PEMODELAN TOPIK DALAM AL-QUR'AN MENGGUNAKAN *LIBRARY* BERTOPIC PADA MODEL BAHASA BERT

Herwinsyah
Fakultas Sains Dan Teknologi
Universitas Islam Negeri Sunan Kalijaga, Yogyakarta
Email: herwinsy@gmail.com

ABSTRAK

Al-Qur'an adalah kitab suci bagi umat beragama Islam Islam yang terdiri dari 114 surah dan 6236 ayat. Para Ulama besar dari seluruh dunia telah mengkategorikan ke dalam berbagai topik diantaranya hukum, adat (perang), kenikmatan, sejarah, dan kehidupan setelah kematian. Dalam penelitian ini penulis melakukan *reserch* dengan tujuan mencari topik-topik dalam Al Qur'an melalui pendekatan metode *Deep Learning* dengan menggunakan *Library* model bahasa BERT *(Bidirectional Encoder Representations from Transformers)*, khususnya BERTopic, untuk seluruh 6236 ayat Al-Qur'an sebagai dataset. Penelitian ini menggunakan pendekatan pemodelan topik, tahapan penelitian dilakukan melalui beberapa tahap yaitu; pengumpulan data, Preprocessing dan pemodelan topik. Metode pemodelan topik menggunakan BERTOPIC. Hasil penelitian adalah pemodelan topik menghasilkan 8 topik utama secara terperinci sebagai berikut; Topic 0 Al Quran dengan prosentase sebesar 6%, Topic 1 Aku (Allah) sebesar 6,5%, Topic 2 Langit 3,8%, Topic 3 Rasul 8%, Topic 4 Malaikat 12,5%, Topic 5 Wanita 5%, Topic 6 Neraka dengan prosentase 13%, serta Topic 7 Dibangkitkan sebesar 5,5%. Kata-kata tersebut dianggap sangat penting dalam mewakili topik-topik yang dihasilkan dan juga termasuk dalam kategori spiritual, moral, dan hukum.

Kata kunci: model bahasa, BERTopic, pemodelan topik, Qur'an

ABSTRACT

The Quran is the holy book for the Islamic faith, consisting of 114 chapters and 6236 verses. Esteemed scholars from around the world have categorized it into various topics, including law, customs (war), pleasures, history, and life after death. In this study, the author conducted research with the aim of identifying topics within the Quran using the Deep Learning approach, utilizing the BERT (Bidirectional Encoder Representations from Transformers) language model library, particularly BERTopic, on the entire dataset of 6236 Quranic verses. This research adopted a topic modeling approach, and the research process encompassed several stages: data collection, preprocessing, and topic modeling. The topic modeling method employed was BERTOPIC. The research findings revealed that the topic modeling generated 8 detailed main topics as follows: Topic 0 - The Quran with a percentage of 6%, Topic 1 - God (Allah) with 6.5%, Topic 2 - Heaven with 3.8%, Topic 3 - Messenger with 8%, Topic 4 - Angels with 12.5%, Topic 5 - Women with 5%, Topic 6 - Hell with a percentage of 13%, and Topic 7 - Resurrection with 5.5%. These words are considered highly significant in representing the generated topics and also fall within the categories of spiritual, moral, and legal matters.

Keywords: *language model, BERTopic, topic modeling, Qur'an*

1. PENDAHULUAN

Al-Qur'an adalah kitab suci umat Islam yang terdiri dari 114 surah yang panjangnya bervariasi. Bab-bab dalam Al-Qur'an merupakan bagian-bagian terpisah yang membentuk kitab suci tersebut. Al-Qur'an disusun berdasarkan panjangnya, dimulai dari bab yang terpanjang hingga yang terpendek. Setiap bab memiliki nama unik yang menggambarkan isi atau tema yang dibahas di dalamnya. Nama bab dalam Al-Qur'an diambil dari kata-kata yang terdapat di awal ayat pertama

dari bab tersebut. Contohnya, bab pertama dalam Al-Qur'an dinamakan "Al-Fatihah" yang berarti "pembukaan" [1] karena kata tersebut merupakan kata pertama di ayat pertama dari bab tersebut. Topik yang dibahas dalam bab-bab Al-Qur'an sangat beragam, termasuk ajaran-ajaran agama seperti keimanan kepada Allah, kewajiban shalat, puasa, dan zakat [2]. Selain itu, ada juga perintah-perintah yang harus dilakukan oleh umat Islam, seperti menjauhi hal-hal yang haram dan bertindak adil terhadap sesama.

Selain itu, Al-Qur'an juga terdiri dari bab-bab yang berisi kisah-kisah nabi dan para sahabatnya, seperti kisah Nabi Ibrahim, Nabi Musa, dan Nabi Muhammad SAW. Ada pula bab yang berisi ayat-ayat yang mengandung hikmah dan pelajaran moral, seperti bab Al-Kahfi yang menceritakan kisah para penghuni gua, dan bab Ar-Rahman yang mengajarkan tentang kebesaran Allah. Tiap bab dalam Al-Qur'an memiliki makna dan pesan yang tersendiri bagi umat Islam. Oleh karena itu, setiap muslim diwajibkan untuk membaca dan memahami isi dari setiap bab dalam Al-Qur'an. Dengan memahami isi dari tiap bab, umat Islam dapat lebih memahami ajaran-ajaran agama dan melaksanakan perintah-perintah yang terkandung dalam Al-Qur'an dengan lebih baik.

Al-Quran adalah kitab suci Islam yang diyakini sebagai firman Allah yang diturunkan kepada Nabi Muhammad SAW melalui malaikat Jibril selama 23 tahun [3]. Al-Quran terdiri dari 114 surah, yang masing-masing memiliki panjang unik dan total terdiri dari 6.236 ayat. Kata "ayah" dalam bahasa Arab berarti "tanda" atau "ayat". Ayat dalam Al-Quran terdapat dalam ayat-ayat tersendiri yang dikelompokkan bersama dalam bab-bab yang masing-masing mengandung pesan dari Allah. Setiap ayat dalam Al-Quran dianggap sebagai petunjuk atau nasehat bagi orang yang beriman, sehingga mampu memperbaiki kehidupan dan akhlak mereka. Al-Quran dipandang sebagai sumber doktrin dan hukum yang paling otoritatif dalam agama Islam.

Mayoritas pembaca Alquran mengandalkan ringkasan para ahli untuk mendapatkan pengetahuan tentang masalah utama yang dibahas dalam kitab suci ini. Setiap surat dalam Al-Qur'an memiliki tema yang berbeda, yang didasarkan pada latar belakang kejadian atau asbabun nuzul-nya [4]. Secara umum, subjek-surat dapat dibagi menjadi tiga kategori, yaitu spiritual, moral, dan hukum. Keimanan, ketauhidan, dan akhirat adalah contoh dari masalah spiritual, sedangkan kebajikan, kesabaran, dan kejujuran adalah contoh prinsip moral. Masalah hukum dalam Islam merupakan subjek hukum. Namun, apakah seluruh isi Al-Qur'an mengikuti tema yang ditentukan oleh ayat-ayatnya? Banyak orang sampai pada kesimpulan bahwa seluruh substansi Al-Qur'an terdiri dari lima masalah utama [5], yaitu: hukum, perselisihan, penyebutan nikmat Allah, penyebutan peristiwa sejarah yang terjadi di masa lalu, dan penyebutan kehidupan setelah kematian.

Cara ntuk mengetahui apakah pokok-pokok bahasan tersebut konsisten dengan prediksi topik-topik pada pembahasan di paragraf sebelumnya, penelitian ini mencoba mengusulkan sebuah metode yang sejalan dengan pengertian *natural language processing* (NLP) [6]. Dalam model bahasa berbasis BERT (*Bidirectional Encoder Representations from Transformers*) [7], terdapat modul Python [8] populer yang disebut BERTopic [9], yang merupakan representasi dari *transformer* [10] canggih dengan teknik *unsupervised* [11], yang dapat mengidentifikasi subjeknya sendiri hanya dengan menggunakan dokumen korpus dataset. Paket ini tidak memerlukan label untuk setiap dokumen.

Pemodelan topik pada model bahasa BERT melibatkan penggunaan metode BERT (*Bidirectional Encoder Representations from Transformers*) untuk mengidentifikasi dan memahami topik atau subyek yang dibicarakan dalam suatu teks. BERT adalah model bahasa yang telah dilatih secara mendalam menggunakan data besar untuk memahami konteks kata dalam teks dengan baik. Dalam pemodelan topik, tujuannya adalah untuk mengelompokkan teks-teks yang berhubungan dengan topik yang sama ke dalam kelompok-kelompok tertentu. Ini dapat membantu mengklasifikasikan dan mengatur teks-teks yang ada dalam koleksi besar, serta memfasilitasi pencarian informasi yang lebih efektif. Berikut adalah beberapa langkah umum yang terlibat dalam pemodelan topik dengan metode BERT:

Pemrosesan Teks: Teks-teks yang akan diolah pertama kali dibersihkan dan diolah untuk menghapus karakter khusus, tanda baca, dan kata-kata umum yang tidak memberikan banyak informasi (*stop words*). Pemahaman Konteks dengan BERT: Setiap kalimat atau dokumen dalam koleksi teks diubah menjadi representasi vektor yang menggambarkan pemahaman konteks kata-

kata di dalamnya. Ini mencakup informasi tentang bagaimana kata-kata tersebut saling berhubungan dalam konteks kalimat.

Klasifikasi atau Pengelompokan: Setelah dokumen-dokumen diubah menjadi representasi vektor dengan BERT, teknik klasifikasi atau pengelompokan diterapkan untuk mengelompokkan dokumen-dokumen ke dalam topik-topik yang sesuai. Ini bisa melibatkan algoritma seperti *k-means clustering* atau algoritma klasifikasi yang lebih kompleks. Penyaringan dan Interpretasi: Hasil keluaran dari pemodelan topik bisa memerlukan penyaringan lebih lanjut dan interpretasi. Ini termasuk mengidentifikasi topik-topik yang muncul, memberi label pada kelompok-kelompok dokumen yang relevan, dan menghapus hasil yang tidak informatif atau ambiguitas. Evaluasi: Evaluasi dilakukan untuk memeriksa sejauh mana hasil pemodelan topik sesuai dengan ekspektasi dan tujuan analisis. Evaluasi bisa melibatkan metrik kualitatif dan kuantitatif untuk mengukur seberapa baik topik-topik yang dihasilkan merepresentasikan konten sebenarnya.

Penelitian yang membahas tentang pemodelan topik seperti pemodelan teks berita berbahasa Indonesia dengan pengambilan data sebanyak 68.537 data, metode yang digunakan pada penelitian ini menggunakan metode LDA dan memperoleh nilai topic coherence sebesar 0.67 [12]. Penelitian selanjutnya pemodelan topik pada teks berita berbahasa Indonesia dengan pengambilan data sebanyak 59.279 data, metode pada penelitian ini menggunakan metode LDA, hasil nilai topic coherence sekitar 0.2 [13]. Kemudian Penelitian lainnya adalah analisis sentimen pada judul teks berita berbahasa Indonesia, metode yang digunakan pada penelitian ini adalah metode LDA dan LSTM, hasil akurasi yang dihasilkan sebesar 71.13% [14].

Berdasarkan hal di atas maka tujuan penelitian ini adalah mencari topik-topik dalam Al Qur'an melalui pendekatan metode *Deep Learning* dengan menggunakan *Library* model bahasa BERT *(Bidirectional Encoder Representations from Transformers)*, khususnya BERTopic, untuk seluruh 6236 ayat Al-Qur'an sebagai dataset.

2. METODOLOGI PENELITIAN

Tahap pertama dalam metode penelitian topic modeling menggunakan BERTopic adalah mengumpulkan data. Data dapat diperoleh dari berbagai sumber, seperti basis data, survei, dan lainnya. Selanjutnya, tahap berikutnya adalah preprocessing data [15]. Preprocessing data melibatkan tahap-tahap seperti membersihkan data, mengubah data menjadi format teks, dan memilih kata-kata yang akan digunakan dalam topic modeling. Setelah itu, tahap selanjutnya adalah pemodelan topik menggunakan BERTopic. BERTopic adalah salah satu algoritma topic modeling yang menggunakan metode embedding untuk mengidentifikasi topik-topik yang terdapat dalam data. Pada tahap ini, kita juga perlu menentukan jumlah topik yang diinginkan. Dengan BERTopic, jumlah topik akan secara otomatis ditentukan. Tahap terakhir dalam metode penelitian topic modeling menggunakan BERTopic adalah interpretasi hasil. Pada tahap ini, kita perlu menganalisis hasil pemodelan topik untuk memahami tema-tema apa yang terdapat dalam data yang telah kita kumpulkan.

2.1. *Pengumpulan Data*

Data dapat diambil dari berbagai sumber, seperti database, survei, dan lain sebagainya. Dalam penelitian ini, dataset yang digunakan adalah terjemahan Al-Quran berbahasa Indonesia. Proses pengumpulan data untuk menerjemahkan Al-Quran ke dalam bahasa Indonesia memerlukan langkah-langkah yang teliti dan terencana. Pertama, orang yang bertanggung jawab untuk menerjemahkan harus memastikan bahwa ia memiliki akses ke sumber data yang akurat dan terpercaya, yaitu teks Al-Quran dalam bahasa Arab asli. Kedua, orang tersebut juga harus memastikan bahwa ia memiliki akses ke berbagai sumber terjemahan Al-Quran ke dalam bahasa Indonesia yang telah ada sebelumnya, untuk memastikan bahwa terjemahan yang dihasilkan akurat dan sesuai dengan konvensi yang telah ditetapkan. Ketiga, orang tersebut harus memastikan bahwa ia memiliki akses ke sumber-sumber lain yang dapat membantu dalam proses menerjemahkan, seperti kamus bahasa Arab-Indonesia atau kamus tata bahasa Indonesia. Dengan melakukan

langkah-langkah tersebut, orang yang bertanggung jawab untuk menerjemahkan dapat memastikan bahwa data yang dikumpulkan akurat dan sesuai dengan tujuan yang ingin dicapai. Untuk mendapatkan dataset berupa terjemahan Al-Quran yang terpercaya, penelitian ini menggunakan dataset terjemahan Al-Quran yang diunduh melalui website resmi Kementerian Agama Republik Indonesia. Setelah itu, data tersebut yang terpecah persurahannya disatukan kembali menjadi satu Al-Quran lengkap. Setelah dataset diproses, tahap selanjutnya adalah preprocessing data.

2.2. *Teks Preprocessing*

Preprocessing data memiliki beberapa fungsi yang sangat penting dalam proses analisis data. Pertama, preprocessing data bertujuan untuk membersihkan data dari kesalahan-kesalahan atau kekurangan-kekurangan yang mungkin terjadi [16]. Hal ini sangat penting karena data yang tidak bersih dapat menyebabkan hasil analisis yang tidak akurat atau bahkan tidak mungkin dilakukan. Kedua, preprocessing data bertujuan untuk mengubah data ke dalam format yang lebih sesuai untuk analisis, seperti mengubah data ke dalam bentuk tabel atau mengubah tipe data dari teks ke numerik. Ketiga, preprocessing data juga bertujuan untuk mengintegrasikan data dari berbagai sumber menjadi satu kesatuan yang lebih terstruktur. Hal ini sangat penting karena data dari sumber yang berbeda mungkin memiliki format yang berbeda, sehingga diperlukan integrasi data untuk memudahkan proses analisis. Preprocessing data memiliki peran yang sangat penting dalam proses analisis data karena dapat membantu menghasilkan hasil yang lebih akurat dan terstruktur. Tahap-tahap dalam preprocessing data meliputi pembersihan data, pengubahan data ke dalam bentuk teks, dan pemilihan kata-kata yang akan digunakan dalam topic modelling [17]. Proses dalam preprocessing data, antara lain:

2.2.1. *Remove Simbol dan Number*

Pada tahap *preprocessing*, langkah yang sering dilakukan adalah menghapus simbol dan angka yang terdapat dalam data. Hal ini dilakukan karena simbol dan angka biasanya tidak memiliki makna yang signifikan dalam proses analisis dan dapat mengganggu pemodelan topik atau analisis lainnya. Untuk menghapus simbol dan angka, kita dapat menggunakan fungsi seperti *re sub ()* atau *string punctuation* pada Python. Dengan menggunakan fungsi tersebut, kita dapat mengidentifikasi dan menghapus simbol dan angka yang terdapat dalam data. Namun, perlu diingat bahwa langkah ini harus dilakukan dengan hati-hati karena beberapa simbol atau angka mungkin memiliki makna yang signifikan dalam konteks tertentu, seperti tanda baca atau angka pada alamat email. Oleh karena itu, sebelum menghapus simbol dan angka, perlu dipastikan bahwa kita tidak menghapus informasi penting dalam proses analisis yang akan dilakukan.

2.2.2. *Remove Duplicate*

Langkah selanjutnya adalah menghapus duplikat atau data yang sama dengan data lain yang telah ada. Tahap ini dilakukan untuk menghindari adanya data yang tidak akurat atau tidak relevan yang mungkin terjadi akibat adanya data yang sama dengan data lain. Untuk menghapus duplikat, kita dapat menggunakan fungsi-fungsi seperti drop duplicates () atau unique () pada Python. Dengan menggunakan fungsi tersebut, kita dapat mengidentifikasi dan menghapus baris-baris data yang sama dengan baris-baris data lain yang telah ada. Namun, perlu diingat bahwa langkah ini harus dilakukan dengan hati-hati karena ada beberapa kasus di mana data yang sama mungkin memiliki makna yang berbeda dalam konteks tertentu. Oleh karena itu, sebelum menghapus duplikat, kita perlu memastikan bahwa kita tidak menghapus informasi yang penting.

2.2.3. *Lower case*

Lower case adalah proses mengubah semua teks menjadi huruf kecil (lower case). Proses ini dilakukan untuk menghindari perbedaan bentuk yang mungkin terjadi akibat adanya huruf besar

(upper case) dalam data. Misalnya, kata "Analisis" dan "analisis" mungkin dianggap sebagai kata yang berbeda oleh beberapa algoritma pemodelan topik atau analisis lainnya. Untuk mengubah semua teks menjadi huruf kecil, kita dapat menggunakan fungsi lower () pada Python. Dengan menggunakan fungsi tersebut, kita dapat dengan mudah mengubah semua teks dalam data menjadi huruf kecil.

2.2.4. *Tokenizing*

Pada tahap preprocessing berikutnya, terdapat tokenisasi, yaitu proses membagi teks menjadi sejumlah token atau kata-kata. Tujuan dari tokenisasi adalah untuk memudahkan proses pemodelan topik atau analisis lainnya dengan memecah teks menjadi kata-kata yang lebih kecil. Untuk melakukan tokenisasi, kita dapat menggunakan library seperti NLTK (Natural Language Toolkit) atau spaCy pada Python. Dengan menggunakan library tersebut, kita dapat memecah teks menjadi token dengan mudah menggunakan metode seperti word_tokenize() atau .sent_tokenize(). Meskipun demikian, perlu diingat bahwa tokenisasi harus dilakukan dengan hati-hati karena ada beberapa kasus di mana token yang dihasilkan mungkin tidak sesuai dengan konteks atau makna yang sebenarnya. Oleh karena itu, pemilihan teknik tokenisasi yang tepat sangat penting dalam memastikan hasil analisis yang akurat.

2.2.5. *Remove Stopword*

Tahap selanjutnya adalah menghapus stopword, yaitu kata-kata yang sering muncul dalam teks tetapi tidak memiliki makna yang signifikan dalam konteks tertentu. Stopword biasanya tidak membawa informasi yang berguna dalam proses pemodelan topik atau analisis lainnya dan dapat mengganggu proses analisis. Untuk menghapus stopword, kita dapat menggunakan *library* seperti NLTK (*Natural Language Toolkit*) atau spaCy pada Python. Dengan menggunakan *library* tersebut, kita dapat dengan mudah mengakses daftar *stopword* yang telah tersedia dan menghapus *stopword* yang terdapat dalam teks. Namun, perlu diingat bahwa dalam beberapa kasus, ada kata *stopword* yang mungkin memiliki makna penting dalam konteks tertentu, sehingga penghapusan *stopword* harus dilakukan secara hati-hati.

2.2.6. *Join Case*

Tahap terakhir dalam proses *preprocessing* pada penelitian ini adalah penggabungan kembali kata-kata yang terpisah menjadi satu kata. Langkah ini dilakukan untuk mengembalikan kata-kata ke dalam bentuk aslinya setelah melalui proses tokenisasi atau pemecahan kata menjadi token. Untuk menggabungkan kembali kata-kata yang terpisah, kita dapat menggunakan fungsi " ".join () pada Python. Dengan menggunakan fungsi tersebut, kita dapat menggabungkan kembali token-token menjadi kata-kata dengan mudah. Namun, perlu diingat bahwa langkah ini harus dilakukan dengan hati-hati karena terdapat kasus di mana kata-kata yang terpisah memiliki makna yang berbeda dalam konteks tertentu. Sehingga sebelum menggabungkan kembali kata-kata yang terpisah, kita harus memastikan bahwa kata-kata yang seharusnya dipisahkan tidak ikut digabungkan.

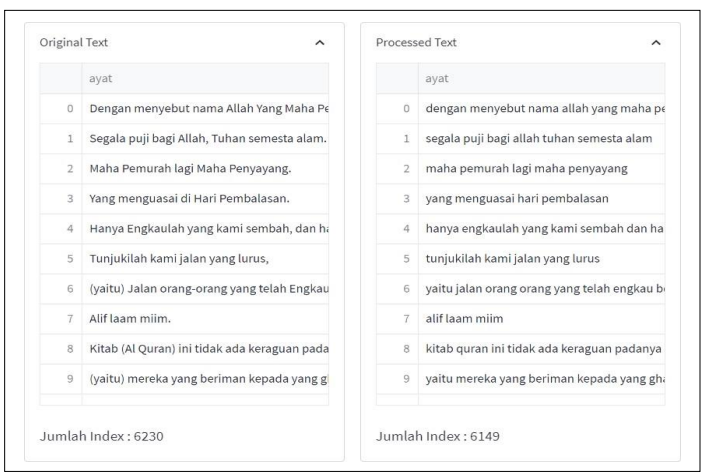
2.3. *Model Generation and Visualization (Model BERTOPIC)*

Setelah itu, tahap selanjutnya adalah pemodelan topik menggunakan BERTopic. BERTopic adalah salah satu algoritma pemodelan topik yang menggunakan metode *embedding* untuk mengidentifikasi topik-topik yang terdapat dalam data. Pada tahap ini, kita juga perlu menentukan jumlah topik yang diinginkan. Tahap terakhir dalam metode penelitian pemodelan topik menggunakan BERTopic adalah interpretasi hasil. Pada tahap ini, kita perlu menganalisis hasil pemodelan topik untuk memahami tema-tema apa yang terdapat dalam data yang telah dikumpulkan. Model generasi dan visualisasi merupakan tahap penting dalam proses pemodelan

topik menggunakan BERTopic. Pada tahap model generasi, kita menggunakan algoritma BERTopic untuk memodelkan topik-topik yang terdapat dalam data yang telah dikumpulkan. Pada tahap ini, kita juga perlu menentukan jumlah topik yang diinginkan dan mengatur parameter-parameter lain yang diperlukan oleh algoritma BERTopic. Setelah model topik tergenerasi, tahap selanjutnya adalah visualisasi. Visualisasi dilakukan untuk membantu kita memahami hasil pemodelan topik dan mengevaluasi seberapa baik model topik tersebut dalam mengidentifikasi topik-topik yang terdapat dalam data. Ada beberapa cara untuk menvisualisasikan hasil pemodelan topik menggunakan BERTopic, seperti dengan menggunakan diagram sankey atau word cloud. Dengan melakukan visualisasi, kita dapat lebih mudah memahami dan mengevaluasi hasil pemodelan topik yang telah dilakukan.

3. HASIL DAN PEMBAHASAN

Pada penelitian ini, digunakan *software* Google Colab yang diakses melalui *browser* Google Chrome dan aplikasi bawaan dari Google yaitu Google Drive sebagai tempat penyimpanan data korpus yang akan diproses. Dataset yang digunakan merupakan terjemahan dari Al-Quran yang diunduh melalui website resmi Kementerian Agama Republik Indonesia. Data ini dijadikan sumber rujukan utama untuk Al-Quran yang diakui dan digunakan oleh masyarakat Indonesia. Setelah dataset diunduh, langkah selanjutnya adalah membersihkan data dari teks, karakter, dan angka yang tidak diperlukan dan tidak bermakna, seperti yang terlihat pada gambar 1.



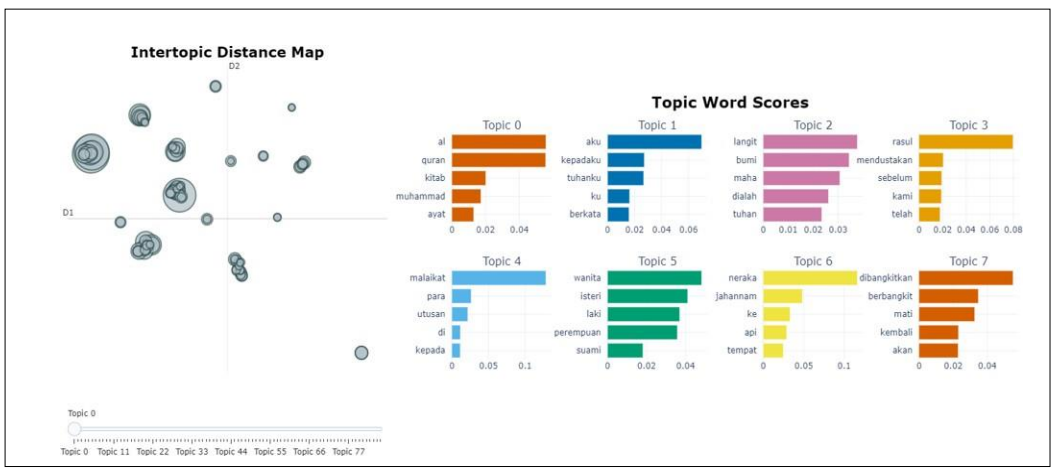
Gambar 1. Preprocessing dataset al-qur'an

Setelah selesai *preprocessing* data, selanjutnya dataset yang telah dibersihkan dijadikan list data yang akan diimplementasikan pada model Bahasa BERTopic dengan Bahasa Indonesia. Proses implementasi model Bahasa BERTopic pada dataset ayat Al-Quran merupakan proses pembuatan model. Proses pembuatan model adalah menjalankan algoritma BERTopic untuk memodelkan topik yang terdapat dalam dataset ayat Al-Quran yang telah disiapkan. Pada tahap ini, algoritma BERTopic akan memproses dataset dan mengelompokkan term-term yang terdapat dalam dataset ke dalam kelompok-kelompok topik seperti pada Gambar 2.

	Topic	Count	Name
0	-1	2341	-1_mereka_kami_dan_yang
1	0	326	0_al_quran_kitab_muhammad
2	1	293	1_aku_kepadaku_tuhanku_ku
3	2	256	2_langit_bumi_maha_dialah
4	3	127	3_rasul_mendustakan_sebelum_kami
5	4	118	4_malaikat_para_utusan_di
6	5	109	5_wanita_isteri_laki_perempuan
7	6	102	6_neraka_jahannam_ke_api
8	7	93	7_dibangkitkan_berbangkit_mati_kembali
9	8	91	8_amal_saleh_pahala_mengerjakan

Gambar 2. Pengelompokan berdasarkan topik

Setelah berhasil meng-*generate* model topik, langkah selanjutnya adalah melakukan visualisasi hasil pemodelan topik agar memudahkan dalam memahami dan mengevaluasi hasil pemodelan yang telah dilakukan. Hasil visualisasi dari pemodelan topik dapat dilihat pada gambar 3.



Gambar 3. Pengelompokan berdasarkan topik

Berdasarkan analisis dari visualisasi, terdapat 84 topik yang muncul dari seluruh ayat Al-Quran yang berjumlah 6236 ayat. Peta jarak intertopik merupakan salah satu alat visualisasi yang dapat digunakan untuk mengevaluasi hasil pemodelan topik. Peta jarak intertopik menampilkan informasi tentang bagaimana topik-topik yang dihasilkan oleh algoritma pemodelan topik saling terkait satu sama lain. Setiap topik ditampilkan sebagai sebuah titik pada peta jarak intertopik dan jarak antar titik menunjukkan tingkat keterkaitan antar topik tersebut. Semakin dekat titik-titik tersebut, semakin terkait topik-topik tersebut. Sebaliknya, semakin jauh titik-titik tersebut, semakin tidak terkait topik-topik tersebut. Peta jarak intertopik yang diterapkan pada hasil pemodelan topik Al-quran memberikan gambaran yang lebih jelas tentang topik-topik yang terdapat dalam Al-Quran dan bagaimana topik-topik tersebut saling terkait satu sama lain, dengan melihat peta jarak intertopik kita dapat mengevaluasi hasil pemodelan topik dan memahami struktur topik yang terdapat dalam Al-quran.

Skor kata topik adalah salah satu metrik yang dapat digunakan untuk mengevaluasi hasil pemodelan topik. Metrik ini memberikan informasi tentang seberapa penting kata-kata tertentu dalam mewakili sebuah topik yang dihasilkan oleh algoritma pemodelan topik. Skor kata topik dihitung dengan menggunakan rumus yang mempertimbangkan jumlah kemunculan kata-kata tertentu dalam topik tersebut dan jumlah kemunculan kata-kata tersebut dalam seluruh dokumen yang telah dianalisis. Semakin tinggi skor kata topik sebuah kata, maka semakin penting kata tersebut dalam mewakili sebuah topik. Dengan demikian, skor kata topik dapat membantu kita dalam mengevaluasi hasil pemodelan topik dan memahami kata-kata yang paling banyak mewakili sebuah topik. Dalam hasil analisis ini, terdapat 8 topik utama, yaitu Al-Quran, Aku (Allah), Langit, Rasul, Malaikat, Wanita, Neraka, dan Dibangkitkan. Kata-kata tersebut dianggap sangat penting dalam mewakili topik-topik yang dihasilkan.

4. KESIMPULAN

Setelah dilakukan pemodelan topik terhadap dataset Al-Quran sebanyak 6236 ayat, ditemukan beberapa topik utama dari masing-masing topik yang dapat dilihat melalui visualisasi topik. Berdasarkan hasil implementasi BERTopic dalam proses pemodelan topik terhadap kitab suci Al-Quran. Dengan menggunakan BERTopic, kita dapat dengan mudah memodelkan topik-topik yang terdapat dalam Al-Quran dan mengekstrak informasi yang bermanfaat dari teks tersebut. Berdasarkan penelitian ini hasil pemodelan topik menghasilkan 8 topik utama yang terwakilkan dengan kata, yaitu Al-Quran, Aku (Allah), Langit, Rasul, Malaikat, Wanita, Neraka, dan Dibangkitkan. Sebanyak 8 topik utama tersebut secara terperinci memiliki prosentase sebagai berikut; Topic 0 Al Quran sebesar 6%, Topic 1 Aku (Allah) sebesar 6,5%, Topic 2 Langit 3,8%, Topic 3 Rasul 8%, Topic 4 Malaikat 12,5%, Topic 5 Wanita 5%, Topic 6 Neraka dengan prosentase 13%, serta Topic 7 Dibangkitkan sebesar 5,5%. Kata-kata tersebut dianggap sangat penting dalam mewakili topik-topik yang dihasilkan dan juga termasuk dalam kategori spiritual, moral, dan hukum.

DAFTAR PUSTAKA

[1] Achmad Chodjim, al-Fatihah. Jakarta, 2017.

[2] E. Agustina, “Kajian Referensi Ayat-Ayat Al Qurâ€TMan Dalam Skripsi Mahasiswa Pendidikan Biologi Fakultas Tarbiyah Dan Keguruan Uin Ar-Raniry,” Biot. J. Ilm. Biol. Teknol. dan Kependidikan, vol. 3, no. 1, p. 69, 2017, doi: 10.22373/biotik.v3i1.994.

[3] S. Sunarsa, “Teori Tafsir,” Al-Afkar, vol. 2, no. 1, pp. 248–260, 2019, doi: 10.5281/zenodo.2561512.

[4] M. I. Helmy, “Kesatuan Tema dalam Al-Qur’an,” Ilmu Ushuludddin, vol. 19, no. 2, 2020, doi: 10.18592/jiiu.v.

[5] F. Rahman, Tema-Tema Pokok Al Qur’an. Bandung: Mizan Media Utama, 2017.

[6] S. Nath, A. Marie, S. Ellershaw, E. Korot, and P. A. Keane, “New meaning for NLP: the trials and tribulations of natural language processing with GPT-3 in ophthalmology,” Br. J. Ophthalmol., vol. 106, no. 7, pp. 889–892, 2022, doi: 10.1136/bjophthalmol-2022-321141.

[7] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 1, pp. 4171–4186, 2019.

[8] A. Nagpal and G. Gabrani, “Python for Data Analytics, Scientific and Technical Applications,”

in 2019 Amity International Conference on Artificial Intelligence (AICAI), 2019, pp. 140–145, doi: 10.1109/AICAI.2019.8701341.

[9] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," 2022, [Online]. Available: http://arxiv.org/abs/2203.05794.

[10] T. Wolf et al., "Transformers : State-of-the-Art Natural Language Processing," pp. 38–45, 2020.

[11] M. Usama et al., "Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges," IEEE Access, vol. 7, pp. 65579–65615, 2019, doi: 10.1109/ACCESS.2019.2916648.

[12] M. Andika, L. Chaerani, K. K. Data, L. D. Allocation, M. Online, and P. Topik, "Pemodelan Topik Berita pada Portal Berita Online Berbahasa Indonesia Menggunakan Latent Dirichlet Allocation (LDA)," J. Ilm. Komputasi, vol. 20, no. 2, pp. 173–180, 2021, doi: 10.32409/jikstik.20.2.2719.

[13] W. Wahyudin, "APLIKASI TOPIC MODELING PADA PEMBERITAAN PORTAL BERITA ONLINE SELAMA MASA PSBB PERTAMA," Semin. Nas. Off. Stat., vol. 2020, no. 1, pp. 309–318, Jan. 2020, doi: 10.34123/SEMNASOFFSTAT.V2020I1.579.

[14] C. Naury, D. H. Fudholi, and A. F. Hidayatullah, "Topic Modelling pada Sentimen Terhadap Headline Berita Online Berbahasa Indonesia Menggunakan LDA dan LSTM," J. MEDIA Inform. BUDIDARMA, vol. 5, no. 1, pp. 24–33, Jan. 2021, doi: 10.30865/MIB.V5I1.2556.

[15] S.-A. N. Alexandropoulos, S. B. Kotsiantis, and M. N. Vrahatis, "Data preprocessing in predictive data mining," Knowl. Eng. Rev., vol. 34, p. e1, 2019, doi: 10.1017/S026988891800036X.

[16] F. A. Muttaqin and A. M. Bachtiar, "Implementasi Teks Mining Pada Aplikasi Pengawasan Penggunaan Internet Anak 'Dodo Kids Browser,'" J. Ilm. Komput. dan Inform., pp. 1–8, 2016.

[17] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "Topic Modeling in Embedding Spaces," Trans. Assoc. Com*put. Linguist.*, vol. 8, pp. 439–453, 2020, doi: 10.1162/tacl_a_00325.