau video lokasi bom di Surabaya dapat ikut menyebarkan teror dan ketakutan yang jadi tujuan pelaku bom dengan nilai probabilitas topik 0.10057.

Kata kunci: Media Sosial, twitter, pemodelan topik, *Latent Dirichlet Allocation, Bayesian*, dan *Gibbs Sampling*

Abstract

This research aims to find out what news is dominantly discussed in the community for a certain period of time. The Latent Dirichlet Allocation (LDA) model, a probability model which will produce several different topics. Beginning by taking data tweet data from twitter. The original vector data is converted into a corpus and then pre-processing the data to form a model. Furthermore, the formation of model on data and parameter estimation used is Bayesian method with Gibbs Sampling estimation. After getting the model from the training data then the model is used in data testing to get the dominant news discussed in the public. The results of this research showed that the highest loglikehood value -1759487 with 10 topics and topics dominantly discussed in the public published @kompascom on May 11, 2018 until May 25, 2018 is spread the image or video of the location of the bomb in Surabaya can participate in spreading terror and fear of the goal bombers with probability value topic 0.10057.

Keywords*: Social Media, twitter, topic modeling, Latent Dirichlet Allocation, Bayesian, and Gibbs Sampling*

PENDAHULUAN

Perkembangan teknologi informasi dan komunikasi di Indonesia semakin meningkat seiring dengan perkembangan zaman. Perkembangan teknologi informasi dan komunikasi di Indonesia membuat masyarakat mulai meninggalkan media tradisional seperti televisi dan koran. Mobilitas aktivitas yang tinggi mengakibatkan masyarakat lebih senang mengakses informasi melalui internet dengan *smartphone* mereka khususnya media sosial.

Media sosial adalah media informasi yang paling diminati masyarakat saat ini. Salah satu *platform* yang semakin popular dari media sosial sebagai sumber dari segala informasi adalah *Twitter*. *Twitter* merupakan sebuah *microblog* penyebar informasi yang sangat cepat dan berbasis *real-time.* Setelah dikeluarkan oleh Jack Dorsey pada tahun 2006 hingga pada awal tahun 2018 pengguna *Twitter* di dunia semakin bertambah begitu pula jumlah teks yang dikeluarkan (*tweet)* perharinya. Pengguna *Twitter* bebas membicarakan apapun dan berargumen positif atau negatif kepada siapapun. Begitu banyaknya opini tersebut, sangat tidak memungkinkan untuk mendapatkan topik yang sedang dominan dibahas di masyarakat dengan cara membaca satu persatu.

Untuk mengantisipasi keadaan tersebut perlu dilakukan pemodelan topik dengan model LDA untuk mendapatkan kalimat utama yang dapat menggambarkan isi keseluruhan data dan topik yang sedang dominan dibahas di masyarakat. Beberapa penelitian terkait dengan pemodelan topik dengan LDA telah dilakukan, diantaranya dilakukan oleh (Blei, 2003), metode yang digunakan adalah algoritma algoritma VEM. (Ponweiser, 2012) melakukan penelitian dengan metode yang digunakan adalah *Gibbs Sampling*. (KB Putra, 2017) pernah melakukan penelitian dengan menggunakan model LDA untuk menganalisis iterasi dan jumlah topuk dengan nilai *perplexity*. (R Kusumaningrum dkk, 2018) pernah melakukan penelitian dengan model yang sama

Berdasarkan hasil penelitian sebelumnya dan sejauh peneliti ketahui, penerapan metode *Bayesian* dalam model LDA dengan algoritma *Gibbs Sampling* untuk mencari topik yang sedang dominan dibahas, khususnya berita pada berbahasa Indonesia *Twitter* yang dikeluarkan oleh salah satu penyedia berita di Indonesia.

METODE PENELITIAN

Deskripsi Data

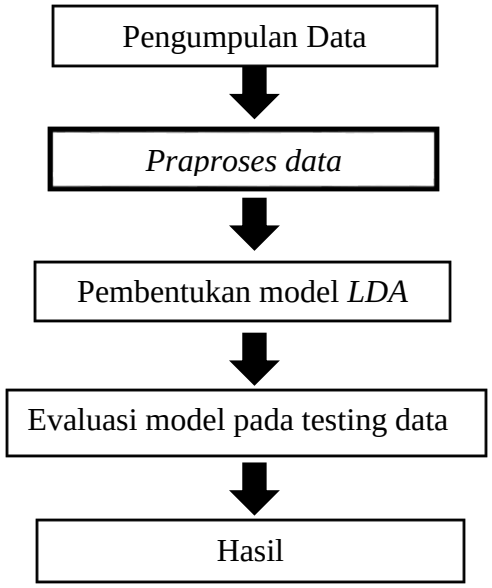
Data yang digunakan berupa data berbentuk teks yang diperoleh dari tweet yang memuat berita di jejaring sosial Twitter. Data *tweet* diperoleh melalui proses *crawling* menggunakan Twitter API (*Application Programming Interface*). Data yang digunakan sebanyak 82.476 *tweet* dengan *keyword* "@kompascom". Pengambilan data *tweet* diperoleh dalam rentang waktu 11 Mei 2018 sampai 21 Mei 2018.

Jenis dan Sumber Data Penelitian

Jenis data yang digunakan untuk penelitian ini adalah data primer yang merupakan data tweet yang berasal dari pengguna jejaring sosial *Twitter*, yaitu data *tweet* yang dikeluarkan merupakan berita dari kompas.

Teknik Analisis Data

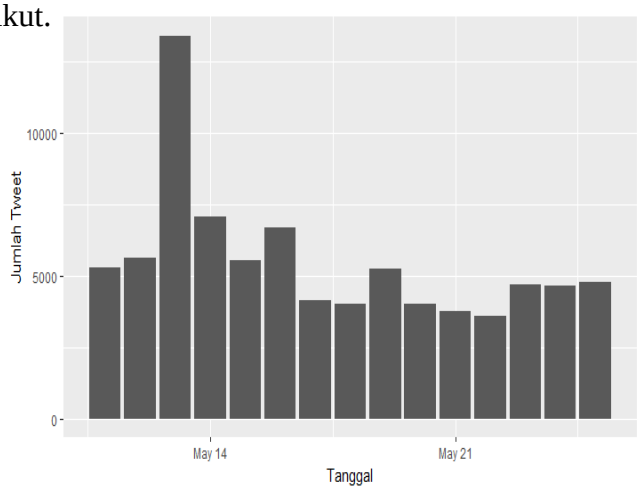
Data *dari* @kompascom yang merupakan data tekstual di twitter dengan menggunakan package "*twitteR*" pada R dalam rentang waktu dua minggu. Agar data yang diolah menjadi data yang lebih bersih dan berarti dilakukan tahap praproses. Setalah tahap praproses, data tersebut direpresentasikan menjadi matriks pasangan kata dokumen untuk dapat dilakukan pembentukan model *Latent Dirichlet Allocation* (LDA) dengan metode *Bayesian* yang dibantu algoritma *Gibbs Sampling*. Setelah mendapatkan model LDA dari data training maka model tersebut digunakan untuk data testing. Kemudian didapatkan interpretasi topik utama yang sedang dominan dibahas di masyarakat yang terbentuk dari kata-kata dari keluaran model tersebut.



GAMBAR 1. Diagram Alir Tahapan Penelitian

HASIL PENELITIAN DAN PEMBAHASAN

Berikut adalah hasil pengumpulan data dari Twitter API API (*Application Programming Interface*) dengan menggunakan *package* "twitteR" pada program R dan menggunakan fungsi *searchTwitter* dapat dilihat pada Gambar 1. berikut.



GAMBAR 2 Hasil Pengambilan Data Twitter

Berdasarkan gambar 2, dapat dilihat bahwa jumlah data yang dikumpulkan sebanyak 82.476 tweet yang merupakan data tekstual yang masih belum bisa dilakukan pembentukan model. Perlu dilakukan tahap selanjutnya, yakni tahap praproses data. Sebelum melakukan tahap praproses, data tekstual tadi diubah menjadi bentuk corpus. Tahap praproses data mencakup dua hal yang paling utama yakni membersihkan data dan penghapusan *stopword*. Pembersihan data dilakukan untuk mengubah penulisan huruf besar menjadi huruf kecil, menghapus tanda baca, angka, *mention*, *hashtag*, *url*, dan menghapus kata dengan jumlah huruf kurang dari. Selanjutnya, untuk penghapusan *stopword* mengacu pada susunan *stopword* yang telah disusun (Talla, 2013) dan terdapat tambahan kata-kata yang membantu pembentukan model. Sampel Hasil sebelum dan sesudah dilakukan tahap praproses data dapat dilihat pada tabel 1. berikut

Tabel 1 Hasil tahap praproses data

Sebelum	Sesudah
RT @kompascom: #kompastv Dimana Keberadaan Kantor Redaksi Obor Rakyat? https://t.co/W03xq01CwB	“dimana” "keberadaan" "kantor" "redaksi" "obor" "rakyat"

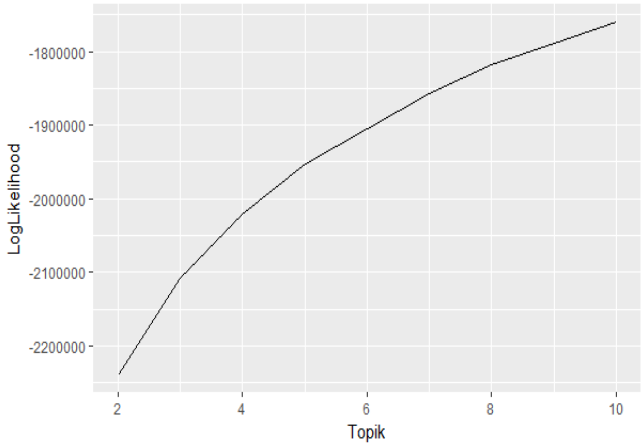
Kemudian hasil dari praproses data diatas dilakukan pemisahan corpus sebesar 70% sebagai training data dan 30% sebagai testing data. Kedua data tersebut kemudia direpresentasikan menjadi bentuk matrik kata dokumen dengan fungsi *DocumentTermMatrix* pada program R. Sampel hasil matriks pasangan kata dokumen pada training data dapat dilihat pada gambar 2. berikut.

```
Sample            :
      Terms
Docs   bom diduga fokus korban ledakan mabes polisi polri surabaya
  6903   1      1     1      1       1     1      1     1        1
      Terms
Docs   terjadi
  6903       1
```

GAMBAR 3 Hasil Pengambilan Data Twitter

Berdasarkan gambar 3, gambar matriks di atas terlihat dokumen ke 6903 pada training data jika isi dari matriks tersebut angka 1 atau selain 1 maka kata yang termuat di dokumen sebanyak angka tersebut. Setelah data tekstual direpresentasikan menjadi matriks pasangan kata dokumen maka dapat dilakukan pembentukan model pada data dengan Latent Dirichlet Allocation. Estimasi yang digunakan adalah Gibbs sampling. Sebelum dilakukan pembentukan model dicari nilai topik yang optimum terlebih dahulu.

Sebaran topik di tiap data berbeda-beda sehingga perlua dilakukan estimasi berapa topik yang mewakili data tersebut. Estimasi topik dicari melalui *loglikelihood* distribusi marginal kata dalam dokumen bersyarat topik yaitu *loglikelihood* $p(w|z)$ dengan estimasi *Gibbs sampling*. Diambil rentangan topik dari 2 sampai 10 topik. Hasil dari estimasi tersebut dapat dilihat pada gambar 4. berikut.



GAMBAR 4 Grafik nilai *loglikelihood* jumlah topik

Berdasarkan hasil gambar 4, jumlah topik dengan nilai *loglikehood* paling besar merupakan nilai topik yang paling optimum dalam menggambarkan data tersebut. Terlihat bahwa nilai topik 10 yang paling tepat dalam menggambarkan data. Setelah nilai topik didapatkan, maka dapat dilakukan pembentukan model pada training data. Kemudian dilakukan pembentukan model dengan algoritma GIBBS sampling dengan jumlah 10 topik. Untuk parameter $\alpha = \frac{10}{k}$ dan $\beta = 0{,}1$ (Griffith, 2004). Selanjutnya dilakukan pembentukan model di program R dengan fungsi LDA pada *package* “topicmodels”. Setelah mendapatkan model LDA dari *training data* maka model tersebut digunakan untuk *testing data*. Data model tersebut didapatkan 10 topik di dalamnya terdapat kelompok kata yang akan membentuk topik tersebut.

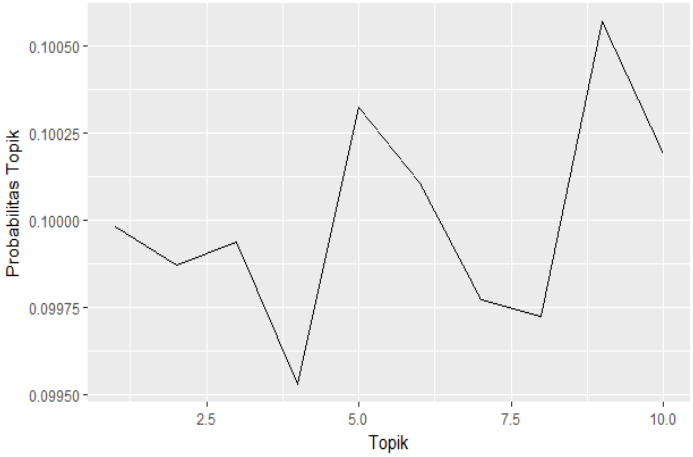
Hasil dari analisis model LDA terhadap testing data dengan 10 topik kemudian diinterpretasikan menjadi suatu kalimat dari 7 kata keluaran dari model yang membentuk setiap topiknya.

Dari hasil interpretasi penulis untuk 10 topik dari model LDA tersebut kemudian dibandingkan dengan berita yang terkait dari @kompascom. Dalam perbandingan tersebut dilihat apakah terdapat kesamaan antara topik berita yang diinterpretasikan dengan berita yang dikeluarkan @kompascom. Hasil dapat dilihat pada tabel 2. berikut.

Tabel 2. Hasil perbandingan interpretasi 10 topik berita

Interpretasi Model	Berita dari @kompascom
Pembunuhan yang dilakukan atas nama Tuhan dan Agama merupakan tindakan yang biadab	RT @kompascom: Tindakan pembunuhan yang dilakukan atas nama Tuhan dan agama merupakan tindakan yang biadab.
BTS dan ARMY cetak rekor twitter dengan tagar	BTS dan ARMY cetak rekor twitter dengan tagar #iVoteBTSBBMAs https:/t.co/OxjDkMexsf
Ledakan di tiga geraja di Surabaya, Jawa Timur, Minggu.	RT @kompascom: Ledakan terjadi di tiga gereja di Surabaya, Jawa Timur, Minggu (13/5/2018) pagi. https://t.co/OfcR2je0lK
Sulit di interpretasikan	
Ledakan di Gereja Katolik Santa Maria Tak Bercela Surabaya	Sebuah ledakan terjadi di Gereja Katolik Santa Maria Tak Bercela di Ngagel, Surabaya, Jawa Timur (13/5/2018)
Jadi habis lebaran makan ikan biar lebih pintar.	@kompascom Tuhhhh kan ... jadi untuk lebih pintar harus makan ikan. Yg tidak makan ikan habis lebaran kita tenggelamkan https://t.co/F0lw8TKGdh
Pernyataan Fadli tentang motif thr, Sri Mulyani heran	Menanggapi pernyataan Fadli yang mempertanyakan motif THR ke-13, Sri Mulyani Indrawati heran.
Sulit di interpretasikan	
Tujuan pelaku untuk menyebarkan terror dengan video lokasi bom	RT @kompascom: Menyebarkan gambar atau video lokasi bom di Surabaya dapat ikut menyebarkan teror dan ketakutan yang jadi tujuan pelaku bom
Teroris mencari surga dengan membunuh orang lain.	RT @kompascom: Jika teroris mencari surga dengan membunuh diri dan membunuh orang lain, dimana sebenarnya alamat surga mereka?

Berdasarkan hasil tabel 4, dapat dilihat bahwa terdapat 8 topik yang bisa di interpretasikan yang terbentuk dari kata hasil keluaran model di atas. Sedangkan ada 2 topik yang sulit di interpretasikan, artinya bahwa dari kata-kata hasil keluaran model tidak dapat membentuk topik berita yang dikeluarkan @kompascom. Dari hasil tersebut model ini layak untuk digunakan dalam mencari topik utama karena dapat mengelompokan topik dengan tepat dan lebih efisien dalam sekumpulan data tekstual yang bersar.Dari hasil interpretasi untuk 10 topik model LDA tersebut kemudian dapat dilihat tren yang dibicarakan masyarakat melalui nilai probabilitas topik terhadap seluruh dokumen. Adapun hasil probabilitas topik dapat dilihat pada gambar 5. berikut..



GAMBAR 5 Grafik nilai *probabilitas* topik

Berdasarkan hasil gambar 5, dapat dilihat topik yang mempunyai probabilitas paling tinggi adalah topik ke-9 dengan nilai 0.10057. Nilai probabilitas yang tinggi tersebut menunjukkan topik tersebut mempunyai mempunyai peluang paling tinggi untuk muncul dalam kumpulan dokumen yang artinya dominan dibicarakan di masyarakat yang berkomentar di @kompascom. Adapun hasil topik ke-9 sebagai berita yang paling dominan dibicarakan di masyarakat dapat dilihat pada tabel 2, berikut.

Tabel 3. Topik berita yang paling dominan (tren)

***Output* model LDA (Kata)**	**Interpretasi**	**Berita terkait @kompascom**
Bom Menyebarkan lokasi Terror Pelaku video tujuan	Tujuan pelaku untuk menyebarkan terror dengan video lokasi bom	Menyebarkan gambar atau video lokasi bom di Surabaya dapat ikut menyebarkan teror dan ketakutan yang jadi tujuan pelaku bom.

SIMPULAN DAN SARAN

Simpulan

Hasil dari model *Latent Dirichlet* (LDA) dengan metode *Bayesian* dengan estimasi parameter *Gibbs Sampling* diperoleh bahwa tren berita di masyarakat yang diterbitkan @kompascom pada tanggal 11 Mei 2018 sampai 25 Mei 2018 adalah *Menyebarkan gambar atau video lokasi bom di Surabaya dapat ikut menyebarkan teror dan ketakutan yang jadi tujuan pelaku bom* dengan nilai probabilitas topik 0.10057.

Saran

Berdasarkan hasil yang diperoleh diharapakan penelitian selanjutnya dapat dilakukan stemming bahasa indonesia pada saat tahap praproses dan rentang penentuan jumlah topik yang lebih besar lagi.

DAFTAR PUSTAKA

Bagus Putra, KB dan Kusumawardani, RP(2017). *Analisis Topik Informasi Publik Media Sosial di Surabaya Menggunakan Pemodelan Latent Dirichlet Allocation (LDA). JURNAL TEKNIK ITS Vol. 6, No. 2*

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

Griffiths, T.L. dan Steyvers, M., 2004, Finding Scientific Topics, Proceeding of the National Academy of Sciences, 101, 5228-

Ponweiser,M., 2012, *Latent Dirichlet Allocation in R, Tesis, Institute for Statistics and Mathematics.* Vienna University of Business and Economics, Vienna.

Retno Kusumaningrum, Satriyo Adhy, Suryono (2018). WCLOUDVIZ: Word Cloud Visualization of Indonesian News Articles Classification based on Latent Dirichlet Allocation. Jurnal TELKOMNIKA UAD. Vol 4 no.16.

Subeno, Bambang and Kusumaningrum, Retno and Farikhin,
Farikhin (2017) *OPTIMASI JUMLAH TOPIK KORPUS MENGGUNAKAN METODE LATENT DIRICHLET ALLOCATION (LDA).* Masters thesis, School of Postgraduate.

Tala, F. (2003). *A study of stemming effects on information retrieval in bahasa indonesia*[Tesis]. Amsterdam (Nl): Universiteit Van Amsterdam.Dou, F. (2013). A train dispacthing model base on fuzzy passenger demand forcesting durring holiday. *0mniascience*.