ANALISIS METODE *LATENT DIRICHLET ALLOCATION* (LDA) UNTUK *CLUSTERING* TERJEMAHAN AL-QUR'AN BAHASA INDONESIA

Skripsi

untuk memenuhi sebagai persyaratan

mencapai derajat Sarjana S-1

Program Studi Teknik Informatika

Disusun oleh:

Ahmad Septiawan

16650053

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS ISLAM NEGERI SUNAN KALIJAGA

YOGYAKARTA

2020

HALAMAN PENGESAHAN

KEMENTERIAN AGAMA
UNIVERSITAS ISLAM NEGERI SUNAN KALIJAGA
FAKULTAS SAINS DAN TEKNOLOGI
Jl. Marsda Adisucipto Telp. (0274) 540971 Fax. (0274) 519739 Yogyakarta 55281

PENGESAHAN TUGAS AKHIR

Nomor : B-1738/Un.02/DST/PP.00.9/07/2020

Tugas Akhir dengan judul : ANALISIS METODE LATENT DIRICHLET ALLOCATION (LDA) UNTUK CLUSTERING TERJEMAHAN AL-QUR'AN BAHASA INDONESIA

yang dipersiapkan dan disusun oleh:

Nama	: AHMAD SEPTIAWAN
Nomor Induk Mahasiswa	: 16650053
Telah diujikan pada	: Senin, 27 Juli 2020
Nilai ujian Tugas Akhir	: A

dinyatakan telah diterima oleh Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta

TIM UJIAN TUGAS AKHIR

Ketua Sidang/Penguji I

Muhammad Didik Rohmad Wahyudi, S.T., MT.
SIGNED

Valid ID: 5f4d8e25c5ce5

Penguji II

Dr. Agung Fatwanto, S.Si., M.Kom.
SIGNED

Valid ID: 5f4845110c7b1

Penguji III

Rahmat Hidayat, S.Kom., M.Cs.
SIGNED

Valid ID: 5f4d661185a2

Yogyakarta, 27 Juli 2020
UIN Sunan Kalijaga
Plt. Dekan Fakultas Sains dan Teknologi

Dr. Murtono, M.Si.
SIGNED

Valid ID: 5f4dab8b50b73

SURAT PERSETUJUAN

Universitas Islam Negeri Sunan Kalijaga FM-UINSK-BM-05-03/R0

SURAT PERSETUJUAN SKRIPSI/TUGAS AKHIR

Hal : Persetujuan Skripsi
Lamp :

Kepada
Yth. Dekan Fakultas Sains dan Teknologi
UIN Sunan Kalijaga Yogyakarta
di Yogyakarta

Assalamu'alaikum wr. wb.

Setelah membaca, meneliti, memberikan petunjuk dan mengoreksi serta mengadakan perbaikan seperlunya, maka kami selaku pembimbing berpendapat bahwa skripsi Saudara:

Nama : Ahmad Septiawan
NIM : 16650053
Judul Skripsi : Analisis Metode Latent Dirichlet Allocation (LDA) Untuk Clustering Terjemahan Al-Qur'an Bahasa Indonesia

sudah dapat diajukan kembali kepada Program Studi Teknik Informatika Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta sebagai salah satu syarat untuk memperoleh gelar Sarjana Strata Satu dalam Program Studi Teknik Informatika

Dengan ini kami mengharap agar skripsi/tugas akhir Saudara tersebut di atas dapat segera dimunaqsyahkan. Atas perhatiannya kami ucapkan terima kasih.

Wassalamu'alaikum wr. wb.

Yogyakarta, 16 Juli 2020
Pembimbing

M. Didik Rohmad Wahyudi, S.T., M.T.
NIP. 19760812 200901 1 015

PERNYATAAN KEASLIAN SKRIPSI

PERNYATAAN KEASLIAN SKRIPSI

Saya yang bertanda tangan dibawah ini :

Nama : Ahmad Septiawan

NIM : 16650053

Jurusan : Teknik Informatika

Fakultas : Sains dan Teknologi

Menyatakan bahwa skripsi saya yang berjudul **"Analisis Metode Latent Dirichlet Allocation (LDA) Untuk Clustering Terjemahan Al-Qur'an Bahasa Indonesia "** merupakan hasil penelitian saya sendiri, tidak terdapat pada karya yang pernah di ajukan untuk memperoleh gelar kesarjana di suatu perguruan tinggi, dan bukan plagiasi karya orang lain kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Yogyakarta, 16 Juli 2020

Ahmad Septiawan

NIM.16650053

KATA PENGANTAR

Alhamdulillahirobbil'alamiin. Segala puji syukur bagi Allah SWT yang telah melimpahkan rahmat dan hidayah-Nya sehingga penulis dapat menyelesaikan skripsi dengan judul **"Analisis Metode Latent Dirichlet Allocation untuk Clustering Terjemahan Al-Qur'an Bahasa Indonesia)"** dengan baik.

Penulis menyadari bahwa penulisan skripsi ini tidak terwujud tanpa adanya bantuan, bimbingan dan dorongan dari berbagai pihak. Oleh karena itu, dengan adanya segala kerendahan hati pada kesempatan ini penulis mengucapkan terimakasih kepada :

1. Bapak Dr. Phil. Sahiron, M.A., selaku Plt. Rektor UIN Sunan Kalijaga Yogyakarta.
2. Bapak Dr. H. Waryono, M.Ag., selaku Wakil Rektor Bidang Kemahasiswaan dan Kerjasama UIN Sunan Kalijaga Yogyakarta.
3. Ibu Dr. Khurul Wardati,M.Si., selaku Dekan Fakultas Sains dan Teknologi.
4. Ibu Maria Ulfa Siregar, S.Kom.,M.I.T, Ph.d., selaku Ketua Program Studi Teknik Informatika UIN Sunan Kalijaga Yogyakarta.
5. Bapak Muhammad Didik Rohmad Wahyudi, S.T., M.T.,selaku Dosen Pembimbing Skripsi dan juga Dosen Pembimbing Akademik yang telah sabar dan meluangkan waktunya untuk memberikan motivasi, koreksi, serta kritik saran kepada penulis sehingga skripsi ini dapat terselesaikan.
6. Bapak Ibu Dosen Program Studi Teknik Informatika UIN Sunan Kalijaga Yogyakarta yang telah memberikan banyak bekal ilmu kepada penulis.

7. Seluruh staf dan karyawan Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta.
8. Mas Ahmad Fathan Hidayatullah, S.Kom., M.Cs., Saudara dan juga tetangga dekat yang telah membimbing dan memberikan arahan dari awal masuk perkuliahan hingga tugas akhir.
9. Ucapan terimakasih yang terdalam kepada kedua orangtua tercinta, Bapak Amat Kalimi dan Ibu Kiptiyah yang senantiasa memberikan doa, kasih sayang serta dukungan yang tak terhingga.
10. Muroby Ruhina K.H Ahmad Imron Rosyidi yang sekaligus Founder dan Trainer Keluarga Besar Ruqyah Aswaja Nasional dan Internasional yang senantiasa mendoakan dan membimbing jejak langkah hidup ini.
11. Guru Besar K.H Ahmad Zabidi, Dr Ahmad Arifin Hidayatulah yang tidak henti hentinya memberikan semangat, motifasi, dan pengalamanya sehingga penulisan skripsi ini dapat terselesaikan
12. Seluruh teman-teman Teknik Informatika 2016 yang tidak dapat disebutkan satu per satu yang telah banyak memberikan bantuan, dukungan, serta motivasi dalam menuntut ilmu.
13. Teman-teman majelis Ratibul Hadad Jonggrangan Yogi Sumarno, Faqih, Ridwan, Fathurahman, Rino, Ali Akbar, Renaldi, Anas yang menambah semangat dan berkah dalam penyelesaian skripsi.
14. Serta semua pihak yang telah memberikan doa, bantuan dan dukungan selama menempuh strata satu Teknik Informatika khususnya dalam penyusunan skripsi ini yang tidak dapat penulis sebutkan satu per satu.

Penulis menyadari tentu masih banyak kekurangan dalam penulisan laporan skripsi ini, sehingga kritik serta saran dari pembaca sangat penulis harapkan. Semoga penelitian skripsi ini dapat dijadikan sebagai dasar penyempurna penelitian sebelumnya.

Yogyakarta, 28 Februari 2020

Penulis

HALAMAN PERSEMBAHAN

Skripsi ini saya persembahkan teruntuk

kedua orang tua saya (Bapak Amat Kalimi & Ibu Kiptiyah)

yang telah bekerja keras untuk mendidik dan mencukupi kehidupan keluarga dan penuh kesabaran dalam membimbing anak tunggalmu ini.

Semoga Allah Limpahkan ampunan, rahmat, dan berkah untuk kedua orang tua saya, diberikan istiqomah dalam ibadah kepada Allah hingga akhir usia diberikan khusnul khatimah.

HALAMAN MOTO

إِنَّ صَلاتِي وَنُسُكِي وَمَحْيَايَ وَمَمَاتِي لِلَّهِ رَبِّ الْعَالَمِينَ

"Sesungguhnya shalatku, ibadahku, hidupku dan matiku hanyalah untuk Allah, Tuhan semesta alam"

DAFTAR ISI

HALAMAN JUDUL i
HALAMAN PENGESAHAN ii
SURAT PERSETUJUAN iii
PERNYATAAN KEASLIAN SKRIPSI iv
KATA PENGANTAR v
HALAMAN PERSEMBAHAN viii
HALAMAN MOTO ix
DAFTAR ISI x
DAFTAR TABEL xiv
DAFTAR GAMBAR xv
DAFTAR PERSAMAAN xvii
INTISARI xviii
ABSTRACT xix
BAB I PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Rumusan Masalah 3
1.3 Batasan Masalah 3
1.4 Tujuan Penelitian 4
1.5 Manfaat Penelitian 4
1.6 Keaslian Penelitian 4
1.7 Sistematika Penulisan 5
BAB II TINJAUAN PUSTAKA DAN LANDASAN TEORI 7

2.1 Tinjauan Pustaka 7
2.2 Landasan Teori 16
2.2.1 Machine Learning 16
2.2.2 Text Mining 17
2.2.3 Topic Modeling 18
2.2.4 Latent Dirichlet Allocation (LDA) 19
2.2.5 Topic Coherence 22
2.2.6 Al-Qur'an 24
2.2.7 Python 25
2.2.8 PyLDAvis 26
BAB III METODE PENELITIAN 28
3.1 Metode Penelitian 28
3.2 Alur Penelitian 28
3.2.1 Studi Pendahuluan 29
3.2.2 Pengumpulan Data 29
3.2.3 Data Preprocessing 29
3.2.4 N Gram Model 30
3.2.5 Membentuk Input Model 30
3.2.6 Topic Modeling 31
3.2.7 Analisis Hasil 31

3.2.8 Luaran Model Topik 31
3.3 Kebutuhan Sistem 32
BAB IV HASIL DAN PEMBAHASAN 33
4.1 Pengumpulan Data 33
4.2 Pengolahan Data 34
4.2.1 Case Folding 34
4.2.2 Stopword Removing 36
4.2.3 Exploratory Data Analysis (EDA) 37
4.2.4 Wordcloud 39
4.2.5 Tokenisasi 40
4.2.6 N Gram Model 41
4.3 Analisis 41
4.3.1 Membentuk Input Model 41
4.3.2 Topic Coherence Measurement 42
4.3.3 *Topic Modeling* 43
4.4 Hasil 44
4.4.1 Analisis Hasil Pemodelan Topik LDA 44
4.4.2 Analisis Hasil Pemilihan Model Terbaik 46
4.4.3 Visualisasi Menggunakan PyLDAvis 48
4.4.4 Analisis Hasil Luaran Model Topik 48
BAB V PENUTUP 57

5.1. Kesimpulan 57
5.2. Saran 58
DAFTAR PUSTAKA 59
LAMPIRAN 61
LAMPIRAN 2 LUARAN MODELTOPIK LDA 68
CURICULUM VITAE 70

DAFTAR TABEL

Tabel 2.1 Tinjauan Pustaka 10

Tabel 4.1 Tabel Indeks 34

Tabel 4.2 Contoh Penerapan Case Folding 35

Tabel 4.3 Contoh Penerapan Stopword Removing 36

Tabel 4.4 Kualitas Model Topik (Sarah, 2019) 47

DAFTAR GAMBAR

Gambar 2.1 Ilustrasi Topic Modeling (Blei 2012) 19

Gambar 2.2 Representasi Graphical LDA Model (Blei et al., 2003) 21

Gambar 2.3 Proses Perhitungan Topic Coherence (Röder & Hinneburg, n.d.) .. 23

Gambar 2.4 Coherence Measurement Score (Mimno, Wallach, Talley, & Leenders, 2011)..... 23

Gambar 3.1 Skema Alur Penelitian 28

Gambar 4.1 Flowchart Proses Analisis..... 33

Gambar 4.2 Source Code Load Data 33

Gambar 4.3 Source Code Data Cleaning..... 35

Gambar 4.4 Source Code Exploratory Data Analysis 37

Gambar 4.5 Diagram Batang EDA..... 38

Gambar 4.6 Source Code Wordcloud..... 39

Gambar 4.7 Wordcloud 40

Gambar 4.8 Source Code Tokenisasi 40

Gambar 4.9 Source Code Membentuk Input Model 41

Gambar 4.10 Segmentasi Cv Coherence Measurement 42

Gambar 4.11 Alur Topik Model..... 43

Gambar 4.12 Gambaran Distribusi Topik dan Kata dalam Dokumen 44

Gambar 4.13 Source Code Pencarian Model Terbaik 45

Gambar 4. 14 Source Code Percobaan Loop..... 46

Gambar 4.15 Rank Hasil Perhitungan Coherence 46

Gambar 4.16 Source Code Input Model Terbaik 47

Gambar 4.17 Visualisasi PyLDAvis 48

Gambar 4.18 Visualissi Topik pertama 49

Gambar 4.19 Visualisasi Topik Kedua 50

Gambar 4.20 Visualisasi Topik Ketiga 51

Gambar 4 21 Visualisasi Topik keempat 52

Gambar 4.22 Visualisasi Topik Kelima 53

Gambar 4.23 Visualisasi Topik Keenam 54

Gambar 4.24 Visualisasi Topik ketujuh 55

Gambar 4.25 Visualisasi Tokens 55

DAFTAR PERSAMAAN

Persamaan 2.1 Inisialisasi Parameter 21

Persamaan 2.2 Distribusi Kata 21

Persamaan 2.3 Memilih Distribusi Topik..... 21

Analisis Metode *Latent Dirichlet Allocation* (LDA) untuk *Clustering*

Terjemahan Al-Qur'an Bahasa Indonesia

Ahmad Septiawan

16650053

INTISARI

Al-Qur'an adalah kitab suci umat islam yang merupakan sumber hukum utama dalam kehidupan umat. Dalam memahami Al-Quran dituntuk untuk berhati-hati, karena pembahasan suatu tema belum tentu terkumpul dalam satu surat, kadang terpisah antar surat yang lain. Maka dari itu dilakukan penelitian guna menemukan tema-tema dari keseluruhan ayat Al-Quran.

Penelitian ini bertujuan untuk mencari topik tersembunyi pada kumpulan data teks terjemahan Al-Quran Bahasa Indonesia secara otomatis guna memudahkan memahami ayat Al-Qura'an secara utuh menggunakan metode *topic modeling* yaitu *Latent Dirichlet Allocation* (LDA). Metode tersebut untuk meng-ekstrak sejumlah 6236 ayat Al-Quran lalu menemukan luaran model topik terbaik.

Hasil evaluasi *Cv coherence measurement* memberikan nilai 0.489256 dengan luaran model sebanyak 7 topik, diantaranya "hubungan manusia dengan Allah", "kebesaran Allah", "ciptaan Allah", "cobaan dan ujian manusia", "ancaman Allah", "firman Allah", "nur".

Kata kunci : *text mining, topic modeling,* LDA, terjemahan Al-Qur'an.

Analysis of the Latent Dirichlet Allocation (LDA) Method for Clustering

Indonesian Al-Quran Translations

Ahmad Septiawan

16650053

ABSTRACT

Al-Qur'an is the Muslim holy book which is the main source of law in the life of the people. In understanding the Koran, it is necessary to be careful, because the discussion of a theme may not necessarily be collected in one letter, sometimes separate from other letters. Therefore research is carried out to find themes from all verses of the Koran.

This study aims to find hidden topics in the Indonesian Koran text translation data collection automatically in order to make it easier to understand the verses of Al-Qura'an in full using the topic modeling method that is Latent Dirichlet Allocation (LDA). The method is to extract 6236 verses from the Koran and then find the best model topic outputs.

Cv coherence measurement evaluation results give a value of 0.489256 with a model output of 7 topics, including "*hubungan manusia dengan Allah*", "*kekuasaan Allah*", "*ciptaan Allah*", "*cobaan dan ujian manusia*", "*ancaman Allah*", "*firman Allah*", "*Nur*".

Keywords: text mining, topic modeling, LDA, Al-Qur'an translation.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Islam adalah agama terbesar kedua di dunia ini setelah Kristen menurut data dari *Association Religion Data Archives (ARDA)* yang menyatakan bahwa terdapat 22,8% penganut Islam di dunia ini dari total populasi manusia di bumi (Johnson, T. M., & Grim, 2015). Menurut hasil sensus penduduk Indonesia 2010, 87% dari 237.641.326 penduduk Indonesia adalah pemeluk Islam, 6,96% Kristen, 2,9% Katolik, 1,69% Hindu, 0,72% Buddha, 0,05% Konghucu, 0,13% agama lainya, dan 0,38% tidak terjawab (Statistics Indonesia, 2010). Menurut data tersebut membuat populasi muslim terbesar adalah Indonesia dengan total 12,7% dari populasi Muslim di dunia (Pew Research Center, 2015).

Al-Qur'an adalah salah satu pedoman hidup bagi Muslim yang di dalamnya terdiri dari 114 surat, 6.236 ayat, dan 77.477 kata yang disusun menggunakan Bahasa Arab (Osman dkk, 2016). Al-Qur'an adalah petunjuk utama bagi umat Muslim yang didalamnya terdapat aturan hukum yang mengatur semua aspek kehidupan manusia, difirmankan oleh Allah SWT dalam QS.Al-Baqarah ayat 2 yang berbunyi :

ذٰلِكَ ٱلْكِتَٰبُ لَا رَيْبَ ۛ فِيهِ ۛ هُدًى لِّلْمُتَّقِينَ

"Kitab (Al-Qur'an) ini tidak ada keraguan padanya petunjuk bagi mereka yang bertaqwa." (QS.Al-Baqarah:2).

Dalam mempelajari Al-Qur'an, seringkali hanya dilakukan dengan memahami setiap ayat pada surat tertentu tanpa memperhatikan ayat-ayat pada surat yang lainya, padahal didalam Al-Qur'an sendiri belum tentu masalah dikemukakan secara urut dalam kumpulan ayat satu surat aja. Sebab dalam pembahasan masalah dalam Al-Qur'an terpencar satu dengan yang lainya, tidak dalam satu bahasan (Choiruddin,2015).

Dewasa ini sering kali kita jumpai dalam berita di media sosial, terjadinya kekisruhan dan kekerasan dengan dalih Al-Qur'an, mereka membenarkan apa yang mereka lakukan dan memakai ayat Al-Qur'an dengan hawa nafsunya saja. Padahal sangat berbahaya jika kita hanya berpegang kepada satu ayat saja tanpa memprtimbangkan ayat-ayat yang lain. Maka pentngnya kita belajar memahami Al-Qur'an secara menyeluruh dan mengetahui tema yang terkandung dalam Al-Qur'an secara menyeluruh. Mengetahui adanya hubungan antara ayat-ayat dan surat-surat dapat membantu untuk memahami dengan tepat ayat-ayat dan surat-surat yang bersangkutan (Yusuf, 2012). Hal tersebut menyebabkan diperlukanya penelitian mengenai pengelompokan ayat Al-Qur'an berdasarkan topiknya, agar lebih mudah dalam memahami Al-Qur'an.

Salah satu cara untuk menemukan topik dalam Al-Quran adalah dengan menggunakan *Topic Modeling. Topic Modeling* adalah kumpulan algoritma yang digunakan untuk menemukan struktur tersembunyi dari tema yang terdapat dalam setiap dokumen (Blei dkk, 2003).

Pada penelitian kali ini, penulis akan melakukan melakukan *Topic Modeling* terjemahan Al-Quran dari Kementrian Agama Versi 02.07.2018 menggunakan

salah satu algoritma *Topic Modeling* yaitu *Latent Dirichlet Allocation.* Hasil penelitian ini diharapkan menemukan topik dan jumlah topik dengan nilai koherensi yang paling tinggi sehingga memudahkan orang awam dalam mempelajari ayat demi ayat melalui topik yang sama.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan, maka permasalahan yang ada adalah bagaimana analisis algoritma *Latent Dirichlet Allocation* dalam melakukan clustering terjemahan Al-Qur'an Bahasa Indonesia versi Kemenag 02.07.2018 dan menemukan luaran model topik yang terbaik.

1.3 Batasan Masalah

Dalam penelitian ini terdapat beberapa batasan masalah yang dibahas agar penyusunan dan pembahasan penelitian dapat dilakukan secara terarah dan tercapai sesuai dengan yang diharapkan. Antara lain sebagai berikut:

1. Penerapan *text mining* menggunakan algoritma *Latent Dirichlet Allocation* dalam melakukan pemodelan topik terjemahan Al-Qur'an Bahasa Indonesia.
2. Data yang akan digunakan adalah 6236 ayat terjemahan Al-Quran Kementrian Agama.
3. Bahasa pemrograman yang akan digunakan adalah bahasa pemrograman *Python* dengan berbagai macam *library* di dalamnya.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah melakukan analisis Metode *Latent Dirichlet Allocation* untuk clustering terjemahan Al-Qur'an Bahasa Indonesia versi kemenag 02.07.2018 dan menemukan luaran model topik terbaik.

1.5 Manfaat Penelitian

Berdasarkan latar belakang dan tujuan di atas, adapuun manfaat dari penelitian ini adalah sebagai berikut:

1. Penelitian ini dapat menambah wawasan dan pengetahuan khususnya di bidang *text mining.*
2. Mengetahui pola topik abstrak pada terjemahan ayat Al-Qur'an Bahasa Indonesia.
3. Mengetahui seberapa jauh permodelan topik menggunakan *Latent Dirichlet Allocation* dapat diterapkan pada text Al-Qur'an.
4. Memberikan manfaat kepada umat Islam untuk menemukan topik topik dalam Al-Qur'an.

1.6 Keaslian Penelitian

Penelitian menggunakan *text mining* sudah banyak dilakukan. Namun, penelitian Analisis Metode *Latent Dirichlet Allocation* (LDA) untuk clustering Terjemahan Al-Qur'an Bahasa Indonesia, yang diajukan sebagai Tugas Akhir strata S1 pada Program Studi Teknik Informatika Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta ini belum pernah dilakukan. Hal ini diketahui berdasarkan dari referensi dan tinjauan pustaka yang dilakukan oleh peneliti

sebelumnya.

1.7 Sistematika Penulisan

Sebagai gambaran dan kerangka yang jelas mengenai pokok bahasan setiap bab dalam penelitian ini, maka diperlukan sistematika penulisan. Penyusunan laporan tugas akhir ini memiliki sistematika penulisan yang diawali dari BAB I dan diakhiri BAB V. Berikut adalah penjelasan pada tiap-tiap bab dalam laporan penelitian ini:

BAB I PENDAHULUAN

Bab pendahuluan berisikan penjelasan mengenai latar belakang dilakukannya penelitian, rumusan masalah penelitian, batasan masalah, tujuan penelitian, manfaat penelitian, keaslian penelitian, dan sistematika penulisan penelitian.

BAB II TINJAUAN PUSTAKA DAN LANDASAN TEORI

Bab tinjauan pustaka dan landasan teori berisikan mengenai penelitian terdahulu dan teori-teori dasar yang terkait dengan penelitian ini. Teori yang digunakan terdiri dari *text mining, topic modeling*, metode *Latent Dirichlet Allocation, text preprocessing, N Gram, Python, dan PyLDAvis*.

BAB III METODE PENELITIAN

Bab metode penelitian berisi tentang penjelasan mengenai metode ataupun algoritma yang digunakan serta tahapan-tahapan yang dilakukan untuk mencapai tujuan dan kesimpulan tugas akhir.

BAB IV HASIL DAN PEMBAHASAN

Bab hasil dan pembahasan membahas analisis data dan hasil dari penelitian yang telah dilakukan.

BAB V PENUTUP

Bab penutup berisi tentang kesimpulan dari hasil penelitian yang telah dilakukan. Selanjutnya, kekurangan yang ada pada penelitian dituliskan pada saran untuk pengembangan penelitian di masa yang akan datang.

BAB V

PENUTUP

5.1 Kesimpulan

Berdasarkan penelitian yang telah dilakukan dapat disimplkan bahwa algoritma *Latent Dirichlet Allocation (LDA)* dapat melakukan pemodelan topik terjemahan Al-Qur'an Bahasa Indonesia. Pada penelitian ini dilakukan percobaan range topik 1-10, untuk alpha dan beta range 0.01-1 lalu dicari koheren palig tinggi. Percobaan yang telah dilakukan menghasilkan nilai koheren paling tinggi yaitu 0.4892566080934023 dengan topik sebanyak 7, alpha 0.01 dan beta 1.

Kesimpulan topik dari ketujuh topik tersebut, yaitu topik 1 membahas hubungan manusia dengan Allah, topik 2 membahas tentang kebesaran Allah, topik 3 membahas tentang makhluk ciptaan Allah, topik 4 membahas tentang ujian dan cobaan manusia, topik 5 membahas tentang ancaman dan siksa Allah, topik 6 membahas tentang firman Allah secara umum, dan topik 7 membahas tentang nur atau cahaya.

Hasil topik tersebut diambil berdasarkan data statistic kata-kata penyusun yang sering muncul saja, walaupun memiliki nilai koheren yang cukup tinggi, tetapi dalam pencarian topik text terjemahan Al-Qur'an ini belum bisa dijadikan sumber hukum dan masih perlu dikaji oleh ahli tafsir, karena didalam ayat Al-Qur'an ada ayat muhkam yang dapat dipahami secara gamblang tanpa perlu takwil dan ada juga ayat mutasyabih yang dapat dipahami dengan penakwilan.

5.2 Saran

Pada penelitian ini masih banyak sekali kekurangan. Maka dari itu penulis menyarankan beberapa hal untuk penelitian selanjutnya, diantaranya:

1. Penelitian selanjutnya untuk mengkombinasi algoritma Topik modeling yang lain.
2. Melakukan validasi dan pertimbangan kepada ahli tafsir.
3. Memakai data teks tafsir Al-Qur'an yang lebih komplek dan sudah dijabarkan ulama.
4. Menggunakan versi terjemahan dalam Bahasa yang lain.
5. Menggunakan versi terjemahan yang terbaru.
6. Memakai library yang lain untuk memperoleh hasil yang lebih baik.

DAFTAR PUSTAKA

Ahmad Wahid Faizin. (2018). *Implementasi K-MEANS Clustering pada Terjemahan Al-Qur'an Berdasarkan Keterkaitan Topik*.

Al-augby, S. H. M. (2020). *LSA & LDA Topic Modeling Classification : Comparison study on E-books LSA & LDA Topic Modeling Classification : Comparison study on E-books*. (January). https://doi.org/10.11591/ijeecs.v19.i1.pp

Alphaidin. (2015). *Introduction to Machine Learning Second Edition*.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). *Latent Dirichlet Allocation*. *3*, 993–1022.

Douven, I., & Meijs, W. (2007). Measuring coherence. Retrieved from https://doi.org/10.1007/s11229-006-9131-z

Institute, M. (2018). Pengenalan terhadap Machine Learning. Retrieved from https://medium.com/@makersinstitute/pengenalan-terhadap-machine-learning-9011fe71d1e4

Johnson, T. M., & Grim, B. J. (2015). Religious Adherents. Retrieved from http://www.thearda.com/internationalData/countries/Country_109_2.asp

Juairiah omar. (2017). *Kegunaan Terjemahan Qur'an Bagi Umat Muslim*.

Kurniawan, W., & others. (2018). *SISTEM MONITORING PERCAKAPAN PADA TOKO ONLINE MENGGUNAKAN METODE LATENT DIRICHLET ALLOCATION (LDA) STUDI KASUS: TOKO ONLINE "BERRYBENKA.*

COM."

Kusnanta, I. M., Putra, B., & Kusumawardani, P. (2017). *Analisis Topik Informasi Publik Media Sosial di Surabaya Menggunakan Pemodelan Latent Dirichlet Allocation (LDA)*. *6*(2), 4–9.

Mimno, D., Wallach, H. M., Talley, E., & Leenders, M. (2011). *Optimizing Semantic Coherence in Topic Models*. (2), 262–272.

Osman, M., Hegazi, A., Hilal, A., & Alhawarat, M. (2016). *Fine-Grained Quran Dataset*. (January). https://doi.org/10.14569/IJACSA.2015.061241

Pew Research Center. (2015). The Future of the Global Muslim Population.

Python – The new generation Language - GeeksforGeeks. (2020). GeeksforGeeks. Retrieved from https://www.geeksforgeeks.org/python-the-new-generation-language/

Röder, M., & Hinneburg, A. (n.d.). *Exploring the Space of Topic Coherence Measures*.

Shafiei, M. M. (2009). *LEVERAGING STRUCTURAL INFORMATION FOR STATISTICAL Faculty of Computer Science*. (August).

Statistics Indonesia. (2010). *Results of 2010 Population Census*. Retrieved from https://sp2010.bps.go.id/

Vijayarani, S., Ilamathi, M. J., & Nithya, M. (n.d.). *Preprocessing Techniques for Text Mining - An Overview*. *5*(1), 7–16.

LAMPIRAN

```
#import package
import numpy as np
import pandas as pd
from IPython.display import display
import tqdm
from collections import Counter
import ast
import matplotlib.pyplot as plt
import matplotlib.mlab as mlab
import seaborn as sb


from sklearn.feature_extraction.text import
CountVectorizer
from textblob import TextBlob
import scipy.stats as stats
import os
from sklearn.decomposition import TruncatedSVD
from sklearn.decomposition import
LatentDirichletAllocation
from sklearn.manifold import TSNE
from bokeh.plotting import figure, output_file, show
from bokeh.models import Label
from bokeh.io import output_notebook
output_notebook()


%matplotlib inline
%matplotlib inline
```

```
#load data
data=pd.read_csv('Indonesian-Bahasa-Indonesia-68.csv')
len(data)


#data cleaning
import re
# Remove punctuation
data['AyahText'] = data['AyahText'].map(lambda x:
re.sub('[,\.!?]():&;', '', x))
# Convert the titles to lowercase
data['AyahText'] = data['AyahText'].map(lambda x:
x.lower())
# Print out the first rows of papers
data['AyahText'].head()


# Import the wordcloud library
from wordcloud import WordCloud
# Join the different processed titles together.
long_string = ','.join(list(data['AyahText'].values))
# Create a WordCloud object
wordcloud = WordCloud(background_color="white",
max_words=5000, contour_width=3,
contour_color='steelblue')
# Generate a word cloud
wordcloud.generate(long_string)
# Visualize the word cloud
```

```
# EDA
# Load the library with the CountVectorizer method
from sklearn.feature_extraction.text import CountVectorizer
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
sns.set_style('whitegrid')
%matplotlib inline
# Helper function
def plot_10_most_common_words(count_data, count_vectorizer):
    import matplotlib.pyplot as plt
    words = count_vectorizer.get_feature_names()
    total_counts = np.zeros(len(words))
    for t in count_data:
        total_counts+=t.toarray()[0]
    count_dict = (zip(words, total_counts))
    count_dict = sorted(count_dict, key=lambda x:x[1], reverse=True)[0:10]
    words = [w[0] for w in count_dict]
    counts = [w[1] for w in count_dict]
    x_pos = np.arange(len(words))
plt.figure(2, figsize=(15, 15/1.6180))
    plt.subplot(title='10 most common words')
    sns.set_context("notebook", font_scale=1.25, rc={"lines.linewidth": 2.5})
    sns.barplot(x_pos, counts, palette='husl')
    plt.xticks(x_pos, words, rotation=90)
    plt.xlabel('words')
    plt.ylabel('counts')
    plt.savefig('top10.png',idi=300)
    plt.show()
```

```
# Initialise the count vectorizer with the English stop words
count_vectorizer = CountVectorizer(stop_words='english')
# Fit and transform the processed titles
count_data = count_vectorizer.fit_transform(data['AyahText'])
# Visualise the 10 most common words
plot_10_most_common_words(count_data, count_vectorizer)


%%time
import gensim
from gensim.utils import simple_preprocess
def sent_to_words(sentences):
    for sentence in sentences:
        yield(gensim.utils.simple_preprocess(str(sentence), deacc=True))  # 
deacc=True removes punctuations
data = data.AyahText.values.tolist()
data_words = list(sent_to_words(data))
print(data_words[:1])


# Build the bigram and trigram models
bigram = gensim.models.Phrases(data_words, min_count=5, threshold=100) # 
higher threshold fewer phrases.
trigram = gensim.models.Phrases(bigram[data_words], threshold=100)
# Faster way to get a sentence clubbed as a trigram/bigram
bigram_mod = gensim.models.phrases.Phraser(bigram)
trigram_mod = gensim.models.phrases.Phraser(trigram)
 #fungsi bigram dan trigram
def make_bigrams(texts):
    return [bigram_mod[doc] for doc in texts]
def make_trigrams(texts):
    return [trigram_mod[bigram_mod[doc]] for doc in texts]
```

```
# Form Bigrams
import spacy
data_words_bigrams = make_bigrams(data_words)
print(data_words_bigrams[:1])


import gensim.corpora as corpora
# Create Dictionary
id2word = corpora.Dictionary(data_words_bigrams)
# Create Corpus
texts = data_words_bigrams
# Term Document Frequency
corpus = [id2word.doc2bow(text) for text in texts]
# View
print(corpus[:1])
#opsional
# supporting function
def compute_coherence_values(corpus, dictionary, k, a, b):
  lda_model = gensim.models.LdaMulticore(corpus=corpus,
        id2word=id2word,
        num_topics=10,
        random_state=100,
        chunksize=100,
        passes=10,
        alpha=a,
        beta=b,
        per_word_topics=True)
        coherence_model_lda =
CoherenceModel(model=lda_model, texts=data_words_bigrams,
dictionary=id2word, coherence='c_v')
    return coherence_model_lda.get_coherence()
```

```
#dipakai
import numpy as np
import tqdm
grid = {}
grid['Validation_Set'] = {}
# Topics range
min_topics = 2
max_topics = 11
step_size = 1
topics_range = range(min_topics, max_topics, step_size)
# Alpha parameter
alpha = list(np.arange(0.01, 1, 0.3))
alpha.append('symmetric')
alpha.append('asymmetric')
# Beta parameter
beta = list(np.arange(0.01, 1, 0.3))
beta.append('symmetric')
hasil = {'Topics': [],'Alpha': [],'Beta': [],'Coherence':
[]}
for k in topics_range:
    for a in alpha:
        for b in beta:
            cv = compute_coherence_values(corpus=corpus,
dictionary=id2word,
k=k, a=a, b=b)
            hasil['Topics'].append(k)
            hasil['Alpha'].append(a)
            hasil['Beta'].append(b)
            hasil['Coherence'].append(cv)
pd.DataFrame(hasil).to_csv('percobaanallloop.csv',
index=False)
```

```
allmodels=pd.read_csv('percobaanallloop.csv')

allmodels.sort_values('Coherence',ascending=False)

#model terbaik

lda_model_final =
gensim.models.LdaMulticore(corpus=corpus,
id2word=id2word,
num_topics=7,
random_state=100,
                                        chunksize=100,
                                         passes=10,
                                         alpha=0.01,
                                         beta=1)

from pprint import pprint

# Print the Keyword in the 7 topics

pprint(lda_model_final.print_topics())

doc_lda = lda_model_final[corpus]

from gensim.models import CoherenceModel

# Compute Coherence Score

coherence_model_lda =
CoherenceModel(model=lda_model_final,
texts=data_words_bigrams, dictionary=id2word,
coherence='c_v')

coherence_lda = coherence_model_lda.get_coherence()

print('\nCoherence Score: ', coherence_lda)

#Pyldavis

import pyLDAvis.gensim

import pickle

import pyLDAvis

# Visualize the topics

pyLDAvis.enable_notebook()

LDAvis_prepared = pyLDAvis.gensim.prepare(lda_model_final,
corpus, id2word)

LDAvis_prepared
```

LAMPIRAN 2 LUARAN MODELTOPIK LDA

Topic: 0	'0.097*"orang" + 0.027*"allah" + 0.015*"hari" + 0.012*"benar" + ' '0.010*"kafir" + 0.009*"beriman" + 0.008*"manusia" + 0.007*"azab" + ' '0.007*"rasul" + 0.006*"datang"'
Topic: 1	'0.029*"maha" + 0.027*"allah" + 0.018*"tuhan" + 0.013*"bumi" + ' '0.013*"langit" + 0.010*"manusia" + 0.010*"malaikat" + 0.010*"benar" + ' '0.009*"dialah" + 0.009*"tuhanmu"'
Topic: 2	[(0, '0.012*"air" + 0.006*"surga" + 0.006*"mata" + 0.005*"gunung_gunung" + ' '0.004*"buah" + 0.004*"minum" + 0.004*"kedua" + 0.003*"binatang" + ' '0.003*"panas" + 0.003*"pohon"'
Topic: 3	'0.007*"wanita" + 0.005*"buhul" + 0.003*"tukang_sihir" + 0.003*"kejahatan" + ' '0.003*"menghembus" + 0.002*"pergilah" + 0.002*"kesulitan" + 0.002*"penuh" + ' '0.002*"bangunan" + 0.002*"nasehat"'
Topic: 4	'0.026*"neraka" + 0.012*"api" + 0.007*"jahannam" + 0.006*"syaitan" + ' '0.005*"menguasai" + 0.005*"berlindung" + 0.004*"masuk" + 0.004*"kayu" + ' '0.003*"tempat" + 0.003*"menyala_nyala"'
Topic: 5	'0.017*"ayat" + 0.003*"makanan" + 0.003*"subuh" + 0.002*"dibacakan" + ' '0.002*"beban" + 0.002*"al_quran" + 0.002*"ditutup" + 0.002*"mendaki" + ' '0.002*"daun" + 0.001*"sukar"'

Topic: 6	'0.004*"bulan" + 0.003*"kanan" + 0.002*"tiang" + 0.002*"digoncangkan" + ' '0.002*"cahayanya" + 0.002*"kemudahan" + 0.001*"muda" + 0.001*"mengira" + ' '0.001*"rencana" + 0.001*"seri"'

CURICULUM VITAE

A. Biodata Diri

Nama Lengkap : Ahmad Septiawan

Jenis Kelamin : Laki-Laki

Tempat, Tanggal Lahir : Sleman, 17 September 1997

Alamat Asal : Jonggrangan, Sumberadi, Mlati, Sleman, DIY

Alamat Tinggal : Jonggrangan, Sumberadi, Mlati, Sleman, DIY

E-mail : septiawan.ahmad96@mail.com

B. Latar Belakang Pendidikan Formal

Jenjang	Nama Sekolah	Tahun
TK	TK ABA Sleman	2003 – 2004
SD	SDN Sleman 1	2004 – 2010
SMP	SMP NEGERI 3 Sleman	2010 – 2013
SMA	SMA NEGERI 1 GODEAN	2013 – 2016
S1	UIN SUNAN KALIJAGA YOGYAKARTA	2016 – 2020