Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 1, No. 1, Januari 2017, hlm. x-x http://j-ptiik.ub.ac.id

Topic Modelling Pada Aktivitas Pengembangan Perangkat Lunak Menggunakan BERTopic

Bagas Raditya Nur Listyawan[1], Nanang Yudi Setiawan[2], Mochamad Chandra Saputra[3]

Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]bagass.rnl@student.ub.ac.id, [2]nanang@ub.ac.id, [3]andra@ub.ac.id

Abstrak

Data merupakan kumpulan fakta atau informasi yang dikumpulkan, diukur, atau dihimpun untuk analisis. Dalam era digital, data pengembangan perangkat lunak menjadi sangat penting karena mencakup aktivitas yang dilakukan programmer untuk mengembangkan aplikasi. Namun, data ini sering kali menumpuk dan sulit untuk dianalisis secara manual. Oleh karena itu, pengelompokan topik menjadi penting untuk memahami tren dan evaluasi aktivitas pengembangan perangkat lunak. Penelitian ini bertujuan untuk mengelompokkan topik pada data aktivitas pengembangan perangkat lunak menggunakan metode BERTopic. Metode tersebut dikembangkan berdasarkan teknik BERT. Data yang digunakan dalam penelitian ini adalah data aktivitas pengembangan perangkat lunak. Metode penelitian meliputi pengumpulan data, *preprocessing data*, pembuatan *corpus* dan *dictionary*, implementasi BERTopic, serta evaluasi model menggunakan matriks *topic coherence* dan *topic diversity*. Selain itu, evaluasi juga dilakukan dengan meminta validasi langsung kepada *stakeholder*. Hasil penelitian menunjukkan bahwa penerapan BERTopic berhasil mengidentifikasi topik dalam data aktivitas pengembangan perangkat lunak. Evaluasi model menunjukkan hasil yang cukup akurat dengan nilai *topic coherence* sebesar 0.625 dan *topic diversity* sebesar 0.828. Selain itu, validasi berdasarkan pernyataan *stakeholder* memberikan respon bahwa hasil dari BERTopic memiliki kekurangan berupa topik yang *overlap* dan topik yang tidak terdeteksi. Penelitian ini menyimpulkan bahwa BERTopic cukup layak dalam mengelompokkan topik pada data aktivitas pengembangan perangkat lunak. Namun, diperlukan penyesuaian parameter untuk memaksimalkan hasil.

Kata kunci: *BERTopic, topic modelling, data mining, pengembangan perangkat lunak, topic diversity, topic coherence*

Abstract

Data is a collection of facts or information gathered, measured, or assembled for analysis. In the digital era, software development data is crucial as it encompasses the activities performed by programmers to develop applications. However, this data often accumulates and becomes difficult to analyze manually. Therefore, topic modeling is essential to understand trends and evaluate software development activities. This research aims to cluster topics in software development activity data using the BERTopic method. The method was developed based on the BERT technique. The data used in this research comprises software development activity data. The research methodology includes data collection, data preprocessing, corpus and dictionary creation, BERTopic implementation, and model evaluation using topic coherence and topic diversity metrics. Additionally, evaluation is conducted by directly seeking validation from stakeholders. The results indicate that applying BERTopic successfully identifies the main topics in the software development activity data. The model evaluation shows reasonably accurate results, with a topic coherence score of 0.625 and a topic diversity score of 0.828. Furthermore, stakeholder validation feedback highlights shortcomings such as overlapping topics and undetected topics. This research concludes that BERTopic is fairly adequate for clustering topics in software development activity data. However, parameter adjustments are necessary to optimize the results.

Keywords: BERTopic, *topic modelling*, *data mining*, pengembangan perangkat lunak, *topic diversity*, *topic coherence*

1. PENDAHULUAN

Data merupakan kumpulan fakta atau informasi yang dikumpulkan, diukur, atau dihimpun untuk analisis. Dalam era digital saat ini, data menjadi sangat penting dalam berbagai bidang, termasuk bisnis, ilmu pengetahuan, dan teknologi. Pengumpulan data dilakukan melalui berbagai metode, termasuk survei, sensor, dan pengukuran. Data dapat berupa angka, teks, gambar, atau suara, dan analisis data dapat memberikan wawasan yang berharga untuk pengambilan keputusan. Namun, pentingnya data juga diiringi oleh tantangan terkait privasi, keamanan, dan etika penggunaannya.

Pada era digital ini, teknologi telah digunakan dimana – mana. Hal ini, menyebabkan perkembangan data yang digunakan maupun dihasilkan menjadi semakin cepat dan besar (Yaqoob et al., 2016). Menurut data pada *Search Engine Badan Pusat Statistik* (BPS), pada awal tahun 2024 terdapat sekitar 820.000 data yang tersedia. Data tersebut merupakan data yang didapat oleh Badan Pusat Statistik (BPS) dan belum termasuk data pribadi yang dimiliki oleZh perusahaan – perusahaan yang ada di Indonesia. Hal ini membuktikan bahwa perkembangan data saat ini menjadi lebih cepat dan besar.

Perkembangan pada era digital ini juga semakin banyak instansi yang memproduksi produk digital berkembang. Dalam proses bisnis instansi tersebut, tentu saja menghasilkan cukup banyak data yang memuat aktivitas pengembangan perangkat lunak yang telah dibuat.

Perkembangan dalam proses bisnis instansi yang cepat ini menghasilkan perkembangan data yang pesat sehingga banyak data dihasilkan. Data yang berkembang menjadi terlalu banyak ini akan menjadi permasalahan baru yang mana akan terjadi kesulitan dalam membaca dan menyimpulkan data tersebut. Oleh karena itu, perlu dilakukan pemrosesan data dengan merangkum semua data dan mengelompokkannya menjadi beberapa jenis berupa daftar topik dari data. Hal ini akan membantu seseorang yang ingin mengetahui mengenai apa saja yang telah dilakukan berdasarkan data tersebut.

Pengelompokkan topik pada data aktivitas pengembangan perangkat lunak dapat dibilang memiliki cukup banyak manfaat maupun tujuan. Tujuan dari pengelompokkan topik adalah untuk mengelompokkan data dengan membagi data menjadi beberapa topik. Hal ini dapat dimanfaatkan untuk mengetahui secara garis besar mengenai apa saja yang dibahas di dalam data melalui daftar topik yang dihasilkan. Pengelompokkan topik pada data aktivitas pengembangan perangkat lunak dapat dimanfaat untuk melakukan evaluasi mengenai apa saja yang telah dilakukan dalam mengembangkan produk. Selain itu, pengelompokkan topik memberikan manfaat lainnya seperti, kita dapat memperikan *job description* yang akan dikerjakan dalam *project* kedepannya yang dapat digunakan untuk membuat persiapan agar kinerja dan hasil menjadi maksimal saat mengerjakan pekerjaan yang berhubungan dengan topik tersebut. Pengelompokkan topik ini akan dilakukan dengan melewati beberapa tahapan Data Mining.

Data Mining merupakan proses pengolahan data untuk mencari suatu informasi yang belum diketahui dan memiliki potensi berguna untuk digunakan (Chen et al., 1996). Topic Modelling adalah salah satu metode *clustering* pada data mining yang biasa digunakan untuk mengatur, memahami, mencari, dan meringkas data secara otomatis(Tong & Zhang, 2016). Topic Modelling dapat digunakan untuk mengelompokkan topik pengembangan perangkat lunak yang sedang populer. Dalam implementasi Topic Modelling dapat menggunakan beberapa cara atau library yang telah disediakan

dalam bahasa pemrograman Python.

BERTopic merupakan sebuah *package* dalam Python yang dibuat berdasarkan pendekatan *clustering* dan memperluasnya dengan menggabungkan metode TF-IDF berbasis kelas untuk membuat gambaran topik (Grootendorst, 2022). BERTopic adalah salah satu teknik Topic Modelling yang dibuat berdasarkan BERT (*Bidirectional Encoder Representations from Transformers)* (Hutama & Suhartono, 2022). BERTopic berguna untuk mengidentifikasi dan memahami topik atau subjek yang terdapat pada sebuah teks atau file. Implementasi BERTopic ditujukan untuk mengelompokkan teks yang berhubungan dengan topik yang sama.

BERTopic merupakan metode baru yang dibuat oleh Maarten Grootendorst pada tahun 2020 dan dipublikasikan secara resmi melalui jurnal pada tahun 2022 (Grootendorst, 2022). Metode ini masih tergolong sebuah metode yang baru sehingga masih sedikit penelitian yang membahas mengenai BERTopic. Sehingga, BERTopic ini cukup menarik untuk digunakan sebagai salah satu contoh pembuktian apakah metode ini layak digunakan atau tidak.

Berdasarkan penjelasan tersebut, penelitian ini bertujuan untuk mengelompokkan topik yang ada di dalam data aktivitas pengembangan perangkat lunak untuk mengetahui daftar topik pengembangan yang telah dilakukan. Selain itu, penelitian ini dilakukan untuk mengetahui apakah metode BERTopic menghasilkan hasil yang akurat dan layak digunakan dalam pengelompokkan topik. Penelitian ini diharapkan memberikan manfaat kepada *stakeholder* agar dapat menggunakan topik yang dihasilkan sebagai referensi untuk monitoring pengerjaan *project* di tahun berikutnya atau dapat memanfaatkan topik tersebut untuk melatih *programmer* baru untuk mengerjakan *project* tersebut.

2. LANDASAN TEORI

2.1. Topic Modelling

Topic Modelling atau pemodelan topik merupakan metode yang cukup umum digunakan untuk segmentasi dan sistem rekomendasi dengan mengelompokkan sekelompok data yang memiliki karakteristik yang sama (An et al., 2023). Topic Modelling sendiri merupakan bentuk dari model statistic yang digunakan pada *machine learning* dan *natural language processing* (NLP) untuk mengidentifikasi struktur tersembunya dari kumpulan data atau teks (Egger & Yu, 2022). Topic Modelling memiliki fungsi untuk mengelompokkan kumpulan data atau teks menjadi beberapa topik. Hal ini dapat dimanfaatkan untuk segmentasi produk dan pembuatan rekomendasi produk untuk *stakeholder* yang akan memberikan keuntungan jangka panjang.

Topic Modelling dapat dibuat dengan berbagai metode seperti Lateng Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), Top2VEC, dan BERTopic (Egger & Yu, 2022). Penilitian ini akan menggunakan BERTopic sebagai metode untuk melaksanakan penelitian.

2.2. Python

Python adalah sebuah bahasa pemrograman tingkat tinggi yang terkenal karena mudah untuk dipelajari (Terragni et al., 2021). Python memiliki perkembangan popularitas yang luar biasa dalam komunitas *scientific computing.* Hal ini dikarenakan banyaknya *library machine learning* dan *deep learning* menggunakan *python* sebagai bahasa utama mereka. Beberapa contoh library yang menerapkan *machine learning* adalah Bidirectional Encoder Representations from Transformer atau BERT yang dikenal sebagai dasar dari beberapa *library* yang digunakan untuk menganalisis data seperty PyABSA dan BERTopic.

2.3. BERTopic

BERTopic merupakan sebuah metode untuk Topic Modelling atau pemodelan topik yang memanfaatkan teknik pengelompokkan

dan variasi TF-IDF berbasis kelas untuk menghasilkan topik yang bersangkutan (Grootendorst, 2022). BERTopic menyediakan *document embedding extraction* atau ekstraksi dokumen menjadi bentuk vector dengan model pengubah kalimat yang mendukung lebih dari 50 bahasa (Egger & Yu, 2022).

BERTopic memiliki 3 langkah utama untuk menghasilkan topik yang telah diolah (Grootendorst, 2022). Langkah pertama yaitu, Document Embeddings yang mana pada tahap ini akan dilakukan konversi document atau biasanya teks menjadi bentuk vektor. Langkah selanjutnya adalah Document Clustering, pada tahap ini akan dilakukan *clustering* atau pengelompokkan data berdasarkan jarak terdekat. Langkah terakhir adalah Topic Representation, pada tahap ini akan ditentukan sebuah kalimat atau kata sebagai perwakilan kluster yang mana penentuan perwakilan kluster tersebut dilakukan dengan prosedur TF-IDF.

2.4. Class-Based TF-IDF

TF-IDF adalah sebuah prosedur yang menggabungkan 2 statistik yaitu, Term Frequency dan Inverse Document Frequency. Class-Based TF-IDF adalah salah satu jenis varian TF-IDF yang biasa dikenal dengan c-TF-IDF. Prosedur c-TF-IDF dilakukan dengan mengeneralisasi rumus TF-IDF ke dalam kluster dokumen. Semua data atau dokumen pada sebuah kluster akan diperlakukan sebagai satu dokumen dengan menggabungkan dokumen. Setelah itu, TF-IDF disesusaikan untuk mengelompokkan dokumen ke kluster dan terbentuklah rumus c-TF-IDF sebagai berikut :

$$W_{t,c} = tf_{t,c} \times \log(1 + \frac{A}{t_{ft}}) \quad (1)$$

$tf_{t,c}$ = Frekuensi kata *t* pada kluster *c*
t_{ft} = Frekuensi kata *t* pada semua kluster
A = Jumlah rata – rata kata per kelas *A*

2.5. OCTIS

Optimizing and Comparing Topic Models is Simple atau biasa dikenal OCTIS merupakan sebuah *library python* berupa *framework* evaluasi yang digunakan untuk melatih, menganalisis, dan membandingkan model topik (Terragni et al., 2021).

OCTIS memiliki beberapa matriks yang dapat digunakan untuk mengevaluasi sebuah model yang telah dibuat. Berikut adalah jenis matriks yang disediakan oleh OCTIS :

1. Classification Metrics
2. Coherence Metrics
3. Diversity Metrics
4. Similiarity Metrics

2.6. Topic Coherence

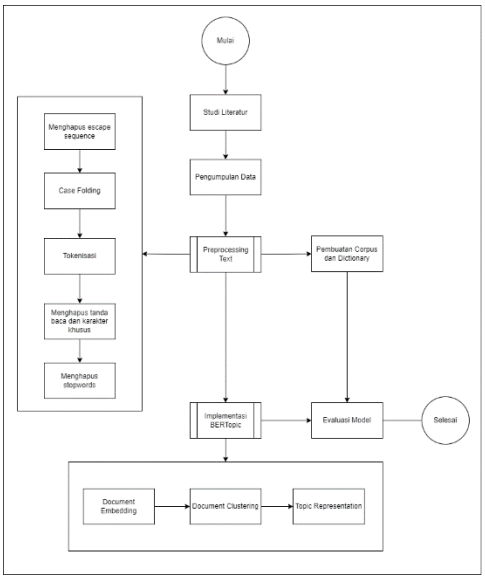
Topic Coherence merupakan salah satu metode evaluasi yang digunakan untuk menemukan hubungan antar 1 kata dengan kata lain pada suatu dokumen (Röder et al., 2015). Evaluasi ini dilakukan dengan mengambil data *top N words* dari sebuah topik dan mengukur hubungannya dengan semua pasangan kata.

2.7. Topic Diversity

Topic Diversity merupakan salah satu jenis evaluasi yang terdapat pada *diversity matrics*. Topic Diversity bertujuan untuk mengukur keberagaman kata – kata teratas dari suatu topik satu sama lain (Terragni et al., 2021). Metode evaluasi ini merupakan salah satu metode umum yang sering digunakan dalam pengembangan Topic Modelling.

3. METODE PENELITIAN

Pada penelitian ini ada beberapa tahapan yang akan dilakukan oleh peneliti, tahapan tersebut dapat dilihat pada Gambar 1 berikut.



Gambar 1. Alur Penelitian

3.1. Studi Literatur

Pada tahap ini, peneliti melakukan studi literatur untuk menambah wawasan dan pengetahuan peneliti untuk melakukan penelitian dengan topik yan relevan. Studi literatur ini dilakukan dengan mempelajari penelitian terdahulu dan mempelajari *library* yang mendukung penelitian.

Studi literatur pada penelitian ini mempelajari terkait pemodelan topik, *text preprocessing*, matriks evaluasi (*topic coherence* dan *topic diversity)*, dan *library* BERTopic dan OCTIS

3.2. Pengumpulan Data

Pengumpulan data dilakukan dengan meminta data kepada *stakeholder*. Data merupakan salah satu data pengembangan aktivitas *project* aplikasi yang dikembangkan oleh *stakeholder*. Data yang didapat memiliki beberapa kolom yaitu tahun_bulan, *issue_id*, *project*, *subject*, dan *description*. Penelitian ini akan menggunakan kolom *subject* sebagai data untuk penelitian.

3.3. Data Preprocessing

Preprocessing Data akan dilakukan untuk memproses data agar menjadi lebih mudah dideteksi saat melakukan proses selanjutnya. Proses *preprocessing data* memuat proses *remove escape sequence* (menghapus kode tab, enter, dan sejenisnya seperti *n*, *t*, dan sejenisnya) lalu melakukan *case folding* atau mengubah huruf menjadi kecil semua atau *lowercase*. Setelah itu akan dilakukan *tokenization* untuk memisahkan setiap kata dan akan dilanjutkan dengan penghapusan tanda baca dan karakter khusus serta *stopwords* seperti *the*, *and*, *is*, *in*, dan lain-lain. Setelah semua langkah *preprocessing data* selesai, *file* baru akan disimpan untuk data yang telah diolah.

3.4. Pembuatan Corpus dan Dictionary

Pembuatan Corpus dan Dictionary akan digunakan untuk mengevaluasi model yang telah dibuat. Dalam pembuatan corpus dibutuhkan sebuah *function* untuk mengubah data menjadi bentuk token dan pembuatan *dictionary* berdasarkan token tersebut.

3.5. Implementasi BERTopic

Implementasi BERTopic dibagi menjadi 3 tahapan yaitu sebagai berikut.

3.5.1. Document Embedding

Document Embedding merupakan proses pengubahan dokumen atau data teks yang tersedia ke dalam bentuk vektor. Proses *embedding* dilakukan dengan menggunakan *sentence*-BERT (SBERT).

3.5.2. Document Clustering

Dokumen atau data teks yang telah diubah menjadi vektor akan dilanjut dengan proses *clustering* atau pengelompokkan. Langkah pertama yang dilakukan adalah pengurangan dimensi atau *dimensionality reduction* dengan Uniform Manifold Approximation and Projection (UMAP). Proses *dimensionality reduction* dilakukan untuk mengurangi dimensi data yang naik setelah proses *embeddings*. Proses ini akan menghasilkan data yang menjadi vektor namun, dengan nilai dimensi yang lebih kecil.

Data hasil *dimensionality reduction* akan dilanjutkan dengan proses *clustering* yang akan diproses menggunakan algoritma Hierarchical Density-Based Spatial Clustering of Application (HDBSCAN). Proses *clustering* dengan menggunakan HDBSCAN akan dilakukan

dengan menggunakan salah satu *function* BERTopic dan akan menghasil daftar berupa nomor *cluster* tempat data berada.

3.5.3.Topic Representation

Topic Representation merupakan proses penentuan sebuah kata atau kalimat yang akan menjadi perwakilan setiap kluster atau dapat dibilang tahap ini adalah tahap untuk menentukan nama topik yang ditemukan. Penentuan perwakilan kata atau kalimat pada setiap kluster akan ditentukan dengan menggunakan metode TF-IDF berbasis kelas (c-TF-IDF). Hasil dari tahapan Topic Representation berupa tabel yang memuat kolom *topic*, *count*, *name*, *representation*, dan *representative_docs*.

3.6. Evaluasi Model

Tahap ini merupakan tahap terakhir berupa evaluasi dari model yang telah dibuat. Evaluasi model akan dilakukan dengan 2 cara yaitu menerapkan matriks evaluasi dan validasi kepada pihak *stakeholder* mengenai aspek keakuratan hasil yang dihasil oleh BERTopic. Evaluasi menggunakan matriks yaitu dengan menerapkan 2 matriks yaitu, *topic coherence* dan *topic diversity*. Proses evaluasi model ini akan dilakukan dengan salah satu *library* pada *python* yaitu Optimizing and Comparing Topic models Is Simple atau biasa dikenal dengan OCTIS.

4. HASIL DAN EVALUASI

Model yang telah dibuat pada pemodelan topik dilakukan tahapan evaluasi untuk mengetahui performa dan keakuratan dari *output* yang dihasilkan oleh model. Evaluasi ini akan dilakukan dengan perhitungan statistik menggunakan *topic coherence* dan *topic diversity*. Selain itu, juga akan dilakukan evaluasi dengan memvalidasi hasil topik pada pemilik data atau *stakeholder* untuk mengetahui keakuratan apakah hasil topik cukup relevan dengan pengembangan yang telah dilakukan.

4.1. Hasil BERTopic

Model yang telah dibuat pada pemodelan topik dilakukan tahapan evaluasi untuk mengetahui performa dan keakuratan dari *output* yang dihasilkan oleh model. Evaluasi ini akan dilakukan dengan perhitungan statistik menggunakan *topic coherence* dan *topic diversity*. Selain itu, juga akan dilakukan evaluasi dengan memvalidasi hasil topik pada pemilik data atau *stakeholder* untuk mengetahui keakuratan apakah hasil topik cukup relevan dengan pengembangan yang telah dilakukan.

Tabel 1. Hasil Topik

No	Count	Topics	Topic Representation
1	789	0_override_siswa_ajuan_sekolah	['override', 'siswa', 'ajuan', 'sekolah', 'pendaftaran', 'akun', 'jalur', 'verifikasi', 'pilihan', 'penyesuaian']
2	291	1_data_migrasi_migrasi data_siswa	['data', 'migrasi', 'migrasi data', 'siswa', 'data siswa', 'ppdb', 'lulusan', 'dnt', 'data dnt', 'kota']
3	89	2_publik_situs publik_situs_publik kota	['publik', 'situs publik', 'situs', 'publik kota', 'publik riil', 'kota', 'riil', 'kab', 'publik situs', 'prov']
4	76	3_engine_closing engine_closing_jalur	['engine', 'closing engine', 'closing', 'jalur', 'tahap', 'tahap closing', 'engine kota', 'kota', 'prestasi', 'engine kab']
5	55	4_deploy_deploy real_real_kota	['deploy', 'deploy real', 'real', 'kota', 'real kota', 'real kab', 'kab', 'real prov', 'prov', 'bontang']
6	48	5_database_database migrasi_migrasi_part database	['database', 'database migrasi', 'migrasi', 'part database', 'migrasi database', 'part', 'siswa', 'database siswa', 'siswa part', 'pendaftar']

4.2. Analisa Hasil

Hasil yang tertera pada Tabel 1. Hasil Topik merupakan hasil yang didapat dari proses pemodelan topik menggunakan BERTopic. Setelah hasil daftar topik didapatkan, selanjutnya akan dilakukan analisis mengenai topik tersebut dengan memastikan pembahasan topik tersebut melalui *file* data yang digunakan dalam penelitian. Proses analisis pembahasan topik dilakukan dengan membaca kolom *description* pada data. Berdasarkan analisis

tersebut didapatkan penjelasan mengenai pembahasan topik sebagai berikut.

1. Topik 1 yaitu *0_override_siswa_ajuan_ sekolah* Topik ini membahas mengenai proses penyesuaian atau perubahan ajuan siswa di sekolah, seperti pendaftaran siswa baru, verifikasi akun, dan penyesuaian jalur pendaftaran. Ini mencakup semua kegiatan administrasi yang terkait dengan pengajuan dan penyesuaian pendaftaran siswa di sekolah.
2. Topik 2 yaitu *1_data_migrasi_migrasi data_siswa*. Topik ini membahas mengenai proses pemindahan atau migrasi data siswa, seperti data pribadi, akademik, dan lainnya dari satu sistem ke sistem lain. Ini mencakup prosedur untuk memastikan data siswa dipindahkan dengan benar, terutama dalam konteks PPDB (Penerimaan Peserta Didik Baru).
3. Topik 3 yaitu *2_publik_situs publik_situs_publik kota*. Topik ini membahas mengenai pengelolaan situs web publik setiap kota yang mencakup akses perubahan, pemeliharaan, dan melakukan aktivasi *website* serta segala hal yang mencakup segala sesuatu yang berhubungan dengan informasi yang disediakan di dalam *website*.
4. Topik 4 yaitu *3_engine_closing engine_closing_jalur*. Topik ini membahas mengenai penggunaan sistem untuk menyelesaikan suatu proses pendaftaran atau seleksi jalur, seperti jalur zonasi, prestasi, dan lain-lain. Ini melibatkan tahapan penutupan pendaftaran dan memastikan semua proses telah diselesaikan dengan benar.
5. Topik 5 yaitu 4_deploy_deploy real_real_kota. Topik ini membahas mengenai implementasi atau peluncuran sistem atau layanan di berbagai kota atau kabupaten. Ini mencakup kegiatan yang memastikan bahwa layanan atau sistem baru diterapkan dengan sukses di wilayah-wilayah yang ditargetkan.
6. Topik 6 yaitu 5_database_database migrasi_migrasi_part database. Topik ini membahas mengenai migrasi atau pemindahan bagian tertentu dari database siswa termasuk data pendaftar. Ini mencakup prosedur untuk memindahkan bagian-bagian data yang spesifik untuk memastikan integritas dan konsistensi data selama proses migrasi.

4.3. Evaluasi Hasil

Evaluasi hasil dilakukan dengan 2 cara yaitu dengan matriks evaluasi dan validasi kepada *stakeholder*. Berikut merupakan hasil evaluasi.

4.3.1. Evaluasi Hasil Menggunakan Matriks Evaluasi

Topic Coherence merupakan salah satu metode yang digunakan untuk mengevaluasi hasil dari hubungan antara satu kata dengan kata lain. Sedangkan Topic Diversity merupakan salah satu metode yang digunakan untuk mengevaluasi hasil dari keberagaman suatu kata. Kedua metode tersebut digunakan dalam evaluasi hasil dari *topic modelling* yang mana menggunakan berbagai keberagaman serta hubungan antar kata.

Tabel 2 Hasil Evaluasi Menggunakan Matriks Evaluasi

Metode Evaluasi	Score
Topic Coherence	0.6250604341955988
Topic Diversity	0.8285714285714286

Dapat dilihat pada Tabel 2 Hasil Evaluasi Menggunakan Matriks Evaluasi bahwa hasil *topic coherence* yang didapat adalah 0.625. Nilai tersebut dapat dibilang cukup baik. Sedangkan hasil dari *topic diversity* pada Tabel 5.2 Hasil Evaluasi Menggunakan Topic Coherence dan Topic Diversity menunjukkan nilai 0.828. Nilai tersebut dapat dibilang sangat baik

4.3.2. Evaluasi Hasil Kepada Stakeholder

Berdasarkan hasil yang didapat, *stakeholder* telah mengisi beberapa pertanyaan yang telah disiapkan oleh penulis. Berdasarkan jawaban *stakeholder*, 3 dari 8 pertanyaan mendapat jawaban positif sedangkan sisanya memiliki jawaban negatif. Hal ini dapat disimpulkan bahwa masih terdapat cukup banyak kekurangan yang dihasilkan dari pemodelan topik menggunakan BERTopic.

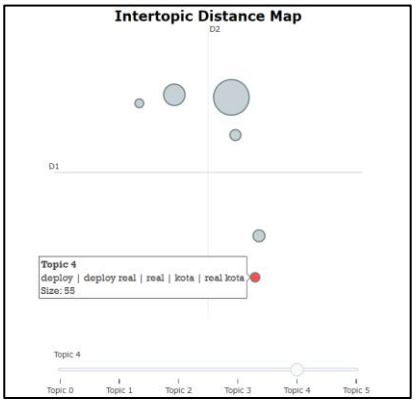
Kekurangan utama yang ditemukan yaitu terdapat topik yang *overlap* (tumpeng tindih) dan terdapat topik yang tidak terdeteksi.

4.4. Visualisasi Hasil

Visualisasi hasil ini dilakukan agar hasil dapat dilihat lebih mudah dan detail dari pengelompokkan topik, kesamaan antar topik, dan *keyword* yang mewakili topik. Visualisasi topik akan dilakukan dengan menampilkan *intertopic distance map* dan *topic word scores*.

4.4.1. Visualisasi Dengan Intertopic Distance Map

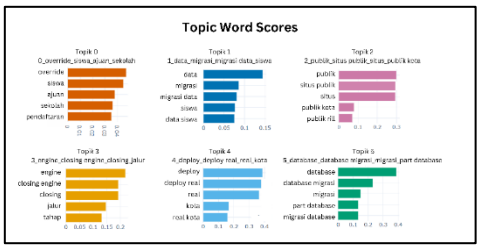
Visualisasi *intertopic distance map* dilakukan untuk mengetahui hubungan dan jarak antar topik. Vektor topik akan diubah menjadi 2 dimensi sehingga dapat divisualisasikan. Titik – titik yang terdapat pada *map* ini akan melambangkan seberapa mirip atau berbedanya topik tersebut. Berikut pada Gambar 5.1 Intertopic Distance Map merupakan hasil visualisasi *output topic modelling* dengan *intertopic distance map*.



Gambar 2. Intertopic Distance Map

4.4.2. Visualisasi Topic Word Scores

Visualisasi *topic word scores* dilakukan untuk mengetahui nilai intensitas dari sebuah perwakilan kata atau *keywords* pada tiap topik. Nilai yang dilihat merupakan nilai yang didapat dari perhitungan skor *c-TF-IDF*. Semakin besar nilai *c-TF-IDF* maka kata atau *keyword* tersebut dapat dikatakan paling sering muncul dalam topik tersebut. Berikut pada Gambar 5.2 Topic Word Scores merupakan hasil dari visualisasi *topic word scores output topic modelling*.



Gambar 3. Topic Word Scores

5. PENUTUP

5.1. Kesimpulan

Berdasarkan penelitian topic modelling menggunakan BERTopic dengan studi kasus aktivitas pengembangan perangkat lunak, terdapat beberapa Kesimpulan yang dapat ditarik, yaitu :

a. Topic Modelling dengan menggunakan BERTopic pada studi kasus aktivitas pengembangan perangkat lunak menghasilkan 6 topik. Topik pertama yang dihasilkan yaitu override_siswa_ajuan sekolah, data_ migrasi_migrasi data_siswa, public_situs, publik_situs_publik, engine_ closing engine_closing_jalur, deploy_deploy real_real_kota, database_database migrasi_migrasi_part database.
b. Hasil dari topic modelling pada penelitian ini sudah cukup akurat dengan menghasil keterkaitan antar kata pada topik dengan cukup baik yang dilambangkan dengan skor topic coherence sebesar 0.625 dan keberagaman kata yang sangat bervariasi yang dilambangkan dengan skor topic diversity sebesar 0.828. Namun, berdasarkan pernyataan oleh *stakeholder* masih terdapat kesalahan berupa topik yang *overlap* (tumpang tindih) dan terdapat beberapa topik yang tidak terdeteksi. Berdasarkan hasil evaluasi menggunakan metriks evaluasi serta pernyataan *stakeholder* dapat disimpulkan bahwa metode BERTopic ini cukup akurat karena dapat menghasilkan keselaran antar kata yang baik, keberagaman kata yang sangat baik, dan topik yang cukup dikenali oleh *stakeholder*. Namun, hasil dari BERTopic ini masih kurang maksimal dikarenakan

terdapat topik yang *overlap* dan juga tidak terdeteksi.

5.2. Saran

Berdasarkan dari hasil penelitian *topic modelling* menggunakan BERTopic yang dilakukan dengan studi kasus aktivitas pengembangan perangkat lunak, terdapat beberapa saran yang diberikan untuk selanjutnya, yaitu:

a. Dalam penelitian ini terdapat salah satu kekurangan yaitu, berdasarkan fakta lapangan yang dialami oleh *stakeholder* terdapat topik yang tidak tercantum pada hasil penelitian ini. Penelitian selanjutnya diharapkan mencoba untuk mengatur parameter dalam BERTopic untuk dapat menghasilkan hasil topik yang lebih maksimal.
b. Hasil akhir dari penelitian ini terdapat cukup banyak data yang dikategorikan *outliers*. Penelitian selanjutnya diharapkan mencoba untuk mencari *function* pada dokumentasi BERTopic untuk mengurangi *outliers* yang dihasilkan. Selain itu, juga dapat mencari referensi atau menambahkan *source code* yang diperuntukan mengurangi *outliers*.
c. Penelitian ini menggunakan metode BERTopic sebagai dasar utama proses pemodelan topik. Penelitian selanjutnya dapat dikembangkan lagi dengan membandingkan hasil dari BERTopic dengan hasil metode yang serupa dan jarang atau belum pernah digunakan seperti, Latent Dirichlet Allocation (LDA), Top2Vec, atau Non-negative Matrix Factorization (NMF).

6. DAFTAR PUSTAKA

An, Y., Oh, H., & Lee, J. (2023). Marketing Insights from Reviews Using Topic Modeling with BERTopic and Deep Clustering Network. *Applied Sciences (Switzerland)*, *13*(16). https://doi.org/10.3390/app13169443

Chen, M. S., Han, J., & Yu, P. S. (1996). Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering*, *8*(6), 866–883. https://doi.org/10.1109/69.553155

Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, *7*(May), 1–16. https://doi.org/10.3389/fsoc.2022.886498

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. http://arxiv.org/abs/2203.05794

Hutama, L. B., & Suhartono, D. (2022). Indonesian Hoax News Classification with Multilingual Transformer Model and BERTopic. *Informatica (Slovenia)*, *46*(8), 81–90. https://doi.org/10.31449/inf.v46i8.4336

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, 399–408. https://doi.org/10.1145/2684822.2685324

Search Engine Badan Pusat Statistik. (n.d.). Retrieved February 15, 2024, from https://searchengine.web.bps.go.id/

Terragni, S., Fersini, E., Galuzzi, B., Tropeano, P., & Candelieri, A. (2021). OCTIS: Comparing and optimizing topic models is simple! *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the System Demonstrations*, 263–270. https://doi.org/10.18653/v1/2021.eacl-demos.31

Tong, Z., & Zhang, H. (2016). *A Text Mining Research Based on LA Topic Modelling*. 201–210. https://doi.org/10.5121/csit.2016.60616

Yaqoob, I., Hashem, I. A. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B., & Vasilakos, A. V. (2016). Big data: From beginning to future. *International Journal of Information Management*, *36*(6), 1231–1247. https://doi.org/10.1016/j.ijinfomgt.2016.07.009