PENGELOMPOKAN TOPIK DOKUMEN BERBASIS TEXT MINING DENGAN ALGORITME K-MEANS: STUDI KASUS PADA DOKUMEN KEDUTAAN BESAR AUSTRALIA JAKARTA

Wishnu Hardi, M.P.[1], **Dr. Eng. Wisnu Ananta Kusuma**[2], **Prof. Dr. Sulistyo Basuki**[3]

Abstrak

Kedutaan Besar Australia di Jakarta menyimpan beragam dokumen rilis media. Menganalisis koleksi dokumen yang berpola khusus dan vital sangatlah penting untuk menghasilkan wawasan baru dan pengetahuan tentang kelompok topik penting dari dokumen. *K-Means* digunakan sebagai metode pengelompokan data non-hirarkis objek data menjadi klaster. Metode ini bekerja dengan meminimalkan variasi data di dalam klaster dan memaksimalkan variasi data di antara klaster. Dari dokumen yang dikeluarkan antara 2006 dan 2016, 839 dokumen diperiksa untuk menentukan frekuensi jangka dan untuk menghasilkan klaster. Evaluasi dilakukan dengan menunjuk seorang ahli untuk memvalidasi hasil klaster. Hasil penelitian menunjukkan bahwa ada 57 istilah bermakna yang dikelompokkan menjadi 3 kelompok. “Hubungan orang-orang”, “kerja sama ekonomi”, dan “pembangunan manusia” dipilih untuk mewakili topik rilis media Kedutaan Besar Australia Jakarta dari tahun 2006 hingga 2016. Penelitian ini menyimpulkan bahwa *text mining* dapat digunakan untuk mengelompokkan topik dokumen. Ini memberikan proses pengelompokan yang lebih sistematis karena analisis teks dilakukan melalui sejumlah tahapan dengan parameter yang ditetapkan secara khusus.

Kata Kunci: *text mining*, analisis konten, klasterisasi dokumen, algoritme *K-Means*

Abstract

The Australian Embassy in Jakarta is storing a wide array of media release document. Analyzing particular and vital patterns of the documents collection is imperative as it will result in new insights and knowledge of significant topic groups of the documents. K-Means was used algorithm as a non-hierarchical clustering method which partitioning data objects into clusters. The method works through minimizing data variation within cluster and maximizing data variation between clusters. Of the documents issued between 2006 and 2016, 839 documents were examined in order to determine term frequencies and to generate clusters. Evaluation was conducted by nominating an expert to validate the cluster result. The result showed that there were 57 meaningful terms grouped into 3 clusters. “People to people links”, “economic cooperation”, and “human development” were chosen to represent topics of the Australian Embassy Jakarta media releases from 2006 to 2016. This research concluded that text mining can be used to cluster topic groups of documents. It provides a more systematic clustering process as the text analysis is conducted through a number of stages with specifically set parameters.

Keywords: text mining; content analysis; document clustering; K-Means algorithm

[1] Koleksi Luar Negeri dan Manajemen Metadata, Perpustakaan Nasional Australia

[2] Program Studi Magister Teknologi Informasi untuk Perpustakaan, Institut Pertanian Bogor

[3] Sekolah Pascasarjana, Fakultas Ilmu Budaya, Universitas Indonesia

Pendahuluan

Perusahaan teknologi informasi, Oracle menyebutkan bahwa hampir setiap organisasi di seluruh dunia menyimpan sekitar 80 persen data tidak terstruktur dalam database-nya terutama dalam bentuk teks dan pertumbuhan tersebut meningkat secara eksponensial dengan konsisten dari waktu ke waktu (Mathew, 2012). Sementara itu, menurut lembaga konsultan Spire, 90 persen pengambilan keputusan pada organisasi mengandalkan 20 persen data terstruktur yang dimiliki (Spire Technologies, 2017).

Situasi ini memunculkan sebuah pertanyaan yaitu bagaimana organisasi dapat memanfaatkan 80 persen data tidak terstruktur menjadi lebih bermakna sehingga dapat mendukung proses pengambilan keputusan. *Text mining* adalah sebuah proses dalam menemukan informasi dari kumpulan koleksi teks dalam jumlah besar, serta mengidentifikasi pola-pola menarik serta keterkaitannya dalam

data tekstual (Feldmen & Sanger, 2007). Adapun tujuh area penerapan *text mining*, adalah temu kembali informasi, klasifikasi, klasterisasi, web mining, ekstraksi informasi, Natural Language Processing (NLP), dan ekstraksi konsep (Miner et al., 2012).

Kajian mengenai analisis teks berbasis *text mining* pernah dilakukan pada disiplin ilmu yang cukup bervariasi. Antara lain, Zade (2017) menjelaskan kerangka dasar proses klasterisasi dokumen melalui pendekatan *text mining;* Allayhari (2017) melakukan studi mengenai teknik dan aspek-aspek fundamental dalam penerapan *text mining* untuk bidang kesehatan dan biomedis; Gurusamy (2017) melakukan penelitian mengenai pengelompokan perilaku pengguna media sosial dengan teknik klasterisasi berbasis algoritme k-means; Prilianti (2014) mengembangkan aplikasi klasterisasi berbasis algoritme k-means untuk mengidentifikasi tren topik skripsi mahasiswa; dan Lama (2013) menerapkan *text mining* untuk melakukan klasterisasi isu-isu utama pada artikel headline surat kabar elektronik dengan algoritme k-means.

Kedutaan Besar Australia Jakarta adalah salah satu kantor pemerintah asing yang menyimpan dokumen teks tidak terstruktur dalam bentuk siaran media. Signifikansi dokumen siaran media secara kelembagaan dapat merujuk pada Konvensi Wina tahun 1961 mengenai hubungan internasional di mana salah satu fungsi lembaga diplomatik adalah untuk mewakili negara pengirim dengan tujuan meningkatkan kerja sama pada sejumlah bidang dengan negara penerima (United Nations, 1961). Dalam konteks ini, dokumen siaran media bernilai strategis karena merupakan merupakan instrumen diplomasi dari kebijakan luar negeri negara pengirim.

Pertumbuhan dokumen siaran media yang dirilis oleh Kedutaan Besar Australia bersifat fluktuatif dalam kurun waktu sepuluh tahun terakhir. Konten dari dokumen ini berisi komunikasi diplomatik yang disampaikan kepada publik dan pemerintah di Indonesia dalam konteks hubungan bilateral kedua negara. Dengan mempelajari pola-pola menarik dan vital dari koleksi dokumen, maka akan diperoleh pengetahuan baru yang dapat digunakan sebagai petunjuk dalam pengambilan keputusan sekaligus bahan evaluasi dalam perumusan strategi komunikasi yang diterapkan.

Sepanjang observasi penulis, belum pernah ada kajian dalam bentuk penelitian yang secara khusus mengaplikasikan *text mining* untuk menganalisis dokumen resmi yang dikeluarkan lembaga setingkat kedutaan. Berdasarkan pemikiran tersebut, maka penelitian ini berupaya untuk mengaplikasikan *text mining* dengan tujuan melakukan klasterisasi topik dokumen siaran media Kedutaan Australia Jakarta tahun 2006 sampai dengan 2016.

Landasan Teori

Text mining

Menurut Feldmen (2007), *text mining* adalah teknik untuk mengekstraksi informasi penting dengan cara mengidentifikasi dan mengeksplorasi pola-pola vital dan menarik dari kumpulan dokumen. Menurut Weiss (2004), metode *text mining* dapat diaplikasikan untuk melakukan klasifikasi dokumen, temu kembali informasi, pengorganisasian dan pengelompokan dokumen, ektraksi informasi, serta evaluasi dan prediksi informasi.

Praproses

Praproses data adalah langkah paling mendasar dalam *text mining* di mana data mentah ditransformasi menjadi bentuk yang lebih bermakna dan dapat dipahami. Hal ini disebabkan karena data tekstual yang diambil dari dokumen bersifat tidak terstruktur, tidak konsisten, dan banyak mengandung noise sehingga diperlukan upaya normalisasi agar data dapat diproses lebih lanjut. Langkah-langkah praproses data terdiri dari *case folding, filtering, stopwords removal,* dan *stemming.*

Case folding adalah mengkonversi penggunaan huruf kapital pada teks menjadi huruf kecil (*lowercase*) untuk tujuan konsistensi dan mempermudah perbandingan teks dalam praproses data. Proses *filtering* menghilangkan angka numerik, punktuasi, *email*, dan *website*, serta karakter lain selain huruf. *Stopwords removal* merupakan proses menghilangkan kata henti dengan menggunakan daftar *stopwords* untuk Bahasa Inggris. Langkah terakhir dalam praproses data adalah *stemming,* yaitu mengembalikan kata ke dalam bentuk morfologi dasarnya.

Pembobotan Istilah dengan *Term Frequency* dan *Inverse Document Frequency* (TFIDF)

Pembobotan istilah adalah sebuah metode pengukuran statistik untuk mengevaluasi seberapa penting sebuah istilah dalam kumpulan dokumen atau korpus. Pembobotan istilah memerlukan dua hal, yaitu *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF). TF adalah pengukuran terhadap frekuensi kemunculan sebuah istilah dalam dokumen dibagi dengan seluruh jumlah istilah dalam dokumen. Sedangkan, IDF digunakan untuk mengukur kekuatan perbedaan yang dihasilkan sebuah istilah. Hal ini disebabkan istilah yang muncul pada seluruh dokumen tidak dapat digunakan untuk

membedakan dokumen untuk topik tertentu. Bobot nilai TFIDF adalah hasil perkalian dari nilai TF dan IDF.

Signifikansi sebuah istilah akan meningkat secara proporsional sejalan dengan frekuensi kemunculannya dalam sebuah dokumen yang diimbangi dengan frekuensi kemunculannnya pada keseluruhan dokumen. Rumus TFIDF yang digunakan adalah sebagai berikut:

$$TFIDF_{ij} = tf_{ij} \; x \log(D / df_j) \quad (1)$$

dengan $TFIDF_{ij}$ adalah bobot istilah i pada dokumen ke j, tf_{ij} adalah jumlah frekuensi kemunculan istilah i pada dokumen j, D adalah jumlah keseluruhan dokumen dan df_i adalah jumlah dokumen yang mengandung istilah i.

Algoritme K-means

Algoritme k-means diperkenalkan oleh MacQueen tahun 1967 sebagai metode klasterisasi data nonhierarki yang mempartisi objek data ke dalam kelompok-kelompok. K-means adalah proses klasterisasi tanpa supervisi di mana objek data ditempatkan secara 'alami' dalam sebuah kelompok dengan tidak mengetahui pola atau pengetahuan yang dimiliki untuk memandu proses klasterisasi (Miner et al., 2012). Dalam metode ini, data yang memiliki karakteristik yang sama dimasukan ke dalam satu kelompok yang sama dan data yang memiliki karakteristik berbeda dimasukan ke dalam kelompok yang berbeda. Proses klasterisasi k-means dilakukan dengan cara meminimalkan variasi objek data dalam kelompok yang sama dan memaksimalkan variasi data antarkelompok yang berbeda. Langkah-langkah klasterisasi dengan metode k-means menurut adalah sebagai berikut:

1. Menentukan jumlah klaster k yang ingin dibentuk
2. Membangkitkan sentroid atau pusat klaster sebanyak k klaster secara acak
3. Menghitung jarak setiap data input terhadap masing-masing sentroid dengan menggunakan formula Cosine Similarity
4. Mengelompokan setiap objek data berdasarkan kedekata (jarak terkecil)
5. Memutakhirkan nilai sentroid dengan menghitung nilai rata-rata klaster
6. Melakukan langkah pengulangan dengan menentukan pusat klaster secara acak agar pusat klaster yang paling merepresentasikan posisi dari sebuah kelompok data terhadap kelompok data lainnya dapat ditemukan atau dengan kata lain sampai nilai jarak anggota tiap klaster tidak berubah
7. Jika langkah 6 terpenuhi, maka nilai rata-rata klaster pada iterasi terakhir akan digunakan sebagai parameter untuk menentukan pengelompokkan data

Cosine Similarity

Cosine Similarity adalah sebuah metode pengukuran kemiripan istilah dengan menghitung sudut antara dua buah vektor. Dalam konsep ini sebuah istilah dianggap memiliki bobot (*magnitude*) dan arah (*direction*) dalam sebuah ruang berdimensi tinggi. Metode ini telah digunakan secara luas untuk proses klasterisasi, termasuk data tekstual. Dengan mengetahui jarak antara dua buah istilah, maka dapat diketahui kemiripan antar istilah berdasarkan perbedaan minimal jarak antara satu istilah dengan istilah lainnya (Davis & Shaw, 2011). Rumus yang digunakan adalah sebagai berikut:

$$Cosine\ (x,y) = \frac{\sum_{i=1}^{n} Xi.Yi}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} Y_i^2}} \quad (2)$$

dengan X adalah vektor; Y adalah vektor; X_i adalah bobot istilah *i* pada blok X_i; Y_i adalah bobot istilah *i* pada blok Y_i; i adalah jumlah istilah dalam kalimat; dan n adalah jumlah vektor.

Within Cluster Sum-of-Square (WSS)

WSS adalah metode pengukuran variabilitas intra klaster. Sebuah klaster yang dengan nilai WSS rendah memiliki tingkat kohesivitas yang lebih baik daripada klaster dengan nilai WSS yang lebih tinggi (Zhang & Franti, 2009). Rumus yang digunakan adalah sebagai berikut:

$$WSS = \sum_{i=1}^{k}\sum_{j=1}^{ni}(X_{ji} - \bar{X_i})^2 \quad (3)$$

dengan WSS adalah jumlah kuadrat dalam, k adalah jumlah klaster, ni adalah ukuran sampel dari klaster i, X_{ij} adalah pengukuran ke-j dari klaster ke-i, dan X adalah nilai rata-rata dari jarak keseluruhan data.

$$TFIDF_{ij} = tf_{ij} \; x \log(D / df_j)$$

klaster dengan bantuan grafis yang diusulkan oleh Rousseeuw pada tahun 1987. Tujuan dari metode ini adalah menentukan jumlah klaster optimal dengan memperhitungkan skala rasio jarak antar objek data (Rousseeuw, 1987). Metode ini mengukur kohesi dan separasi dalam rentang nilai koefisien Silhouette untuk menentukan kemiripan objek data dalam sebuah klaster. Rumus yang digunakan adalah sebagai berkut:

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \quad (4)$$

dengan $a(i)$ adalah rata-rata jarak data ke-i terhadap semua data lainnya dalam satu klaster, $b(i)$ adalah hasil rata-rata jarak data ke-i terhadap semua data dari klaster lain, kemudian diambil data yang paling kecil. Langkah-langkah menghitung koefisien *Silhouette* adalah sebagai berikut:

1. Menghitung rata-rata jarak dari jarak data ke-*i* dengan seluruh objek data yang berada dalam satu klaster sehingga akan diperoleh nilai $a(i)$
2. Menghitung rata-rata jarak dari jarak data ke-*i* dengan objek data yang berada di klaster lainnnya. Dari semua jarak rata-rata tersebut ambil nilai yang paling kecil sehingga diperoleh nilai $b(i)$
3. Maka setiap data ke-*i* memiliki nilai koefisien *Silhouette*

Metode

Metode yang digunakan pada penelitian ini memodifikasi teknik penerapan *text mining* yang dikemukakan oleh Solka (2008) yang diawali dari pengumpulan dan mempelajari karakteristik data. Data yang sudah dikumpulkan kemudian dinormalisasi melalui praproses data. Data teks hasil praproses kemudian ditransformasi ke dalam bentuk numerik dengan melakukan pembobotan istilah (term weighthing) yang direpresentasikan dalam bentuk *Term Document Matrix* (TDM) sebagai input data untuk algoritme klasterisasi. Penentuan tendensi klaster dilakukan sebagai upaya memperoleh jumlah klaster optimal. Setelah hasil klaster terbentuk dilakukan evaluasi dalam bentuk intepretasi data oleh pakar. Alat bantu yang digunakan dalam penelitian ini adalah bahasa pemprograman R versi 3.4.4.

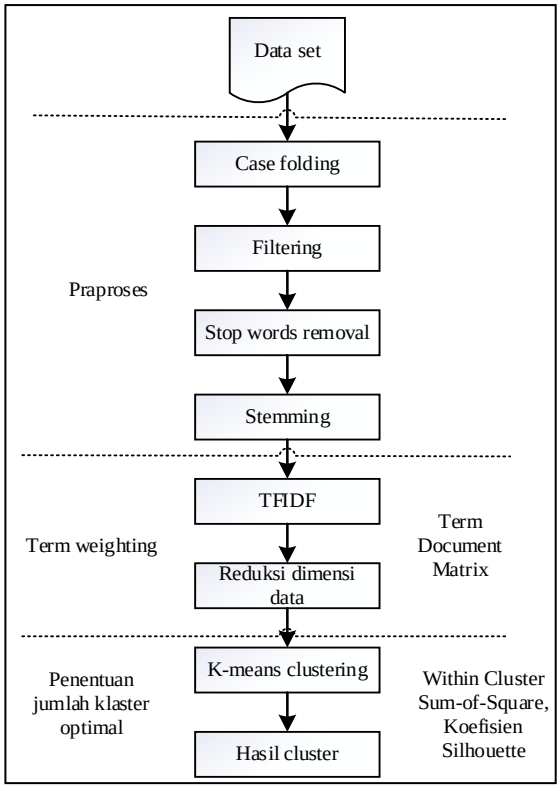
Data Penelitian

Data yang digunakan dalam penelitian ini adalah dokumen siaran media dalam Bahasa Inggris yang bersumber dari website resmi kedutaan Australia Jakarta dengan alamat http://indonesia.embassy.gov.au/jakt/MediaRelease.html. Proses analisis teks dilakukan pada keseluruhan judul dan isi tanpa menyertakan gambar, tabel, dan bentuk ilustrasi lainnya.

Alur Proses Klasterisasi

Pengolahan data dimulai dengan melakukan praproses yang meliputi seluruh kegiatan persiapan data mentah untuk pemrosesan lebih lanjut. Praproses data terdiri dari case folding, filtering, stopwords removal, dan, stemming. Proses ini menghasilkan teks yang telah dikonversi menjadi huruf kecil, hilangnya karakter selain huruf, segmentasi kata pada rangkaian kalimat dan paragraph menjadi individu teks atau token, dan pengembalian kata pada bentuk morfologi dasar, yang kemudian disusun dalam sebuah daftar istilah.

Transformasi data tekstual ke dalam bentuk numerik ditempuh dengan memberikan bobot nilai TFIDF pada setiap istilah. Tahapan ini menghasilkan sebuah korpus baru yang akan digunakan untuk algoritme klasterisasi. Penentuan jumlah klaster optimal dilakukan dengan parameter *Within Cluster Sum-of-Square* (WSS) dan koefisien Silhouette. Pembentukan klaster dilakukan dengan mengunakan algoritme k-means di mana perhitungan jarak antar objek data dihitung berdasarkan formula Cosine Similarity. Alur proses klasterisasi yang digunakan dalam penelitian dapat dlihat pada gambar 1.



Gambar 1. Alur proses klasterisasi

Analisis klaster

Evaluasi pada penelitian ini proses interpretasi data terhadap hasil klaster oleh pakar yang berfokus pada dua hal. Pertama menentukan label klaster sebagai topik dokumen. Kedua, penjelasan mengenai alasan pemilihan label. Kriteria pakar dalam penelitian ini adalah:

1. memiliki kompetensi akademis
2. memiliki kompetensi dan pengetahuan mendalam mengenai konsep dan

implementasi kebijakan luar negeri Australia terhadap Indonesia
3. memiliki kapasitas secara kelembagaan untuk melakukan interpretasi data

Pakar yang ditunjuk dalam penelitian ini adalah Wijaya Kusuma, MBA, analis senior pada Bidang Politik dan Ekonomi, Kedutaan Australia Jakarta.

Hasil Dan Pembahasan

Pengumpulan Data

Tahap pengumpulan data dilakukan dengan mengunduh dokumen siaran media yang dirilis mulai tanggal 1 Januari 2006 sampai 31 Desember 2016 dari *website* resmi Kedutaan Australia yang beralamat di http://indonesia.embassy. gov.au/jakt/ MediaRelease.html. yang pada observasi awal berjumlah 898. Dalam proses pengunduhan, terdapat 69 dokumen yang tidak berhasil diunduh karena file tidak tersedia pada *server* sehingga diperoleh hasil 839 dokumen sebagai data penelitian. Keseluruhan dokumen yang berekstensi html tersebut kemudian dikonversi sebagai data teks dalam format csv untuk kebutuhan penelitian.

Praproses

Praproses dalam penelitian ini adalah normalisasi data tekstual agar menghasilkan output yang lebih konsisten untuk diolah lebih lanjut. Praproses terdiri dari *case folding*, *filtering, stopwords removal,* dan, *stemming*. Ilustrasi praproses dapat dilihat pada tabel 1.

Tabel 1. Tahapan praproses data

Praproses	Hasil
Teks awal	Exhibition to Support Bali Rehabilitation Effort Today to commemorate the closure of the Bali Rehabilitation Fund (BRF)
Case folding	exhibition to support bali rehabilitation effort today to commemorate the closure of the bali rehabilitation fund (brf)
Filtering	exhibition to support bali rehabilitation effort today to commemorate the closure of the bali rehabilitation fund brf
Stopwords removal	exhibition support bali rehabilitation effort commemorate closure bali rehabilitation fund brf
Stemming	exhibit support bali rehabilit effort commemor closur bali rehabilit fund brf

Tabel 1 memperlihatkan tahapan demi tahapan praproses data. Proses penghapusan *stopwords* menggunakan daftar *stopwords* untuk Bahasa Inggris yang yang terdiri dari 571 stopwords (Salton & Buckley, 1988). Proses *stemming* menggunakan algoritme *stemming* berbasis afiks yang dikembangkan M.F Porter (1980). Keseluruhan praproses data mampu menurunkan jumlah istilah dari 12598 menjadi 8250 atau berkurang sebesar 39%. Hasil sejalan dengan penelitian yang pernah dilakukan oleh Kannan (2014) bahwa penghapusan *stopwords* dapat menurunkan jumlah istilah sebesar 20-30%, sedangkan menurut Dolamic (2009) 571 stopwords dapat mengurangi jumlah istilah sebesar 30-50%.

Pembobotan Istilah (*Term Weighting*)

Proses perhitungan bobot nilai TFIDF dilakukan dengan terlebih dahulu menentukan nilai TF dan IDF untuk setiap istilah. Hasil perkalian nilai TF dan IDF kemudian akan menghasilkan nilai TFIDF sebagaimana dapat dilihat pada cuplikan *Term Document Matrix* (TDM) TFIDF pada gambar 2.

Terms							
Docs	addit	advers	advis	affect	alloc	altern	announc
1	0.04419437	0.04896123	0.03441281	0.02264055	0.03125883	0.07243479	0.02714573
10	0.00000000	0.00000000	0.00000000	0.02009803	0.00000000	0.00000000	0.00000000
2	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.03665907	0.02747677
3	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
4	0.00000000	0.00000000	0.00000000	0.00000000	0.03164003	0.00000000	0.00000000
5	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.01706890
6	0.00000000	0.00000000	0.00000000	0.02646712	0.00000000	0.00000000	0.00000000
7	0.05277889	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
8	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
9	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000

Gambar 2. Term Document Matrix TFIDF

Reduksi Dimensi Data

Reduksi dimensi data dilakukan untuk mepertahankan istilah-istilah yang paling signifikan sebagai variabel data yang menentukan tren dari topik dokumen. Nilai batas (threshold) yang ditetapkan adalah sebesar 0.79 yang berarti membuang istilah-istilah yang memiliki frekuensi kemunculan di bawah 21

persen. Proses ini menghasilkan 57 istilah yang akan diproses lebih lanjut. Angka 0.79 diperoleh dari observasi terhadap beberapa kali proses iterasi yang mengkombinasikan penurunan persentase *sparsity* (variabel data dengan nilai 0 pada sel-sel matriks) dan nilai koefisien Silhouette.

Klasterisasi dengan Algoritme K-means

Tahap klasterisasi dilakukan berdasarkan data pada matriks TFIDF yang telah dibuat sebelumnya dengan menggunakan tahapan algoritme k-means. Penentuan jumlah klaster optimal dilakukan dengan menguji jumlah klaster sebanyak k=3 sampai k=6. Pengujian adalah bagian dari observasi awal untuk mendapatkan nilai Within Cluster Sum of Squares (WSS) dan koefisien Silhouette sebagai parameter untuk menentukan jumlah klaster yang akan dibentuk. Klaster dengan nilai WSS yang mendekati 0 mengindikasikan semakin rendah tingkat variabilitas pada intra klaster tersebut, atau dengan kata lain semakin rendah nilai WSS, maka tingkat kemiripan anggota klaster semakin baik. Nilai koefisien Silhouette berada pada rentang 0 sampai 1, di mana klaster dengan nilai yang paling mendekati 1 mengindikasikan tingkat kohesi intra klaster dan separasi inter klaster yang semakin baik. Tabel 2 menunjukan hasil pengujian jumlah klaster k berikut nilai WSS dan koefisien Silhouette yang diperoleh.

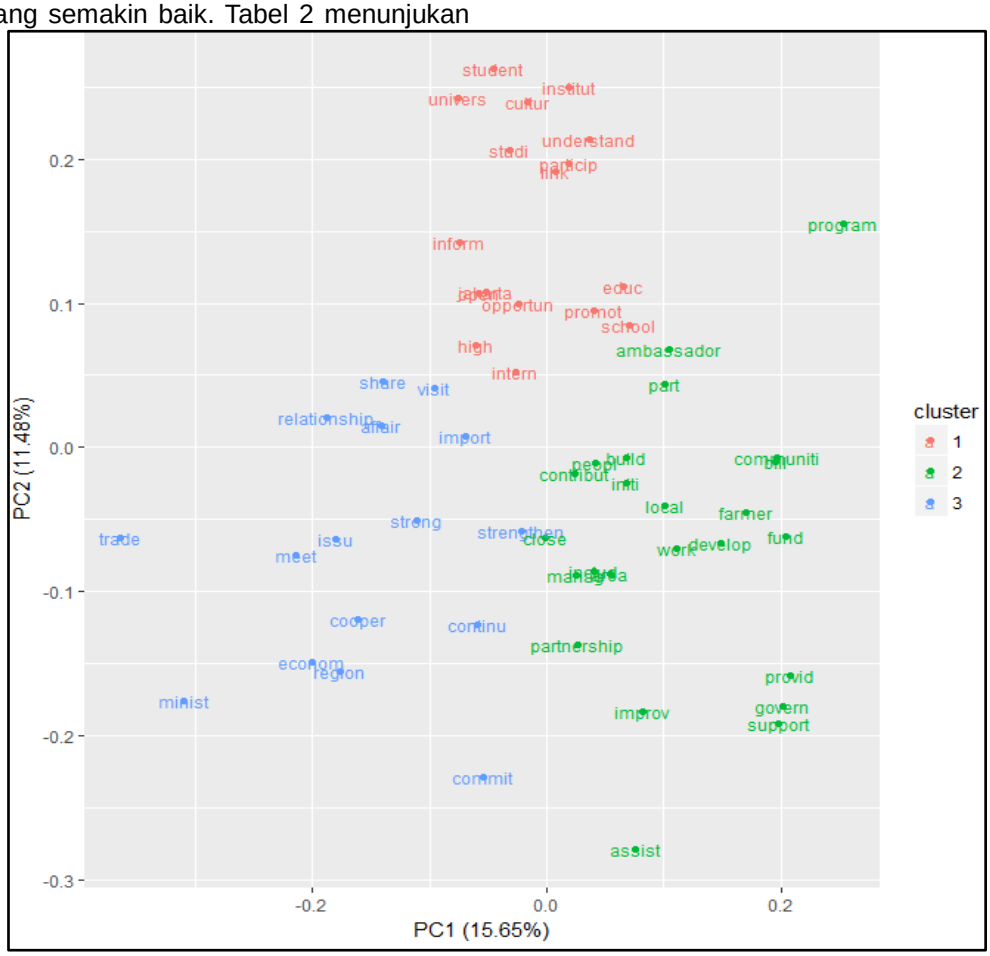
Tabel 2. Nilai WSS & Silhouette

klaster (*k*)	WSS	Si
3	19.6%	0.10
4	25.0%	0.09
5	29.2%	0.09
6	32.7%	0.09

Tabel 2 memperlihatkan bahwa dari hasil pengujian terhadap klaster *k=3* sampai dengan *k=6,* nilai WSS terkecil adalah *k=3* sebesar 19.6% dan nilai koefisien Silhoutte yang paling mendekati 1 adalah k=3 sebesar 0.10. Dengan demikian dapat disimpulkan bahwa jumlah klaster *k=3* adalah jumlah klaster optimal yang akan dibentuk.

Visualisasi Principal Component Analysis (PCA)

Gambar 3 memperlihatkan pengelompokan 57 istilah ke dalam 3 klaster di mana setiap klaster terdiri dari populasi anggota klaster dengan jumlah yang bervariasi. Setiap variabel data atau istilah memiliki koordinat berbeda-beda dalam sebuah ruang 3 dimensi.



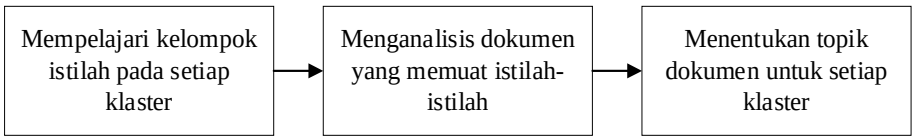
Gambar 3. Hasil plotting klaster k=3

Hasil dari plotting tersebut menggambarkan kecenderungan pola pengelompokan yang dihitung berdasarkan jarak setiap istilah terhadap pusat klaster. Proses selanjutnya adalah melakukan inventarisasi istilah pada masing-masing klaster untuk mengetahui jumlah anggota kelompok setiap klaster, sebagaimana yang dapat dilihat pada tabel 4.

Tabel 3. Item istilah pada klaster

Id klaster	Jumlah anggota	Item istilah
1	17	{opportun, student, studi, high, school, institut, understand, jakarta, link, intern, cultur, educ, univers, particip, promot, open, inform}
2	16	{issu, relationship, continu, commit, trade, strengthen, cooper, meet, affair, econom, minist, visit, import, share, region, strong}
3	24	{assist, govern, close, contribut, part, build, improv, provid, local, communiti, includ, develop, initi, support, peopl, partnership, ambassador, farmer, program, fund, bill, work, manag, area}

Tabel 4 memberikan informasi jumlah anggota pada setiap klaster. Klaster dengan *id* 1 memiliki anggota terbanyak dengan jumlah 17, diikuti klaster dengan *id* 2 sebanyak 16, dan klaster dengan *id* 3 sebanyak 24. Item istilah yang berada pada setiap klaster adalah komponen linguistik yang akan dianalisis lebih oleh pakar. Hal yang juga perlu dipertimbangkan adalah pembentukan klaster bersumber dari frekuensi kemunculan istilah sehingga terdapat kemungkinan ditemukan istilah yang kurang dapat dimaknai sebagai sebuah unsur pembangun topik.



Gambar 4 Tahapan interpretasi data

Interpretasi data

Tahapan terakhir dari penelitian adalah melakukan evaluasi terhadap hasil klasterisasi. Evaluasi dilakukan dengan tujuan agar hasil klasterisasi lebih dapat dimaknai melalui interpretasi kualitatif oleh pakar. Pakar yang ditunjuk dalam penelitian ini adalah Bapak Kusuma Wijaya, MBA, seorang analis senior pada Political and Economic Branch, Kedutaan Australia Jakarta. Dalam melakukan evaluasi, pakar melakukan tahapan interpretasi data pada Gambar 4.

Gambar 4 adalah tahapan proses interpretasi data oleh pakar. Pada tahap Pertama, pakar mempelajari 57 istilah yang dikelompokan ke dalam 3 klaster. Kecenderungan istilah-istilah yang memiliki kemiripan berkelompok dalam sebuah klaster mengindikasikan adanya pola atau tema yang akan diungkap lebih lanjut oleh pakar. Kemudian pada tahap yang kedua, dokumen-dokumen yang memuat istilah dianalisis oleh pakar untuk mendapatkan pemahaman yang lebih utuh mengenai konteks istilah pada klaster. Tahap yang ketiga adalah penentuan topik dari masing-masing klaster berdasarkan analisis awal yang diperoleh pada kedua tahap sebelumnya. Pada tahap ini pula, pakar diminta untuk memberikan penjelasan dari alasan pemilihan sebuah topik. Prinsip dari penentuan topik adalah upaya menemukan konsep dapat merepresentasikan kelompok istilah setiap klaster.

Hasil intrepretasi oleh pakar

Klaster 1 *(People to people links)*

Pakar menentukan topik "*people to people links*" atau "hubungan antar masyarakat" sebagai label data untuk klaster 1. Pakar menjelaskan bahwa topik ini sebagai salah satu misi utama diplomasi pemerintah Australia di Indonesia yang dibuktikan dengan adanya intensitas dialog dan diskusi interaktif komunitas bisnis, akademik, dan media kedua negara yang bertujuan membangun rasa saling memahami. Para pemimpin kedua negara juga menyadari peran masyarakat sipil dalam membangun hubungan bilateral dan oleh karena itu menyambut baik berbagai masukan sebagai bagian dari solusi untuk merespon tantangan kedua negara. Pakar juga memberikan contoh pendirian pusat Bahasa Indonesia di Kota Darwin, Brisbane, dan Sydney bertujuan untuk mempromosikan budaya dan Bahasa Indonesia di seluruh Australia. Hal ini dasari sebuah

pemikiran bahwa dengan memahami bahasa yang digunakan satu sama lain akan meningkatkan rasa hormat, saling memahami, dan sikap menghargai pada masyarakat kedua negara.

Pada sektor akademik, pakar berpendapat bahwa para pemimpin negara juga mengakui bahwa inovasi, ilmu pengetahuan dan tekonologi, media dan pertukaran budaya adalah sarana vital dalam mempromosikan sikap saling menghormati masyarakat di kedua negara. Di bawah kebijakan New Colombo Plan, 2000 pelajar Australia telah berkesempatan menuntut ilmu di Indonesia. Pakar menambahkan, dalam aspek hubungan antar masyarakat, inovasi dan ilmu pengetahuan disorot sebagai area baru dalam kolaborasi bilateral yang berperan penting dalam perkembangan ekonomi. Bentuk nyata dari kolaborasi ini adalah berkumpulnya 100 ilmuwan ternama dari Australia dan Indonesia yang membahas penelitian inovatif di bidang kesehatan, kelautan, pertanian, dan *big data* dalam sebuah simposium di Canberra, Australia pada tahun 2016 lalu.

Klaster 2 (*Human development*)

Pakar menentukan topik "*human development"* atau "pembangunan kualitas hidup manusia" karena pemerintah Australia mendukung secara konsisten usaha-usaha yang dilakukan pemerintah Indonesia dalam meningkatkan akses masyarakat kepada layanan sosial yang lebih baik, termasuk masyarakat miskin yang tinggal di kawasan timur Indonesia. Program bantuan pendidikan yang diberikan kepada Indonesia bertujuan untuk meningkatkan kualitas guru dan proses pembelajaran di sekolah, termasuk dukungan terhadap pendekatan-pendekatan inovatif sebagai upaya pemerintah Indonesia dalam memperoleh *output* yang lebih baik lagi. Pada sisi yang lain, pemerintah Australia juga menyediakan beasiswa program master dan doktoral sebagai upaya untuk menciptakan pemimpin Indonesia masa depan. Aspek lain yang juga ditekankan dalam sektor pendidikan adalah terciptanya hubungan yang kuat antar alumni dan masyarakat kedua negara.

Pakar juga menjelaskan, dalam bidang kesehatan, pemerintah Australia bekerja sama dengan pemerintah Indonesia dalam peningkatan kesehatan masyarakat dan sistem kesehatan hewan yang bertujuan mengurangi ancaman penyakit menular. Selain itu, sejumlah program bantuan juga diberikan untuk perbaikan nutrisi bagi perempuan, anak, dan bayi. Sementara itu, program manajemen bencana menyediakan paket dukungan kebijakan dan teknis untuk meningkatkan kesiapan dan sistem penanggulangan yang dapat diterapkan ketika terjadi bencana yang mengancam keselamatan manusia. Secara singkat, pada aspek kemanusiaan, pemerintah Australia berupaya untuk memberikan kontribusi untuk meningkatkan pelayanan kesehatan dan pendidikan di Indonesia.

Klaster 3 (*Economic cooperation*)

Pakar memilih topik "*economic cooperation"* atau "kerja sama ekonomi" untuk klaster 2 dengan karena salah satu fokus kebijakan pemerintah Australia terhadap Indonesia adalah memperkuat kerja sama ekonomi dan membuka kesempatan-kesempatan baru dalam rangka memaksimal seluruh potensi ekonomi demi keuntungan bersama di masa mendatang. Melalui forum Indonesia-Australia Comprehensive Economic Partnership (IA-CEPA), pemimpin kedua negara berkomitmen untuk saling mendukung kepemimpinan regional untuk meningkatkan pertumbuhan dan kesejahteraan bersama. Forum yang dibangun bersama oleh kedua negara ini adalah upaya untuk mentransformasi kerja sama ekonomi tradisional yang telah dibangun untuk lebih dapat menjawab tantangan-tantangan ekonomi baru secara lebih terfokus. Sejalan dengan itu, forum IA-CEPA adalah sebuah mekanisme kerja sama untuk meningkatkan perdagangan, investasi, dan kemitraan ekonomi yang lebih luas lagi. Pakar menambahkan, pengelolaan tata kelola investasi bertujuan untuk mendukung Indonesia dalam mempercepat pertumbuhan dan mencapai keuntungan bersama dari perdagangan internasional. Untuk mendukung tujuan tersebut, pemerintah Australia menyediakan bantuan teknis pada tingkat pemerintahan dengan memprioritaskan program reformasi ekonomi yang meliputi, supervisi pada sektor finansial, anggaran, perdagangan dan kompetisi, kebijakan perpajakan, serta manajemen administrasi pemerintahan dan ekonomi makro. Pemerintah Australia juga bekerja sama dengan Indonesia secara intensif untuk mengatasi hambatan disinsentif pada investasi di bidang infrastruktur, memberikan masukan pada regulasi perencanaan proyek, demi menjamin hasil terbaik dari pembangungan infrastruktur di Indonesia. Dengan melihat fakta bahwa dua pertiga masyarakat miskin di Indonesia tinggal di pedesaan, pemerintah Australia juga memberikan program bantuan secara berkelanjutan pada pembangunan sektor pertanian. Program bantuan yang diberikan bertujuan untuk mendorong pertumbuhan ekonomi inklusif dengan cara memaksimalkan manfaat pasar pertanian bagi masyarakat miskin, meningkatkan keamanan pangan, produktivitas pertanian, dan membantu petani

dalam memperoleh pendapatan tinggi dengan mengatasi hambatan akses pada pinjaman.

Kesimpulan

Penelitian ini menunjukan bahwa teknik penambangan teks dapat digunakan untuk klasterisasi topik dokumen siaran media. Dengan metode tersebut, proses klasterisasi dapat dilakukan lebih sistematis karena proses analisis teks dilakukan melalui beberapa tahapan dengan parameter-parameter yang telah ditentukan. Penelitian ini juga menyimpulkan bahwa algoritme k-means dapat digunakan untuk proses penemuan pola klaster yang menghasilkan berkelompoknya istilah-istilah yang memiliki kemiripan dalam satu klaster. Metode validasi klaster yang mengukur tingkat kohesi dan separasi klaster terbukti dapat membantu penentuan jumlah klaster optimal. Evaluasi oleh pakar juga membuktikan bahwa klasterisasi topik yang dihasilkan telah sesuai dengan diplomasi Kedutaan Australia di Indonesia yang menekankan aspek “hubungan antar masyarakat”, “kerja sama ekonomi”, dan “pembangunan kualitas hidup manusia” sebagai tema utama yang ingin disampaikan kepada pemerintah dan masyarakat di Indonesia.

Ucapan Terima Kasih

Terima kasih penulis ucapkan kepada Bapak Dr Eng Wisnu Ananta Kusuma, ST MT dan Bapak Prof Dr Sulistyo Basuki, MA MSLS selaku supervisor dalam penelitian ini. Penghargaan penulis sampaikan kepada Ibu Kestrilia Rega Prilianti, M.Sc yang telah banyak memberi masukan dan saran serta Bapak Wijaya Kusuma, MBA analis senior pada Political and Economic Branch, Kedutaan Besar Australia Jakarta selaku evaluator.

Referensi

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: classification, clustering and extraction techniques. Diakses 17 Desember 2017, dari *ArXiv Preprint ArXiv:1707.02919*.

Davis, C.H., & Shaw, D. (2013). *Introduction to information science and technology*. Medford, N.J.: American Society for Information Society.

Dolamic, L., & Savoy, J. (2010). When stopword lists make the difference. *Journal of the American Society for Information Science and Technology*, *61*(1), 200–203. Diakses 17 Desember 2017, dari https://doi.org/10.1002/asi.21186.

Feldman, R., & Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. *Imagine*. Diakses 15 November 2017, dari https://doi.org/10.1179/1465312512Z.00000000017.

Gurusamy, V., Kannan, S., & Prabhu, J. R. (2017). Mining the attitude of social network users using K-Means clustering. *International Journal of Advanced Research in Computer Science and Software Engineering*, *7*(5), 226–230. Diakses 15 November 2017, dari https://doi.org/10.23956/ijarcsse/SV7I5/0231.

Kannan, S., & Gurusamy, V. (2014). Preprocessing techniques for text mining. Diakses 18 Februari 2018, dari https://www.researchgate.net/publication/273127322_Preprocessing_Techniques_for_Text_Mining.

Lama, P. (2013). *Clustering system based on text mining using the K-means algorithm : news headlines clustering*. Turku University of Applied Sciences. Diakses 20 November 2017, dari http://www.theseus.fi/handle/10024/69505.

Mathew, S. (2012). *Financial services data management : big data technology in financial services*. *Oracle Financial Services*. Diakses 4 November 2017, dari http://www.oracle.com/us/industries/financial-services/bigdata-in-fs-final-wp-1664665.pdf.

Miner, G. D., Elder, J., & Nisbet, R. A. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Diakses 20 Desember 2017, dari https://doi.org/10.1016/C2010-0-66188-8.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137. Diakses 11 Januari 2018, dari https://doi.org/10.1108/eb046814.

Prilianti, K.R., & Wijaya, H. (2014). Aplikasi text mining untuk automasi penentuan tren topik skripsi dengan metode K-Means Clustering. *Jurnal Cybermatika*, *2*(1), 1–6. Diakses 15 Oktober 2017, dari http://cybermatika.stei.itb.ac.id/ojs/index.php/cybermatika/article/view /58/28.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*(November), 53–65. Diakses 20 Februari 2018, dari https://doi.org/10.1016/0377-0427(87)90125-7.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523. Diakses 10 Februari 2018, dari https://doi.org/10.1016/0306-4573(88)90021-0.

Solka, J. L. (2008). Text data mining: theory and methods. *Statistics Surveys*, *2*, 94–112. Diakses 15 Februari 2018, dari https://doi.org/10.1214/07-SS016.

Spire Technologies. (2016). Making sense of unstructured data with Spire. Diakses February 25, 2018, dari http://spiretechnologies.com/making-sense-unstructured-hr-data-spire/.

Sulistyo-Basuki. (2014). *Senarai pemikiran Sulistyo Basuki : profesor pertama ilmu perpustakaan dan informasi di Indonesia.* Jakarta : Ikatan Sarjana Ilmu Perpustakaan dan Informasi Indonesia.

United Nations. (1961). Vienna Convention on Diplomatic Relations. *International and Comparative Law Quarterly*. Diakses 6 November 2017, dari https://doi.org/10.1093/iclqaj/10.3.600.

Zade, J., Bamnote, D., & Agrawal, P. (2017). Text document clustering using K-Means algorithm with its analysis and implementation. *Imperial Journal of Interdisciplinary Research*, *3*(2), 1528–1531. Diakses 16 Desember 2017, dari http://www.imperialjournals.com/index.php /IJIR/article/view/ 4259/4079.

Zhao, Q., Xu, M., & Fränti, P. (2009). *Adaptive and natural computing algorithms*. (M. Kolehmainen, P. Toivanen, & B. Beliczynski, Eds.) (Vol. 5495). Berlin, Heidelberg: Springer Berlin Heidelberg. Diakses 23 Januari 2018, dari https://doi.org/10.1007/978-3-642-04921-7.