Pembangunan *Prototype* Aplikasi Deteksi Topik Pada Petisi

(Studi kasus: *Change.org* Indonesia)

Rifki Adhitama[1], Aditya Wijayanto[2]
[1,2]Rekayasa Perangkat Lunak, Fakultas Teknilogi Industri dan Informatika, Institut Teknologi Telkom Purwokerto
[1,2]J.L. D. I. Panjaitan 128 Purwokerto
[1]rifki@ittelkom-pwt.ac.id, [2]aditya.wijayanto@ittelkom-pwt.ac.id

Abstrak – Petisi elektronik merupakan salah satu alat yang menyediakan mekanisme yang memfasilitasi publik untuk mengekspreskan pendapatnya baik kepada pemerintah maupun pihak terkait. Petisi elektronik telah terbukti dapat mempengaruhi kebijakan publik. Dilain sisi petisi elektronik yang populer di indonesia yaitu change.org tidak memiliki mekanisme pengkategorian khusus, atau pengkategorian populer yang dinamis. Penelitian ini bertujuan untuk membangun sebuah prototype aplikasi yang mampu mendeteksi topik tersembunyi pada petisi secara dinamis. Penelitian ini menggunakan 100 dokumen petisi elektronik yang memiliki jumlah tandatangan elektronik minimal 10.000 tandatangan. Ekeperimen bertujuan untuk mengetahui apakah penggunaan KBBI dan Latent Dirichlet Allocation (LDA) mampu mendeteksi topik secara relevan pada dokumen petisi berbahasa Indonesia. Cohen Kappa Coefficient digunakan untuk mengukur relevansi dari hasil deteksi topik, dimana hasil dari penelitian ini menunjukkan bahwa nilai mean relevance rate dari hasil metode yang digunakan adalah sebesar 50%.

Kata kunci – Latent Dirichlet Allocation, Kappa, Petisi, KBBI

Abstract—The electronic petition is a tool that provides mechanism and facilitate public to express their opinions to both government or related parties/ stakeholder. Electronic petition has been proved, that is can influenced public policy. In the other hand, change.org is the most popular electronic petition in Indonesia does not have a special categorizing mechanism or dynamic popular categorization. This study aims to build prototype that able to detect hidden topic within electronic petition documents dynamically. This study uses 100 electronic petition documents that have minimum 10.000 electronic signatures. The experiment aims to determine whether the use of KBBI dictionary and *Latent Dirichlet Allocation* (LDA) is able to detect hidden topic that relevant to Indonesian language petition documents. *Cohen Kappa Coefficient* is used to measure the relevance of the topic detection, where the results of this study indicate that the relevance rate is 50%.

Keywords-Latent Dirichlet Allocation; Kappa; Petition; KBBI

I. PENDAHULUAN

Pemerintah di berbagai negara telah mulai mempertimbangkan petisi elektronik (*e-petition)* untuk menilai *feedback* dari publik mengenai saran kebijakan atau sesuatu yang perlu menjadi perhatian dari pemerintah [7]. Menurut pendapat di kalangan akademik, petisi elektronik menyediakan sebuah mekanisme yang memfasilitasi publik untuk mengekspresikan pandangan mereka terhadap pemerintah [8] dan petisi elektronik ini telah terbukti memberikan pengaruh yang cukup signifikan terhadap kebijakan pemerintah [9]. Beberapa kebijakan berubah dikarenakan adanya petisi elektronik salah satunya adalah peraturan FIBA yang pada akhirnya melegalkan atlet basket wanita memakai hijab, dimana sebelumnya FIBA melarang atlet basket wanita memakai hijab yang ditandatangani secara elektronik lebih dari 130.000 orang.

Sosial media saat ini berperan besar dalam penyebaran sebuah petisi elektronik, setiap orang dapat mengekspresikan apa yang menjadi perhatiannya atau menyuarakan solusi terhadap suatu masalah baik politik, lingkungan, kesehatan dan lain-lain. Berdasarkan data dari TechInAsia, Change.org merupakan situs petisi online yang paling popular di Indonesia. Walaupun begitu sampai saati ini Change.org tidak memiliki klasifikasi kategorisasi dari petisi yang dibuat oleh masyarakat. Adapun Change.org hanya membuat daftar petisi popular hanya tiap akhir tahun, sehingga menyulitkan pihak-

pihak yang bersangkutan untuk menganalisa topik apa saja yang popular dan harus menjadi perhatian pada jangka waktu tertentu misal 5 tahun terakhir, 10 tahun terakhir dan seterusnya Meskipun petisi elektronik sudah dangat popular, tetapi penelitian tentang klasifikasi dan clustering text petisi masih sedikit.

Metode klasifikasi yang umum digunakan adalah *Naïve Bayes* [10] dan metode clustering yang sedang berkembang saat ini adalah *Latent Dirichlet Allocation* (LDA), tetapi LDA menunjukkan performa yang lebih baik untuk dokumen text dibandingkan dengan naïve bayes [11]. Disisi lain Ontology dapat digunakan untuk memberikan pelabelan dari topik/ cluster yang telah terbentuk tanpa harus menggunakan seorang ahli untuk melabeli topik yang telah terbentuk [12].

Penelitian ini bertujuan menghasilkan sebuah aplikasi yang mampu memberikan analisa topik-topik apa saja yang muncul dalam petisi-petisi yang telah dibuat dalam jangka waktu tertentu dengan menggunakan LDA dan mengetahui apakah skema ontology dalam KBBI yang dibagi berdasarkan 71 label topik dari Kamus Besar Bahasa Indonesia (KBBI) (https://kbbi.kemdikbud.go.id/) masih relevan jika digunakan dalam dokumen petisi dibandingkan dengan penggunaan dalam dokumen berita. Mengingat meskipun dokumen petisi ditulis dengan menggunakan kata kata formal, tetapi format penulisannya tidak se-formal penulisan berita.

Susunan dari dokumen ini disajikan sebagai berikut. Bagian 2 menjelaskan metodologi penelitian dan pemodelan yang digunakan. Bagian 3 mendeskripsikan tentang persiapan eksperiman , scenario eksperimen yang akan dilakukan dan hasil dari eksperimen. Kesimpulan akan disajikan pada bagian 4.

II. METODE PENELITIAN

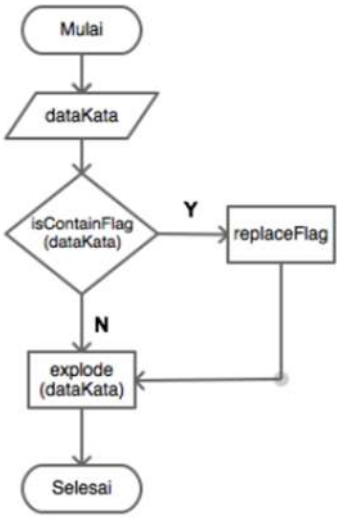
Pada penelitian ini penulis menggunakan metode LDA untuk melakukan klastering dari petisi pada daring Change.org. Petisi yang digunakan pada penelitian ini adalah petisi yang diambil antara tahun 2012 – 2017 dengan penandatangan petisi minimal 10.000 tandatangan elektronik.

Tahapan pertama dalam penelitian ini adalah proses tokenisasi yang membagi data text mentah menjadi token untuk setiap kata. Tahapan selanjutnya adalah *stopword removal* untuk menghilangkan kata-kata yang kurang bermakna. Setelah *stopword removal* tahapan selanjutnya adalah melakukan *N-Gram splitting* untuk memunculkan serangkaian makna pada token token yang telah dibentuk sebelumnya. Setelah proses *N-Gram* selesai, selanjutnya data siap diolah untuk dikelompokkan kedalam beberapa topik dengan LDA dan melabeli klaster yang terbentuk dengan ontology yang berbasis kamus Bahasa indonesia. Secara umum proses klastering dokumen petisi ditunjukkan pada Gambar 1.

Gambar 9. Kerangka umum penelitian

A. Tokenisasi

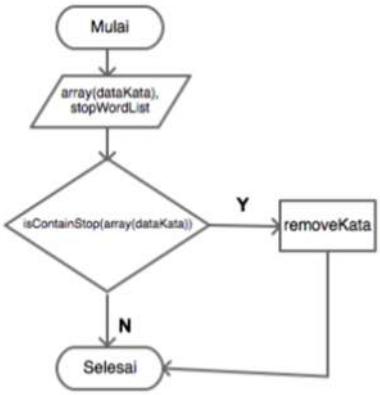
Tokenisasi adalah proses pembagian keseluruhan dokumen ke dalam bentuk kata tunggal (*term)* dengan mengeliminasi spasi dan tanda baca (titik, koma, tanda petik, *underscore, slash,* dan sebagainya). Diagram alir untuk proses stopword removal ditunjukkan pada Gambar 2.



Gambar 10. Diagram Alir Proses Tokenisasi

B. Stop-word Removal

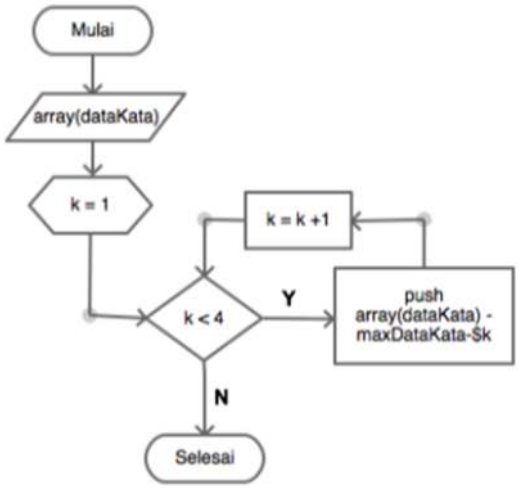
Setelah proses tokenisasi selesai dilakukan, proses selanjutnya adalah *stop word removal*. Proses ini menghilangkan kata-kata yang tidak memiliki makna khusus dan kata sambung seperti (“jika”, “hanya”, “akan”, “ada”, “adalah”, dan sebagainya). Diagram alir untuk proses stopword removal ditunjukkan pada Gambar 3.



Gambar 11. Diagram Alir Proses Stopword Removal

C. N-Gram Splitting

Proses selanjutnya adalah N-Gram splitting yaitu penggabungan sejumlah N *term* untuk mencari sebuah makna baru sesuai dengan kamus KBBI. Penelitian ini mengelompokkan *term* mulai dari satu *term* sampai dengan empat *term*. Hal ini dikarenakan pada KBBI terdapat kata yang sampai dengan empat *term* untuk menghasilkan satu makna. Diagram alir proses N-Gram splitting ditunjukkan pada Gambar 4.



Gambar 12. Diagram Alir N-Gram Splitting

D. Latent Dirichlet Allocation

LDA adalah sebuah probabilistic topic model yang menganggap pada setiap dokumen terdapat topik topik tersembunyi, dan setiap topik direpresentasikan berdasarkan sebaran kata pada tiap dokumen [13]. LDA menggunakan distribusi dirichlet untuk membentuk sekumpulan kata dalam suatu topik tertentu [14]. dasarnya LDA tidak dapat memberikan label secara otomatis kepada topik yang telah terbentuk sehingga membutuhkan pemberian label secara manual. Penggabungan antara ontology dan *similarity measure* memberikan hasil yang cukup baik dalam pelabelan secara otomatis terhadap topik-topik yang telah terbentuk. Algoritma inferensi yang digunakan dalam LDA adalah *collapsed gibbs sampling*, dimana formula untuk *collabsed gibbs sampling* adalah sebagai berikut [15].

$$p(z_i = k | z^{(-i)}, w) = \frac{n_{k,-i}^{(w)} + \beta}{n_{k,-i}^{(*)} + V * \beta} * \left(n_{d,-i}^{(k)} + \alpha\right) \quad (1)$$

$$\theta_j^{(m)} = \frac{n_j^{(m)} + \alpha}{n_\cdot^{(m)} + k\alpha} \quad (2)$$

$$\varphi_j^{(w)} = \frac{n_j^{(w)} + \beta}{n_j^{(\cdot)} + w\beta} \quad (3)$$

Dimana:
m: Dokumen
z: Topik
w: Kata
α: hyperparameter distribusi dokumen terhadap topik
β: hyperparameter distribusi kata terhadap topik

Penelitian ini menggunakan LDA untuk mengklaster topik topik tersembunyi yang terdapat dalam dokumen petisi beserta labelnya.

III. EKSPERIMEN DAN HASIL

Pada bagian ketiga ini dibagi kedalam tiga sub bab yang pertama adalah persiapan eksperimen, scenario eksperimen dan yang terakhir adalah hasil eksperimen.

A. Persiapan Eksperimen

Dataset yang digunakan dalam penelitian ini adalah data petisi berbahasa indonesia sebanyak 100 buah petisi yang diambil antara tahun 2012-2017, dimana setiap dokumen petisi tersebut minimal telah ditandatangani secara elektronik lebih dari 10.000 penandatangan. Skenario pertama adalah membagi dokumen menjadi 5 topik agar hasil pengelompokan tidak terlalu sparse dan lebih spesifik menyesuaikan dengan perilisan topik pada change.org, dikarenakan belum ada standar khusus atau kategorisasi yang banyak dari petisi itu sendiri. dimana setiap topik berisi 30 kata untuk setiap topik. Karena pada penelitian sebelumnya nilai paling tinggi didapatkan pada kumpulan 30 dan 5 kata untuk setiap topik [12], tetapi penggunaan 5 kata untuk setiap topik tidak digunakan karena topik akan menjadi terlalu bersifat general. scenario kedua adalah menghitung nilai kappa dari hasil scenario ekperimen pertama.

Penelitian ini diimplementasikan dengan menggunakan *framework CodeIgniter* yang dijalankan pada google chrome di atas komputer dengan spesifikasi sebagai berikut.

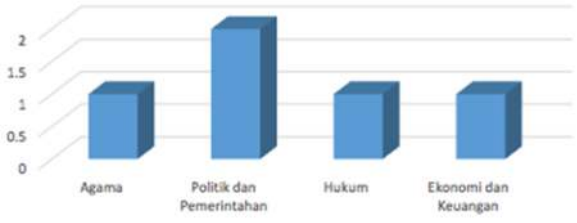
- Intel Core i5 2.5 GHz
- 10 Gb RAM
- 500 GB HDD

B. Skenario Eksperimen

Seperti yang telah dijelaskan pada tahapan persiapan eksperimen, scenario pertama membagi dokumen ke dalam 5 topik dimana masing-masing topik memiliki 30 kata. Skenario ekperimen pertama bertujuan untuk mengetahui topik apa saja yang populardalam rentang waktu antara tahun 2012-2017.

Skenario eksperimen kedua adalah dengan menghitung nilai kappa dari hasi scenario ekperimen pertama. Skenario eksperimen kedua bertujuan untuk mengetahui relevansi dari topik topik yang muncul tersebut.

C. Hasil Eksperimen



Gambar 13. Data hasil eksperimen skenario pertama

Gambar 5 menunjukkan hasil dari scenario eksperimen pertama yaitu topik-topik yang popular dari tahun 2012-2017 berdasarkan kata bidang dari KBBI dengan kriteria *dataset* yang telah ditentukan

sebelumnya. Topik-topik tersebut antara lain Agama, Politik dan Pemerintahan, hukum, dan Ekonomi dan Keuangan.

Tabel 1. Hasil scenario ekperimen kedua

Penilai	Hasil penilaian				
	Nilai relevan	*Nilai tidak relevan*	*Nilai Kappa*	*Relevance rate*	*Mean relevance rate*
P1	0.6	0.4	0.615	0.6	0.5
P2	0.4	0.6		0.4	

Tabel 1 menunjukkan hasil scenario ekperimen kedua dengan nilai kappa 0.615 dengan nilai mean relevance rate 0.5. pembahasan ekperimen akan dijelaskan pada bagian selanjutnya.

IV. PEMBAHASAN

Berdasarkan hasil skenario ekperimen pertama, pembagian ke dalam lima topik utama menghasilkan empat topik unik, yaitu agama, politik dan pemerintahan, hukum dan ekonomi dan keuangan. Hal ini terkait sifat dari LDA yang mengelompokkan topik berdasarkan probabilitas kata dan tidak adanya kelas khusus yang dapat mengklasifikasikan topik secara unik. Topik yang terbagi dalam LDA bersifat dinamis, sehingga memungkinkan ada beberapa topik yang memiliki label topik yag sama.

Skenario eksperimen kedua yaitu menghitung nilai kappa untuk menentukan relevansi dari hasil yang didapatkan. Nilai kappa dari skenario eksperimen kedua adalah 0.615. Dimana hasil ini sudah melewati threshold minimal yaitu 0.41, sehingga selanjutnya adalah menghitung nilai mean relevance rate dari perhitungan kappa yang didapatkan hasil sebesar 0.5.

V. PENUTUP

Kesimpulan dari hasil penelitian ini didapatkan bahwa penggunaan KBBI dengan menggunakan metode *Latent Dirchlet Allocation* (LDA) untuk melakukan deteksi topik pada petisi menghasilkan nilai mean relevance rate sebesar 0.5. Hasil nilai dari mean relevance rate yang tidak terlalu tinggi dikarenakan adanya perbedaan pendapat yang berimbang antara kedua penilai pada perhitungan metode kappa, atau dapat pula pengaruh jumlah topik yang terbentuk dan KBBI yang digunakan. Sehingga dapat disimpulkan bahwa skema penggunaan KBBI dan LDA untuk mendeteksi topik tersembunyi pada dokumen petisi masih perlu dilakukan penelitian lebih lanjut untuk mengetahui faktor apa saja yang berpengaruh dalam penilaian relevansi dalam dokumen petisi.

DAFTAR PUSTAKA

[1] M. Muslihudin and D. Hartini, "Perancangan Sistem Pendukung Pengambilan Keputusan Untuk Penerimaan Beasiswa Di SMA PGRI 1 Talang Padang Dengan Model Fuzzy Multiple Attribute Menggunakan Metode Simple Additive Weighting (SAW)," *Journal Technology Acceptance Model,* vol. 4, pp. 34 - 40, 2015.

[2] Depdiknas, "Peraturan Menteri Pendidikan Nasional No. 20, Tahun 2003," Jakarta, 2003.

[3] A. Nuraziq, Kusrini and E. T. Luthfi, "Rancang Bangun Sistem Pendukung Keputusan Penerimaan Beasiswa Menggunakan Metode Clustering K-Means," *Seminar Nasional Multi Disiplin Ilmu,* vol. 1, 2017.

[4] J. Aranda and W. A. G. Natasya, "Penerapan Metode K-Means Cluster Analysis Pada Sistem Pendukung Keputusan Pemilihan Konsentrasi Untuk Mahasiswa International Class STMIK AMIKOM Yogyakarta," *Seminar Nasional Teknologi Informasi dan Multimedia,* vol. IV, no. 2, pp. 1 - 6, 2016.

[5] A. Masruro, Kusrini and E. T. Luthfi, "Sistem Penunjang Keputusan Penentuan Lokasi Wisata Menggunakan K-Means Clustering dan TOPSIS," *Jurnal Ilmiah DASI,* vol. 15, no. 4, pp. 1 - 5, 2015.

[6] A. Ristyawan, Kusrini and Andi Sunyoto, "Pemanfaatan Algoritma FCM Dalam Pengelompokan Kinerja Akademik Mahasiswa," in *Konferensi Nasional Sistem dan Informatika 2015*, Bali, 2015.

[7] C. L. Dumas, D. LaManna, T. M. Harrisson, S. Ravi, C. Kotfila, N. Gervais, L. Hagen and F. Chen, "Examining political mobilization of online communities through e- petitioning behavior in We the People," *Big Data & Society,* vol. 2, no. 2, 2015.

[8] C. Bochel, "Petitions Systems: Contributing to Representative Democracy?," *Parliamentary Affairs,* vol. 66, no. 4, pp. 798-815, 2013.

[9] C. Bochel, "Petitions: Different Dimensions of Voice and Influence in the Scottish Parliament and the National Assembly for Wales," *Social and Policy Administration,* vol. 46, no. 2, pp. 142-160, 2012.

[10] I. Saputra, Y. Sibaroni and A. P. Kurniati, "Analysis and Implementation of Classification Indonesian News with Naive Bayes Method," Telkom University, Bandung, 2008.

[11] R. Kusumaningrum, S. Adhy, M. I. A. Wiedjayanto and Suryono, "Classification of Indonesian News Articles based on Latent Dirichlet Allocation," in *International Conference on Data and Software Engineering*, Denpasar, Bali, 2016.

[12] R. Adhitama, R. Kusumaningrum and R. Gernowo, "Topic labeling towards news document collection based on Latent Dirichlet Allocation and ontology," in *2017 1st International Conference on Informatics and Computational Sciences (ICICoS)*, Semarang, 2017.

[13] D. M. Blei, *Probabilistic Topic Models,* vol. 55, New York: ACM, 2012, pp. 77-84.

[14] Z. Liu, "High Performance Latent Dirichlet Allocation for Text Mining," Brunel University, London, 2013.

[15] G. Heinrich, "Parameter Estimation for Text Analysis," Fraunhofer IGD, Darmstadt, Germany, 2009.

[16] M. jazzuli, "e-tarnsparancy budgeting award," 2014. [Online]. Available: https://www.kompasiana.com/jazzmuhammad/. [Accessed 8 Januari 2018].