METODE LATENT DIRICHLET ALLOCATION UNTUK EKSTRAKSI TOPIK DOKUMEN

Putu Manik Prihatini, I Ketut Suryawan, I Nyoman Mandia
Politeknik Negeri Bali
Bukit Jimbaran, P.O. Box 1064 Tuban Badung – BALI
Phone:+62-361-701981, Fax:+62-361-701128 E-mail: manikprihatini@pnb.ac.id

Abstrak: Proses ekstraksi dalam temu kembali informasi menghasilkan fitur yang akan menjadi ciri unik dari suatu dokumen, sehingga dokumen tersebut dapat dinyatakan relevan atau tidak relevan terhadap kata kunci yang diketikkan oleh pengguna. Salah satu metode ekstraksi berbasis topik yang mampu menemukan hubungan struktural internal dalam dokumen adalah Latent Dirichlet Allocation, karena mampu bekerja pada level kata, dokumen dan korpus. Akan tetapi, penelitian-penelitian terkait ekstraksi topik dokumen dengan metode Latent Dirichlet Allocation lebih banyak dikembangkan untuk teks berbahasa asing, dan sampai saat ini, sangat jarang ditemukan untuk teks dokumen berbahasa Indonesia. Untuk itu, pada penelitian ini, diimplementasikan metode ekstraksi topik Latent Dirichlet Allocation untuk aplikasi teks dokumen berbahasa Indonesia. Tahapan penelitian terdiri dari akusisi data, tokenisasi, filtering, stemming, re-filtering, inisialisasi, sampling topik, perhitungan parameter final dan evaluasi. Hasil implementasi metode ekstraksi topik Latent Dirichlet Allocation untuk teks dokumen berbahasa Indonesia diuji dengan metriks pengukuran Precision, Recall dan F-Measure. Hasil penelitian ini nantinya diharapkan dapat menjadi referensi dalam melakukan penelitian-penelitian tentang metode ekstraksi topik untuk teks dokumen berbahasa Indonesia.

Kata Kunci: ekstraksi topik, latent dirichlet allocation, teks Indonesia

LATENT DIRICHLET ALLOCATION METHOD FOR DOCUMENT TOPIC EXTRACTION

Abstract: *The extraction process in the information retrieval results in features that will be the unique characteristics of a document, so that the document can be declared relevant or irrelevant to the keyword typed by the user. One of the topic-based extraction methods that can find internal structural relationships in documents is Latent Dirichlet Allocation, because it is able to work at word, document and corpus levels. However, studies related to document topic extraction with the Latent Dirichlet Allocation method are more developed for foreign texts, and very rarely found for Indonesian text. Therefore, in this study, Latent Dirichlet Allocation topic extraction method was implemented for Indonesian documents text. The study consists of data acquisition, tokenization, filtering, stemming, re-filtering, initialization, topic sampling, final parameter calculation and evaluation. The results were tested with Precision, Recall and F-Measure measurement metrics. The results of this study will be expected to be a reference in conducting research on topic extraction for Indonesian documents text.*

Key words*: topic extraction, latent dirichlet allocation, Indonesian texts*

I. PENDAHULUAN

Pencarian informasi, saat ini, tidak hanya dilakukan melalui media cetak, melainkan lebih banyak dilakukan dengan memanfaatkan teknologi internet, dalam bentuk teknologi sistem temu kembali informasi, atau *Information Retrieval System* (IRS). Salah satu implementasi dari IRS adalah mesin pencari informasi, yang dikenal sebagai *search enginee*. Mesin ini memudahkan pengguna dalam melakukan pencarian informasi, dimana pengguna cukup mengetikkan kata kunci pada kotak teks yang tersedia, kemudian mesin menampilkan sekumpulan dokumen yang dianggap sesuai dengan kata kunci tersebut.

Proses temu kembali informasi yang dilakukan oleh mesin pencari informasi terbagi menjadi proses *offline* dan *online*. Proses *offline* dilakukan untuk mengolah sekumpulan teks dokumen digital yang tersimpan dalam server mesin pencari, sedangkan proses *online* dilakukan untuk mengolah kata kunci yang diketikkan oleh pengguna. Proses pengolahan teks dokumen digital melibatkan proses pra pengolahan teks dan ekstraksi fitur. Proses pra pengolahan teks terdiri dari proses penguraian teks dokumen menjadi daftar kata (*tokenization*), penghapusan kata-kata yang tidak penting (*filtering*), dan jika diperlukan dilanjutkan dengan proses pemisahan kata dasar dari imbuhan (*stemming*). Proses ekstraksi fitur terdiri dari proses pembobotan dan pengindeksan

teks pada dokumen. Proses pengolahan kata kunci melibatkan proses pencocokan antara kata kunci yang diketikkan pengguna dengan fitur-fitur hasil ekstraksi yang dimiliki oleh setiap dokumen. Hasil pencocokan ini berupa dokumen yang relevan dan tidak relevan. Dokumen yang dianggap relevan diurutkan berdasarkan nilai kemiripannya dan ditampilkan sebagai hasil pencarian informasi.

Dilihat dari mekanisme proses temu kembali informasi diatas, maka proses ekstraksi fitur memegang peranan penting dalam proses temu kembali informasi. Hal ini disebabkan karena hasil ekstraksi menjadi ciri unik dari suatu dokumen, sehingga dokumen tersebut dapat dinyatakan relevan atau tidak relevan terhadap kata kunci yang diketikkan oleh pengguna. Proses ekstraksi fitur dapat dilakukan dengan beberapa metode. Salah satu metode ekstraksi fitur yang banyak digunakan dalam pengolahan teks dokumen adalah Term Frequency-Inverse Document Frequency (TF-IDF), seperti yang dilakukan pada penelitian-penelitian [1-5]. Akan tetapi metode ini memiliki kelemahan, yaitu tidak mampu menemukan hubungan struktur internal dalam dokumen, seperti sinonim dan polisemi.

Sinonim mengacu pada beberapa kata yang berbeda tetapi memiliki makna yang sama, misalnya, kata "mobil" memiliki sinonim dengan kata "kendaraan". Polisemi mengacu pada suatu kata yang sama tetapi dapat memiliki konteks makna yang berbeda, misalnya kata "model" bisa mengacu pada makna "bentuk" atau "profesi" sesuai dengan konteks dimana kata tersebut digunakan. Metode Latent Semantic Indexing (LSI) dikembangkan untuk menangani masalah sinonim dan polisemi, seperti yang dilakukan pada beberapa penelitian [6-8]. Metode LSI masih memiliki kelemahan karena hanya bekerja sampai pada level kata dan dokumen. Dalam kenyataannya, jumlah dokumen dalam server yang disebut dengan korpus sangatlah besar, sehingga hubungan antar dokumen dalam korpus tersebut harus dianalisis juga. Untuk itu, dikembangkanlah metode ekstraksi topik Latent Dirichlet Allocation (LDA) yang mampu menangani level kata, dokumen dan korpus [9].

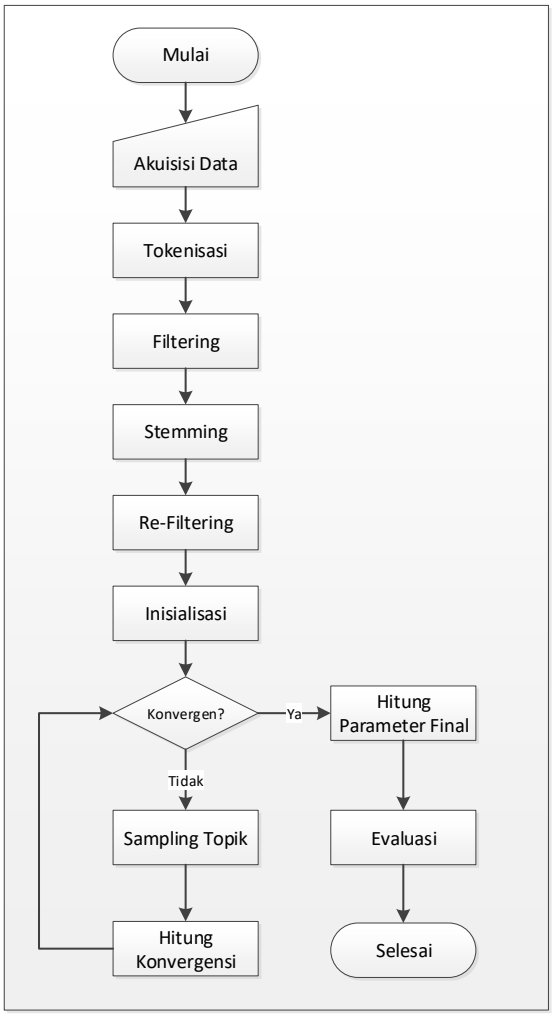
LDA telah banyak dikembangkan dalam penelitian terkait pengolahan teks dokumen [10-12]. Mekanisme kerja LDA terbagi menjadi dua bagian yaitu penalaran dan implementasi. Penalaran merupakan proses LDA untuk menentukan bobot dari setiap kata yang ada pada setiap dokumen dalam korpus. Implementasi merupakan tahap penerapan LDA untuk kebutuhan temu kembali informasi selanjutnya. Ada beberapa metode penalaran untuk LDA, namun yang paling banyak digunakan adalah Gibbs Sampling [13, 14]. Penelitian-penelitian terkait pengolahan teks dokumen dengan metode LDA yang disebutkan diatas lebih banyak dikembangkan untuk teks berbahasa asing, dan sampai saat ini, sangat jarang ditemukan untuk teks dokumen berbahasa Indonesia.

Berdasarkan uraian diatas, maka pada penelitian ini, diimplementasikan metode ekstraksi topik LDA untuk aplikasi teks dokumen berbahasa Indonesia. Penelitian ini diharapkan dapat menjadi referensi dalam melakukan penelitian-penelitian tentang metode ekstraksi topik untuk teks dokumen berbahasa Indonesia.

II. METODE PENELITIAN

Metode pembangunan aplikasi ekstraksi topik teks dokumen Bahasa Indonesia dengan metode LDA digambarkan dalam bentuk rancangan penelitian seperti pada Gambar 1.

Tahap akuisisi data merupakan proses yang dilakukan untuk mengumpulkan data yang digunakan dalam penelitian. Data penelitian yang digunakan berupa data file berita yang diperoleh melalui proses pengambilan data secara manual pada situs media berita Indonesia. Data penelitian yang dikumpulkan sebanyak 500 berita dari 5 kategori.



Gambar 1. Rancangan Penelitian

Tahap tokenisasi merupakan proses untuk memecah teks dokumen menjadi bentuk terkecil seperti paragraf, kalimat atau kata. Proses ini dilakukan pada 500 file teks berita yang telah dikumpulkan.

Tahap filtering merupakan proses untuk menghilangkan kata-kata yang tidak memiliki makna terhadap isi teks seperti kata depan, kata sambung, dan sejenisnya. Proses ini dilakukan pada 500 file hasil tokenisasi.

Tahap stemming merupakan proses untuk memperoleh kata dasar dari setiap kata pada teks dokumen, dengan membuang imbuhan pada kata. Proses ini dilakukan pada 500 file hasil filtering menggunakan aturan prefix dan suffix , serta kamus kata dasar hasil penelitian sebelumnya [15].

Tahap re-filtering merupakan proses untuk menghilangkan kembali kata-kata yang tidak bermakna terhadap hasil teks, dimana proses ini dilakukan terhadap 500 file hasil stemming. Proses re-filtering dilakukan dengan langkah yang sama dengan langkah filtering.

Tahap inisialisasi merupakan proses untuk menentukan frekuensi kemunculan kata dari setiap kata pada setiap file teks. Proses ini dilakukan pada 500 file hasil re-filtering. Proses inisialisasi dilakukan dengan langkah: menghitung frekuensi kemunculan setiap kata pada setiap file teks, menentukan topik setiap kata dengan distribusi multinomial berdasarkan nilai frekuensi kemunculan kata, menentukan matriks kata-topik dan dokumen-topik, menghitung jumlah total dari distribusi kata-topik dan dokumen-topik, dan menyimpan hasil matriks.

Tahap sampling topik merupakan proses untuk menentukan topik baru dari setiap kata pada setiap file teks. Proses ini dilakukan pada 500 file hasil re-filtering. Proses sampling topik dilakukan dengan langkah: mengurangi nilai matriks kata-topik dan dokumen-topik untuk setiap kata, menghitung probabilitas kata, menentukan topik baru dari setiap kata dengan distribusi multinomial berdasarkan nilai probabilitas kata, menambahkan nilai matriks kata-topik dan dokumen-topik untuk setiap kata sesuai dengan topik baru, dan menyimpan hasil matriks. Langkah-langkah ini dilakukan sebanyak *n* iterasi/pengulangan sampai mencapai kondisi konvergen.

Tahap perhitungan parameter final merupakan proses untuk menghitung jumlah dokumen untuk setiap topik dan jumlah kata untuk setiap topik berdasarkan matriks kata-topik dan dokumen-topik yang telah konvergen. Hasil perhitungan digunakan untuk tahap evaluasi.

Tahap evaluasi merupakan proses untuk menguji hasil metode ekstraksi fitur Latent Dirichlet Allocation yang telah diimplementasikan pada aplikasi teks bahasa Indonesia. Evaluasi dilakukan dengan membandingkan hasil ekstraksi fitur dengan metode Latent Dirichlet Allocation dengan metode Term Frequency-Inverse Document Frequency (TF-IDF). Metode evaluasi yang digunakan adalah metriks pengukuran Precision, Recall dan F-Measure [16]. Precision (*P*) merupakan nilai pembagian dari jumlah dokumen relevan yang diperoleh terhadap jumlah seluruh dokumen yang diperoleh, seperti pada (1)

$$P(relevant|retrieved) = \frac{\#(relevant\ items\ retrieved)}{\#(retrieved\ items)} \quad (1)$$

Recall (*R*) merupakan nilai pembagian dari jumlah dokumen relevan yang diperoleh terhadap jumlah seluruh dokumen yang relevan, seperti pada (2)

$$R(retrieved|relevant) = \frac{\#(relevant\ items\ retrieved)}{\#(relevant\ items)} \quad (2)$$

F-Measure (*F*) merupakan bobot rata-rata harmonik dari nilai P dan R, seperti pada (3)

$$F = \frac{2PR}{P+R} \quad (3)$$

III. HASIL DAN PEMBAHASAN

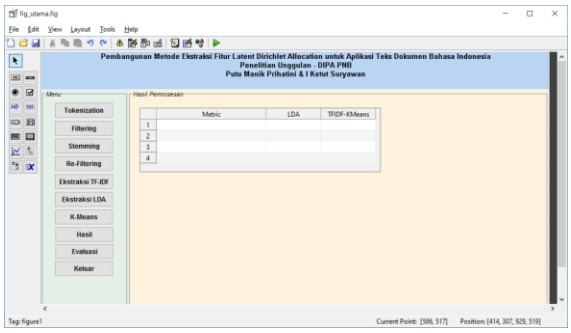
Pembangunan aplikasi ekstraksi topik teks dokumen Bahasa Indonesia dengan metode LDA sesuai dengan rancangan penelitian membutuhkan basis pengetahuan, rancangan algoritma dan aplikasi berbasis komputer.

Basis pengetahuan dalam penelitian ini menggunakan basis pengetahuan yang sudah dibangun pada penelitian sebelumnya yaitu file katadasar.txt memuat 28.527 kata dan file katahubung.txt memuat 907 kata [15].

Algoritma untuk aplikasi yang dibangun terdiri dari 21 fungsi yaitu 10 fungsi utama dan 11 subfungsi. Fungsi-fungsi utama terdiri dari Tokenisasi, Filtering, Stemming, ReFiltering, Ekstraksi TFIDF, Ekstraksi LDA, KMeans, Hasil, Evaluasi dan Keluar. Sub-sub fungsi terdiri dari CekImbuhanSalah, CekKamus, HapusPartikel, Hapus Possesive, HapusPrefixDua, HapusPrefixSatu, HapusSuffix, HitungNdkNkw, HitungSum, KonversiData, dan Stemming.

Antarmuka untuk aplikasi berbasis komputer dalam penelitian ini dibangun dengan menggunakan perangkat lunak Matlab R2010a seperti pada Gambar 2.

Untuk menguji aplikasi yang dibangun, dalam penelitian ini digunakan data masukan berupa file teks yang diperoleh melalui media berita digital. Jumlah file yang digunakan adalah 500 file. File-file ini dikelompokkan menjadi lima kategori yaitu berita, otomotif, olahraga, teknologi dan bisnis. Dari hasil Tokenisasi, Filtering, Stemming dan Refiltering diperoleh daftar teks unit seperti pada Tabel 1.



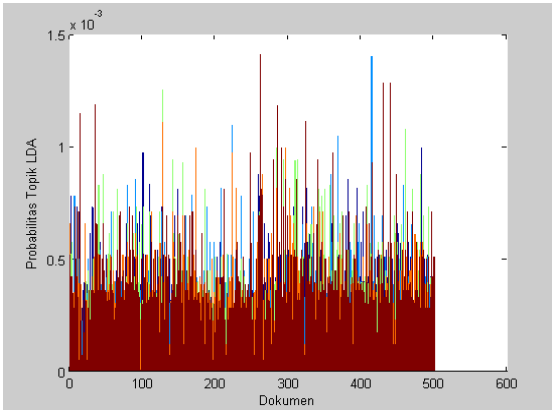
Gambar 2. Rancangan Antarmuka Aplikasi

Tabel 1. Hasil Pra Pengolahan Teks

Tahap	Hasil Teks Unit
Tokenisasi	306.870
Filtering	192.934
Stemming	119.022
ReFiltering	113.968

Untuk melakukan ekstraksi fitur dengan metode LDA membutuhkan parameter yaitu jumlah topik, nilai alpha dan nilai beta. Pada penelitian ini, jumlah topik ditentukan sebanyak lima topik, nilai alpha sebesar 5,0, nilai beta sebesar 0,01 dan kondisi konvergen tercapai pada iterasi kedelapan. Hasil yang dicapai pada kondisi konvergen ditunjukkan dalam bentuk grafik seperti pada Gambar 3.

Untuk grafik, sumbu *x* menunjukkan nomor dokumen (1-500) dan sumbu *y* menunjukkan probabilitas dokumen terhadap lima topik yang telah ditentukan. Warna pada grafik menunjukkan topik. Dari grafik tersebut terlihat bahwa ekstraksi topik LDA mengarahkan suatu dokumen pada seluruh topik seperti yang ditunjukan dengan jelas oleh distribusi warna topik. Hasil ini menguatkan kelebihan dari metode LDA yang sesuai dengan kondisi dalam dunia nyata dimana suatu dokumen belum tentu membahas satu topik saja, melainkan dapat juga membahas banyak topik.



Gambar 3. Hasil Ekstraksi Topik LDA

Tabel 2. Hasil Pengujian Ekstraksi Topik LDA

Metrik Pengujian	Hasil (%)
Precision (P)	94
Recall (R)	91
F-Measure (F)	91

Pada penelitian ini, pembandingan dilakukan antara data pengelompokan yang dihasilkan melalui ekstraksi topik LDA dengan data manual berupa data teks dokumen yang telah dikelompokkan oleh media digital. Proses pembandingan dilakukan dengan cara menghitung nilai Precision, Recall dan F-Measure. Nilai rata-rata P, R, dan F dari 500 file dokumen teks ditampilkan pada Tabel 2 yang dihitung berdasarkan jumlah hasil pencarian dokumen, jumlah dokumen manual, dan jumlah hasil pencarian dokumen yang relevan.

Nilai rata-rata tersebut menunjukkan hasil pengujian ekstraksi topik LDA memiliki nilai P, R dan F yang tinggi karena mendekati 100%. Hal ini disebabkan karena metode LDA mampu menemukan lebih banyak dokumen yang relevan dengan pengelompokan manual. Dengan demikian, dapat disimpulkan bahwa metode LDA method memiliki kinerja yang sangat baik dalam melakukan ekstraksi fitur untuk dokumen teks berbahasa Indonesia.

IV. SIMPULAN DAN SARAN

4.1 Simpulan

Pembangunan metode ekstraksi topik LDA untuk aplikasi teks dokumen bahasa Indonesia dalam penelitian ini dilakukan dengan menggunakan perangkat lunak Matlab R2010a yang menghasilkan dua basis pengetahuan yaitu katadasar.txt dan katahubung.txt; 10 fungsi utama yaitu Tokenisasi, Filtering, Stemming, Refiltering, Ekstraksi TFIDF, Ekstraksi LDA, KMeans, Hasil, Evaluasi dan Keluar; serta 11 subfungsi yaitu cek kamus, cek imbuhan salah, hapus partikel, hapus possesive, hapus prefix dua, hapus prefix satu, hapus sufiks, hitungNdkNkw, hitungsum, konversidata dan stemming. Dalam penelitian ini, metode LDA dinisialisasi dengan parameter lima topik, nilai alpa 5,0 dan nilai beta 0,01 dimana konvergensi dicapai pada iterasi kedelapan. Pengujian terhadap 500 file data berita dari media berita digital menghasilkan nilai rata-rata kinerja metode LDA sebesar 94% untuk Precision, 91% untuk Recall dan 91% untuk F-Measure. Hasil pengujian ini menunjukkan metode LDA memiliki kinerja sangat baik dalam melakukan ekstraksi topik untuk dokumen teks berbahasa Indonesia. Oleh karena itu, metode LDA dapat menjadi referensi untuk melakukan ekstraksi topik dokumen teks berbahasa Indonesia.

4.2 Saran

Untuk mengetahui unjuk kerja lebih lanjut dari metode LDA dalam melakukan ekstraksi topik dokumen teks berbahasa Indonesia, disarankan untuk membandingkan metode ini dengan metode berbasis non topik lainnya.

UCAPAN TERIMA KASIH

Penulis mengucapkan terimakasih kepada unit P3M Politeknik Negeri Bali yang telah membantu pendanaan penelitian ini melalui program penelitian unggulan DIPA tahun 2017.

DAFTAR PUSTAKA

[1]. Zhao, Z., X. He, L. Zhang, W. Ng, and Y. Zhuang, "Graph regularized feature selection with data reconstruction", *IEEE Transactions on Knowledge and Data Engineering*, 28, 689-700, 2016.

[2]. Tutkan, M., M.C. Ganiz, and S. Akyokuş, "Helmholtz principle based supervised and unsupervised feature selection methods for text mining", *Information Processing & Management*, 52, 885-910, 2016.

[3]. Hai, Z., K. Chang, J.-J. Kim, and C.C. Yang, "Identifying features in opinion mining via intrinsic and extrinsic domain relevance", *IEEE Transactions on Knowledge and Data Engineering*, 26, 623-634, 2014.

[4]. Liu, K., L. Xu, and J. Zhao, "Co-extracting opinion targets and opinion words from online reviews based on the word alignment model", *IEEE Transactions on Knowledge and Data Engineering*, 27, 636-650, 2015.

[5]. Noh, H., Y. Jo, and S. Lee, "Keyword selection and processing strategy for applying text mining to patent analysis", *Expert Systems with Applications*, 42, 4348-60, 2015.

[6]. Wu, C.-H., H.-P. She, and C.-S. Hsu, "Code-Switching Event Detection by Using a Latent Language Space Model and the Delta-Bayesian Information Criterio", *IEEE/ACM Transaction on Audio, Speech, and Language Processing*, 23, 1892-1903, 2015.

[7]. Gong, L., R. Yang, Q. Yan, and X. Sun, "Prioritization of Disease Susceptibility Genes Using LSM/SVD", *IEEE Transaction on Biomedical Engineering*, 60, 3410-3417, 2013.

[8]. Cosma, G. and M. Joy, "An Approach to Source-Code Plagiarism Detection and Investigation Using Latent Semantic Analysis", *IEEE Transaction on Computers*, 61, 379-394, 2012.

[9]. Blei, D.M., A.Y. Ng, and M.I. Jordan, "Latent dirichlet allocation", *Journal of Machine Learning Research*, 3, 993-1022, 2003.

[10]. Gao, Y., Y. Xu, and Y. Li, "Pattern-based Topics for Document Modelling in Information Filtering", *IEEE Transactions on Knowledge and Data Engineering*, 27, 1629-1642, 2015.

[11]. Chien, J.-T., "Hierarchical Pitman–Yor–Dirichlet Language Model", *IEEE/ACM Transaction on Audio, Speech, and Language Processing*, 23, 1259-1272, 2015.

[12]. Archambeau, C., B. Lakshminarayanan, and G. Bouchard, "Latent IBP Compound Dirichlet Allocation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37, 321-333, 2015.

[13]. Li, Y., X. Zhou, Y. Sun, and H. Zhang, "Design and implementation of weibo sentiment analysis based on LDA and dependency parsing", *China Communications*, 91-105, 2016.

[14]. Lau, R.Y.K., Y. Xia, and Y. Ye, "A probabilistic generative model for mining cybercriminal networks from online social media", *IEEE Computational intelligence magazine*, 31-43, 2014.

[15]. Prihatini, P.M. and I.K. Suryawan, "Text processing application development for Indonesian documents clustering", *The 1st International Joint Conference on Science and Technology (IJCST)*, 2016, Bali, Indonesia.

[16]. Prihatini, P.M., I.K.G.D. Putra, I.A.D. Giriantari, and M. Sudarma, "Fuzzy-gibbs latent dirichlet allocation model for feature extraction on Indonesian documents", *Contemporary Engineering Sciences*, 10, 403-421, 2017.