TUGAS UJIAN AKHIR SEMESTER

MATAKULIAH DATA MINING

Dosen Pengampu

Dr. SAJARWO ANGGAI S.ST., M.T

Oleh

ASEP RIDWAN HIDAYAT

231012050036

TI 01MKME001 REGULAR C

PROGRAM STUDI MAGISTER TEKNIK INFORMATIKA
UNIVERSITAS PAMULANG
TANGERANG SELATAN
2024

PEMODELAN TOPIK TERJEMAH AL-QUR'AN BERBAHASA INDONESIA DENGAN MENGGUNAKAN BERTOPIC

1. PENDAHULUAN

Al-Quran adalah kitab suci orang muslim dengan jumlah surat 114 surat dan 6127 ayat. Sebagai kitab pedoman muslim tentunya banyak makna yang bisa diambil dari apa yang ada pada setiap ayatnya dan banyak topik yang bisa digali dari Al-Quran. Atas dasar itu penulis ingin mencoba memodelkan terjemah Al-Quan berbahasa indonesia dengan salah satu topik modeling pada *text mining* yaitu BERTopik model, selain juga sebagai salah satu pemenuhan tugas, tapi menjadi salah satu sarana untuk penerapan teori yang diajarkan pada matakuliah dataming.

BERTopic modeling adalah kerangka pemodelan topik yang memungkinkan pengguna membuat model topik versi mereka sendiri. Dengan banyaknya variasi pemodelan topik yang diterapkan, idenya adalah bahwa pemodelan tersebut harus mendukung hampir semua kasus penggunaan.

Pada pemodelan kali ini selain untuk pemenuhan tugas ujian akhir semerter, sekaligus untuk menambah pemahaman penulis terhadap terjemah ayat-ayat Al-Qur'an, juga agar menjadi referensi tambahan bagi penulis pribadi untuk penerapan topik modeling dengan model pada Al-Quran dalam terjemah Bahasa Indonesia.

2. METODOLOGI

2.1 Pengumpulan data

Dalam penelitian ini dataset yang digunakan adalah Al-Qur'an terjemah bahasa indonesia yang didapat dari quranenc.com [1].

Selanjutnya setelah pengambilan data, ada beberapa kolom data yang disesuaikan seperti hanya kolom text terjemah bahasa indonesia saja yang digunakan untuk penomoran surat dan ayat tidak diikut sertakan. Setelah pemilihan kolom text dipilih maka dilakukan preprocesing terhadap data.

Berikut contoh sebagian dataset yang didapat:

surah	ayah	text
1	1	Dengan menyebut nama Allah Yang Maha Pemurah lagi Maha Penyayang.
1	2	Segala puji bagi Allah
1	3	Maha Pemurah lagi Maha Penyayang.
1	4	Yang menguasai di Hari Pembalasan.
1	5	Hanya Engkaulah yang kami sembah
1	6	Tunjukilah kami jalan yang lurus
1	7	(yaitu) Jalan orang-orang yang telah Engkau beri nikmat kepada mereka; bukan (jalan) mereka yang dimurkai dan bukan (pula jalan) mereka yang sesat.
2	1	Alif laam miim.
2	2	Kitab (Al Quran) ini tidak ada keraguan padanya; petunjuk bagi mereka yang bertakwa

2	3	(yaitu) mereka yang beriman kepada yang ghaib
2	4	dan mereka yang beriman kepada Kitab (Al Quran) yang telah diturunkan kepadamu dan Kitab-kitab yang telah diturunkan sebelummu
2	5	Mereka itulah yang tetap mendapat petunjuk dari Tuhan mereka
2	6	Sesungguhnya orang-orang kafir
2	7	Allah telah mengunci-mati hati dan pendengaran mereka
2	8	Di antara manusia ada yang mengatakan: "Kami beriman kepada Allah dan Hari kemudian
2	12	Ingatlah
2	13	Apabila dikatakan kepada mereka: "Berimanlah kamu sebagaimana orang-orang lain telah beriman". Mereka menjawab: "Akan berimankah kami sebagaimana orang-orang yang bodoh itu telah beriman?" Ingatlah

Tabel 1.1 Contoh sebagian Dataset Al-Qur'an

Untuk pemodelan yang digunakan yaitu pemodelan BERTopik dengan menggunakan pemograman python dan tools *Google Colabs.*

2.2 *Procesing* dan *Preprocesing* Teks

2.2.1 Procesing Teks

Berikut scrypt pegolahan data dengan menggunakan Bahasa pemograman Python .

```
import pandas as pd
import numpy as np
from bertopic import BERTopic
import re,string

# 1 select data
df = pd.read_csv('alquran_terjemah_indonesian.csv')
text = df['text']

# 2. PREPROCESSING
# Drop nilai yang kosong
text = text.fillna('')

# 3. Fungsin untuk prepocessing
def clean_text(text):
    text = text.lower()  # Lower
    text = re.sub(r'\d+', '', text)  # Menghapus angka
    text = re.sub(r'\W', ' ', text)  # Menghapus karakter non-alfanumerik
    text = re.sub(r'\s+', ' ', text)  # Menghapus spasi berlebih
    # text = re(str.maketrans("","",string.punctuation))
    return text.strip()

# 4 panggil fungsi clean text dan simpan divariabel docs
```

```
docs = text.apply(clean_text)

# 5. Deklarasikan variabel untuk Membuat model BERTopic
model = BERTopic(verbose=True)

# 6.Melakukan fit dan transformasi dokumen
topics, probabilities = model.fit_transform(docs.tolist())

# 7. Fungsi untuk get topic yang dihasilkan
model.get_topic_info()
# 8 Menampilkan topik array ke 0 indeks 1
model.get_topic(0)

# 9 visualisasi data
# visualisasi topic (intertopic Distance Map)
model.visualize_topics()

# 10 model.visualize_barchart() (topic wors score)
model.visualize_barchart()

# 11 model visualisai hirarki
model.visualize_hierarchy()

# 12 Ctf Idf score
model.visualize_term_rank()
#13.  similiarity map
model.visualize_heatmap()
```

2.2.3 *Preprocesing* Teks

Proses dalam preprocessing data, antara lain sebagai berikut:

1) *Remove symbol* dan *Number*

   Langkah yang sering dilakukan sebelum pengolahan dataset 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.

   Berikut salah satu contoh script yang digunakan dalam pengolahan data:

```
text = text.lower()  # Lower
    text = re.sub(r'\d+', '', text)  # Menghapus angka
    text = re.sub(r'\W', ' ', text)  # Menghapus karakter non-
alfanumerik
    text = re.sub(r'\s+', ' ', text)  # Menghapus spasi berlebih
```

2) *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.

```
text = text.lower()  # Lower
```

3) *Tokenezing*

Pada tahap preprocessing berikutnya, terdapat tokenisasi, yaitu proses membagi teks menjadi sejumlah token atau kata-kata. Banyak library yang bisa digunakan untuk tokenizing seperti tokenizing English atau Bahasa Indonesia dengan library sastrawi.

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

```
topics, probabilities = model.fit_transform(docs.tolist())
```

Setelah dilakukan preprosesing dataset, maka dilakukan install library bert, selain dari library yang biasa digunakan dalam pemodelan dataset pada pemograman python.

3. HASIL DAN PEMBAHASAN

Setelah dilakukan pengolahan data dengan menggunakan software *Google Colab*, dan dataset seperti dijelaskan sebelumnya, selanjutnya dataset dirapihkan dan dirubah menjadi datalist, setalah data menjadi datalist data dilakukan tramsform data *Model Generation and Visualization* (Model BERTOPIC).

Berikut ini gambaran data sebelum preposesing dan setelah siap dilakukan pengolahan pada model BERTopic.

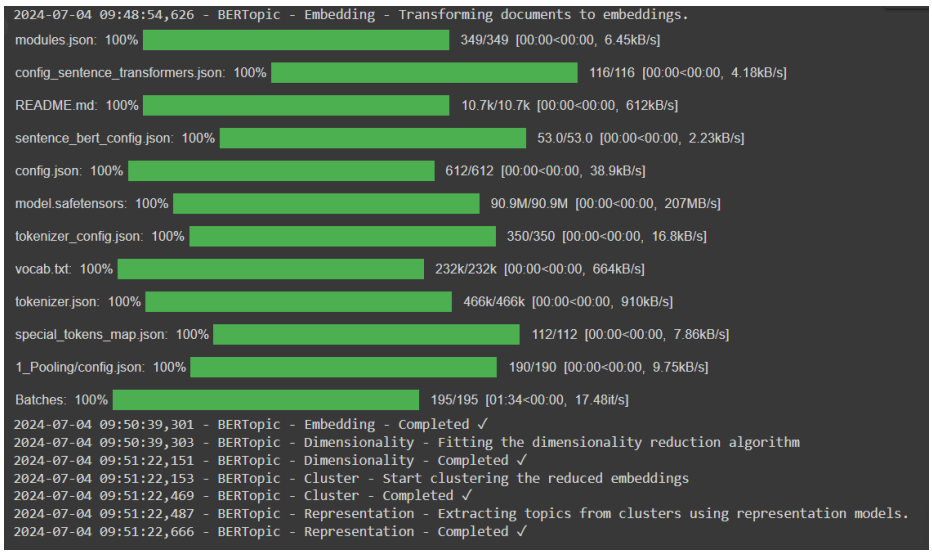
```
0    Dengan menyebut nama Allah Yang Maha Pemurah l...
1                        Segala puji bagi Allah
2                  Maha Pemurah lagi Maha Penyayang.
3                  Yang menguasai di Hari Pembalasan.
4                      Hanya Engkaulah yang kami sembah
Name: text, dtype: object
```

Gambar 1.1 dataset sebelum preprocessing dan datalist

```
['dengan menyebut nama allah yang maha pemurah lagi maha penyayang',
 'segala puji bagi allah',
 'maha pemurah lagi maha penyayang',
 'yang menguasai di hari pembalasan',
 'hanya engkaulah yang kami sembah']
```

Gambar 1.2 Dataset setelah preprocessing dan list

Setelah selesai preprocessing data, selanjutnya data akan diimplementasikan juga penyesuaian (*fit transformation*) sekaligus proses embeding pada model bahasa
Berikut proses dan output dari dataset yang dikelompokan dalam kelompok-kelompok topik.



Gambar 1.3 Proses fit dan transformasi, *embedding* dokumen BERTopic

Setelah proses fit transformasi dan *embedding* dokumen, maka akan terbentuk beberapa topik juga representasi dari pemodelan BERTopic ini. Pembentukan topik pada penelitan disini masih menggunakan *default* dari pemodelan BERTopic itu sendiri, selanjutnya bisa juaga digunakan library Bert yang lainnya agar bisa ditentukan jumlah topik yang diharapkan. Tetapi untuk kali ini pembentukan topik tidak ditentukan.

Hasil dari pembentukan topik untuk dataset Al-quran terjemah Bahasa Indonesia ini bisa dilihat dari data output dibawah ini.

	Topic	Count	Name	Representation	Representative_Docs
0	-1	1653	-1_kamu_dan_dia_yang	[kamu, dan, dia, yang, allah, mereka, tidak, i...	[sesungguhnya pada langit dan bumi benar benar...
1	0	781	0_orang_kafir_yang_beriman	[orang, kafir, yang, beriman, allah, dan, ayat...	[sesungguhnya orang orang kafir, sesungguhnya ...
2	1	700	1_telah_kami_sesungguhnya_mereka	[telah, kami, sesungguhnya, mereka, rasul, kep...	[dan sesungguhnya kami telah mengutus rasul ra...
3	2	300	2_mereka_tidak_azab_maka	[mereka, tidak, azab, maka, di, isteri, apakah...	[maka apakah mereka tidak melihat langit dan b...
4	3	211	3____	[, , , , , , , , ,]	[, ,]
...	...	...	...	...	...
69	68	12	68_sekali_kali_jangan_tidak	[sekali, kali, jangan, tidak, curang, tambah, ...	[sekali kali tidak, sekali kali tidak, sekali ...
70	69	12	69_menurunkan_angan_menurut_menurunkannya	[menurunkan, angan, menurut, menurunkannya, me...	[dan kami tidak menurunkan kepada kaumnya sesu...
71	70	11	70_alif_laam_miim_mim	[alif, laam, miim, mim, shaad, raa, , , ,]	[alif laam miim, alif laam miim, alif laam miim]
72	71	11	71_nun_shaad_demi_allah	[nun, shaad, demi, allah, dan, , , , ,]	[shaad, demi allah, demi allah]
73	72	10	72_kiranya_kemah_moga_rumah	[kiranya, kemah, moga, rumah, karun, pemimpinp...	[isa putera maryam berdoa ya tuhan kami turunk...

Gambar 1.4 Model Get Topic Informasi

Dari pembentukan informasi diatas dapat dilihat pembentukan topik pada pemodelan sebanyak 72 topik dengan jumlah pengelompokan text data, juga informasi refrensentasi datanya. selanjutnya dari pembentukan topik ini kita bisa memilih topic mana yang akan kita tampilkan.

Pada topik bernilai -1 artinya pada dokumen atau dataset terdapat *outlier* nya, dan nilai *outlier* nya secara otomatasi ada penanganan dari model bert itu sendiri baik dikurangi atau dihilangkan[5].

Dari pembentukan topik diatas untuk nilai -1 itu sebanyak 1.653, kemudian dengan menggunakan soft-clustering seperti yang dilakukan oleh HDBSCAN akan dipilih topik mana yang paling cocok untuk setiap dokumen pada seitap *outlier* nya [5].
Berikut *summary* pembentukan topik 10 teratas dari 72 topik.

	Topic	Count
1	-1	1653
6	0	781
2	1	700
5	2	300
22	3	211
4	4	200
18	5	152
19	6	143
21	7	94
34	8	91
9	9	81

Gambar 1.5 *Summary* 10 top topic

Dari top 10 topic diatas ada beberapa topik yang dapat kita pilih dari banyaknya kata yang dikelompokan, berikut ini kita coba memilih topik diatas dengan mengambil nomor topik, berikut salah satu contoh topik nomor 4 yang dipilih outputnya seperti ini.

```
[('quran', 0.11408784011390871),
 ('al', 0.10968833701583154),
 ('kitab', 0.029954547997825737),
 ('ini', 0.02225287157902865),
 ('ayat', 0.019665328372164854),
 ('menurunkan', 0.01777476886934314),
 ('turunkan', 0.017566122839953596),
 ('diturunkan', 0.016029491966925304),
 ('telah', 0.014215376037013807),
 ('benar', 0.013693632881734l9)]
```

Gambar 1.6 tampilan topik ke-4

Pada gambar 1.6 adalah salah satu hasil topik ke-4, pada output bisa diliah keterang text topik dikuti dengan nilai c-Tf-Idf nya. dari pengelompokan topik diatas terdapat kata ‘Qur’an’, ’kitab’, ’ini’,’ayat’,’menurunkan’, ’turunkan’, ’telah ’, ’benar’ dijadikan menjadi satu topik, kalau kita bisa simpulkan topik ini berfokus pada pembahasan **Al-Quran kitab benar yang diturunkan**.

Untuk memudahkan memvisualisasikan topik pemodelan diatas, terdapat juga visualisasi yang bisa memudahkan kita untuk memilih topik yang sudah terbentuk. Contoh visualisasi data bisa dilihat dari beberapa grafik dibawah ini:

Gambar 1.7 *Intertopic distance map*

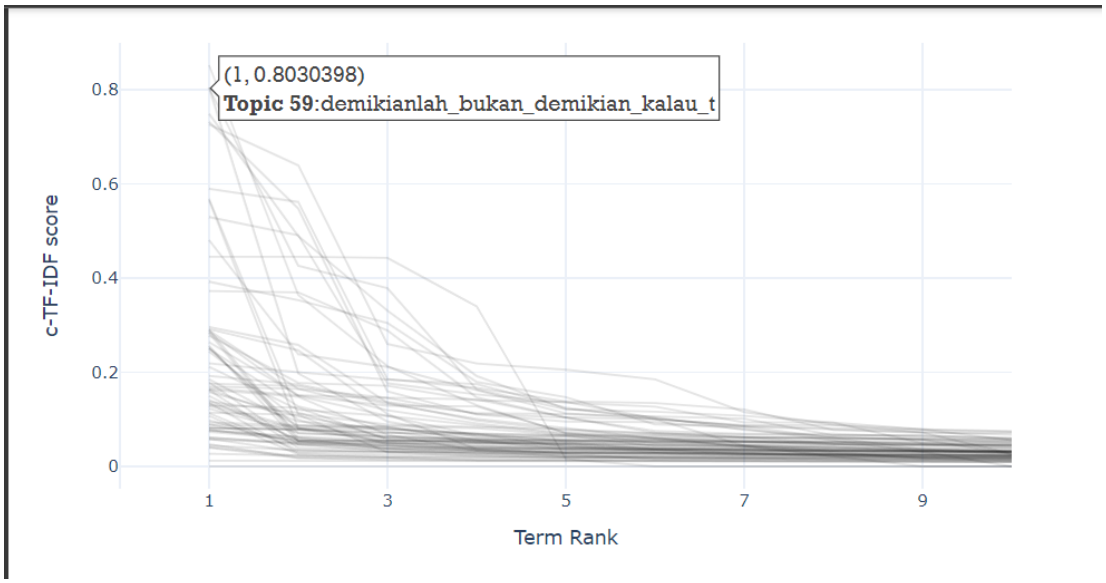
Intertopic distance map merupakan visualisasi dari pengelompokan topik pada dataset, bisa dilihat dari gambar 1.7 topik yang dibuat sebanyak 70 topik dari 72 topik yang terbentuk,

kemungkinan ada beberapa topik *outlier* yang tidak diikut sertakan pada *Intertopic distance map.*

Berikut ini visualisasi lainnya dari pembentukan topik Bert dibawah ini :

Gambar 1.8 Topic Word Score

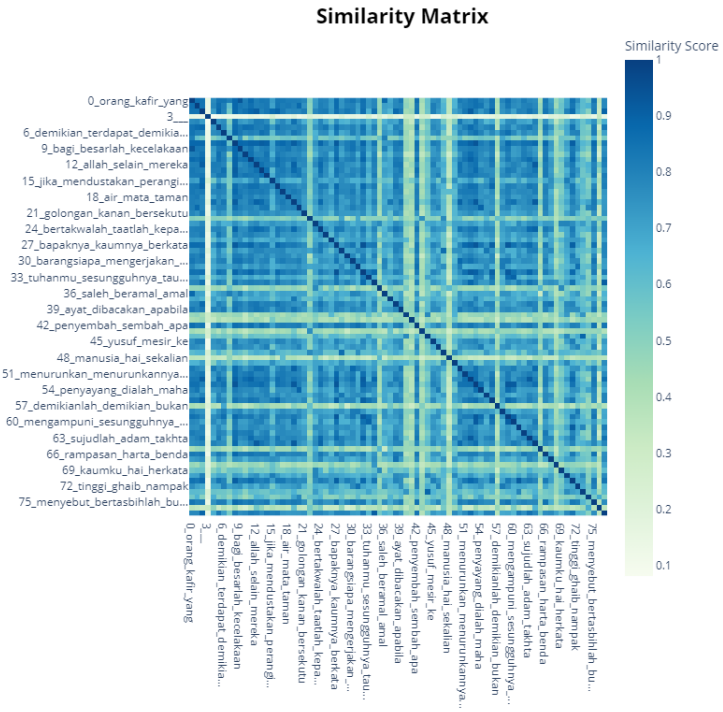
Topic Word Score merupakan visualisasi beberapa topik lengkap dengan range scorenya, dan ada 7 topik yang bisa tergambar dalam gambar 1.8 dengan nilai range berbeda-beda. Dari sini kita bisa menganalisa lebih lanjut untuk pemodelan selanjutnya, apakah ada topik yang sekiranya kita bisa sortir kembali agar terbentuk pemodelan topik yang lebih optimal.Visualisasi lainnya seperti dibawah ini.



Gambar 1.9 Term Rank

Pada grafik *Term Rank* ketika kita mengarahkan kursor pada grafik ini, akan muncul informasi yang dapat kita lihat langsung dengan nilai c-TF-IDF dan indeks topiknya.

Seperti contoh pada gambar 1.9 terlihat topic 59 dengan nilai c-TF_IDF sebesar 0.8. diagram diatas membantu kita memutuskan apakah kita dapat memangkas jumlah istilah yang ingin kita bedakan. Misalnya, kita mungkin menganggap bahwa hanya 11 istilah teratas yang cukup penting untuk kita pertimbangkan[6]. Berikut visualisasi lainnya yang ada pada BERTopik.



Gambar 1.8 *Similiarity Matrix*

Similiarity Matrix pada gambar 1.8 digunakan untuk melihat seberapa besar kemiripan setiap topik dan membantuk kita untuk mengevaluasi setiap topik.

Pada intinya beberapa visualisasi diatas memudahkan kita melihat topik yang terbentuk, dengan dibantu visualisasi diatas memudahkan kita untuk memberikan lebel pada topik yang terbentuk dari pemodelan dengan BETopik ini.

4. KESIMPULAN DAN SARAN

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. Pada pemodelan kali ini terbentuk secara otomatis oleh model Bert ini dikelompokan menjadi 72 topik, dan kita dapat dengan mudah memodelkan topik-topik yang terdapat dalam Al-Quran dan mengekstrak informasi yang bermanfaat dari teks tersebut, dengan memilih beberapa indeks topik yang bisa kita pilih.

Berdasarkan pemodelan kali ini hasil pemodelan 10 top topik utama. Salah satu contoh topik yang kita tampilkan yaitu topik ke-4 dengan pengelompokan kata, ‘Qur’an’, ’kitab’, ’ini’,’ayat’,’menurunkan’, ’turunkan’, ’telah ’, ’benar’ dari pengelompokan topik ke-4 ini bisa dilihat kecendrungan topik ini mengarah ke Qur’an kitab yang benar diturunkan.

Tentunya banyak kekurangan pada pengolahan pemodelan BERTopik di dataset kali ini, salah satunya dengan nilai outlier data cukup tinggi sebanyak 1653 artinya 20% dari data terdapat outlier, salah satu hal ini bisa ditimbulkan dari *preprocessing* yang kurang optimal, juga keterbatasannya penulis dalam pemahaman model BERTopik ini, sehingga banyak hal yang belum optimal pada pemodelan topik yang terbentuk. Sarannya perlu banyak bimbingan dan pengkajian lebih dalam untuk mengolah data topik Al-Qur'an terjamah Bahasa Indonesia, baik dari ahli Bahasa dan penafsiran Al-Quran agar bisa menyimpulkan setiap topik yang terbentuk.

5. DAFTAR PUSTAKA

[1] https://quranenc.com/id/browse/indonesian_complex/1
[2] https://maartengr.github.io/BERTopic/getting_started/manual/manual.html
[3] https://towardsdatascience.com/topics-per-class-using-bertopic-252314f2640
[5] https://maartengr.github.io/BERTopic/api/bertopic.html#bertopic._bertopic.BERTopic.reduce_outliers
[6] https://cees-roele.medium.com/a-term-score-matrix-for-bertopic-821e78e198ee
[7] https://www.kaggle.com/code/yhirakawa/bertopic-visualization-of-topic-modeling#Bertopic
[8] https://maartengr.github.io/BERTopic/getting_started/clustering/clustering.html#agglomerative-clustering