ISSN: 2528-4630

PEMODELAN TOPIK DENGAN LATENT DIRICHLET ALLOCATION

Zulhanif
Departemen Staistika FMIPA UNPAD
dzulhanif@yahoo.com

ABSTRAK. *Latent Dirichlet Allocation* (LDA), model probabilistik generatif pada sekumpulan data teks (*corpus*) . LDA adalah model Bayesian Hirarki , di mana sekumpulan data teks dimodelkan sebagai model campuran dari berbagai topik. Dalam kontek pemodelan teks, Pengembangan pemodelan LDA senduru merupakan pengembangan pada model topik sebelumnya yang dikenal sebagai *Probabilistic Latent Semantic Analysis* (PLSA), PLSA sendiri memiliki suatu keterbatasan dalam menentukan suatu topik dari sekumpulan data teks dikarenakan model PLSA tidak memperhatika urutuan kata sehingga suatu teks dengan jumlah kata yang sama akan bermakna lain jika memperhatikan urutannya. Salah satu pemodelan topik yang memperhatikan urutan dari suatu kata adalah model LDA. Model LDA pada penelitian ini merupakan model LDA yang dikembangkan oleh Blei (2003). Model LDA merupakan model probabistik dari sekumpulan *latent* (Topik) dari sekumpulan data teks (*corpus*) atau dikatakan model probabilitas topik yang memberikan representasi eksplisit dari sebuah dokumen. Pada penelitian ini menyajikan teknik inferensi berdasarkan algoritma Gibbs, untuk mengestimasi parameter Bayes dalam pemodelan pengelompokkan dokumen teks.

***Kata kunci**: Text Mining; LDA; Gibbs Sampling*

1. PENDAHULUAN

Perkembangan analisis teks pada pemodelan topik sendiri pada dasanya bersumber pada matriks *term frequency–inverse document frequency (tf-idf* . Selanjutnya berlanjut pada perkembangan pereduksian mariks *tf-idf* dengan menggunakan metode pereduksian dimensi seperti *Latent Semantic Analysis*(LSA) dan *Probabilistic Latent Semantic Analysis* (PLSA). *Latent Semantic Analysis*(LSA) metode yang dipatenkan pada tahun 1988 (US Patent 4,839,853) oleh Scott Deerwester, Susan Dumais, George Furnas, Richard Harshman, Thomas Landauer, Karen Lochbaum dan Lynn Streeter. Dalam konteks aplikasinya ke pencarian informasi, metode LSA ini juga disebut sebagai *Laten Semantic Indexing* (LSI). LSA dapat ditafsirkan sebagai cara yang cepat dan praktis untuk mendapatkan perkiraan perkiraan *substitutability* kontekstual penggunaan kata-kata dalam segmen teks yang besar yang belum ditentukan makna kesamaan antara kata-kata dan segmen teks yang mungkin mencerminkan suatu hubungan tertentu. Sebagai metode praktis untuk mengkarakterisasi arti dari kata, LSA menghasilkan ukuran hubungan kata-kata, bagian kata dan bagian-bagian yang berkorelasi dengan beberapa fenomena kognitif manusia yang melibatkan asosiasi atau kesamaan semantik. Konsekuensi praktis dari metode LSA ini, memungkinkan kita untuk sangat mendekati penilaian manusia untuk menilai kesamaan makna antara kata dan secara objektif memprediksi konsekuensi dari keseluruhan kata berdasarkan kesamaan antara bagian-bagian kata serta perkiraan yang kata yang sering muncul. Sedangkan *Probabilistic Latent Semantic Analysis* (PLSA) adalah sebuah algoritma yang diterapkan untuk memperkirakan makna sekumpulan teks menjadi suatu *cluster* atau kelompok (kategori) tertentu sehingga mempermudah para

analis untuk menarik suatu kesimpulan dari pengelompokkan yang terbentuk. Secara umum metode PLSA menggabungkan teori klasik tentang *vector space model*, *Singular Value Decompo*sition (SVD) serta model variabel latent, yang diformulasikan kedalam suatu bentuk model peluang dengan tujuan untuk mendapatkan suatu kelompok (*latent*) dari sekumpulan teks (*bag of words*). Aplikasi PLSA ini dapat diterapkan dalam analisis *sentiment* pada pasar saham, analisis kemiripan dokumen untuk mendeteksi plagiarism, analisis trending topik pada media sosial, Permasalahan yang timbul dalam penggunaan metode LSA ini adalah adanya faktor polysemi dalam pengelompokkan kata (Hofmann, 2001). Permasalahan polysemi pada kata dapat diatasi dengan mengunakan varian dari LSA yang dikenal sebagai *Probabilistic Latent Semantic Analysis* (PLSA). *Latent class* Metode PLSA pada dasarnya merupakan model campuran dari model *latent class* dengan kata lain model *latent class* untuk data teks, PLSA sendiri merupakan salah satu model *based clustering* yang bertujuan untuk membentuk *cluster* berdasarkan model peluang statistik, berbeda dengan metode *cluster* yang konvensional metode ini dapat dievaluasi berdasarkan ukuran statistik tertentu. Keutamaan metode PLSA sendiri dapat mereduksi dimensi matriks *term* yang terbentuk dari sekekumpulan teks yang direpresentasikan dalam sebuah variabel latent, sehingga ukuran dimensi matriks kemunculan *term* pada metode LSA dapat direduksi Hofmann[7]. Baik model LSA dan PLSA mengabaikan urutan kata (word ordering) dalam proses analisisnya, hal ini menjadi masalah dikarenakan beberapa term tertentu akan memiliki makna yang jauh, perkembangan selanjutnya pada analisis teks dalam pemodelan topik adalah seperti yang dikemukan oleh Blei[10] yang mana Blei[10] menggunakan model latent variabel dengan pendekatan bayesian dalam penaksiaran parameter modelnya.

2. METODE PENELITIAN

Latent Dirichlet Allocation (LDA) adalah model probabilistik generatif dari sekumpulan *corpus*, Ide dasarnya adalah bahwa dokumen dapat direpresentasikan sebagai model campuran dari berbagai topik yang disebut juga laten, di mana setiap topik dikarakateristikan oleh kata. LDA mengasumsikan proses generatif berikut untuk setiap dokumen $\boldsymbol{w}$ dalam sebuah corpus $\boldsymbol{D}$ adalah sbb :

1. Pilih $N \sim Pissson(\xi)$
2. Pilih $\boldsymbol{\theta} \sim Dir(\alpha)$
3. Untuk setiap *N* kata w_n :
   a. Pilih Topik $z_n \sim Multinomial(\boldsymbol{\theta})$
   b. Pilih sebuah kata w_n dari $p(w_n \mid z_n, \beta)$

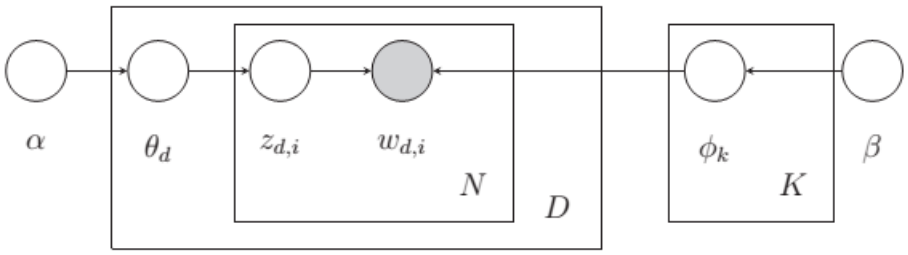
Beberapa asumsi penyederhanaan yang dibuat dalam model dasar LDA. Pertama, distribusi dari topik (latent) diketahui mengikuti *k* distribusi Dirichlet. Kedua, probabilitas kata adalah matriks $\boldsymbol{\beta}$ berukuran $k \times V$ yang mana $\beta_{ij} = p(w^j = 1 \mid z^i = 1)$. Sedangkan *k* distribusi Dirichlet memiliki fungsi densitas sbb:

$$p(\theta \mid \alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \ldots \theta_k^{\alpha_k - 1} \tag{2.1}$$

Bentuk distribusi bersama dari Topik mixture θ dari *N* topik ***z*** dan *N* kata ***w*** bersyarat α dan β adalah :

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \tag{2.2}$$

Representasi model LDA jika digambarkan dalam sebuah diagram dapat digambarkan sbb:



Gambar 2.1 Representasi Model LDA

Berdasarkan gambar 1 didapat distribusi bersama yang dari parameter pada model LDA sbb:

$$p(w, z, \theta, \phi \mid \alpha, \beta) = p(\theta \mid \alpha) p(z \mid \theta) p(\phi \mid \beta) p(w \mid z, \phi) \tag{2.3}$$

Pada Persamaan(2.3) diasumsikan bahwa topik setiap dokumen mengikuti distribusi Dirichlet sbb:

$$p(\theta \mid \alpha) = \frac{\Gamma(\alpha)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \tag{2.4}$$

Sedangkan distribusi peluang dari *z* untuk semua dokumen dan topik dalam terms dinotasikan $n_{d,k}$ yang merupakan berapa banyak topik *k* dikelompokan dengan kata dalam dokumen *d*:

$$p(z \mid \theta) = \prod_{d=1}^{D} \prod_{k=1}^{K} \theta_{d,k}^{n_{d,k}} \tag{2.5}$$

Distribusi peluang bersyarat untuk semua corpus ϕ_k juga mengikuti distribusi Dirichlet dengan parameter β. $\phi_{k,v}$ yang diformulasikan pada Persamaan (2.6)

$$p(\phi \mid \beta) = \prod_{k=1}^{K} \frac{\Gamma(\beta_{k,\cdot})}{\prod_{v=1}^{V} \Gamma(\beta_{k,v})} \prod_{v=1}^{V} \phi_{k,v}^{\beta_{k,v}-1} \tag{2.6}$$

Selengkapnya distribusi peluang corpus *w* bersyrat *z* dan ϕ direpresentasikan sbb:

$$p(w \mid z, \phi) = \prod_{k=1}^{K} \prod_{v=1}^{V} \phi_{k,v}^{n_{\cdot,k,v}} \tag{2.7}$$

$$p(w, z, \theta, \phi \mid \alpha, \beta) = p(\theta \mid \alpha)\, p(z \mid \theta)\, p(\phi \mid \beta)\, p(w \mid z, \phi) =$$

$$= \left(\prod_{d=1}^{D} \frac{\Gamma(\alpha)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{d,k}^{\alpha_k + n_{d,k,\cdot} - 1} \right) \times$$

$$= \left(\prod_{k=1}^{K} \frac{\Gamma(\beta_k)}{\prod_{v=1}^{V} \Gamma(\beta_{k,v})} \prod_{k=1}^{K} \phi_{k,v}^{\beta_{k,v} + n_{\cdot,k,v} - 1} \right) \tag{2.8}$$

Dalam proses perhitungannnya dilakukan pengintergrasian (Persamaan 2.8) sehingga bentuk Persamaan (2.8) menjadi sbb:

$$p(w, z \mid \alpha, \beta) = \iint p(z, w, \theta, \phi \mid \alpha, \beta) d\theta d\phi$$

$$= \iint \left(\prod_{d=1}^{D} \frac{\Gamma(\alpha)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{d,k}^{\alpha_k + n_{d,k,\cdot} - 1} \right) \times \left(\prod_{k=1}^{K} \frac{\Gamma(\beta_k)}{\prod_{v=1}^{V} \Gamma(\beta_{k,v})} \prod_{k=1}^{K} \phi_{k,v}^{\beta_{k,v} + n_{\cdot,k,v} - 1} \right) d\theta d\phi \tag{2.9}$$

Estimation

Proses komputasi pada Persamaan (2.9) menggunakan pendekatan *Gibbs sampling*, Gibbs Sampling sendiri merupakan pendekatan simulasi untuk mengkonstruk distribusi bersama berdasarkan distribusi marginal, pada proses estimasi parameter LDA, Gibbs sampling untuk LDA memerlukan nilai peluang dari topik *z* yang diasosiasikan untuk sebuah kata

(term) w_i sbb:

$$p(z_i \mid z_{-i}, \alpha, \beta, w) \tag{2.10}$$

yang z_{-i} dinotasikan sebagai semua topik kecuali topik z_i. Formulai model statistiknya adalah sbb :

$$p(z_i \mid z_{-i}, \alpha, \beta, w) = \frac{p(z_i, z_{-1}, w \mid \alpha, \beta)}{p(z_{-1}, w \mid \alpha, \beta)} \propto p(z_i, z_{-1}, w \mid \alpha, \beta) = p(w, z \mid \alpha, \beta) \quad (2.11)$$

$$\begin{aligned} p(z_i|\mathbf{z}^{(-i)}, \mathbf{w}) &= \frac{p(\mathbf{w}, \mathbf{z})}{p(\mathbf{w}, \mathbf{z}^{(-i)})} = \frac{p(\mathbf{z})}{p(\mathbf{z}^{(-i)})} \cdot \frac{p(\mathbf{w}|\mathbf{z})}{p(\mathbf{w}^{(-i)}|\mathbf{z}^{(-i)})p(w_i)} \\ &\propto \prod_d \frac{B(n_{d,\cdot} + \alpha)}{B(n_{d,\cdot}^{(-i)} + \alpha)} \prod_k \frac{B(n_{k,\cdot} + \beta)}{B(n_{k,\cdot}^{(-i)} + \beta)} \\ &\propto \frac{\Gamma(n_{d,k} + \alpha_k)\Gamma(\sum_{k=1}^K n_{d,k}^{(-i)} + \alpha_k)}{\Gamma(n_{d,k}^{(-i)} + \alpha_k)\Gamma(\sum_{k=1}^K n_{d,k} + \alpha_k)} \cdot \frac{\Gamma(n_{k,w} + \beta_w)\Gamma(\sum_{w=1}^W n_{k,w}^{(-i)} + \beta_w)}{\Gamma(n_{k,w}^{(-i)} + \beta_w)\Gamma(\sum_{w=1}^W n_{k,w} + \beta_w)} \\ &\propto (n_{d,k}^{(-i)} + \alpha_k) \frac{n_{k,w}^{(-i)} + \beta_w}{\sum_{w'} n_{k,w'}^{(-i)} + \beta_{w'}} \end{aligned} \quad (11)$$

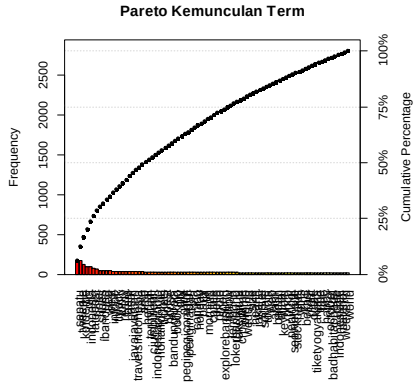
$$\begin{aligned} p(z_i \mid z^{(-i)}, w) &= \frac{p(w, z)}{p(w, z^{(-i)})} = \frac{p(z)}{p(z^{(-i)})} \cdot \frac{p(w \mid z)}{p(w^{(-i)} \mid z^{(-i)})\, p(w_i)} \\ &\propto \prod_d \frac{B(n_{d,\cdot} + \alpha)}{B(n_{n_{d,\cdot}}^{(-i)} + \alpha)} \prod_k \frac{B(n_{k,\cdot} + \alpha)}{B(n_{n_{k,\cdot}}^{(-i)} + \alpha)} \\ &\propto \frac{\Gamma(n_{d,k} + \alpha_k)\Gamma(\sum_{k=1}^{K} n_{d,k}^{(-i)} + \alpha_k)}{\Gamma(n_{d,k}^{(-i)} + \alpha_k)\Gamma(\sum_{k=1}^{K} n_{d,k} + \alpha_k)} \cdot \frac{\Gamma(n_{k,w} + \beta_w)\Gamma(\sum_{w=1}^{W} n_{k,w}^{(-i)} + \beta_w)}{\Gamma(n_{k,w}^{(-i)} + \beta_w)\Gamma(\sum_{w=1}^{W} n_{k,w} + \beta_k)} \quad (10) \\ &\propto (n_{d,k}^{(-i)} + \alpha_k) \frac{n_{k,w}^{(-i)} + \beta_w}{\sum_{w} n_{k,w}^{(-i)} + \beta_w} \end{aligned}$$

dengan $B(\alpha) = \dfrac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)}$

3. HASIL PENELITIAN DAN PEMBAHASAN

Pada penelitian ini akan dipergunakan data dari twitter dengan mengambil sampel sebanyak 1500 tweet dengan kata kunci #Bandung, Pada proses awal dilakukan tahapan pembersihan data teks dengan cara melakukan pembuangan kata yang tidak penting melalui tahapan *stopword*, dilanjutkan dengan proses *steamming* . Setelah tahapan *cleaning* data text sudah dilakukan langkah selanjutnya adalah dengan membuat matriks kemunculan kata berdasarkan tweet yang terjadi, proses ini dilakukan dengan bantuan software R. Hasil analisis awal menunjukkan *term* data memiliki tingkat kejadian yang

paling tinggi jika dibandingan dengan *term* kata lainnya hal ini dapat dilihat pada Gambar 3.1 sbb:

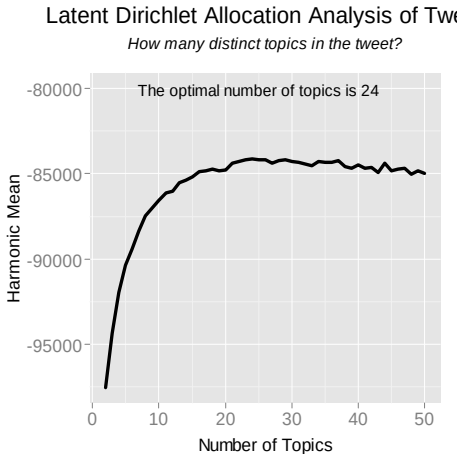


Gambar. 3.1 Pareto *Terms*

Kemunculan *term-term* yang sering muncul juga dapat dilihat dari gambar *wordcloud* pada Gambar 3.2 sbb

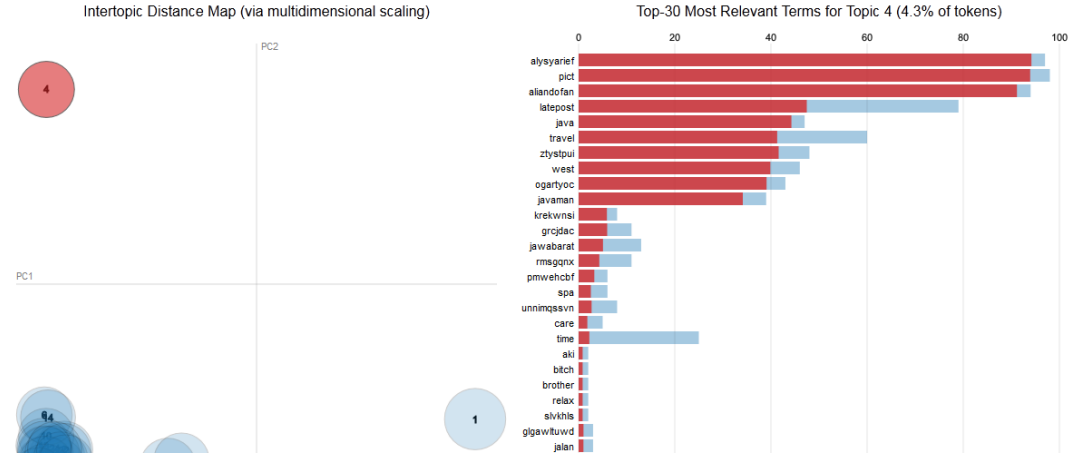
Gambar. 3.2 *Wordcloud Terms*

Penggunaan LDA sendiri akan dipergunakan pada pengidentifikasian Topik pada tweet #RDatamining. Adapun langkah awal dari proses LDA adalah menetapkan jumlah topik awal yang selanjutnya akan diproses lebih lanjut dengan Algoritma LDA. Jumlah topik ditentukan berdasarkan plot antara banyaknya topik dengan nilai *likelihood* dari model LDA sbb:



Gambar 3.3 Plot jumlah topik terhadap *Likelihood*

Bentuk Visualisasi dari untuk masing-masing topik dan kata dapat dilihat dalam diagram sbb:



Gambar 3.4 Plot Visualisai LDA

4. SIMPULAN

Metode pengklasteran LDA mengelompokkan data twitter dengan kata kunci bandung menghasilkan 24 buah Topik, Metode LDA yang diimplementasikan pada R masih terbatas dalam hal jumlah dimensi dari *term* yang terbentuk serta jumlah *term* yang dapat *diretrieve* dari server twitter, Keterbatasan lainnya bahwa metode yang ada saat ini hanya berlaku untuk bahasa inggris, untuk dapat diterapkan pada bahasa Indonesia diperlukan algoritma dan database kata dasar dalam bahasa Indonesia. Sehingga hal ini menjadi saran bagi peneliti lainnya untuk dapat mengimplementasikan metode ini dalam bahasa Indonesia.

DAFTAR PUSTAKA

[1] Anglin, J. M. (1970) The growth of word meaning. Cambridge, MA.: MIT Press.

[2] Anglin, J. M., Alexander, T. M., & Johnson, C. J. (1996). Word learning and the growth of potentially knowable vocabulary. Submitted for publication.

[3] Dumais, S. T. (1994). Latent semantic indexing (LSI) and TREC-2. In D. Harman (Ed.), The Second Text Retrieval Conference (TREC2) (National Institute of Standards and Technology Special Publication 500-215, pp. 105-116).

[4] Dumais, S. T. & Nielsen, J. (1992). Automating the assignment of submitted manuscripts to reviewers. In N. Belkin, P. Ingwesen, & A. M. Pejtersen (Eds.)

[5] Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, Association for Computing Machinery.

[6] Hutomo, A. & Zulhanif. 2013. Analisis Keluhan Penumpang PT. Kereta Api Indonesia (Persero) Menggunakan LSA dan Analisis Korespondensi. Univesitas Padjadjaran.

[7] Hofmann. T. 1999, Probabilistic Latent Semantic Indexing, Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR-99).

[8] Hofmann, T., Puzicha, J., & Jordan, M. I. (1999). Unsupervised learning from dyadic data. In Advances in Neural Information Processing Systems, Vol. 11, MIT Press

[9] Saul, L. & Pereira, F. (1997). Aggregate and mixed order Markov models for statistical language processing. In Proceedings of the 2nd International Conference on Empirical Methods in Natural Language Processing, pp. 81-89.

[10] Blei, David M.; Ng, Andrew Y.; Jordan, Michael I (January 2003). Lafferty, John, ed. "Latent Dirichlet Allocation". Journal of Machine Learning Research **3** (4–5): pp. 993–1022