IMPLEMENTASI TOPIC MODELING DAN TRENDING TOPIC MENGGUNAKAN GENSIM PADA PHILOIT

PROPOSAL PENELITIAN

Diajukan untuk Memenuhi Persyaratan Akademik dalam

Menyelesaikan Pendidikan pada Program Studi

S1 Teknik Informatika Universitas Kristen Maranatha

Oleh

Edward – 1872002

Michael Sebastian – 1872005

Juan David - 1872008

Rolando Vieri – 1872010

Anthony Halim - 1872027

PROGRAM STUDI S1 TEKNIK INFORMATIKA

FAKULTAS TEKNOLOGI INFORMASI

UNIVERSITAS KRISTEN MARANATHA

BANDUNG

2021

DAFTAR ISI

DAFTAR GAMBAR

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Trending topic adalah satu hal yang sedang hangat dibicarakan pada saat itu Kami memilih kasus topic modeling karena dengan adanya topic modeling, data yang memiliki kesamaan dapat dikelompokkan menjadi topik. Topik merupakan kumpulan kata kunci yang terdapat dalam dokumen. Dengan topik, kita dapat mengetahui garis besar dari sekumpulan data yang ada dengan lebih mudah. Apabila modeling topik berhasil diimplementasikan, kita bisa dengan mudah menentukan trending topic. Dari sisi pengelola website, topik akan mempermudah dalam mengekstraksi data juga meningkatkan traffic pengunjung.Sedangkan dari sisi pengguna, dengan adanya topik, pengguna dapat mencari kumpulan informasi tentang suatu hal dengan mudah. Trending topik akan memudahkan pengguna untuk mengetahui trend yang terjadi pada masa itu. Lalu dengan adanya topik juga pengguna dapat mendapat rekomendasi Q&A dan writing. Berdasarkan hal ini, penulis mengajukan proposal "Implementasi Topic Modeling dan Trending Topic menggunakan Gensim pada PhiloIT".

1.2 Rumusan Masalah

Berdasarkan latar belakang yang sudah dipaparkan, didapatkan beberapa rumusah masalah yaitu:

1. Apa saja data yang diperlukan untuk melakukan topic modeling?
2. Bagaimana cara menerapkan topic modeling?
3. Bagaimana menemukan trending topic secara efektif?

1.3 Tujuan Pembahasan

Tujuan pembahasan yang didapat berdasarkan rumusan masalah yang dibuat yaitu:

1. Menentukan data yang diperlukan untuk memodelkan topic.

2. Melakukan *topc modeling* menggunakan clustering untuk mengelompokkan data berdasarkan kesamaan-kesamaan
3. Menerapkan algoritma pada data kelompok clustering untuk memperoleh trending topic

1.4 Ruang Lingkup dan Batasan Penelitian

Dalam pembahasan ini yang menjadi batasan kami bagaimana cara menemukan trending topic secara karena setelah kami riset di situs philoit.id. Adapun batasan yang dimiliki yaitu:

1. Data yang digunakan untuk *topic modeling* yaitu *questions* dan *answers* yang diambil dari situs PhiloIT.
2. Penggunaan *Python 3.8* untuk pemodelan topik, *PHP* versi 7.2.34, dan *MySQL* 10.4.14 untuk database penyimpanan data.
3. Pemodelan topik menggunakan jumlah cluster atau topik sebanyak 20.

BAB 2

KAJIAN LITERATUR

2.1 Introduksi *Topic Modeling* Menggunakan *Python*

Saxton (2018) [1], *A Gentle Introduction to Topic Modeling Using Python*, berisi tentang pegenalan tahapan-tahapan dalam melakukan *topic modeling* menggunakan Python. Topic modeling adalah sebuah metode *data mining* yang digunakan untuk mempelajari dan mengkategorikan data yang sangat besar. Topic modeling dapat dilakukan menggunakan bahasa pemrograman *Python* untuk pemula yang belum mempunyai pemahaman dasar akan topik. Sebuah model topik akan memodelkan sebuah *corpus of document*. Asumsi dalam membuat model topik yaitu topik itu terpendam dalam dokumen. Jadi, topik dibuat oleh kata-kata yang memiliki semantik yang mirip. Kata-kata digabungkan akan menjadi topik, dan topik-topik bila digabungkan akan menjadi dokumen.

Pada artikel ini, *corpus* yang digunakan untuk topik model yaitu *Theological Librarianship (TL)* yang relatif kecil (dibawah 350 artikel). Pada bahasa pemrograman *python*, Genism adalah *topic modeling package* yang sedang populer. Genism adalah sebuah *tool* yang melakukan *computational work* untuk mengkonstruksikan model, yang hanya diperlukan yaitu sebuah *corpus document*. Tahap pertama ***create a document corpus***, apabila sumber data belum memiliki format *corpus document*, maka harus dikonversi terlebih dahulu.

Tahap kedua dilakukan ***preprocess the text***, untuk melakukan *cleaning up* seperti normalisasi *misspelled words*, mengubah semua huruf menjadi *lowercase*, membuang "*stop words" ("a", "the", "in", "because", "and").* Lemmatization yaitu proses mengurangi kata menjadi bentuk *lexical*, seperti mengubah kata *"reduce", "reducing",* dan *"reduced"* menjadi bentuk *lexical* yaitu "*reduce"*. Lalu tahap yang terakhir yaitu *tokenization*, yaitu proses memecah teks polos menjadi unit individual pada level kata.

The unprocessed plain text looks like this:

"˜ousands of volumes were contributed to the library by retiring pastors and seminary professors."

Account for errors from text extraction:

"thousands of volumes were contributed to the library by retiring pastors and seminary professors."

Remove stop words:

"thousands volumes contributed library retiring pastors seminary professors."

Lemmatize:

"thousand volume contribute library retire pastor seminary professor."

Tokenize:

["thousand", "volume", "contribute", "library", "retire", "pastor", "seminary", "professor"]

Tahap ketiga yaitu, ***Process the Text into a Gensim Corpus*** yaitu dengan membuat "*dictionary",* yang *primary function* dari *dictionary* itu adalah memberikan integer ID untuk setiap kata *unique* dan kemudian di-*map* dengan kata tersebut. Kata yang terlalu sering digunakan tidaklah informatif karena tidak bisa membedakan antar dokumen, juga kata yang terlalu sedikit tidak informatif. Jadi kata yang digunakan lebih dari 50% dari seluruh *corpus* dan kata yang kurang muncul kurang dari 5 document akan dihapus.

Tahap keempat yaitu, ***Initialize the Topic Model***, Gensim menyediakan beberapa algoritma, tapi disini akan digunakan LDA. Parameter yang dibutuhkan yaitu banyaknya topic yang ingin dicari pada *corpus*, yang ditentukan oleh human user. Tahap Kelima yaitu, ***Analyze the Topic Model***, topik model sudah terbentuk dan tinggal tugas user untuk mengevaluasi apakah model tersebut berguna atau tidak.

2.2 Pemodelan Topik Pengguna Twitter

Arianto, et al (2020) [2], *Pemodelan TopikPengguna Twitter Mengenai Aplikasi "Ruangguru"*. Ruangguru adalah perusahaan teknologi yang menangani education-based services. Tujuan dari penelitian ini yaitu untuk mengelompokkan opini dari pengguna Ruangguru mengenai pelayanan yang diberikan. Data yang digunakan diambil dari tweets dari pengguna Twitter menggunakan Twitter API. Akun Twitter yang digunakan pada studi ini yaitu @ruangguru. Dari hasil analisis menggunakan latent dirichlet allocation, ditemukan 28 topik.

Terdapat 3 tipe clustering yaitu hard clustering, hierarchical clustering, dan soft/fuzzy clustering. Pemodelan topik termasuk dalam soft/fuzzy clustering yang mana setiap objek dapat dimiliki lebih dari satu cluster dengan tingkat tertentu (Doig, 2015). Variabel penelitian yang digunakan yaitu (X1) atau bobot kata kunci dari setiap tweet yang diperoleh dari hasil term frequency-inverse document frequency (tf-idf).

Langkah analisis yang dijalankan yaitu pengambilan data tweet menggunakan Twitter API, melakukan text preprocessing, mengubah data tweet ke dalam bentuk frekuensi kemunculan kata menggunakan tf-idf, melakukan pemodelan topik menggunakan LDA, dan menarik kesimpulan dan saran. Pada tahap text processing dilakukan 4 tahap yaitu cleansing, case folding (mengubah karakter menjadi huruf kecil), stemming (mengubah kata menjadi kata dasar), dan melakukan stopword. Untuk metode LDA, yang dilakukan yaitu menentukan jumlah topik dan jumlah iterasi kemudian melakukan pemodelan topik berdasarkan jumlah topik dengan nilai coherence.

Untuk menampilkan hasil dapat digunakan wordcloud untuk memunculkan kata dalam analisis.Lalu nilai kelompok optimum didapatkan dengan mengambil nilai coherence terbesar pada seluruh iterasi. Kemudian dari 28 topik dapat diambil 5 topik lagi yang memiliki persentase muncul terbesar. Dengan menganalisa topik yang ada didapat kesimpulan bahwa topik yang sering diperbincangkan adalah diskon ruangguru.

2.3 *Topic Modeling Menggunakan Latent Dirichlet Allocation*

Nurlayli, et al (2019) [3], *Topik Modeling Penelitian Dosen JPTEI UNY Pada Google Scholar Menggunakan Latent Dirichlet Allocation*. Penelitian ini bertujuan untuk mengimplementasikan *topic modeling* pada judul publikasi dari dosen Jurusan Pendidikan Teknik Elektronika dan Informatika (JPTEI UNY). yang diambil dari *Google Scholar*. Metode yang digunakan adalah Latent Dirichlet Allocation (LDA), yaitu model probabilistik generatif untuk mencari struktur semantik dari kumpulan korpus yang berdasarkan *hierarchical bayesian analysis.*

Metode-metode yang digunakan dalam penelitian ini yaitu: (1) *Data Retrieval*, mengambil dataset dari akun masing-masing dosen pada *Google Scholar*;

(2) *Data Pre-Processing*, menghasilkan data *clustering yang tepat* dengan beberapa langkah: (a) Mengubah kalimat menjadi kata; (b) Menghilangkan beberapa kata yang tidak punya arti misalnya: *using, of, the, in, on, as, and, based*; (3) Mengubah susunan kalimat menjadi bentuk bigram.; (3) *Topic Modeling*, metode *unsupervised machine learning* yang menerapkan pengelompokan untuk menemukan variable laten dari data teks yang besar. Metode yang paling populer untuk pemodelan topik adalah Latent Dirichlet Allocation (LDA).

LDA merupakan kumpulan dokumen sebagai topik campuran yang berisi kata-kata dengan probabilitas tertentu. Prosedur cara kerja LDA adalah sebagai berikut: (1) Inisialisasi beberapa parameter, termasuk jumlah dokumen, topik, dan iterasi. Dalam LDA, parameter yang paling penting adalah jumlah topik k. (2) Menetapkan kata untuk topik tertentu secara acak sesuai dengan distribusi dirichlect. (3) Mengulangi masing-masing alur proses untuk semua kata dalam korpus. Parameter yang digunakan ketika proses perhitungan LDA sebagai berikut: (1) Random state: 100; (2) Update Every: 1; (3) Chunk Size: 10; (4) Passes: 10; (5) Alpha: Symmetric; (6) Iterations: 100; (7) Per Word Topics: True.

Berdasarkan hasil penelitian, hasil yang paling optimum adalah pengelompokan data judul penelitian menjadi empat klaster atau empat topik. Sehingga dapat disimpulkan bahwa topik penelitian dosen JPTEI UNY adalah tentang pendidikan vokasi (Klaster ke-1/Topic 0), pengembangan system (Klaster ke-2/Topic 1), media pembelajaran (Klaster ke-3/Topic 2),
dan sistem pembelajaran di SMK (Klaster ke-4/Topic 3).

2.4 *Preprocessing Text* untuk Meminimalisir Kata yang Tidak Berarti

H. (2015) [4], *Preprocessing Text untuk Meminimalisir Kata yang Tidak Berarti dalam Proses Text Mining*. *Preprocessing text* dalam proses *text mining* diperlukan untuk mengurangi text dengan menghilangkan kata – kata yang tidak perlu atau tidak mempunyai arti. Tujuannya adalah untuk
meringankan proses selanjutnya dalam rangka menambang informasi yang berada dalam
dokumen tanpa mengurangi arti dari informasi yang dikandung dalam dokumen tersebut. Beberapa langkah diperlukan dalam melakukan *text-processing*.

Pertama, *Transform Cases*, yaitu mengubah semua huruf menjadi kecil semua atau bisa juga huruf besar semua. Kedua *Filter stop word*, menghilangkan teks yang tidak berhubungan dengan analisa sentimen sehingga teks akan berkurang tanpa mengubah arti dari sentimen tersebut. Ketiga *Filter Tokenize*, yaitu menghilangkan karakter tertentu seperti tanda baca. Dan yang terakhir yaitu merubah teks menjadi matrix dimana sebuah kalimat akan diubah menjadi atribut teks dan teks akan diubah menjadi matrix yang berisi numerik untuk dapat dimasukan kedalam algoritma menggunakan *TF-IDF (term frequency-inverse document frequency).*

Kesimpulan yang didapat dengan adanya proses *preprocessing text* ini maka data berlebih yang tidak terpakai akan tereliminasi terlebih dahulu sebelum dataset dikenakan metode penelusuran sentiment analisis yang ada. Untuk optimasi *preprocessing text*, dapat digunakan berbagai kombinasi pengurangan kata, maupun stopword dengan bahasa yang lain atau campuran apabila teks mengandung lebih dari satu bahasa.

2.5 Pendekatan Praktis *Topic Modeling* pada Konten Media Sosial

Rohani, et al (2016) [5], *Topic modeling for social media content: A practical approach*. Media sosial memberikan informasi yang sangat besar dan mencakup berbagai bidang, seperti *marketing*, keamanan, pendidikan, dan manajemen. *Topic modeling* memberikan metode yang *powerful* dalam memproyeksikan teks dokumen menjadi topik. Penelitian ini kana membahas *unsupervised topic modeling* menggunakan algoritma LDA untuk menemukan topik-topik yang tersedia di konten media sosial.

Percobaan dilakukan menggunakan 90.527 baris dataset sosial media, dan hasilnya pendekatan ini cukup efektif dalam mendeteksi *topic facet* dan *extracting dynamic over time*. Setelah studi analisa dataset dilakukan, lima buah topik ditemukan dengan akurat. Model yang diajukan cukup umum dan bisa diaplikasikan pada berbagai jenis domain untuk melakukan *mining* topic secara otomatis dari media sosial.

LDA adalah suatu teknik otomatisasi pencarian topik yang mengandung kata tertentu. Dari dataset yang ada, peneliti membuang 645 *common words* berbahasa Inggris dan Malaysia untuk meningkatkan akurasi. Setelah dilakukan

eksperimen, ditemukan 5 buah topik beserta *related keyword*-nya. Lalu setiap *related keyword* memiliki nilai probabilitas kemunculannya masing-masing.

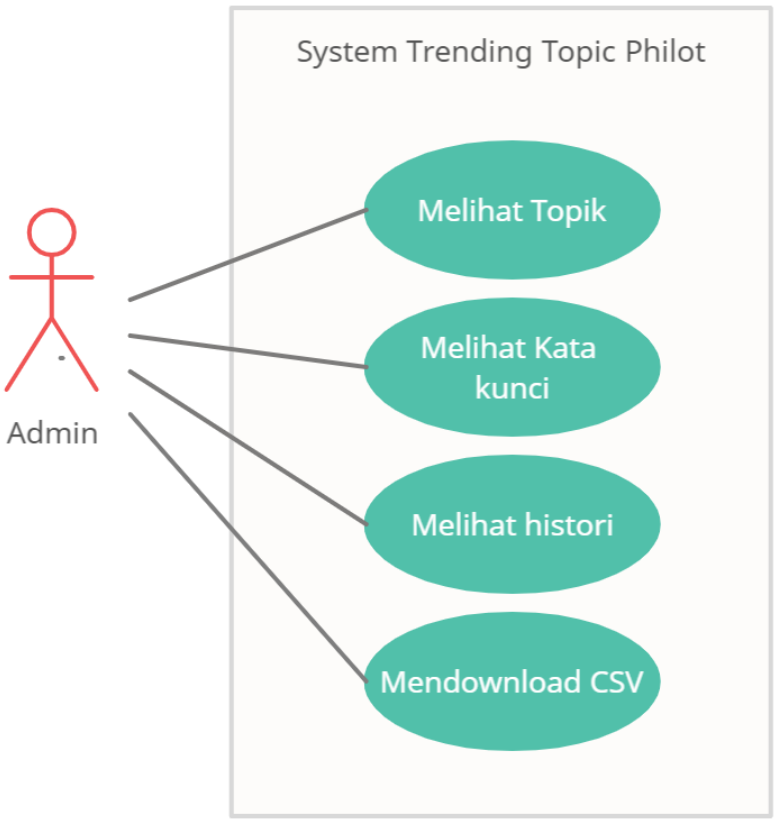
BAB 3
ANALISIS DAN DESAIN

3.1 Cara Kerja Aplikasi

Pemodelan topik dan pencarian trending topic akan dilakukan menggunakan bahasa pemrograman *Python.* Kemudian data yang telah diproses akan ditampilkan melalui *website* interaktif.

3.1.1 Use Case Diagram

Pada gambar 3.1 menunjukkan *use case* diagram dari sistem *trending topic* PhiloIT yang kita ajukan. Aktor dari *use case* ini adalah Admin. Admin dapat melakukan fitur melihat topik, melihat kata kunci, melihat histori, dan mendownload CSV.



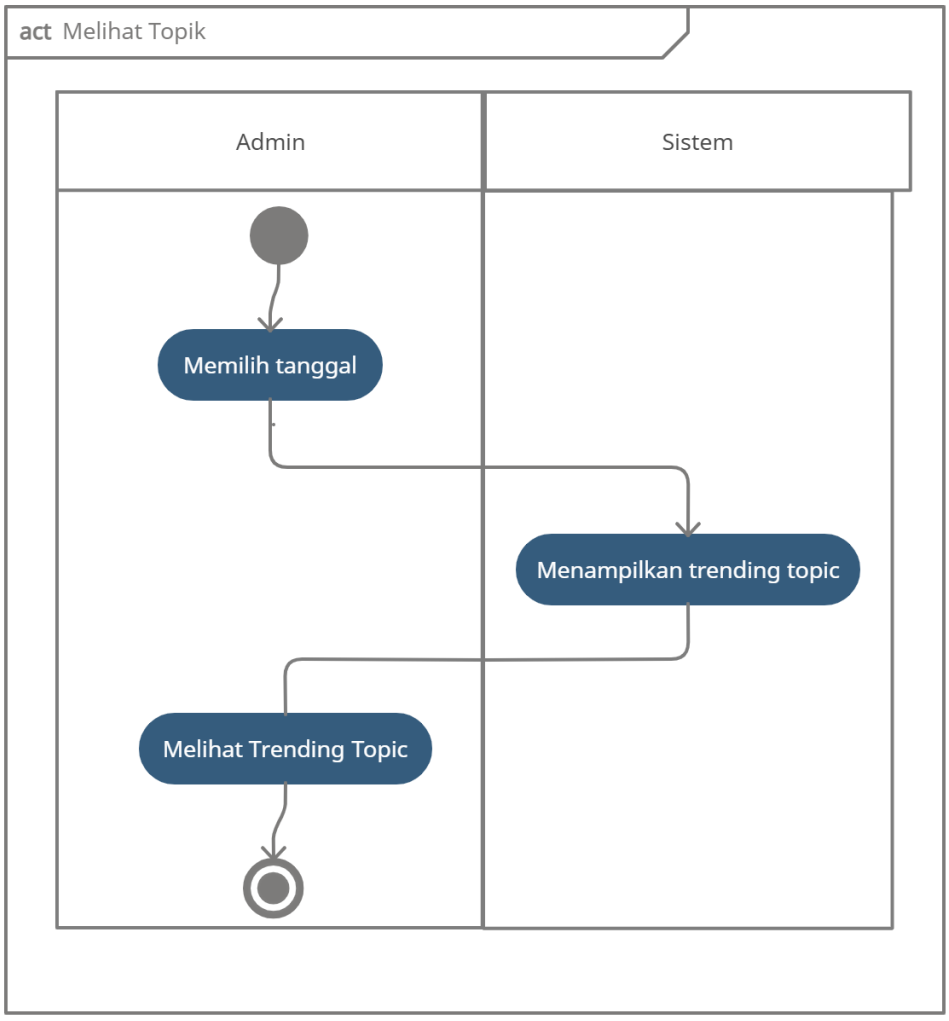
Gambar 3.1 Rancangan *Use Case* Diagram

3.1.2 *Activity* Diagram

Activity diagram digunakan untuk menjelaskan alur dari setiap *use case* yang ada, berikut *activity* diagram dari aplikasi ini.

3.1.2.1 Activity Diagram Melihat Topik

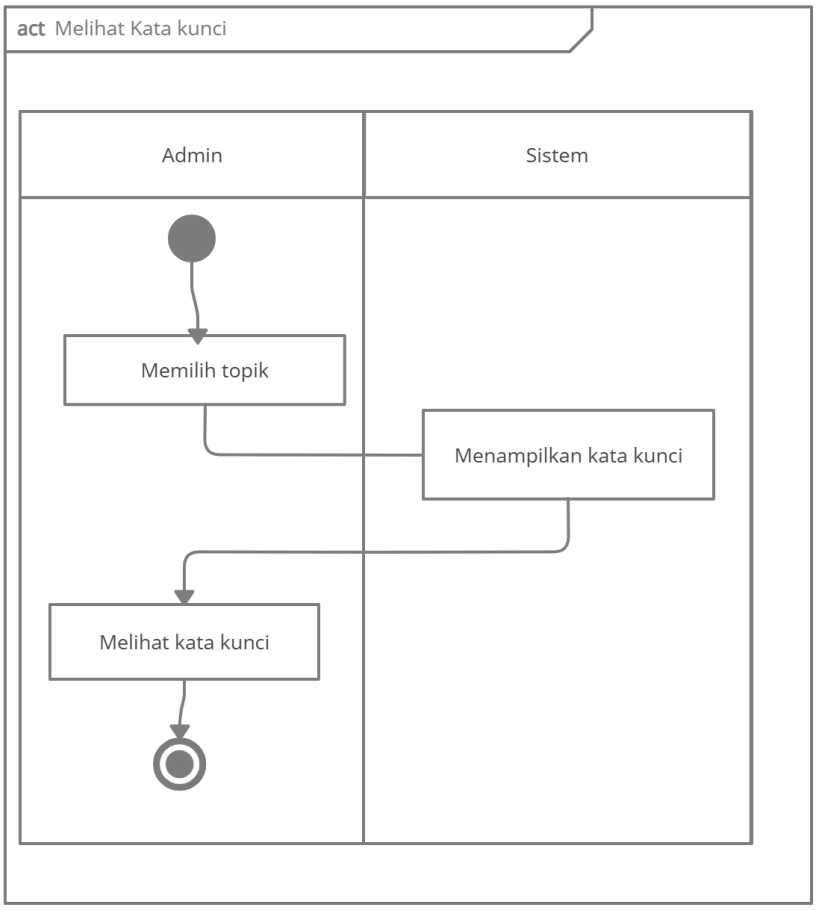
Gambar 3.2 merupakan *activity* diagram dari melihat topik. Aktor admin memilih tanggal pada aplikasi lantas sistem akan menampilkan trending topic sesuai tanggal yang dipilah lalu admin dapat melihatnya.



Gambar 3.2 *Activity* Diagram Melihat Topik

3.1.2.2 *Activity* Diagram Melihat Kata Kunci

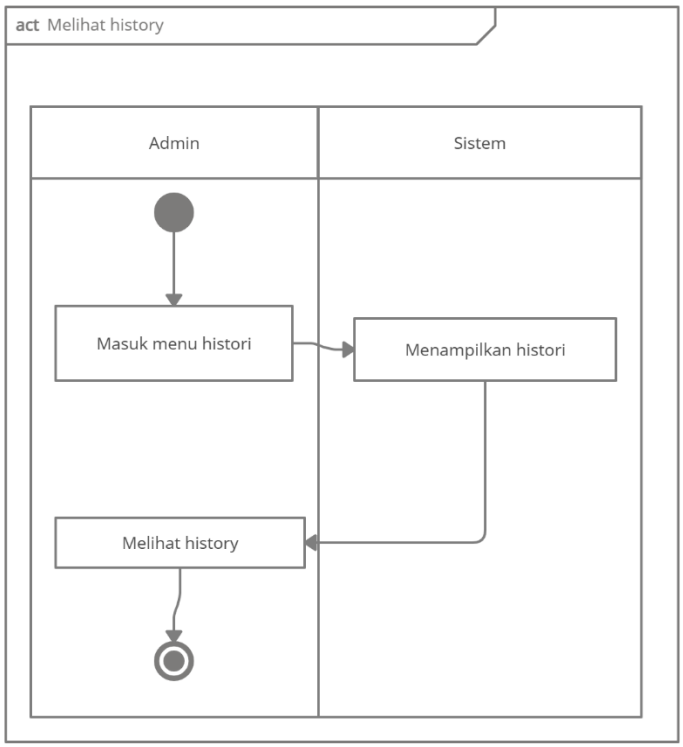
Gambar 3.3 merupakan gambar *activity* diagram melihat topik. Admin memilih topik,lalu sistem akan menampilkan daftar kata kunci, lalu admin dapat melihat kata kunci yang admin minta



Gambar 3.3 *Activity* Diagram Melihat Kata Kunci

3.1.2.3 *Activity* Diagram Melihat History

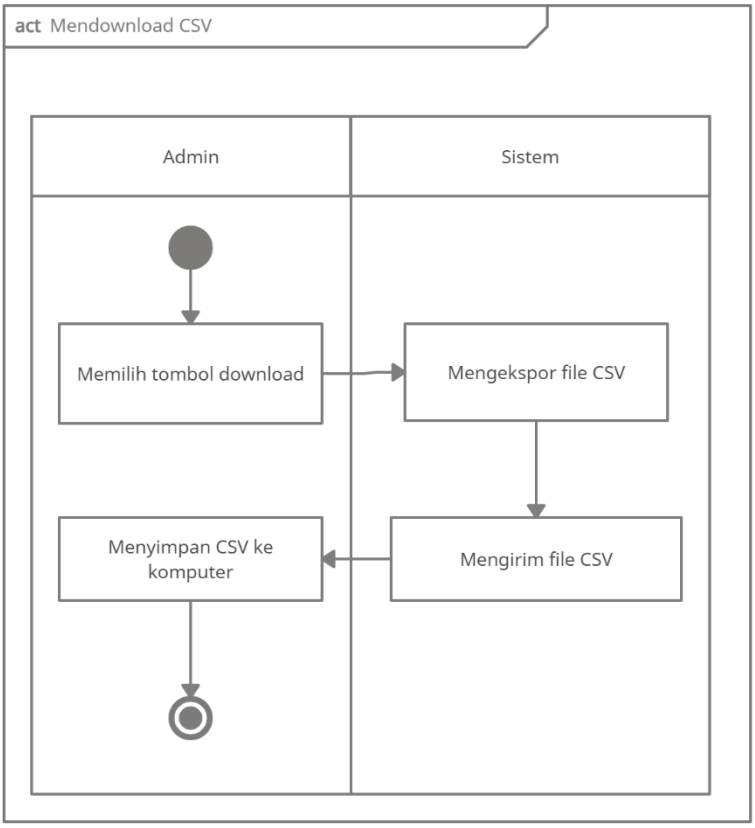
Gambar 3.4 menunjukkan *activity* melihat history. Admin akan memilih menu melihat history lalu sistem akan menampilkan riwayat pencarian trending topic.



Gambar 3.4 *Activity* Diagram Melihat History

3.1.2.4 *Activity* Diagram Mendownload CSV

Gambar 3.5 menunjukkan *activity* diagram mendownload CSV. Aktor admin memijit tombol download pada aplikasi alhasil sistem akan mengekspor lalu mengirim file csv yang dipilih lalu admin dapat menyimpan file tersebut ke komputer.

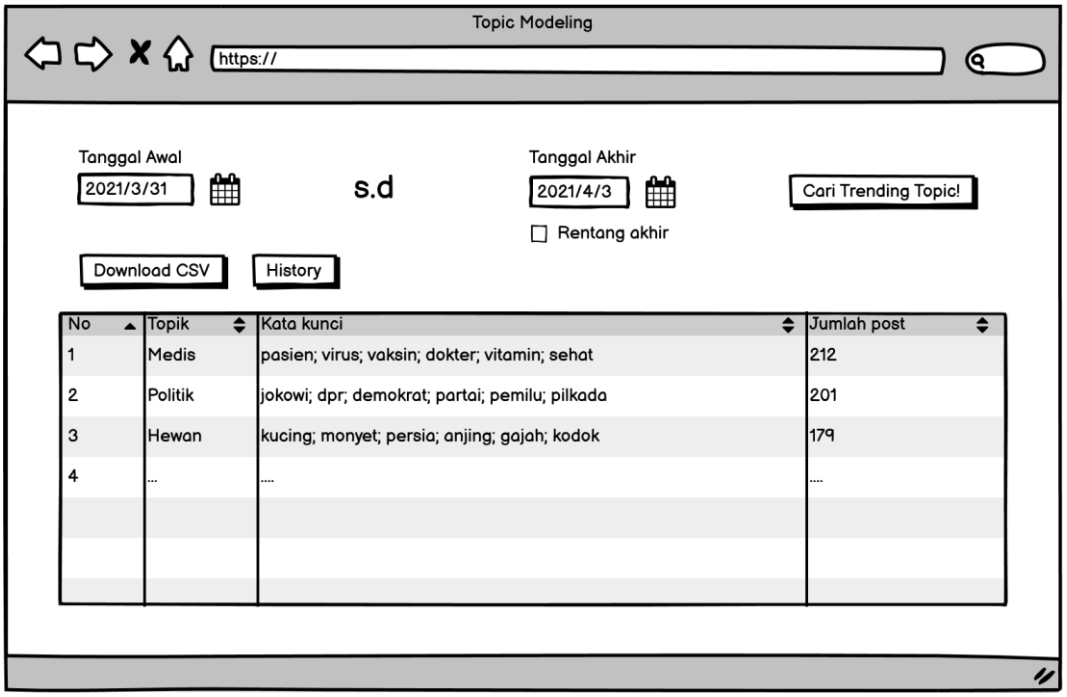


Gambar 3.5 *Activity* Diagram Mendownload CSV

3.2 Rancangan Antarmuka

3.2.1 Rancangan Antarmuka Halaman Utama

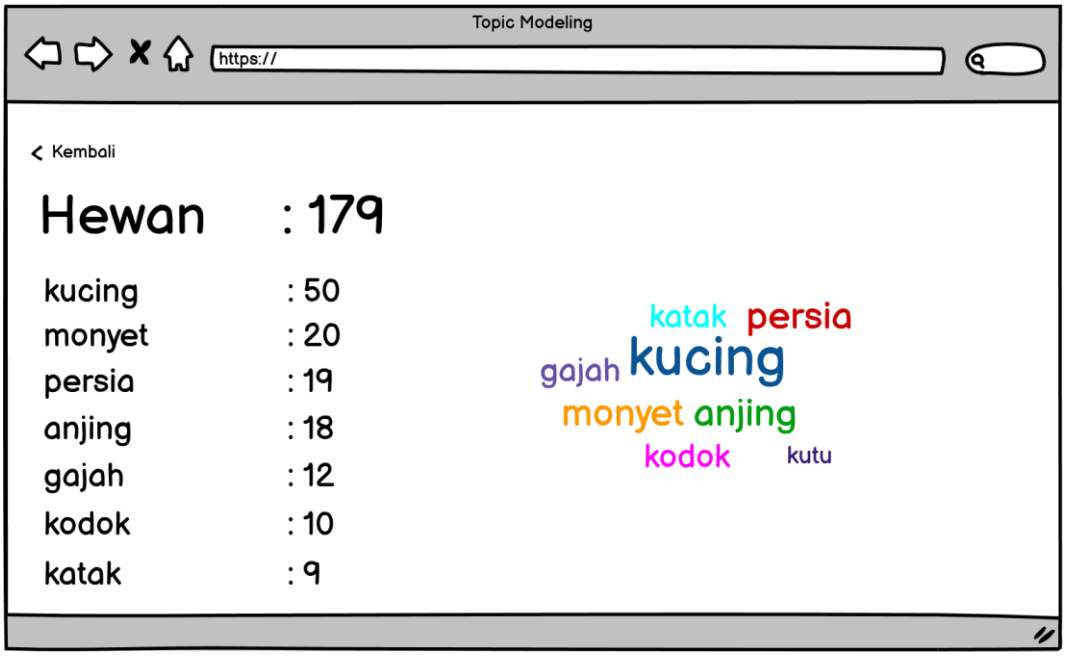
Pada gambar 3.6 menunjukkan halaman utama dari website, halaman ini akan muncul ketika alamat website pertama kali dibuka. Pada halaman ini, terdapat dua buah input tanggal untuk tanggal awal dan tanggal akhir. Untuk mencari trending topic dapat menekan tombol “Cari Trending Topic!”. Tabel digunakan untuk melihat trending topic yang berisi *ranking*, topic, kata kunci, dan jumlah post. Tombol “Download CSV” digunakan untuk menyimpan trending topic yang telah dicari ke penyimpanan komputer.



Gambar 3.6 Rancangan Antarmuka Halaman Utama

3.2.2 Rancangan Antarmuka Halaman Kata Kunci

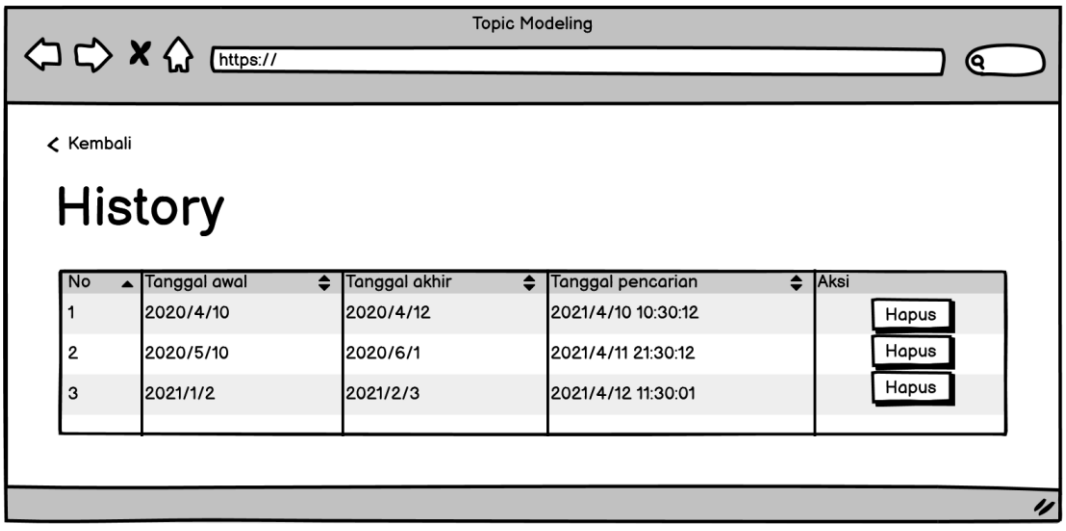
Pada gambar 3.7 menunjukkan halaman kata kunci. Halaman ini dapat diakses dengan memilih topic yang ada di halaman utama. Dalam halaman ini terdapat nama topik dan kata kuncinya beserta jumlah post dari masing-masing topik atau kata kunci. Pada bagian kanan terdapat *word cloud* yang memvisualisasikan kata kunci berdasarkan kemunculan kata tersebut pada rentang waktu yang dipilih.



Gambar 3.7 Rancangan Antarmuka Halaman Kata Kunci

3.2.3 Rancangan Antarmuka Form Soal

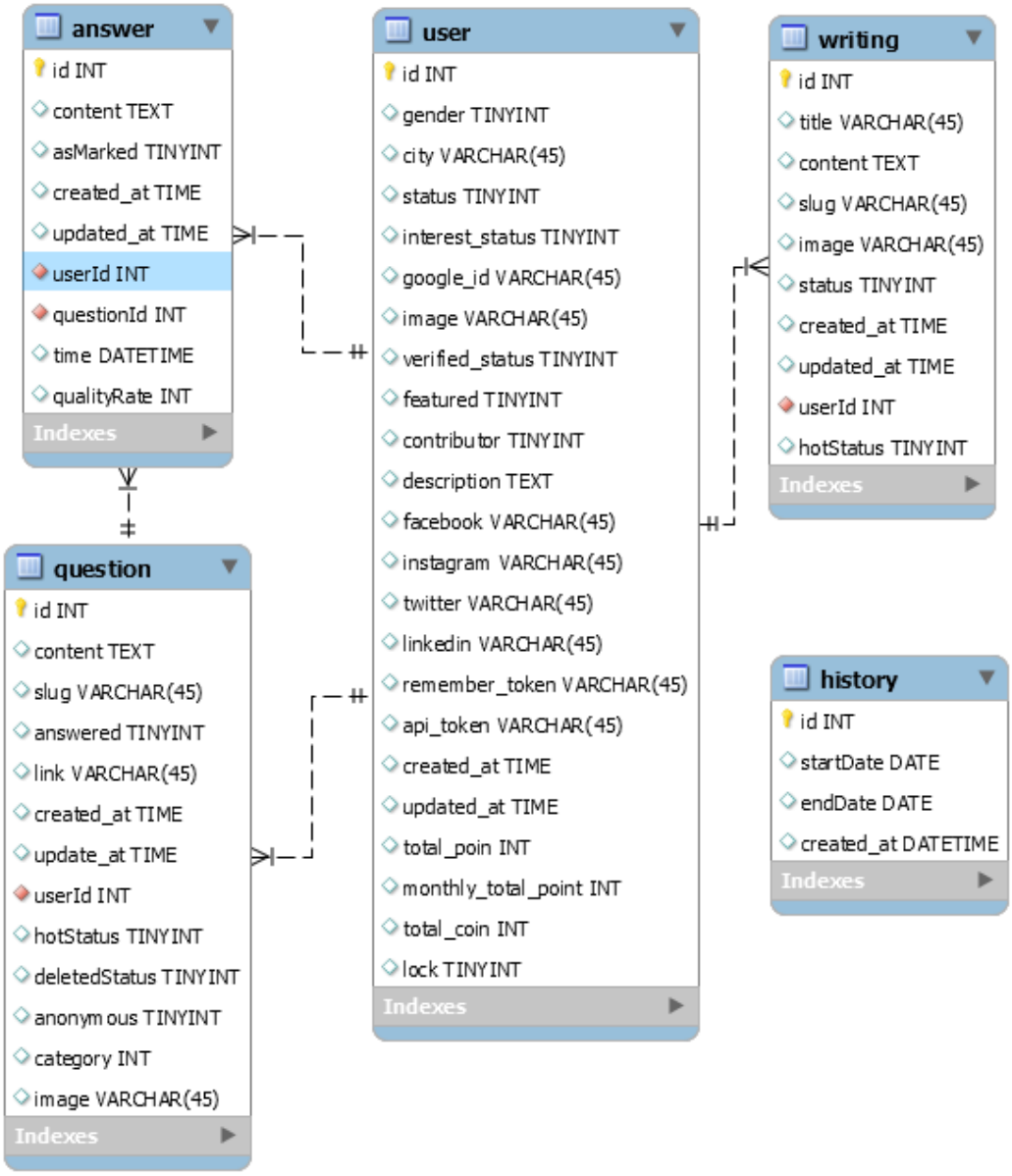
Gambar 3.8 menunjukkan halaman history berisi daftar *history* pencarian *trending topic* yang pernah dilakukan. Ketika salah satu history ditekan maka, halaman akan dipindahkan ke halaman utama yang menampilkan pencarian *trending topic* pada rentang waktu tersebut. Tombol hapus digunakan untuk menghapus *history*.



Gambar 3.8 Rancangan Antarmuka Halaman History

3.3 *Entity Relationship* Diagram

Gambar 3.9 menunjukkan *Entity Relationship* Diagram (ERD) dari website yang akan kami buat. ERD digunakan sebagai rancangan basis data yang nanti akan diimplementasikan untuk membuat tabel beserta atribut, tipe data, dan relasinya.



Gambar 3.9 Rancangan Entity Relationship Diagram

DAFTAR PUSTAKA

[1] M. D. Saxton, "A Gentle Introduction to Topic Modeling Using Python," *THEOLOGICAL LIBRARIANSHIP,* vol. XI, no. 1, pp. 18-26, 2018.

[2] B. W. Arianto dan G. Anuraga, "Pemodelan Topik Pengguna Twitter Mengenai Aplikasi "Ruangguru"," *Jurnal ILMU DASAR,* vol. XII, no. 2, pp. 149-154, 2020.

[3] A. Nurlayli dan M. A. Nashichuddin, "Topic Modeling Penelitian Dosen JPTEI UNY pada Google Scholar Menggunakan Latent Dirichlet Allocation," *ELINVO (Electronics, Informatics, and Vocational Education),* vol. IV, no. 2, pp. 154-161, 2019.

[4] A. T. J. H, "Preprocessing Text untuk Meminimalisir Kata yang Tidak Berarti," *Jurnal Informatika UPGRIS,* vol. I, no. Juni, pp. 1-9, 2015.

[5] V. A. Rohani, S. Shayaa dan G. Babanejaddehaki, "Topic Modeling for Social Media Content: A Practical Approach," *2016 3rd International Conference On Computer And Information Sciences (ICCOINS),* pp. 397-402, 2016.