# Pengukuran Topik pada LinkedIn Telkom University dengan Metode *Probabilistic Latent Semantic Analysis* (PLSA)

1st Nanda Firmansyah
*Fakultas Informatika*
*Universitas Telkom*
Bandung, Indonesia
nandafirmansyah@student.telkomuniversity.ac.id

2nd Donni Richasdy
*Fakultas Informatika*
*Universitas Telkom*
Bandung, Indonesia
donnir@telkomuniversity.ac.id

3rd Siti Sa'adah
*Fakultas Informatika*
*Universitas Telkom*
Bandung, Indonesia
sitisaadah@telkomuniversity.ac.id

***Abstrak*— Saat ini, media sosial sangat diharapkan manusia untuk mencari atau memberi info. LinkedIn merupakan salah satu media sosial yang digunakan untuk saling bertukar informasi secara terbuka. LinkedIn adalah suatu wadah media sosial yang menampung identitas asal pengguna tersebut, sehingga orang yg membutuhkan info tadi dapat mencari ataupun mengunjungi akun profil yg diperlukan sesuai dengan informasinya. Universitas Telkom menggunakan media sosial LinkedIn untuk memberitahu apa saja berita yg terdapat di Universitas Telkom. Terdapat banyak informasi serta topik yang dibahas pada profil LinkedIn Universitas Telkom. Dipenelitian ini dilakukan pengukuran topik pada LinkedIn Universitas Telkom menggunakan metode Probabilistic Latent Semantic Analysis (PLSA). Pengukuran topik ini dilakukan supaya mengetahui efektifitas pengukuran topik di LinkedIn Universitas Telkom dan dengan adanya PLSA pula diketahui apa saja topik yg dibahas pada profil LinkedIn Universitas Telkom. Kemudian dilakukan juga pengukuran perbandingan pengukuran dengan adanya stemming dan stopword. Dari analisis yg dilakukan semakin banyak topik yang diterapkan semakin bagus nilai dari perhitungan log-likehood yg didapat. Dan dari pengukuran yang dilakukan, adanya penerapan stopword dan stemming adalah kondisi terbaik daripada tidak adanya stemming ataupun stopword.***

***Kata kunci*— Pengukuran topik, universitas telkom, LinkedIn, probabilistic latent semantic.***

## I. PENDAHULUAN

### A. Latar Belakang

Saat ini, media sosial sangat diperlukan manusia untuk mencari ataupun memberi info. LinkedIn merupakan media sosial yang digunakan agar dapat mencari ataupun memberikan info secara terbuka. Universitas Telkom mengunakan LinkedIn untuk memberikan informasi apa saja tentang Universitas Telkom, oleh karena itu orang yang mencari suatu info tersebut bisa mencari info yang dibutuhkan pada profil LinkedIn Universitas Telkom.

Banyak topik atau info yang tersedia berkaitan dengan Universitas Telkom pada profil LinkedIn Universitas Telkom. Salah satu metode yang cepat dan efisien untuk mengetahui topik-topik apa saja yang sedang tren di suatu media sosial adalah menggunakan pemodelan topik. Pemodelan topik berfungsi menyampaikan gambaran topik bersifat informatif yang bisa langsung diterima oleh pengguna. Pemodelan topik diharapkan bisa membantu pengguna dengan praktis dan cepat memahami perkembangan berita terkini [1]. Pemodelan topik ialah analisis teks yg bermanfaat pada pemodelan data tekstual dengan tujuan menemukan topik yg tersembunyi di dalamnya [2].

Pada penelitian ini dilakukan pemodelan topik menggunakan metode *Probabilistic Latent Semantic Analysis* (PLSA). PLSA memiliki dasar statistik yang kuat karena di penerapannya memakai prinsip *likelihood*. Metode ini memakai nilai probabalitas asal setiap kata, sehingga istilah-istilah tadi mampu dikelompokkan sebagai topik tertentu. PLSA umumnya digunakan dalam menganalisis pengolahan bahasa alami, sistem pembelajaran dengan data teks, serta lain-lain [3]. Selanjutnya dilakukan pengukuran analisis menggunakan pengukuran log-likelihood. Pengukuran log-likelihood, nilai yg didapat berupa negatif serta hasil pengukuran yang semakin mendekati nilai nol adalah hasil terbaik [4].

### B. Topik dan Batasannya

Topik pada penulisan ini adalah melakukan analisis pengukuran log-likelihood pada profil LinkedIn Universitas Telkom menggunakan metode *Probabilistic Latent Semantic Analysis* (PLSA). Dan dari metode PLSA yang diterapkan didapatkan juga topik apa saja yang paling bahas pada profil LinkedIn Universitas Telkom. Batasan pada penelitian ini adalah data yang diperoleh berasal dari profil LinkedIn Universitas Telkom sebanyak 139 data postingan, dan data yang diperoleh sudah tersedia.

### C. Tujuan

Penelitian ini dilakukan memiliki tujuan untuk mengetahui hasil dari pengukuran topik dan topik apa saja yang banyak dibahas pada profil LinkedIn Universitas Telkom, dengan menggunakan metode *Probabilistic Latent Semantic Analysis* (PLSA). Dalam proses penelitian dilakukan preprocessing data yang didapat lalu diterapkan pemodelan topik menggunakan PLSA untuk mendapatkan tren topik yang dibicarakan pada Linkedin Telkom University. Selanjutnya dilakukan perbandingan nilai log-likelihood antar jumlah topik dan pengaruh adanya penerapan stopword dan stemming dari hasil pengukuran nilai *log-likelihood.*

Dengan dilakukannya penelitian ini dapat diketahui kondisi apa terbaik pada hasil pengukuran topik dan

diketahui topik apa saja yang banyak dibahas dalam suatu informasi yang didapat.

## II. KAJIAN TEORI

### A. LinkedIn

LinkedIn adalah jaringan profesional terbesar di dunia di internet. Kita dapat menggunakan LinkedIn untuk menemukan pekerjaan atau magang yang tepat, terhubung dan memperkuat hubungan profesional, dan mempelajari keterampilan yang kita butuhkan untuk berhasil dalam karier kita. Kita dapat mengakses LinkedIn dari desktop, aplikasi seluler LinkedIn, pengalaman web seluler, atau aplikasi seluler.

### B. Probabilistic Latent Semantic Analysis (PLSA)

*Probabilistic Latent Semantic Analysis (PLSA),* memiliki pengelompokan data berdasarkan modelnya [9]. *Probabilistic Latent Semantic Analysis (PLSA)* juga merupakan teknik *information retrieval* untuk menganalisis dua keterhubungan kejadian data yang berdasarkan model statistik yang disebut dengan *aspect model* yang merupakan model variabel laten untuk data *co-occurrence* yang mengasosiasikan variabel kelas laten $z \in Z = \{z1, z2, \dots, zK\}$ dengan setiap pengamatan kejadian kata dalam sebuah dokumen tertentu [10].

$$P(w_j d_j) = \sum_{k=1}^{K} P(w_j|z_k) P(d_i|z_k) \quad (3)$$

Simbol pada perumusan (3) yaitu:

1. $P(d_i)$ : Menunjukkan probabilitas suatu kejadian kata yang akan diamati dalam dokumen $(d_i)$ tertentu.
2. $P(w_j|z_k)$ : Menunjukkan probabilitas bersyarat dari sebuah kata $(w_j)$ tertentu yang dikondisikan terhadap variabel laten $(z_k)$.
3. $P(d_i|z_k)$ : Menunjukkan distribusi probabilitas dokumen tertentu terhadap ruang variabel laten.

Pada penelitian ini PLSA menggunakan algoritma *Expectation Maximization* untuk memperkirakan nilai *maximum likelihood* dalam model variabel laten.

Expectation Step (E-step), menghitung posterior probabilities untuk variabel z berdasarkan estimasi saat ini.

Rumus E-Step:

$$P(z_k|d_i, w_j) = \frac{P(w_j|z_k).P(d_i|z_k)}{\sum_{k=1}^{K} P(w_j|z_k).P(d_i|z_k)} \quad (4)$$

Maximization Step (M-step), memaksimalkan M-step dengan memperbaharui parameter yang digunakan untuk menghitung probabilitas posterior variabel z untuk memaksimalkan nilai *likelihood*.

$$P(w_j|z_k) = \frac{\sum_{i=1}^{I} n(d_i, w_j).P(z_k|d_i, w_j)}{\sum_{i=1}^{I} \sum_{j=1}^{J} n(d_i, w_j).P(z_k|d_i, w_j)} \quad (5)$$

dan

$$P(z_k|d_i) = \frac{\sum_{i=1}^{I} n(d_i, w_j).P(z_k|d_i, w_j)}{\sum_{j=1}^{J} \sum_{k=1}^{K} n(d_i, w_j).P(z_k|d_i, w_j)} \quad (6)$$

Algoritma EM akan mengusahakan memperoleh nilai error yang semakin kecil. Jika nilai error masih tinggi, nilai bobot akan dioptimasikan agar nilai error yang ada semakin kecil. sehingga akan sangat memungkinkan dilakukan semakin banyak *training*, maka topik dan probabilitas yg dibentuk akan semakin baik [11].

Langkah-langkah menjalankan algoritma EM tersebut adalah sebagai berikut:

1. Menentukan variabel laten *(z).*
2. Menginisialisasi secara random nilai $P(w_j|z_k)$ dan $P(d_i|z_k)$.
3. Normalisasi nilai yang telah diberikan pada parameter $P(w_j|z_k)$.
4. Menghitung probabilitas untuk masing-masing parameter menggunakan EM.
5. Menghitung nilai *likelihood L* berdasarkan nilai parameter terakhir.

### C. Preprocessing

Data diolah terlebih dahulu sebelum diproses untuk analisis berikutnya. Preprocessing adalah langkah sebelum analisis data untuk mengolah kata-kata yang digunakan. Ada banyak langkah preprocessing yang bisa dilakukan, antara lain case folding, data cleaning, tokenizing, stopword dan stemming. Tahapan ini mencakup beberapa langkah pra-pemrosesan yang dijelaskan di bawah ini:

1.Data cleaning

Pada tahap ini dilakukan standarisasi pada data, seperti perubahan semua huruf menjadi huruf kecil, menghapus angka, menghapus tanda baca dan simbol unik [5].

2. Tokenizing

Pada tahap ini dilakukan proses perubahan teks menjadi potongan kata yang selanjutnya digunakan pada analisis berikutnya [6].

3.Stopword

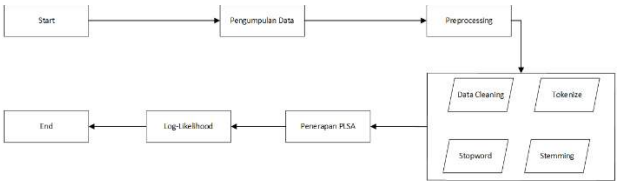
Dalam tahapan ini dilakukan proses penghilangan kata–kata memiliki informasi yang rendah atau yang tidak memiliki makna [7]. Pada tahap ini digunakan library dari NLTK.

4.Stemming

Tahapan stemming ini dilakukan perubahan suatu kata menjadi kata baku [7]. Pada tahap ini digunakan library dari sastrawi.

## III. METODE

Berikut rancangan sistem yang dilakukan dalam menganalisis pengukuran topik dengan metode *Probabilistic Latent Semantic Analysis (PLSA)* pada LinkedIn Universitas Telkom.




Gambar 1
(Tahapan penelitian)

Pada gambar 1, tahap pertama yang dilakukan dilakukan pengumpulan data yang berasal dari profil LinkedIn Universitas Telkom. Tahap selanjutnya dilakukan tahap

*preprocessing* yang melewati proses *data cleaning, tokenize, stopword,* dan *stemming*. Selanjutnya dilakukan penerapan PLSA yang akan menghasilkan topik apa saja yang banyak muncul pada profil LinkedIn Universitas Telkom. Dan pada tahap akhir dilakukan pengukuran topik menggunakan *log-likelihood* dan dilakukan perbandingan penerapan *stopword* dan *stemming.*

Dalam tahapan ini data yang digunakan dalam analisis sudah tersedia, data tersebut diperoleh dari profil LinkedIn Univeritas Telkom. Pada data yang belum diproses tersebut dilakukan analisis dengan pembuatan word cloud untuk mendapatkan kata apa saja yang muncul data tersebut.

Berikut beberapa contoh data yang digunakan, terdapat pada tabel 1. Pada gambar 2 hasil dari *word cloud* terdapat kata terbanyak yang ada pada data profil LinkedIn Universitas Telkom.

Tabel 1
(Beberapa data yang diperoleh dari LinkedIn Universitas Telkom)

| No | Postingan |
|---|---|
| 1 | Selamat Hari Raya Idul Adha 1442 H<br>Semoga Idul Adha kali ini mampu meningkatkan keimanan dan ketakwaan kita dalam beragama dan semakin bersabar dalam menghadapi musibah Covid-19<br>Semoga kita semua dalam lindungan Allah Swt |
| 2 | Tel-U News Highlight Eps.26 \| Simak berita terkini dari kampus Telkom University https://lnkd.in/gqAHWwm |
| 3 | Tel-U Kampus Pertama Dengan Program Studi Cyber Security and Digital Forensic https://lnkd.in/gA2C3SY |
| 4 | Telkom University Bersama Alumni Gelar Doa Bersama dan Galang Donasi https://lnkd.in/gd95jQd |

Dari contoh beberapa data yang ditampilkan dapat dilihat data tersebut belum bersih, masih banyak karakter, tanda baca yang tidak dibutuhkan oleh karena setelah ini dilakukan tahapan *data cleaning.*


Gambar 2
(Word Cloud data LinkedIn Universitas Telkom)

## IV. HASIL DAN PEMBAHASAN

Dalam penerapan PLSA ini terdapat 3 skenario, pada skenario pertama dilakukan pembagian topik yang ada pada postingan profil LinkedIn Universitas Telkom. Tahap ini menggunakan 10 topik dengan iterasi 30 kali dan *threshold* 10. Lalu pada skenario ke 2 diterapkan pengujian berdasarkan banyaknya topik yang diterapkan dan skenario ke 3 dilakukan pengujian terhadap adanya penerapan stopword dengan stemming. Semua pengujian ini dilakukan dengan metode pengukuran *log-likelihood*.

A. Hasil Pembagian Topik

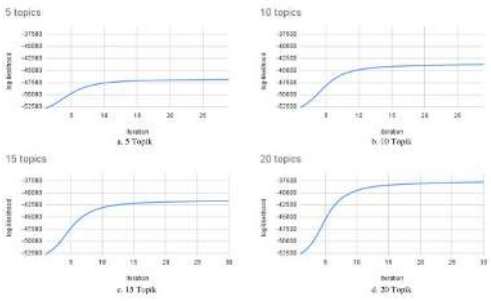
Dari penerapan metode PLSA yang dilakukan dengan metode *EM (Expectation-maximization)* maka didapatkan hasil topik pada tabel 5 dengan topik yang digunakan sebanyak 10. Topik tersebut dibagi dalam 10 pembagian topik, 10 topik yang tersedia pada tabel 5 merupakan pembagian topik berdasarkan pembahasan paling banyak muncul pada tiap masing-masing topik pada profil LinkedIn Universitas Telkom.

Tabel 2
(Hasil topik yang dihasilkan)

| Topic ID | Term |
|---|---|
| Topik 1 | creative, th, university, talk, telkom, rp, engineering, series, bandung, movement |
| Topik 2 | university, telkom, indonesia, tip, telu, lab, covid, dokter, alat, masker |
| Topik 3 | university, telkom, gemastik, webinar, live, wisuda, rumah, tuan, program, xiii |
| Topik 4 | indonesia, telkom, batik, university, dr, webometrics, direktur, bangsa, prof, budaya |
| Topik 5 | telkom, university, indonesia, telu, choir, career, kerja, virtual, hasil, teknologi |
| Topik 6 | selamat, nasional, alumni, didik, indonesia, bangsa, moga, sharing, digital, raya |
| Topik 7 | university, telkom, telu, kerjasama, raih, qs, universitas, hasil, program, green |
| Topik 8 | university, telkom, peringkat, guru, indonesia, masuk, dunia, capai, rankings, universitas |
| Topik 9 | telkom, program, university, studi, mahasiswa, daftar, selamat, informasi, buka, telkomsel |
| Topik 10 | telkom, university, inovasi, indonesia, telu, teknik, dosen, riset, fakultas, elektro |

B. Hasil Pengukuran Log-Likelihood

Dilakukan beberapa perbandingan dengan pengukuran antara *log-likelihood* dengan iterasi. Iterasi yang digunakan pada setiap pengujian sebanyak 30 iterasi.

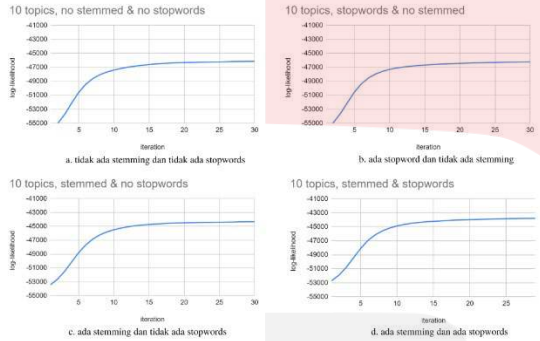



Gambar 3
(Hasil perbandingan topik)

Nilai yang dihasilkan dari pengukuran *log-likelihood* bernilai negatif, nilai yang dihasilkan semakin bagus jika nilai pengukurannya mendekati nilai 0

Dari percobaan pengukuran berdasarkan 5 topik (gambar 4a), 10 topik (gambar 4b), 15 topik (gambar 4c), *dan* 20 topik (gambar 4d), memiliki hasil pengukuran yang berbeda, pengukuran tersebut dilakukan berdasarkan perbandingan antara banyaknya topik dengan iterasi, disini nilai iterasi yang digunakan 30, dengan *threshold* 10. Berdasarkan tujuan penelitian yang dilakukan dilakukan pengukuran topik yang diukur menggunakan *log-likelihood* dapat disimpulkan dari pengukuran topik tersebut, maka dari percobaan tersebut disimpulkan semakin banyak topik semakin bagus pengukuran yang didapat.

Lalu dilakukan juga perbandingan antara penerapan stopword dan stemmed dengan topik 10 dan iterasi sebanyak 30.




Gambar 4
(Hasil perbandingan penerapan stopword dan stemming)

Sama halnya pengukuran yang dilakukan pada topik, pengukuran pada penggunaan *stopword* dan *stemming* nilai yang didapat pada pengukuran ini dengan 10 topik, pengukuran ini dilakukan perbandingan tidak adanya *stopword* dan *stemming* (gambar 5a), adanya stopword tetapi tidak ada stemming (gambar 5b), adanya stemming tetatpi tidak ada stopword (gambar 5c), dan adanya *stemming* dan *stopword* (gambar 5d). Dari pengukuran tersebut menghasilkan nilai terbaik pada adanya penggunaan *stopword* dan *stemming*.

C. Analisis Hasil Pengujian

Hasil pengujian menggunakan *log-likelihood* bernilai negatif. Nilai terbaik pada pengukuran metode ini hasil yang pengukurannya nilainya yang mendekati nilai nol, untuk hasil terburuk yang didapatkan nilai yang menjauhi dari nilai nol.

Pada skenario pertama dilakukan pengukuran terhadap banyaknya topik yan digunakan pada pengukuran ini menggunakan 5 topik, 10 topik, 15 topik dan 20 topik dengan nilai *threshold* 10 dan iterasi 30. Hasil pengukuran yang didapat bisa dilihat pada tabel 6.

Tabel 3
(Hasil pengukuran topik)

| condition | log-likelihood |
|---|---|
| 20 topics | -37839.03833 |
| 15 topics | -41709.40459 |
| 10 topics | -43790.33199 |
| 5 topics | -46906.33606 |

Pengukuran yang didapat berdasarkan dari perbandingan nilai *log-likelihood* topik dengan banyaknya iterasi yang dilakukan. Dan dari hasil pengukuran yang didapatkan dapat dilihat kondisi dengan 20 topik menghasilkan nilai terbaik, karena nilai yang didapat mendekati nilai 0.

Lalu pada skenario kedua dilakukan pengukuran terhadap pengaruh adanya penerapan *stopword* dan *stemming*. Pada pengukuran ini menggunakan 10 topik dengan nilai *threshold* 10 dan iterasi 30. Hasil pengukuran ada pada tabel 7.

Tabel 4
(Hasil pengukuran stopword dan stemming)

| condition | log-likelihood |
|---|---|
| stemmed & stopwords | -43790.33199 |
| stemmed & no stopwords | -44343.40239 |
| no stemmed & no stopwords | -46160.38634 |
| stopwords & no stemmed | -46219.3147 |

Pengukuran ini berdasarkan dari perbandingan nilai *log-likelihood* dari setiap kondisi yang diaplikasikan dengan iterasi yang dilakukan. Dan dari hasil pengukuran yang didapatkan dapat dilihat kondisi dengan adanya penerapan stopword dan stemming menghasilkan nilai terbaik, karena nilai yang didapat mendekati nilai 0.

## V. KESIMPULAN

Dari pengujian yang dilakukan diperoleh hasil analisis sebagai berikut. Pengukuran pada *log-likelihood* menghasilkan nilai yang bernilai negatif. Nilai yang dihasilkan dari pengukuran semakin bagus apabila nilai tersebut mendekati nilai 0. Lalu dari penerapan metode PLSA yang diterapkan, didapatkan berbagai macam topik yang paling banyak dibahas pada profil LinkedIn Univeritas Telkom. Selanjutnya pada pengukuran *log-likelihood* terhadap topik disimpulkan bahwa semakin banyaknya topik yang diolah semakin bagus nilai yang didapatkan. Dan pengukuran pada penerapan stemming dan stopword diperoleh hasil dengan adanya penerapan stemming dan stopword merupkan kondisi yang terbaik karena dari pengukuran *log-likelihood* menghasilkan nilai yang paling bagus daripada tidak adanya penerapan stemming ataupun stopword.

Dari analisis yang dilakukan tujuan dari dilakukan analasis ini sudah tercapai, kita sudah mengetahui bagaimana hasil dari pengukuran topik yang diaplikasikan pada profil LinkedIn Universitas Telkom dan topik apa saja yang dibahas pada profil LinkedIn Universitas Telkom juga telah didapatkan. Pada penerapan stemming dan stopword dari analisis yang dilakukan dapat disimpulkan dengan adanya stemming dan stopword tersebut merupakan kondisi terbaik.

Saran untuk penelitian yang akan datang dilakukan pengukuran topik dengan metode yang lain agar dari hasil pengukuran yang didapatkan dapat dibandingkan hasil setiap metode tersebut, mana yang lebih efisien dalam melakukan pengukuran topik tersebut.

## REFERENSI

[1] Rezza, Mohammad. (2019). Pemodelan Topik Pada Portal Berita Online Menggunakan Latent Dirichlet Allocation (LDA). Tesis. Universitas Gadjah Mada.

[2] Kengken, Ruske Illa. (2014). Pemodelan Topik Untuk Media Sosial Menggunakan Latent Dirichlet Allocation. Skripsi. Universitas Gadjah Mada.

[3] Kumar, A., Sanyal, S. (2010). Efect of Pronoun Resolution on Document Similarity. International Journal of Computer Application (0975-8887) volume 1-No. 16. India: Indian Institut of Information Technology Allahabad.

[4] Jiang, Bo., Jianjun,Wu, And., Yi, Feng.(2020). Topic Modeling for Short Texts via Word Embedding and Document Correlation. IEEE, 30692-30705.

[5] M. P. Arsya, C. Imam, and A. P. Putra. (2019). Tampilan Analisis Sentimen Tentang Opini Maskapai Penerbangan pada Dokumen Twitter Menggunakan Algoritme Support Vector Machine (SVM).

[6] S.Vijayarani and R.Janani, "Text Mining: open Source Tokenization Tools –An Analysis," Adv. Comput. Intell. An Int. J., vol. 3, no. 1, pp. 37–47, 2016, doi: 10.5121/acii.2016.3104.

[7] E. B. Setiawan, D. H. Widyantoro, and K. Surendro, "Feature expansion using word embedding for tweet topic classification," Proceeding 2016 10th Int. Conf. Telecommun. Syst. Serv. Appl. TSSA 2016 Spec. Issue Radar Technol., no. 2011, 2017, doi: 10.1109/TSSA.2016.7871085.

[8] M. N. Dr. S. Vijayarani, Ms. J. Ilamathi, "Preprocessing Techniques for Text Mining Preprocessing Techniques for Text Mining," Int. J. Comput. Sci. Commun. Networks, vol. 5, no. October 2014, pp. 7–16, 2015

[9] T. Hofmann. (1999). Probabilistic latent semantic analysis. In UAI'99, pages 289–296. Morgan Kaufmann.

[10] Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. Machine Learning, 42, 177–196. USA: Department of Computer Science, Brown University.

[11] Suhartono, D. (2014). Probabilistic Latent Semantic Analysis (PLSA) untuk Klasifikasi Dokumen Teks Berbahasa Indonesia.