TESIS

PENERAPAN TOPIC MODELING TREN PENELITIAN SKRIPSI MAHASISWA MENGGUNAKAN LATENT DIRICHLET ALLOCATION (LDA) PADA UNIVERSITAS DIPA MAKASSAR

Disusun oleh:

Nama	**: Isra Andika Bakhri**
NIM	**: 21.55.1028**
Konsentrasi	**: Business Intelligence**

PROGRAM STUDI S2 PJJ TEKNIK INFORMATIKA

PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2023

TESIS

PENERAPAN TOPIC MODELING TREN PENELITIAN SKRIPSI MAHASISWA MENGGUNAKAN LATENT DIRICHLET ALLOCATION (LDA) PADA UNIVERSITAS DIPA MAKASSAR

APPLICATION OF TOPIC MODELING RESEARCH TREND OF STUDENT THESIS USING LATENT DIRICHLET ALLOCATION (LDA) AT DIPA MAKASSAR UNIVERSITY

Diajukan untuk memenuhi salah satu syarat memperoleh derajat Magister

Disusun oleh:

Nama : **Isra Andika Bakhri**
NIM : **21.55.1028**
Konsentrasi : **Business Intelligence**

PROGRAM STUDI S2 PJJ TEKNIK INFORMATIKA
PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2023

HALAMAN PENGESAHAN

PENERAPAN TOPIC MODELING TREN PENELITIAN SKRIPSI MAHASISWA MENGGUNAKAN LATENT DIRICHLET ALLOCATION (LDA) PADA UNIVERSITAS DIPA MAKASSAR

APPLICATION OF TOPIC MODELING RESEARCH TREND OF STUDENT THESIS USING LATENT DIRICHLET ALLOCATION (LDA) AT DIPA MAKASSAR UNIVERSITY

Dipersiapkan dan Disusun oleh

Isra Andika Bakhri

21.55.1028

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 PJJ Teknik Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Jumat, 3 Februari 2023

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 3 Februari 2023
Rektor

Prof. Dr. M. Suyanto, M.M.
NIK. 190302001

HALAMAN PERSETUJUAN

PENERAPAN TOPIC MODELING TREN PENELITIAN SKRIPSI MAHASISWA MENGGUNAKAN LATENT DIRICHLET ALLOCATION (LDA) PADA UNIVERSITAS DIPA MAKASSAR

APPLICATION OF TOPIC MODELING RESEARCH TREND OF STUDENT THESIS USING LATENT DIRICHLET ALLOCATION (LDA) AT DIPA MAKASSAR UNIVERSITY

Dipersiapkan dan Disusun oleh

Isra Andika Bakhri

21.55.1028

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 PJJ Teknik Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Jumat, 3 Februari 2023

Pembimbing Utama

Prof. Dr. Ema Utami S.Si, M.Kom
NIK. 190302035

Anggota Tim Penguji

Dr. Andi Sunyoto, M.Kom
NIK. 190302052

Pembimbing Pendamping

Hanafi S.Kom, M.Eng, P.hD
NIK. 190302024

Anggit Dwi Hartanto, M.Kom
NIK. 190302163

Prof. Dr. Ema Utami, S.Si., M.Kom
NIK. 190302035

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 3 Februari 2023
Direktur Program Pascasarjana

Prof. Dr. Kusrini, M.Kom.
NIK. 19030210

HALAMAN PERNYATAAN KEASLIAN TESIS

Yang bertandatangan di bawah ini,

Nama mahasiswa : Isra Andika Bakhri
NIM : 21.55.1028
Konsentrasi : Business Intelligence

Menyatakan bahwa Tesis dengan judul berikut:
Penerapan Topic Modeling Tren Penelitian Skripsi Mahasiswa Menggunakan Latent Dirichlet Allocation (LDA) Pada Universitas Dipa Makassar

Dosen Pembimbing Utama : Prof. DR. Ema Utami S.Si, M.Kom
Dosen Pembimbing Pendamping : Anggit Dwi Hartanto M.Kom

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Tim Dosen Pembimbing
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi

Yogyakarta, 3 Februari 2023
Yang Menyatakan,

Isra Andika Bakhri

KATA PENGANTAR

Alhamdulillah, segala puji dan syukur penulis panjatkan kehadirat Allah SWT karena atas segala karunia dan ridho-Nya, sehingga tesis yang berjudul "Penerapan Topic Modeling Tren Penelitian Skripsi Mahasiswa Menggunakan Latent Dirichlet Allocation (Lda) Pada Universitas Dipa Makassar" dapat diselesaikan dengan tepat waktu.

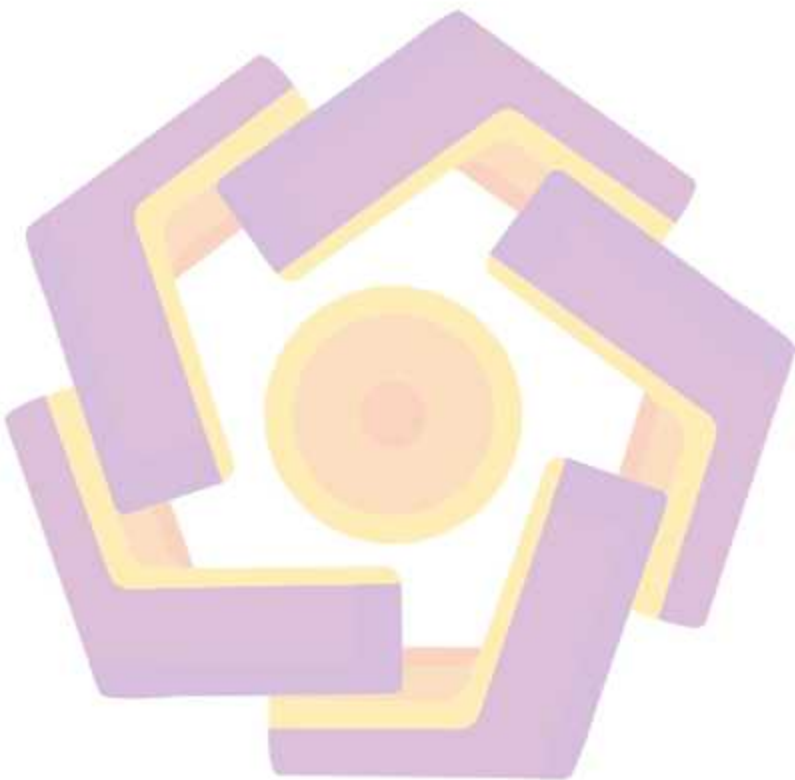
Penyelesaian tesis yang sangat berharga ini tidak lepas dari bantuan dan dukungan dari berbagai pihak. Pada kesempatan ini, penulis mengucapkan rasa syukur dan terima kasih kepada:

1. Orang Tua terutama orang tua, saudara dan teman-teman yang senantiasa memberikan do'a semangat, dan dukungan kepada penulis agar senantiasa semangat dalam menuntut ilmu.
2. Ibu Prof. Dr. Ema Utami, S.Si., M.Kom. selaku pembimbing utama yang telah membimbing dan memotivasi dalam penulisan tesis ini agar dapat terselesaikan dengan baik.
3. Bapak Anggit Dwi Hartanto M.Kom. selaku pembimbing yang telah membimbing dalam penulisan tesis ini agar dapat terselesaikan dengan baik.
4. Dosen Penguji yang telah memberikan saran yang baik demi kemajuan tesis ini.

Yogyakarta, 18 Desember 2022

Penulis

DAFTAR ISI



DAFTAR TABEL

DAFTAR GAMBAR

INTISARI

Salah satu yang menjadi permasalahan dalam pengajuan judul penelitian skripsi mahasiswa adalah topik yang diajukan terlalu monoton, juga dari algoritma yang diajukan. Didalam penelitian ini akan dilakukan pemecahan masalah tersebut dengan melibatkan konsep *topic modeling* dan *information retrieval* untuk mencegah mahasiswa mengajukan dengan topik yang monoton berdasarkan tren penelitian dikampus mereka. Pada penelitian ini akan digunakan algoritma *topic modeling* yang bernama *latent Dirichlet allocation* (LDA) sedangkan untuk algoritma *information retrieval* menggunakan *vector space model* (VSM). Kedua algoritma itu akan diujikan menggunakan 4 skenario preprocessing dengan melibatkan *stemmer truncating* menggunakan library sastrawi dan *stemmer statistical* menggunakan N-gram untuk mendapatkan scenario preprocessing terbaik untuk kedua algoritma tersebut. Selanjutnya akan dikomparasikan hasil dari model tersebut yaitu LDA menghasilkan *latent topics* berisi topik-topik paling banyak disebut didalam dataset sedangkan VSM akan mencari tau dari topik tersebut dataset yang seperti apa yang dianggap monoton dan tren penelitian. Sehingga dataset berjumlah 430 akan difilter sedangkan hasil filternya akan dijadikan dataset baru, selanjutnya akan digunakan algoritma VSM baru untuk dilatih dengan dataset baru, maka model VSM baru inilah yang akan berinteraksi dengan mahasiswa yang mengajukan judul skripsi apakah judulnya masuk dalam kategori monoton atau tidak.

Dari hasil penelitian model LDA dan VSM akan lebih optimal menerapkan *stemmer truncating* dan *stemmer statistical*, masing-masing menghasilkan *coherence score* 0.30663692761114086 dan *mean average precision (mAP)* sebesar 0.6438269428983715. Pada model LDA berhasil menyaring 12 topik mononton dari data selanjutnya dilanjutkan VSM untuk mencari data tersebut sehingga dari 430 dataset tersisa 58 dataset yang kemudian dimasukkan dalam model VSM baru dan dilakukan testing dengan 10 data pengajuan judul dan model tersebut menjawab dengan tepat hasil cek topik monoton berdasarkan tren penelitian mahsiswa pada kampus tersebut.

Maka didapatkanlah metode tepat yang dapat mencegah judul dengan topik yang masuk dalam topik yang monoton. Dampaknya dari hasil penelitian ini membuat mahasiswa memilih mengajukan topik yang lebih bervariasi dan tidak monoton.

Kata kunci: *topic modeling, information retrieval, LDA, VSM, tren penelitian*

ABSTRACT

One of the problems in submitting student thesis research titles is that the topics proposed are too monotonous, which is also evident in the proposed algorithm. This research will solve the problem by involving the concepts of topic modeling and information retrieval to prevent students from submitting monotonous topics based on research trends on their campuses. In this study, a topic modeling algorithm called latent Dirichlet allocation (LDA) will be used, while the information retrieval algorithm will use a vector space model (VSM). The two algorithms will be tested using four preprocessing scenarios involving stemmers truncating using a literary library and statistical stemmers using N-grams to determine the best preprocessing scenario for the two algorithms. Next, the results of the model will be compared, namely, LDA will produce latent topics containing the topics most frequently mentioned in the dataset, while VSM will find out from these topics which datasets are considered monotonous and research trends. So the 430 dataset will be filtered, and the filtered results will be used as a new dataset. The new VSM algorithm will then be used to train with the new dataset, and this new VSM model will interact with students who submit thesis titles, regardless of whether the title falls into the monotone category or not.

From the results of the research, the LDA and VSM models will be more optimal in applying stemmer truncating and stemmer statistical, respectively, producing a coherence score of 0.30663692761114086 and a mean average precision (mAP) of 0.6438269428983715. The LDA model succeeded in filtering 12 viewing topics from the data, then continued with VSM to search for these data so that out of the 430 datasets, the remaining 58 datasets were then included in the new VSM model, and testing was carried out with 10 title submission data, and the model correctly answered the results of the monotonic topic check based on student research trends on the campus.

Then the right method is obtained, which can prevent titles with topics that fall into monotonous topics. The impact of the results of this study made students choose to submit topics that were more varied and not monotonous.

Keywords: topic modeling, information retrieval, LDA, VSM, research trends

BAB I

PENDAHULUAN

1.1. LATAR BELAKANG MASALAH

Salah satu yang menjadi permasalahan bagi mahasiswa yang akan mengajukan judul skripsi adalah tidak adanya rekomendasi untuk tidak mengikuti judul yang topik penelitiannya monoton berdasarkan judul yang diajukan dan penelitian mahasiswa dibeberapa angkatan sebelumnya. Saat ini ketua jurusan di universitas Dipa Makassar dari program studi teknik informatika dan sistem informasi tidak memiliki upaya dalam menetapkan judul dan topik penelitian yang selalu diajukan dan diteliti sebelumnya. Tujuannya adalah agar mahasiswa tidak mengajukan judul pada topik penelitian yang monoton, sehingga dari mahasiswa dapat memilih topik penelitian lain.

Proses ekstraksi informasi dalam dokumen abstrak mahasiswa sebenarnya bisa dijadikan solusi untuk menemukan pola tersembunyi antara dokumen yang berada pada dokumen pengajuan judul dan dokumen skripsi yang telah diteliti. Salah satu tekniknya adalah *topic modeling* dan *information retrieval* yang diharapkan dapat menemukan topik yang sering disebut dalam kumpulan dokumen pengajuan judul dan dokumen skripsi yang telah diteliti. Sehingga akan tampak topik yang sering disebut dan juga dapat dilakukan proses klustering per topik yang berisi beberapa kata kunci.

Menurut I. Vayansky and S.A.P. Kumar (2020) banyak metode yang dapat digunakan dalam penerepan *topic modeling* seperti *latent semantic analysis (LSA), Probabilistic latent semantic analysis (PLSA), latent direchlet allocation (LDA)* dan *correlated topic model (CTM).* Dalam penerapannya LSA lebih cocok digunakan untuk penerapan *information retrieval* (temu balik informasi) karena pada umumnya LSA bekerja dengan melakukan representasi vector untuk teks dengan tujuan membuat konten semantik untuk pencocokan antara query dan representasi vector yang merupakan cara kerja model *information retrieval* namun tidak menutup kemungkinan dapat juga diterapkan untuk *topic modeling* sederhana karena keterbatasan penggunaan rumus statistika yang kurang optimal. Sedangkan PLSA yang merupakan pengembangan dari LSA yang hasil pengembangannya memiliki tujuan dalam mengidentitifkasi dan membedakan konteks penggunaan kata yang berbeda tanpa bantuan kamus/tesaurus. Cara kerjanya adalah dengan melakukan representasi pengelompokan penggunaan kata yang sejenis seperti nama kota, nama makanan dan lain sebagainya, metode PLSA lebih memungkinkan penerapan *topic modeling* karena memiliki representasi vector yang lebih baik namun sayangnya dalam level dokumen PLSA memiliki keterbatasan. Sedangkan untuk CTM tidak sesuai dengan tujuan penelitian ini karena CTM menitikberatkan pada pencarian topik beserta dengan korelasi antar topik yang dihasilkan dari model sedangkan LDA tidak memberikan informasi korelasi topik yang dengan itu akan membantu peneliti dalam menemukan sisi perbedaan tren pada penelitian mahasiswa yang berangkat dari sudut pandang yang berbeda-beda.

Adapun LDA menurut Baris Ozyurt & M. Ali Akacayol (2021) merupakan model yang umum digunakan dalam beberapa paper dalam dekade ini yang sejatinya adalah pengembangan 2 model sebelumnya yaitu LSA dan PLSA dengan cara menangkap pertukaran kata dan dokumen untuk menemukan topic tersembunyi. Bahkan tidak memerlukan sistem pengawasan (*unsupervised)* dalam menangkap topik tersembunyi itu. Hal itu dikarenakan LDA adalah algoritma untuk *text mining* yang didasarkan pada model statistik (Bayesian).

Hasil perhitungan LDA akan menghasilkan anotasi topik yang bersifat inkrimental tanpa label yang jelas karena algoritma ini termasuk *unsupervised learning*, adapun dari tiap topik menghasilkan beberapa kata kunci. Permasalahan dari hasilnya adalah belum tergambar secara nyata judul seperti apa yang mesti dihindari untuk diajukan oleh mahasiswa. Maka dari proses *topic modeling* ini perlu dilakukan *improvement* untuk melakukan proses pencarian judul di dalam kumpulan dokumen terkait kata kunci tiap topik yang dihasilkan.

Konsep tersebut lebih cocok diterapkan dalam konsep sistem *information retrieval* (IR). Ada beebrapa algoritma yang data digunakan dalam membuat sistem IR yaitu algoritma *Boolean Retrieval, Wildcard Query, Binary Tree dan Vector Space Model* (VSM).

Menurut Bucher (2010) algoritma *Boolean Retrieval* hanya cocok pada dokumen yang memiliki konsep terstruktur sehingga jika konsepnya tidak demikian maka akan melakukan pencocokan dokumen yang terlalu sedikit dan kadang terlalu banyak yang disebabkan kakunya proses *index* dalam algoritma ini, disisi lain

karena penggunaan *boolean* pada *query* mengakibatkan hasilnya dianggap sama-sama berbobot.

Menurut Wildcard (2008) algoritma *wildcard query* lebih cocok pada penerapan *spill checker* atau pengecekan ejaan apakah ejaan tersebut sudah tepat atau belum. Menurut Budiharto (2016) algoritma *Binary tree* memilliki struktur pemecahan yang sederhana maka jika diterapkan pada *corpus* yang besar maka proses *query* akan menjadi lambat.

Menurut P. Mahalakshmi & N. Sabiyath Fathima (2021) salah satu teknik pendekatan berbasis metode statistik yang dapat diterapkan pada *information retrieval (IR)* yang peneliti pada umumnya memilih algoritma *vector space model (VSM)* dalam temu balik informasi yang relevan, dibandingkan dengan algortma yang lain maka peneliti memilih algoritma VSM untuk menerapkan IR dalam penelitian ini. Sehingga diharapkan akan didapatkan hasil judul-judul relevan berdasarkan kata kunci yang dihasilkan tiap topik dari teknik *topic modeling*.

Sehingga berdasarkan studi literatur yang dilakukan pada penelitian sebelumnya maka penulis menggunakan algoritma LDA dalam menemukan tren topik penelitian yang diharapkan dapat dengan mudah untuk mendapatkan *latent topic* yang dalam Sedangkan dalam penerapan *information retrieval* menggunakan algoritma VSM.

1.2. RUMUSAN MASALAH

Dari latar belakang diatas maka penulis melakukan penulisan rumusan masalah sebagai berikut.

a. Seberapa besar performa model LDA menemukan topik-topik inti dalam penelitian mahasiswa?
b. Seberapa besar performa model VSM dapat menemukan judul dan abstrak yang relevan berdasarkan *query* yang merupakan topik-topik inti dari model LDA?
c. Bagaimana hasil model LDA dan VSM dapat membuat pembobotan model VSM baru menjadi lebih baik untuk melakukan verifikasi pengajuan judul penelitian baru mahasiswa?

1.3. BATASAN MASALAH

Bagian ini memuat penjelasan tentang:

a. Bentuk data yang digunakan dalam penelitian ini adalah dalam bentuk text.
b. Dataset yang digunakan adalah teks dokumen judul dan abstrak dari *softcopy* penelitian skripsi yang diambil secara manual *(copy paste)* lalu dimasukkan dalam format excel pada jurusan teknik informatika Universitas Dipa Makassar dari tahun 2019 sampai dengan 2021 sebanyak 430 sample yang dimana mewakili 6 periode skripsi.
c. Dalam penelitiaan ini sinonim tidak dijadikan objek dalam hitungan.
d. Program studi yang dijadikan sample adalah program studi teknik informatika.

e. Bahasa yang digunakan adalah bahasa Indonesia untuk tulisan formal pada judul dan abstrak.

f. Data yang digunakan adalah judul penelitian dan abstraknya saja.

g. *Typo* pada penulisan judul dan abstrak tidak diabaikan.

h. Pemprosesan algoritma menggunakan bahasa pemograman python pada platform google yaitu *google colab*.

i. Pada proses stopwords menggunakan *corpus* NLTK.

j. Pada proses stemming menggunakan *corpus* Sastrawi.

k. Mendapatkan data topik yang dihasilkan dari *topic modeling* untuk diolah kembali dalam sistem *information retrieval*.

l. Sistem *information retrieval* mengembalikan dalam bentuk judul yang relevan yang dihasilkan oleh *topic modeling*.

m. Judul yang dihasilkan dalam sistem *information retrieval* adalah judul yang disarankan untuk tidak diajukan oleh mahasiswa.

n. Model *information retrieval* akan dibuat lagi untuk keperluan menjalankan hasil model *information retrieval* sebelumnya untuk mengecek kemiripan dari judul yang diajukan oleh mahasiswa dengan judul yang tidak disarankan hasil dari model sebelumnya.

1.4. TUJUAN PENELITIAN

Bagian ini memuat penjelasan secara spesifik:

a. Untuk mengetahui topik-topik inti dalam penelitian mahasiswa secara akurat menggunakan model LDA.
b. Untuk mengindeks judul dan abstrak berdasarkan model terbaik VSM berdasarkan *query* yang merupakan input model VSM berdasarkan topik-topik inti yang ditemukan model LDA.
c. Melakukan penggabungan hasil model LDA dan VSM untuk memperbaiki pembobotan model VSM baru menjadi lebih baik untuk melakukan verifikasi pengajuan judul penelitian baru mahasiswa.

1.5. MANFAAT PENELITIAN

Bagian ini memuat penjelasan tentang:

a. Diharapkan berkontribusi dalam pengembangan pengetahuan yang berkaitan tentang *topic modeling* dan *information retrieval*.
b. Mahasiswa mendapatkan referensi mencari topik penelitian jika topik penelitian yang diajukan ada dalam tren penelitian.
c. Penelitian skripsi mahasiswa di Universitas Dipa Makassar diharapkan semakin bervariasi.

BAB II

TINJAUAN PUSTAKA

2.1. TINJAUAN PUSTAKA

Pada penelitian yang dilakukan oleh rubiyya dan Khalid (2015) membahas tentang algoritma yang digunakan dalam pembuatan model *topic modeling* seperti *latent semantic analysis (LSA), probabilistic latent semantic analysis (PLSA), latent direchlet allocatin (LDA)* dan *correlation topic model (CTM).* Dalam makalahnya dibandingkan ke empat metode tersebut mulai dari defenisi, karakteristik, keterbatasan, kelebihan sampai dari penerapannya sehingga bisa menjadi rujukan untuk memilih algoritma yang cocok dalam penelitian ini berdasarkan subjek dan objek penelitian dengan memilih LDA.

Pada penelitian yang dilakukan oleh Eka sabna (2020) yang bertujuan untuk mengelompokkan penelitian-penelitian yang telah dilakukan oleh dosen menggunakan algoritma K-Means agar hasil pengelompokan dapat dijadikan rujukan tren penelitian tiap dosen. Selain itu dapat dijadikan referensi oleh jurusan untuk menentukan dosen pembimbing dan penguji. Dari hasil klustering didapatkan kata yang menjadi tren penelitian dosen program studi Ilmu kesehatan masyarakat (IKM) tahun 2020 yang kebanyakan mengangkat topik yang terkait dengan Covid dan HIV. Namun dalam penelitian ini tidak tergambar secara nyata sesuai dengan tujuan dari penelitian ini seperti diuraikan pada latar belakang penelitian. Seharusnya dalam penelitian ini digambarkan

tren penelitian tiap dosen 3 tahun sebelumnya hingga sekarang sehingga tujuan penelitian dapat sesuai dengan rumusan masalah. Serta kesimpulan hanya menghasilkan tren berdasarkan objek penelitian bukan topik penelitian. Sedangkan untuk perbedaanya adalah dalam melakukan pengelompokan topik penelitian meskipun sama-sama menggunakan *unsupervised learning* namun dalam penelitian ini menggunakan K-Means serta hasil dari k-means tidak diolah sedemikian ruap untuk mengekstrak informasi lebih jauh sebagai *method improvement*-nya.

Bagus Wicaksono Arianto & Gangga Anuraga (2020) melakukan penelitian dengan tujuan melakukan pengelompokan persepsi pengguaan aplikasi ruangguru di twitter menggunakan LDA. Kesimpulan yang didapatkan dalam penelitian tersebut adalah mendapatkan 28 buah topik hasil dari klasterisasi topik yang dihasilkan dari metode LDA dan topik yang paling banyak ialah berkaitan dengan diskon ruang guru. Namun penulis melihat kelemahan dari penelitian ini tidak memiliki tinjauan pustaka yang cukup untuk membuktikan hipotesisnya. Sedangkan perbandingan penelitian ini adalah sumber data nya berasal dari twitter yang menulis twet tentang aplikasi ruangguru serta hasil dari topik yang dihasilkan tidak diolah lebih lanjut, peneliti menyimpulkan sendiri hasil topik yang dihasilkan model sedangkan dalam penelitian ini menjadi berbeda karena melibatkan sistem temu kembali informasi (*information retrieval)* dalam menemukan data yang akan dicari.

Pada penelitian yang dilakukan wahyudin (2020) yaitu memetakan topik pemberitaan tentang PSBB pertama dari berbagai portal berita online. Wahuyudin menyimpulkan bahwa pemberitaan 9 media online nasional dapat dikelompokkan kedalam 4 kelompok besar pemberitaan: perkembangan kasus Covid-19, penanganan dampak pademi, seputar Ramadhan dan kegiatan belajar di rumah, serta berita lain-lain diantaranya seputar: PSBB, kriminal, hiburan, dan politik. Namun kelemahan penelitian ini adalah tidak ditampilkannya nilai koherensi terbaik dalam tiap iterasi yang sebenarnya bisa menjadi informasi penting dalam penelitian ini. Sedangkan perbedaan penelitian yang dilakukan wahyudin ini yaitu sumber data berasal dari portal pemberitaan online dan hasil dari topik modeling menggunakan LDA tidak diolah lebih lanjut hanya sekedar menafsir maksud *term-term* per topik yang dihasilkan.

Alif Iffan Alfanzar, Khalid & , Indri Sudanawati Rozas (2020) melakukan penelitian sehingga penulis menyimpulkan dari sebanyak 584 abstrak yang dikumpulkan dilakukan percobaan sebanyak 5 iterasi percobaan dengan total iterasi 100, 500, 1000 dan 5000 dan setiap uji dimasukkan nilai k yaitu 2,3,4 dan 5. Dan didapatkan nilai k terbaik adalah 3 dan hasilnya terkonfirmasi oleh pihak stakeholder program studi bahasa inggris. Dalam penelitian ini terdapat kekurangan yaitu tidak menampilkan table hyperparameter tuning hasil percobaan berdasarkan jumlah k dan iterasi dan juga menampilkan nilai perplexity dan koherensi untuk melengkapi informasi dalam penelitiaan ini. Sedangkan perbedaan penelitiannya adalah sumber

datanya menggunakan abstrak penelitian bahasa inggris sedangkan dalam penelitian ini menggunakan bahasa indonesia.

P. Mahalakshmi & N. Sabiyath Fathima (2020) menjelaskan berbagai representasi teks yang bertanggung jawab atas mengambil hasil pencarian yang relevan, pendekatan serta evaluasi yang dilakukan di dalam pengambilan informasi konseptual di sebuah sistem *information retrieval*. Dalam kesimpulannya penulis mengungkapkan didapatkan pemetaan yang jelas terkait perkembangan teknologi *information retrieval* berdasarkan strategi yang disusun atas 4 kriteria yaitu *representation of text* dalam bentuk apa, *approaches* atau pendekatan perhitungan yang digunakan, *source* atau sumber data dan terakhir *evaluation* atau evaluasi untuk mengukur akurasi dan performa model.

Lalu yang terakhir adalah penelitian yang dilakukan oleh Kabil Boukhari & Mohamed Nazih Omri (2020) tujuannya adalah menerapkan sistem temu balik informasi yang dapat mengindeks lebih baik pada dokumen biomedis dengan menggunakan VSM dan DL (Descripstion Logics). Penelitian ini menyimpulkan telah berhasil melakukan ekstraksi konsep dari semua dokumen biomedis dengan menggabungkan statistika dan metode semantic dari dua metode yaitu VSM dan *description logics* (DL). Sedangkan perbandingan dari penelitiannya adalah Penerapan information retrieval menggunakan penggabunagan konsep VSM dan DL sedangkan pada penelitian ini menggunakan VSM saja.

2.2 LANDASAN TEORI

2.2.1. Penentuan Sample

Dalam penelitian dibutuhkan pengambilan sampel dari populasi sebagai representasi populasi sehingga digunakan untuk mengangkat kesimpulan penelitian sebagai suatu yang berlaku bagi populasi. (Arikunto & Suharsimi, 2010) mengatakan bahwa sampel adalah sebagian atau reresentasi populasi yang diteliti.

Dalam penelitian ini penentuan jumlah sampel dilakukan dengan cara perhitungan statistik yaitu dengan menggunakan Rumus Slovin. Rumus tersebut digunakan untuk menentukan ukuran sampel dari populasi yang telah diketahui jumlahnya yaitu sebanyak 417 judul. Menurut (Sugiyono, 2017). Untuk tingkat presisi yang ditetapkan dalam penentuan sampel adalah 5 %.

Rumus Slovin : $n = N / (1+(N \text{ x } e\ 2))$

Dimana :

n : ukuran sampel

N : ukuran populasi

e : Kelonggaran ketidak telitian karena kesalahan pengambilan sampel yang dapat ditolerir, setelah itu dikuadratkan.

Berdasarkan Rumus Slovin, maka besarnya penarikan jumlah sampel penelitian adalah :

n= N / (1+(417x0,05))
n= 417 / (1+(417 X 0,0025)
n= 417 / (1+2,5)

n= 417 / 3,5
n= 119.142
Maka besar sample pada penelitian adalah sebanyak 119 setelah dibulatkan.

2.2.2. Text Mining

Text mining adalah proses menemukan informasi tersembunyi atau tidak diketahui yang berasal dari berbagai sumber menggunakan metode tertentu. Istilah ini sejatinya melakukan penggalian fakta yang berasal dari ektraksi dengan menggunakan eksperimen yang lebih konvensional (Heidarian dan Dinneen, 2016). Pada umumnya ketika melihat penjelasan singkat tentang *text mining* maka akan tampak sama dengan cara kerja sebuah mesin pencari. Namun hal tersebut adalah sesuatu yang berbeda sama sekali. Mesin pencari memiliki konsep mencari apa yang telah diketahui tetapi *text mining* adalah menemukan sesuatu yang belum diketahui sebelumnya. Bahkan menggunakan *text mining*, informasi dapat diekstraksi untuk memperoleh abstrak kata-kata yang ada dalam dokumen atau untuk menghitung abstrak untuk dokumen berdasarkan kata-kata yang terkandung di dalamnya. *Text mining* sebagai jenis penambangan data yang berbeda yang mencoba untuk menemukan pola yang menarik dari kumpulan data yang sangat besar. Masalah utama yang membedakan penambangan data reguler dari *text mining* adalah bahwa dalam *text mining* pola-pola diekstraksi dari teks bahasa alami dan bukan dari database fakta yang terstruktur.

2.2.3. Text Preprocessing

Text preprocessing adalah sebuah proses yang digunakan untuk merapikan data yang tadinya tidak terstruktur, tujuannya adalah agar data tersebut dapat diproses lebih lanjut sesuai kebutuhan pemprosesan. (Okfalisa & Harahap, 2016). Pada umumnya terdapat beberapa tahap yang dilakukan pada proses *text processing* yaitu *case folding, lexical analysis, stop-removal* dan *stemming* (Fhadli, Fauzi & Afirianto, 2017).

2.2.3.1. *Case Folding*

Case folding adalah proses untuk mengubah huruf pada dataset menjadi huruf kecil secara keseluruhan. Tujuannya ialah mencegah perbedaan interpretasi kata yang sama diakibatkan salah satu dari 2 kata yang sama tersebut memiliki huruf kapital, maka untuk mencegah itu maka akan dilakukan proses ini.

2.2.3.2. *Lexical Analysis*

Proses lexical analysis digunakan untuk menghapus karakter yang tidak diperlukan pada dataset seperti angka, tanda baca, karakter kosong. Setelah itu, karakter tadi akan dipisah per huruf dinamakan token dan prosesnya disebut sebagai proses *tokenizing*. Namun dari keseluruhan proses tersebut harus di support oleh corpus yang *support* terhadap bahasa Indonesia seperti corpus sastrawi.

2.2.3.3. *Stop removal*

Setelah melakukan proses case folding dan lexical analysis maka akan masuk tahap stop removal dimana akan dibentuk stoplist yang akan digunakan untuk memfilter kata pentinfg saja. Kata yang akan dihapus seperti kata hubung seperti dan, yang serta, setelah dan lainnya serta kata-kata yang belum jelas interpretasinya misalkan aku,saya, kami dan sebagainya. Maka akan tersisa kata yang penting-penting saja dari keluaran proses ini. Tentunyadalam penelitian ini aka nada penambahan manual kata yang akan di eleminasi seperti kata penerapan, perancangan dan lainnya karena fokus penelitian ini melakukan ekstraksi topik penelitian saja.

2.2.3.4. *Stemming*

Proses selanjutnya adalah melakukan penghapusan akata yang memiiki imbuhan suffix dan prefix sehingga kata akan kembali kedalam bentuk dasarnya tanpa imbuhan, proses ini akan mudah dikarenakan data yang digunakan hampir seluruhnya menggunakan kata baku, tidak seperti dataset yang menggunakan data tweet yang memerlukan normalisasi kata lagi untuk membuat kata semakin berkualitas sehingga proses inilah akhir dari tahap *preprocessing* (Jelita, 2009).

2.2.4. *Stemmer Truncating* Menggunakan *Sastrawi*

Library sastrawi merupakan pengembangan dari algoritma stemming nazief & adriani yang semula hanya tersedia untuk bahasa pemograman PHP (I Made AA, 2018). Sebelumnya ada pula percobaan yang tujuan penelitiannya adalah meneliti keunggulan beberapa stemmer truncating dalam Bahasa Indonesia seperti algoritma porter, NLTK, nazief&adriani, namun dalam penelitian mengatakan bahwak stemer satrawi lebih baik kinerjanya ketimbang algoritma lainnya. Stemmer Sastrawi adalah sebuah alat yang digunakan untuk mengekstrak kata dasar (stem) dari kata yang berubah-ubah (inflected) dalam bahasa Indonesia. Alat ini dikembangkan oleh Tom De Smedt dan Yusuke Shinyama dari Universitas Indonesia. Stemmer Sastrawi menggunakan algoritma pemotongan akhiran yang sederhana untuk mengekstrak kata dasar dari kata yang berubah-ubah. Algoritma ini menghapus akhiran yang dikenal seperti -i, -an, -kan, -kah, -lah, -tah, -nya, dan -pun dari kata.

Contohnya, kata "berjalan" akan diubah menjadi "jalan" setelah di-stem oleh alat ini. Kata "menulis" akan diubah menjadi "tulis" setelah di-stem. Kata dasar yang dihasilkan dapat digunakan dalam berbagai aplikasi seperti Information Retrieval, Natural Language Processing, dan Text Mining.

Stemmer Sastrawi juga memiliki fitur yang disebut "removal of derivational suffixes" yang dapat digunakan untuk menghapus akhiran yang digunakan untuk membuat kata turunan (contoh: -isasi, -is, -asi, -kan, -i) yang akan menghasilkan kata dasar yang lebih spesifik. Stemmer Sastrawi juga dapat digunakan secara

efektif dengan menggunakan kombinasi beberapa algoritma seperti Case folding, stop word removal, dan stemming.

Secara umum, Stemmer Sastrawi dianggap sebagai salah satu stemmer yang efektif dan cepat untuk bahasa Indonesia. Alat ini tersedia dalam beberapa bahasa pemrograman seperti Python, Java, dan PHP.

2.2.5. *Stemmer Statistical* Menggunakan *N-Gram*

N-gram adalah metode yang digunakan untuk memecah teks menjadi sekumpulan kata yang disusun secara berurutan (Juldta Priess, 2019). Algoritma ini mengambil sekumpulan kata yang terdiri dari N kata dan menganggapnya sebagai satu unit. Beberapa perbedaan antara N-gram dengan algoritma lain dari segi akurasi dan ketepatan adalah sebagai berikut:

Akurasi: N-gram memiliki tingkat akurasi yang lebih tinggi dibandingkan dengan algoritma lain seperti stemming dan lemmatization. Hal ini dikarenakan N-gram mempertahankan konteks kata yang berdekatan, sehingga dapat mengenali kata yang berubah-ubah dengan lebih baik.

Ketepatan: N-gram memiliki tingkat ketepatan yang lebih rendah dibandingkan dengan algoritma lain seperti deep learning model seperti BERT, GPT-2, GPT-3. N-gram tidak dapat mengenali relasi antar kata dengan baik, sehingga dapat menghasilkan kesalahan dalam pengenalan kata.

Scalability: N-gram dapat digunakan untuk menangani data yang besar dan tidak terstruktur dengan baik. Namun, jika digunakan dengan besar N, maka akan membutuhkan memori yang besar dan proses yang lama.

Generality: N-gram lebih general dibandingkan dengan algoritma lain seperti deep learning model karena dapat digunakan untuk berbagai jenis bahasa dan tidak memerlukan data yang besar. Namun, algoritma ini tidak dapat mengenali relasi yang kompleks antar kata.

Secara umum, N-gram dapat digunakan sebagai alat yang efektif untuk pengenalan kata dan analisis teks, namun memiliki keterbatasan dalam hal ketepatan dan relasi antar kata. Namun, digunakan dengan baik, N-gram dapat memberikan hasil yang baik.

2.2.6. *Sentence Embedding* Menggunakan *Word2BoW*

Word2BoW (Word to Bag-of-Words) adalah metode yang digunakan untuk mengubah teks menjadi sekumpulan frekuensi kata. Metode ini mengambil sekumpulan kata dari teks dan menganggap setiap kata sebagai satu unit. Word2bow lebih baik dibandingkan dengan algoritma lain dalam kasus topic modeling karena memiliki beberapa keunggulan yakni Word2bow memiliki interpretabilitas yang lebih baik dibandingkan dengan algoritma lain karena setiap vektor dapat diinterpretasikan sebagai kata yang mewakili topik. Word2bow juga lebih mudah digunakan dan diimplementasikan dibandingkan dengan algoritma lain seperti Word2Vec. Word2bow dapat digunakan dengan baik pada data yang

tidak terstruktur dan besar. Word2bow dapat digunakan dengan berbagai metode seperti LDA, LSA, dan CTM untuk menemukan topik. Secara umum, Word2bow dianggap sebagai metode yang efektif dan mudah digunakan dalam topic modeling. Namun, metode ini memiliki keterbatasan dalam hal relasi antar kata dan konteks kata. Namun, digunakan dengan baik, Word2bow dapat memberikan hasil yang baik.

2.2.7. Topic Modeling

Topic modeling adalah proses pemodelan topik menggunakan pendekatan *unsupervised learning*, hasil yang didapatkan adalah hasil proses clustering yang terdiri dari hasil klasifikasi tanpa label. Dalam konsep topic modeling akan dikelompokkan dokumen berdasarkan kemiripannya masing-masing (Steyvers & Griffiths T, 2007). Tujuannya adalah menemukan pola topik dari sekian banyak dokumen yang dianalisis. Sehingga konsep ini dapat diterapkan dalam mencari topik tersembunyi dari banyak dokumen sekaligus. Contohnya dapat diterapkan pada pengelompokan review suatu produk pada aplikasi *playstore*. *Topic modeling* membantu melihat topik review mana yang hangat dibicarakan oleh para pengguna aplikasi sehingga memudahkan pihak developer maupun owner layanan untuk melakukan evaluasi maupun peningkatan layanan.

2.2.8. Latent Dirichlet Allocation (LDA)

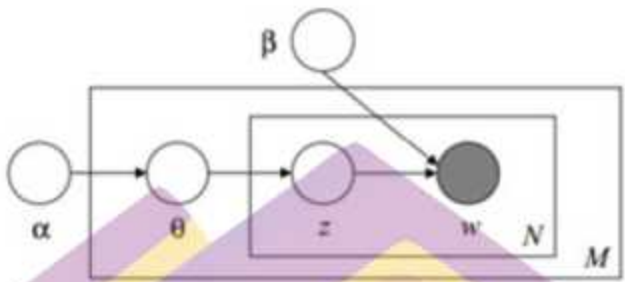
LDA adalah sebuah model yang digunakan untuk melakukan proses representatif eksplisit pada sebuah dokumen. Dikarenakan model LDA ini adalah model probabilistik generatif pada kumpulan data teks atau *corpus*, sehingga data teks dapat dimodelkan sebagai model campuran dari berbagai topik. Sedangkan untuk tipe model LDA ini merupakan tipe model Bayesian herarki (Steyvers & Griffiths T, 2007).

LDA mengasumsikan proses generative berikut untuk setiap dokumen ***w*** dalam sebuah corpus ***D*** adalah sebagai berikut:

1. Pilih N ~ *Pissson* (ξ)
2. Pilih θ ~ Dir(α)
3. Untuk setiap N kata wn :
   a. Pilih Topik z ~ Multinomial(θ)
   b. Pilih sebuah kata wn dari (| ,β)

Beberapa asumsi penyederhanaan yang dibuat dalam model dasar LDA. Pertama, distribusi dari topik (latent) diketahui mengikuti k distribusi Dirichlet. Kedua, probabilitas kata adalah matriks β berukuran k ×V yang mana = (=1| =1) j i ij β p w z (Blei et al, 2003).

Dalam konsep LDA sebagai model *probabilistic* secara visual pada **gambar 2.1** berikut.



Gambar 2.1 Model Representasi LDA (Sumber Blei et al, 2003)

Dapat dilihat gambar 2.1 diatas terdapat tingkatan pada pemodelan dengan LDA dimana ada 2 parameter yaitu parameter α dan β sebagai parameter distribusi topik yang berada pada tingkatan corpus ialah kumpulan dan M sebagai dokumen.

Penjelasan terkait parameter alpha atau α yang digunakan untuk menentukan distribusi topik dokumen yaitu jika nilai alpha atau α semakin besar menandakan dalam suatu dokumen terdpat campuran topik yang dibahas dalam dokumen makin banyak. sedangkan parameter beta atau ß yang digunakan untuk menentukan distribusi kata dalam topik. Jika nilai beta semakin tinggi, maka semakin banyak pula kata atau *term* yang terdapat di dalam topik tetapi begitupun sebaliknya jika nilai beta atau ß semakin kecil, maka menandakan semakin sedikit kata atau *term* yang terdapat di dalam topik yang berarti topik tersebut mengandung kata atau *term* yang lebih spesifik pada variabel θ_m yaitu variabel yang terdapat

pada tingkat dokumen (M) yang dapat dilihat pada gambar. Variabel θ atau *theta* menggambarkan suatu distribusi topik untuk dokumen-dokumen. Jika nilai θ atau *theta* semakin tinggi, maka menandakan semakin banyak topik yang terdapat di di dokumen, jika nilai θ atau *theta* semakin kecil, maka menandakan semakin spresifik pada topik tertentu. Pada variabel Z_n dan W_n yaitu variabel tingkat kata (N) Variabel Z menggambarkan suatu topic pada dokumen yang terdapat kata pada dokumen tersebut, pada variabel W menggambarkan kata yang berhubungan dengan suatu topik tertentu didalam dokumen. Berdasarkan penjelasan diatas maka proses generatif pada LDA akan berkorespondensi kepada joint distribution dari suatu *hidden variabel* dan *obsessed variable*. Berikut ini adalah formula perhitungan probabilitas dari sebuah corpus berdasarkan notasi.

$$P(\theta|\alpha,\beta) = \prod_{d=1}^{N^d} \int p(\theta_d|\alpha)\,(\prod_{d=1}^{N^d} \Sigma_{Z_{dn}} p(Z_{dn}|\theta_d)\; p(W_{dn}|Z_{dn},\beta))\, d\theta_d \ldots (1)$$

Dapat dilihat diatas pada notasi ß menggambarkan topik pada setiap ß yang merupakan distribusi dari sejumlah kata atau *term*. Sedangkan variabel θ_d adalah variabel level dokumen dengan 1 kali sampel/dokumen yang menggambarkan proporsi topik untuk dokumen ke *d*. Pada notasi Z_{dn} dan W_{dn} yakni representasi variabel di level kata atau term dengan 1x sampel untuk masing-masing kata pada setiap dokumennya.

2.2.9. Arsitektur Gensim

Arsitektur Gensim adalah sebuah kerangka kerja yang digunakan untuk melakukan topic modeling dan analisis teks dalam bahasa Python. Beberapa alasan mengapa arsitektur Gensim dianggap lebih baik dibandingkan dengan arsitektur lain dalam penerapan topic modeling.

Gensim memiliki arsitektur yang fleksibel dan dapat digunakan dengan berbagai metode topic modeling seperti Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), dan Correlated Topic Models (CTM). Dari segi performa, Gensim dikembangkan dengan menggunakan teknologi parallel computing dan memiliki performa yang baik dalam menangani data yang besar dan tidak terstruktur. Gensim memiliki arsitektur yang modular, sehingga dapat digunakan dengan mudah dalam berbagai aplikasi seperti Information Retrieval, Natural Language Processing, dan Text Mining. Gensim menggunakan teknologi memory-efficient untuk menyimpan matriks dalam memori, sehingga dapat digunakan dengan efisien pada komputer dengan memori terbatas. Secara umum, Gensim dianggap sebagai kerangka kerja yang efektif dan mudah digunakan dalam topic modeling dan analisis teks. Namun, kerangka kerja ini memiliki keterbatasan dalam hal relasi antar kata dan konteks kata. Namun, digunakan dengan baik, Gensim dapat memberikan hasil yang baik.

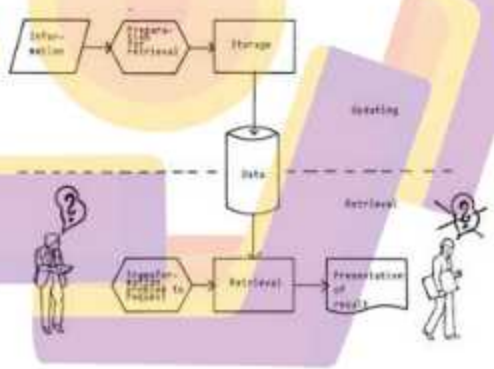
2.2.10. *Information Retrieval* (IR)

Information retrieval (IR) atau sistem temu kembali informasi adalah sebuah metode yang digunakan untuk temu balik informasi yang efisien dengan 3 tujuan yang akan dicapai yakni proses indexing/pencarian dokumen, pemprosesan query atau kata kunci user dengan kecepatan tinggi serta melakukan sebuah proses perankingan berdasarkan pada hasil pengukuran relevansi antara dokumen dak kata kunci/*query* yang dikeluarkan. Pada konsep ini dapat diterapkan pada data yang *unstructured* dan *semi-unstructered. Data unstructured* adalah data memiliki bentuk yang abstrak atau tidak terorganisir atau tidak memiliki model yang bak. Hal itu dikarenakan data tersebut bisa saja memiliki data berupa angka, tanda baca, tanggal dan lainnnya. Sedangkan data *semi-unstructered* adalah data yang memiliki format tetap atau baku namun pada isi data tersebut terdapat jenis data yang berdbeda contoh dalam satu tabel namun dalam data tersebut ada data yang berbeda meskipun dalam 1 tabel (M.Geetha, N.Vimala, 2020).

Ada beberapa karakteristik *information retrieval* (IR) atau sistem temu kembali informasi yaitu:

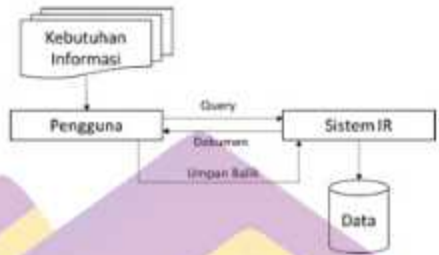
1. Dokumen korpus, bentuk data yang akan diimplemntasikan dalam sistem mesti ditentukan sejak awal, pilihannya bisa saja dalam bentuk paragraf, halaman / teks.
2. Query User, kata kunci yang diminta oleh user berupa query yang dimasukkan dalam suatu kolom pencarian yang bisa saja berbentuk 1 kata atau frasa.

3. Kumpulan Hasil Ranking, sebuah sistem temu kembali informasi (IR) memiliki output hasil berupa perankingan dokumen yang relevan dengan apa yang terdapat pada kata kunci yang dimasukkan oleh pengguna atau user.
4. Presentasi hasil yang ditetapkan, selain dari mengeluarkan hasil ranking dokumen sistem temu kembali informasi (IR) juga mempresentasikan dalam bentuk judul dokumen yang diranking sebagai representasi vector dokumen hal tersebut merupakan pelabelan pada dokumen yang akan diranking.



Gambar 2.2 Proses dari Information Retrieval. (Sumber: Suhartono 2013)

Maka dari itu dalam implementasinya sebuah sistem temuk meblai informasi (IR) sangat erat kaitannya antara kebutuhan informasi dan sistem index dokumen. Hubungan itu dapat ilustrasikan dalam gambar berikut ini:



Gambar 2.3. Arsitektur Sistem *Information Retrieval*. (Sumber: Budiharto 2016)

2.2.11. *Vector Space Model* (VSM)

VSM sering digunakan untuk menginterpretasikan sebuah dokumen dalam sebuah ruang *vector*. VSM membangun sebuah model *information retrieval* yang memasukkan dokumen dalam sebuah *vector* yang sebelumnya diolah menggunakan teknik *preprocessing*, bukan cuma itu tetapi *query* atau kata kunci yang digunakan sama-sama dimasukkan dalam *vector* ruang multidimensi. (Venkat N. Gudivada, Dhan L Rao & Amogh R, 2018)

Ciri Khas model ruang *vector* multidimensi adalah :

1. *Model vector* dibentuk menggunakan *keyterm*
2. *Model vector* berbentuk partial matching atau sebagian sesuai dan menjadi penentuan peringkat dokumen.

3. Prinsip utama *model vector* adalah baik dokumen dan *query* atau kata kunci sama-sama direpresetasikan menggunakan *vector keyterm,* sedangkan ruang dimensi ditentukan oleh *keyterm.* Maka dari itu untuk menghitung kesamaan atau kemiripan dokumen *keyterm* dihitung menggunakan jarak *vector.*

4. Agar model ruang *vector* dapat bekerja maka diperlukan bobot *keyterm* untuk vector dokumen, normalisasi *keyterm* untuk vector dokumen dan kata kunci/*query* dan yang terkahir adalah perhitungan jarak untuk *vector* dengan *keyterm.*

5. Dalam pengukuran kinerja model ruang vector maka yang dilihat adalah efisiensi, bentuk model ruang vector yang dapat dengan mudah melakukan suatu representasi serta dapat mengimplementasikan pada pengindeksan dkumen dan pembobotan indeks agar menghasilkan dokumen yang relevan.

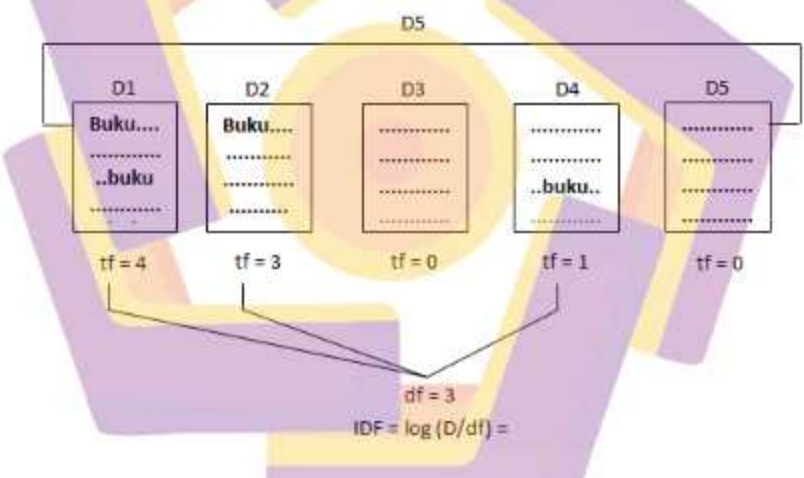
Dalam implementasi yang digunakan dalam metode *TF-IDF (Term Frequency Inverse Document Frequency)* yaitu dengan melakukan suatu pembobotan hubungan antar kata (*term)* terhadap suatu dokumen. Metode ini menggunakan pendekatan perhitungan berdasarkan 2 yaitu pertama frekuasi kemunculan kata dalam sebuah dokumen tertentu dan kedua *inverse* frekuensi dokumen yang terdapat kata tersebut (Budiharto, 2016).

Rumus untuk menghitung *term frequency (tf)* yang merupakan perhitungan jumlah kemunculan suatu kata di dokumen *idf* yaitu menggunakan formula berikut:

$$IDF = \log \frac{D}{Df} \quad \text{..... (2)}$$

Keterangan :

IDF = Invers dokumen frekuensi (IDF)
D = Total Dokumen
Df = Frequensi dokumen dari term
Log = Untuk memperkecil pengaruhnya relative terhadap term frequency (tf)



Gambar 2.4. Ilustrasi Algoritma Penamban

D1, D2, D3, D4, D5= dokumen.

Tf = banyaknya kata yang dicari pada sebuah dokumen.

D = total dokumen.

df = banyak dokumen yang mengandung kata yang dicari.

Dalam hal pembobotan kata yang terjadi didalam proses perhitungan ada factor yang memegang peranan penting yaitu sebagai berikut:

Pendekatan dalam pembobotan dokumen paling banyak diimplementasikan adalah *term frequency* (tf). Faktor ini melihat banyaknya kemunculan suatu kata dalam suatu sebuah dokumen. Kemunculan kata yang semakin sering dalam dokumen, maka kata tersebut dianggap kata yang penting. Ada 4 cara yang bisa digunakan untuk menghasilkan nilai *TF* (Budiharto. W, 2016):

a. *Raw Tf*

Nilai *Term frequency (tf)* sebuah kata atau *term dapat* dihitung berdasarkan kemunculan kata atau *term* tersebut dalam sebuah dokumen.

b. *Logarithmic Tf*

Dalam memperoleh nilai *term frequency (tf)* cara ini menggunakan pendekatan berbeda yaitu memanfaatkan fungsi logaritmik dalam dunia matematika.

$$TF = 1+\log(TF) \quad \ldots\ldots (3)$$

c. *Binnary Tf*

Pendekatan berbeda ditawarkan dengan cara memanfaatkan nilai *Boolean* yang seperti kita ketahui hanya terdiri dari 0 dan 1, jika terdapat dalam dokumen maka aka direpresentasikan bernilai *true* atau 1 begitupun sebaliknya. Sehingga banyaknya kemunculan *term* pada dokumen tidak berpengaruh.

d. *Augmented Tf*

$$TF = 0.5 + 0.5 \times TFmax(TF) \quad \ldots\ldots (4)$$

Nilai *term terquency (tf)* adalah jumlah kemunculan kata atau *term* pada dokumen. Nilai dari max(Tf) menandakan total kemunculan terbanyak akta atau *term* pada sebuah dokumen yang sama. Perhitungan *term frequency (tf)* yang akan digunakan dalam implementasi *information retrieval* (sistem temu kembali informasi) pada sebuah sistem yang dibangun adalah *Raw Tf.*

Rumus yang dipakai dalam menghitung bobot (w) pada masing-masing dokumen terhadap kata kunci atau *query* yaitu:

$$W_{dt} = tf_{dt} * IDF \quad \text{..... (5)}$$

Keterangan :
W = bobot dokumen ke-d terhadap kata ke-t
Tf = kata ke-t dari kata kunci
IDF = Invers dokumen frekuensi
d = dokumen ke-d
t = kata ke-t dari kata kunci

Maka untuk melihat tingkat kemiripan atau similaritas dokumen terhadap kata kunci atau *query* maka dapat dilihat nilai bobot (w). bobot (w) harus diketahui terlebih dahulu lalu diurutkan berdasarkan nilai W terbesar ke terkecil, maka nilai bobot kemiripan yang terbesarlah yang memiliki kemiripan antara dokuen dan kata kunci begitupun sebaliknya.

Pada proses pengurutan maka dokumen dirutukan berdasarkan ukuran kesamaan (*similarity measure).* Perhitungan similaritas antara *vector* kata kunci/*query* dan *vektor* dokumen dilihat dari sudut yang paling kecil. Sudut tersebut ada karena oleh dua buah *vektor* dapat dihitung dengan cara *inner product*

(Graham L Giler, 2012). Menghitung Log a akan dikalkulasikan dengan menggunakan pendekatan perhitungan *Cosine Similarity* dengan nilai kemiripan *(similarity)* antara kata kunci/*query* dengan dokumen ke-j adalah :

$$CosSim(dj, q) = \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}| \cdot |\vec{q}|} = \frac{\sum_{i=1}^{t}(w_{ij} \cdot w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{ij}^{2} \cdot \sum_{i=1}^{t} w_{iq}^{2}}} \quad \ldots\ldots(6)$$

Dimana :

$\vec{d_j}$ = Bobot dokomen j
$\vec{q}$ = Bobot *query*
$|\vec{d_j}|$ = Panjang Bobot Dokumen j
$|\vec{q}|$ = Panjang Bobot *query*
W_{ij} = Bobot *term* pada dokumen j
W_{iq} = Bobot *term* pada *query*
T = Teks

2.3. KEASLIAN PENELITIAN

Tabel 2.1 Matriks literatur review dan posisi penelitian

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
1	A Review of Topic Modeling Methods	I. Vayansky and S.A.P. Kumar, Elsevier Ltd, 2020	Membandingkan penerapan berbagai metode dalam membuat model *topic modeling*	Pengguaan metode bergantung dari kebutuhan penggunaan *topic modeling* itu sendiri, sehingga dalam pembuatan model tersebut dapat efisien.	Kurang ditampakkan penerapan secara luas tentnag penggunaan tiap metode serta dalam tiap sub pembahasan terkait metode tidak ada proses hitungan dan rumus yang ditampilkan tetapi hanya sebagian saja.	Dalam makalah ini menjadi pembanding saja dalam menentukan metode yang tepat dalam memilih yang cocok dalam penerapan *topic modeling*
2	Penerapan *Text Mining* Untuk Pengelompokan Penelitian Dosen	Eka Sabna, Jurnal Ilmu Komputer, 2020	Mengelompokkan penelitian-penelitian yang telah dilakukan oleh dosen	Dari hasil klustering didapatkan kata yang menjadi tren penelitian dosen program studi Ilmu	Dalam penelitian ini tidak tergambar secara nyata sesuai dengan tujuan dari penelitian ini seperti.	Dalam melakukan pengelompokan topik penelitian meskipun sama-sama menggunakan unsupervised learning

Tabel 2.1. (Lanjutan) Matriks literatur review dan posisi penelitian

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
			menggunakan algoritma K-Means agar hasil pengelompokan dapat dijadikan rujukan tren penelitian tiap dosen. Selain itu dapat dijadikan referensi oleh jurusan untuk menentukan dosen pembimbing dan penguji.	keschatan masyarakat (IKM) tahun 2020 yang kebanyakan mengangkat topik yang terkait dengan Covid dan HIV	diuraikan pada latar belakang penelitian. Seharusnya dalam penelitian ini digambarkan tren penelitian tiap dosen 3 tahun sebelumnya hingga sekarang sehingga tujuan penelitian dapat sesuai dengan rumusan masalah. Serta kesimpulan hanya menghasilkan tren berdasarkan objek penelitian bukan topik penelitian.	namun dalam penelitian ini menggunakan K-Means serta hasil dari k-means tidak diolah sedemikian ruap untuk mengekstrak informasi lebih jauh sebagai *method improvement*-nya.

Tabel 2.1. (Lanjutan) Matriks literatur review dan posisi penelitian

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
3	Pemodelan Topik Pengguna Twitter Mengenai Aplikasi "Ruangguru"	Bagus Wicaksono Arianto & Gangga Anuraga, Jurnal ILMU DASAR, 2020	Melakukan pengelompokan persepsi pengguaan aplikasi ruangguru di twitter menggunakan LDA.	Kesimpulan penelitian ini mendapatkan 28 buah topic hasil dari klasterisasi topik yang dihasilkan dari metode *Latent Direchlet Allocation (LDA)* dan topik yang paling banyak ialah berkaitan dengan diskon ruang guru.	Sumber rujukan penelitian masih kurang serta tidak ada sama sekali yang berasal dari jurnal internasional.	Sumber data nya berasal dari twitter yang menulis twet tentang aplikasi ruangguru serta hasil dari topik yang dihasilkan tidak diolah lebih lanjut, peneliti menyimpulkan sendiri hasil topik yang dihasilkan model.
4	Aplikasi Topic Modeling Pada Pemberitaan Portal Berita Online Selama Masa Psbb Pertama	Wahyudin, Seminar Nasional Official Statitstic, 2020	Memetakan topik pemberitaan tentang PSBB pertama dari berbagai portal berita online.	pemberitaan 9 media online nasional dapat dikelompokkan kedalam 4 kelompok besar pemberitaan:	Tidak ditampilkan nilai koherensi terbaik dalam tiap iterasi yang sebenarnya bisa menjadi informasi	Sumber data berasal dari portal pemberitaan online dan hasil dari topic modeling menggunakan LDA tidak diolah lebih lanjut hanya sekedar menafsir maksud

Tabel 2.1. (Lanjutan) Matriks literatur review dan posisi penelitian

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
				Perkembangan perkembangan kasus Covid-19, penanganan dampak pandemi, seputar Ramadhan dan kegiatan belajar di rumah, serta berita lainnya, diantaranya seputar PSBB, Kriminal, Hiburan dan Politik		Trem-term per topik yang dihasilkan, yang artinya tidak melakukan evaluasi dari hasil yang didapatkan misalkan menggunakan metode koherensi.
5	Topic Modelling Skripsi Menggunakan Metode Latentdiriclhet Allocation	Alif Iffan Alfanzar, Khalid & , Indri Sudanawati Rozas, Jurnal Sistem Informasi (JSil), 2020	Memetakan minat penelitian pada program studi bahasa inggris menggunakan LDA.	Dari 584 abstrak yang dikumpulkan dilakukan percobaan sebanyak 5 iterasi percobaan dengan total iterasi 100, 500, 1000 dan 5000.	Lebih baik lagi jika menampilkan table hyperparameter tuning hasil percobaan berdasarkan jumlah k dan iterasi dan juga menampilkan	Dataset menggunakan abstak bahasa inggris. Sedangkan dala penelitian ini menggunakan abstrak bahasa Indonesia. Dilain sisi ada penambahan metode *information retrieval*.

Tabel 2.1. (Lanjutan) Matriks literatur review dan posisi penelitian

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
				Dan setiap uji dimasukkan nilai k yaitu 2, 3, 4 dan 5. Dan didapatkan nilai k terbaik adalah 3 dan hasilnya terkonfirmasi oleh pihak stakeholder program studi Bahasa inggris.	Nilai perplexity dan koherensi untuk melengkapi informasi dalam penelitian ini.	
6	An Art of Review on Conceptual based Information Retrieval	P. Mahalakshmi & N. Sabiyath Fathima, Webology, 2020	Menjelaskan berbagai representasi teks yang bertanggung jawab atas mengambil hasil	Didapatkan pemetaan yang jelas terkait perkembangan teknologi *information retrieval* berdasarkan strategi yang disusun	Tidak ditampilkan contoh penelitian dan juga jika bisa ditampilkan diagram berbentuk decession tree untuk menentukan metode yang tepat / metode	Pada penelitian menggambarkan 4 kriteria penelitian tentang *information retrieval* yang digunakan dan bisa dijadikan pijakan untuk mengetahui pendekatan seperti apa yang

Tabel 2.1. (Lanjutan) Matriks literatur review dan posisi penelitian

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
			Pencarian yang relevan, pendekatan serta evaluasi yang dilakukan di dalam pengembilan informasi konseptual	Atas 4 kriteria yaitu representation of teks dalam bentuk apa, pendekatan perhitungan yang digunakan sumber data terakhir atau evaluasi untuk mengukur akurasi dan performance model	Pendekatan yang tepat dalam suatu contoh kasus	Digunakan oleh pembaca makalah ini, sedangkan dalam penelitian ini diputuskan untuk menggunakan salah satu, metode yang digunakan yaitu menggunakan VSM dengan rumus perhitungan *cosine similarity*,
7	DL-VSM based document indexing approach for information retrieval	Kabil Boukhari & Mohamed Nazih Omri, Journal of Ambient Intelligence and Humanized Computing, 2020	Menerapkan sistem temu balik informasi yang dapat mengindeks lebih baik pada dokumen biomedis dengan menggunakan VSM dan DL	Berhasil melakukan ekstraksi konsep dari semua dokumen biomedis dengan menggabungkan statistika dan metode semantic dari dua metode	Dalam penelitian ini tidak ditampilkan problem normalisasi kata dalam kamus biomedis yang dimaksud dalam penelitian ini agar pembaca mendapat gambaran dari maksud peneliti.	Pada penelitian ini Penerapan information retrieval menggunakan penggabunagan konsep VSM dan DL sedangkan dalam penelitian ini hanya menggunakan VSM.

BAB III

METODE PENELITIAN

3.1. Jenis, Sifat dan Pendekatan Penelitian

3.1.1. Jenis Penelitian

Jenis penelitian yang digunakan oleh peneliti merupakan penelitian kuantitatif, dimana peneliti melakukan perhitungan matematis untuk menemukan hasil yang di inginkan berupa sistem yang padat mengetahui tren topik yang akurat menggunakan model *topic modeling* dengan metode LDA serta mengetahui relevansi hasil *coherence score* model LDA dengan model IR dengan pendekatan metode VSM.

3.1.2. Sifat Penelitian

Sifat dari penelitian yang akan dilakukan adalah ekspremental dimana peneliti melakukan sebuah eksperimen guna mengetahui tren topik yang akurat menggunakan model *topic modeling* serta mengetahui relevansi hasil *coherence score* model LDA dengan model IR.

3.1.3. Pendekatan Penelitian

Pendekatan kuantitatif di gunakan oleh peneliti dimana penelitian akan melakukan penelitian sesuai alur yang telah peneliti buat.

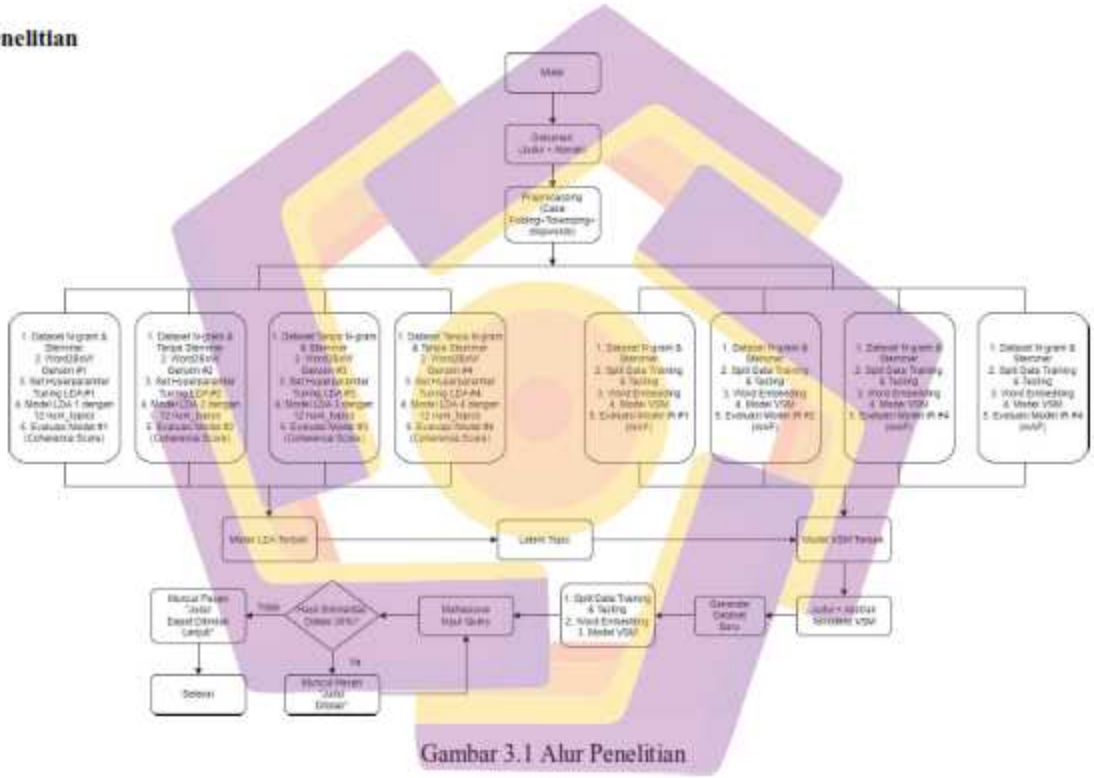
3.2. Metode Pengumpulan Data

Metode yang digunakan adalah studi dokumen, data akan dikumpulkan dari perpustakaan Universitas Dipa Makassar. Skripsi yang dikumpulkan dari 1 program studi yakni jurusan teknik informatika saja. Jumlah data yang akan diambil sebagai sample berjumlah 430 judul beserta abstrak dalam bentuk teks.

3.3. Metode Analisis Data

Analisis data dimulai dari memproses data penelitian berupa data teks abstrak penelitian dalam proses *text preprocessing*. Tahap ini terdiri dari *case folding, lexical analysis* serta *stemming*. Data dokumen lalu dimasukkan dalam model LDA untuk diolah dengan 36 kali percobaan berdasarkan jumlah variabel yang berbeda. Nantinya nilai model dengan *score* koherensi terbesar akan diambil sebagai model yang akan menghasilkan *latent topic*. Sedangkan *latent topic* inilah yang akan menjadi *query* atau kata kunci pada model *information retrieval* yang akan dibuat sehingga menghasilkan data judul yang berhubungan dengan *latent topic* yang dihasilkan tadi.

3.4. Alur Penelitian



Gambar 3.1 Alur Penelitian

3.4.1. Pendekatan Penelitian

Pada tahap ini yang terdiri dari studi literatur, identifikasi masalah serta rumusan masalah. Tahap tersebut melakukan validasi penelitian tentang permasalahan serta studi literatur untuk mengetahui masalah yang akan di analisa serta penggunaan algoritma yang tepat berdasarkan studi literatur sehingga dalam pemilihan algoritma tersebut didasarkan pada hasil ilmu yang dianalisa oleh penelitian sebelumnya

3.4.2. Pengumpulan Data

Data akan dikumpulkan dalam bentuk teks yang terdiri dari judul dan abstrak penelitian yang data akan dikumpulkan dari perpustakaan Universitas Dipa Makassar dan juga data-data pengajuan judul skripsi yang dikumpulkan dari jurusan teknik informatika dengan jumlah data 430 sample data. Data tersebut diambil dari periode tahun 2019 sampai dengan periode 2021.

3.4.3. Analisis Data

Pada tahap analisis data ini akan dijelaskan tentang preprocessing awal sebelum data tersebut di proses preproessing data kedua untuk 4 skenario dilakukan.

3.4.4. Case Folding

Proses case folding yaitu merubah semua karakter huruf pada sebuah kalimat menjadi huruf kecil dan menghilangkan karakter yang dianggap tidak valid seperti angka, tanda baca, dan Uniform Resources Locator (URL) ataupun nama akun pada media sosial seperti twitter.

3.4.5. Tokenezing

Proses tokenization berguna untuk memecah setiap kalimat dari seluruh dokumen pengetahuan ke dalam kata-kata (term) dengan menggunakan pembatas tab dan karakter spasi. Hal yang perlu dilakukan juga adalah menjadikan kata menjadi huruf kecil menghilangkan karakter tanda baca seperti tanda titik(.), koma (,), petik("'), kurung(()), tanda tanya(?), tanda seru(!) dan tanda baca lainnya. Data hasil dari proses tokenization

3.4.6. Stopremoval

Proses stopremoval adalah proses yang untuk menghapus data yang tidak memiliki interpretasi seperti kata ganti orang seperti saya, dia, kamu, mereka dan lain-lain serta kata hubung atau kata pertentangan seperti dan, atau, ketika, namun dan sebagainya. Proses

ini akan dilakukan menggunakan bantuan library NLTK dalam melakukan proses tersebut

3.4.4. Pembuatan Model LDA dan Menemukan Latent Topics

Pembuatan model menggunakan bantuan library gensim untuk membuat word2bow model. Dalam percobaan akan dilaksanakan sebanyak 48 percobaan yang menghasilkan 36 model LDA dimana akan ada 4 skenario utama dan dalam 1 skenario terdiri dari 12 percobaan yang mengikuti berdasarkan jumlah num_topics yang dipilih oleh peneliti yaitu 5, 6, 7, 8, 9, 10, 11, 12 dan 13. Skenario pertama ada 9 percobaan (Berdasarkan num_topics) dengan dataset yang menerapkan model n-gram serta stemmer sastrawi, scenario kedua ada 9 percobaan (Berdasarkan num_topics) dengan dataset yang menerapkan model n-gram tapi tanpa stemmer sastrawi, skenarion ketiga ada 9 percobaan (Berdasarkan num_topics) dengan dataset yang tidak menerapkan n-gram tetapi menggunakan stemmer sastrawi dan scenario keempat ada 9 percobaan (Berdasarkan num_topics) dengan dataset yang tidak menerapkan model n-gram dan juga stemmer sastrawi. Nantinya akan dipilih model yang memiliki nilai koherensi terbesar yang mewakili model terbaik dalam percobaan ini. Pengujian model sendiri akan menggunakan metode koherensi yang mengukur tingkat relevansi pada model yang dihasilkan.

3.4.5.. Pembuatan Model VSM Untuk Membuat Dataset Baru Berdasarakan Latent Topics

Model IR akan dibuat menggunakan algoritma VSM dengan dataset yang sama dengan yang digunakan oleh model VSM sebelumnya. Serta memiliki skenario yang sama dengan algoritma VSM sebelumnya dengan menerapkan dataset yang menerapkan n-gram dan stemmer, n-gram dan tanpa stemmer, tanpa n-gram dan tanpa stemmer. Sehingga total percobaan untuk menemukan model terbaik untuk IR adalah berjumlah 4 percobaan dengan mengevaluasi atau melihat hasil mean average precision (maP). Nilai maP terbesar dari keempat model tadi akan digunakan untuk menguji relevansi hasil yang diperoleh dari 4 model LDA berdasarkan coherence score-nya.

3.4.6. Membuat Dataset Baru Berdasarkan Model LDA Terbaik dan Model VSM Terbaik

Pada alur penelitian ini akan di load ulang model LDA dan juga hasil yang didapatkan berupa *latent topics* dengan jumlah topik yang merupakan hasil yang terbaik. Sedangkan skema scenario yang digunakan untuk memberlakukan penggunaan *stemmer truncating* dalam hal ini adalah library sastrawi dan *stemmer statistical* menggunakan n-gram dilihat dari hasil penelitian. Setelah itu akan di load model VSM untuk melakukan proses *indexing* terhadap *latent topics* tadi yang dianggap sebagai *query*

untuk disandingkan dalam vector yang dibuat bersamaan dengan dataset awal. Maka setelah proses perhitungan menggunakan VSM (*Cosine Similarity*) maka akan tersisa judul-judul yang tidak terindeks. Setelah itu data set yang terindeks akan digunakan untuk membuat dataset baru.

3.4.7. Membuat Model VSM Baru Untuk Berinteraksi Dengan Mahasiswa

Pada alur penelitian ini akan kita gunakan dataset baru yang sebelumnya menjadi output dari alur penelitian ketiga, setelah itu kita membuat model VSM baru dengan alur pembuatan yang tergantung dari hasil penelitian kedua apakah memilih penggunaan preprocessing yang melibatkan *stemmer truncating* menggunakan libarary sastrawi dan *stemmer statistical* menggunakan n-gram agar diharapkan scenario inilah yang bisa menghasilkan mAP tertinggi pada model VSM. Selanjutnya akan dilakukan split data traning dan testing lalu memeasukkan pada *word embedding*, dan mahasiswa selanjutnya memasukkan kata kunci / *query* dan akan dihitung hasil similaritasnya menggunakan VSM (*Cosine similarity*) jika dalam pedetekesian similaritasnya melebihi 20 persen maka akan dimunculkan pesan "Judul Ditolak" dan jika tidak melebih maka akan memunculkan pesan "Judul Dapat Ditindak Lanjuti".

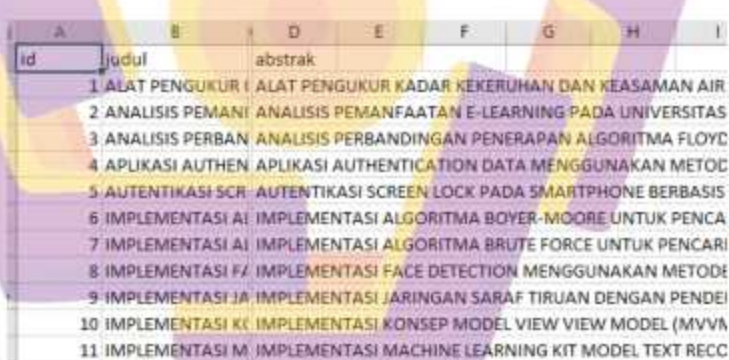
BAB IV

HASIL PENELITIAN DAN PEMBAHASAN

4.1 Persiapan Data

4.1.1 Pengumpulan Data

Hasil pengumpulan data dari jurusan Teknik Informatika Universitas Dipa Makassar dengan mengambil sampel dari 6 periode skripsi mulai dari tahun 2019 sampai dengan 2021 dengan jumlah sampel 430 data yang kemudian dilakukan proses akuisisi dan dipindahkan kedalam file excel agar mudah untuk diolah lebih lanjut. Berikut hasil akuisisi data setelah dipindahkan kedalam file excel.



	A	B	D	E	F	G	H	I
	id	judul	abstrak					
	1	ALAT PENGUKUR I	ALAT PENGUKUR KADAR KEKERUHAN DAN KEASAMAN AIR					
	2	ANALISIS PEMANI	ANALISIS PEMANFAATAN E-LEARNING PADA UNIVERSITAS					
	3	ANALISIS PERBAN	ANALISIS PERBANDINGAN PENERAPAN ALGORITMA FLOYD					
	4	APLIKASI AUTHEN	APLIKASI AUTHENTICATION DATA MENGGUNAKAN METOD					
	5	AUTENTIKASI SCR	AUTENTIKASI SCREEN LOCK PADA SMARTPHONE BERBASIS					
	6	IMPLEMENTASI AI	IMPLEMENTASI ALGORITMA BOYER-MOORE UNTUK PENCA					
	7	IMPLEMENTASI AI	IMPLEMENTASI ALGORITMA BRUTE FORCE UNTUK PENCARI					
	8	IMPLEMENTASI FA	IMPLEMENTASI FACE DETECTION MENGGUNAKAN METODE					
	9	IMPLEMENTASI JA	IMPLEMENTASI JARINGAN SARAF TIRUAN DENGAN PENDEI					
	10	IMPLEMENTASI KO	IMPLEMENTASI KONSEP MODEL VIEW VIEW MODEL (MVVM					
	11	IMPLEMENTASI M	IMPLEMENTASI MACHINE LEARNING KIT MODEL TEXT RECC					

Gambar 4.1. Foto Hasil Akuisisi Data Dari Perpustakaan

4.1.2 *Case Folding*

Pada proses ini bertujuan untuk melakukan penyegaraman karakter pada data, selain membuat huruf dari kecil menjadi besar juga mengahpus

karakter yang tidak memliki interpretasi seperti tanda baca dan angka. Hal ini menguntungkan secara komputasi.

```
kalimat = "Berikut ini adalah 5 negara dengan
Pendidikan terbaik di dunia adalah korea Selatan,
jepang, singapura, hongkong dan finlandia"

lower_case = kalimat.lower()
Print(lower_case)

# output
# berikut adalah 5 negara dengan Pendidikan
terbaik di dunia adalah korea Selatan, jepang,
singapura, hongkong dan finlandia
```

Proses diatas adalah proses penghapusan beberapa karakter yang tidak boleh masuk pada proses selanjutnya agar proses traning model tidak mengalami masalah.

4.1.3 *Tokenizing*

```
kalimat = "rumah idaman adalah rumah yang bersih"
pisah = kalimat.split()
# output
# ['rumah', 'idaman', 'adalah', 'rumah', 'yang']
```

Proses diatas merupakan proses setelah case folding dilakukan guna untuk memisahkan antar kata menjadi term agar dapat dimasukkan kedalam *word embedding* yang sebelumnya akan melalui tahap *stopremoval* terlebih dahulu.

4.1.4 *Stopremoval*

```
kalimat = "Andi kerap melakukan transaksi rutin
secara daring atau online. Menurut andi belanja
online lebih praktis & murah"

tokens = word_tokenize(kalimat)
listStopword = set(stopwords.words('indonesia'))

removed = []
for t in tokens:
    if t not in listStopwords:
        removed.append()
print(removed)

# output
# ['andi', 'kerap', 'transaksi', 'rutin',
'daring', 'online', 'andi', 'belanja', 'online',
'prantis', 'murah']
```

Proses diatas merupakan proses penghapusan kata yang tidak memiliki interpretasi yang sangat tidak dibutuhkan dalam proses traning model sehingga perlu dihapus agar proses traning berjalan lancar.

4.2 Pembuatan Model LDA dan Menemukan Latent Topics

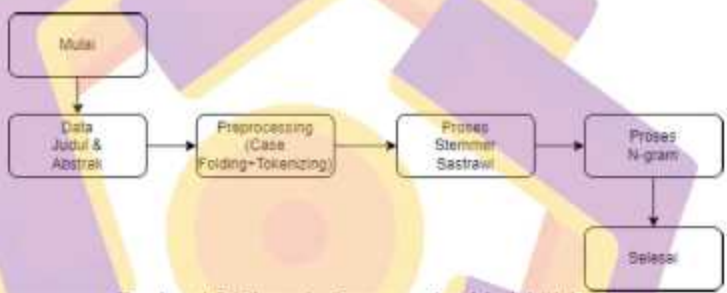
Pada alur penelitian 1 akan melibatkan pengggunaan bersilang pada tahap preprocessing tambahan antara library sastrawi yang mewakili *stemmer truncating* dan library N-gram yang mewakili *stemmer statistical* sebelum dimasukkan pada model 1, 2, 3, dan 4. Penggunaan library sastrawi dalam Bahasa Indonesia menemukan kata dasar pada sebuah kata yang memiliki imbuhan afiks, infiks, sufiks serta konfiks. Imbuhan affiks adalah imbuhan yang terletak didepan kata seperti kata "ber-sambung" adapun contoh lain tentang penggunaan afiks pada kata berbahasa indoensia seperti di-, ter-, pen-, per-, se-, me-, dan ke-. Sedangkan untuk

infiks contoh katanya adalah "t-em-ali" yang kata dasarnya tali sedangkan infiksnya adalah -em- contoh lain dari imbuhan ini adalah -el, -er, -e, dan -in. Sedangkan untuk sufiks adalah imbuhan yang terletak pada akhiran kata dasar contoh untuk jenis ini adalah "sambung-kan" disini -kan merupakan imbuhan dan sambung merupakan kata dasar tersebut. Adapun konfiks adalah penggunaan kata dasar beserta imbuhan/afiks depan dan imbuhan belakang/sufiks yang digunakan dalam satu kata atau bisa dikatakan sebagai perpaduan 2 imbuhan dan 1 kata dasar.

N-gram yang mewakili *stemmer statistical* merupakan model yang digunakan untuk memprediksi kata berikutnya berdasarkan kecocokan penyandingan kata sebeumnya atau N-1. Makanya konsep N-gram sendiri disebut sebagai model probabilistic dikarenakan dapat melakukan pemodelan kata dan membuat keterhubungan antara kata yang saling bersandingan baik dalam posisi belakang kata atau depan kata maka model ini akan menyendingkan kata tersebut lalu memisahkannya dengan *underscore*. Misalkan dalam corpus akan kita temukan nama kota yang selalu diikuti dengan nama kotanya misalkan kota makassar maka model N-gram akan menyandingkan dua kata itu menjadi kota_makassar dengan tanda *underscore*. Sedangkan dalam penelitian ini akan digunakan N = 2 dan N = 3 pada pemodelan N-gram dimana yang dimaksud tersebut adalah menangkap keterhubungan antara kata dengan jumlah kata 2 dan maksimal 3. Dalam model LDA urutan kata tidak begitu penting dan model ini memfokuskan pada kata per kata maka untuk melihat pengaruh proyeksi hasil menggunakan kedua stemmer tersebut maka akan di uji silang. Sedangkan detail skenarionya penggunaan library sastrawi dan N-gram dapat kita lihat dalam penjelasan dibawah ini.

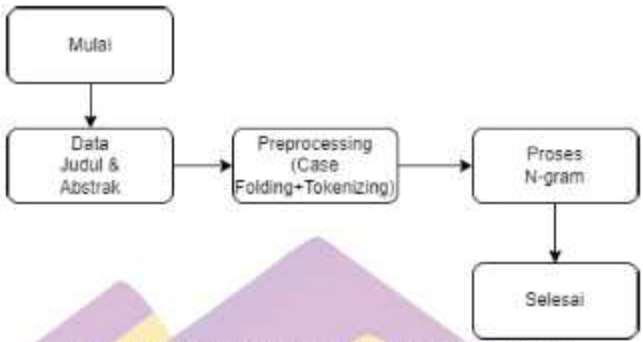
4.2.1 Pembagian *Preprocessing* Data 4 Skenario

Pada Preprocessing yang dilakukan pada data skenario pertama akan diterapkan proses stemmer sastrawi dan proses N-gram sebelum dilanjutkan pada proses pembuatan model, seperti yang tampak pada gambar dibawah ini. Di model pertama ini akan duji pengaruh kedua stemmer tersebut seperti petunjuk sebelumnya untuk melihat hasil *coherence score* yang akan dihasilkan.



Gambar 4.2. Skenario Preprocessing Model #1 LDA

Pada Preprocessing yang dilakukan pada data skenario kedua akan diterapkan proses stemmer sastrawi namun tanpa proses stemmer N-gram sebelum dilanjutkan pada proses pembuatan model, seperti yang tampak pada gambar dibawah ini. Di model kedua ini akan duji pengaruh jika stemmer N-gram saja tanpa melibatkan proses stemmer sastrawi yang hal tersebut seperti petunjuk sebelumnya untuk melihat hasil *coherence score* yang akan dihasilkan.

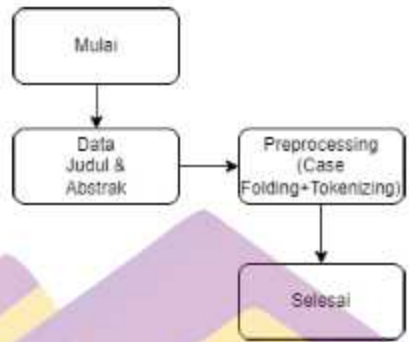


Gambar 4.3. Skenario Preprocessing Model #2 LDA

Pada Preprocessing yang dilakukan pada data skenario ketiga akan diterapkan proses stemmer Sastrawi tapi tanpa proses stemmer N-gram sebelum dilanjutkan pada proses pembuatan model, seperti yang tampak pada gambar dibawah ini.

Gambar 4.4. Skenario Preprocessing Model #3 LDA

Pada Preprocessing yang dilakukan pada data skenario keempat tidak dilakukan sama sekali proses stemmer sastrawi dan proses N-gram sebelum dilanjutkan pada proses pembuatan model, seperti yang tampak pada gambar dibawah ini.



Gambar 4.5. Skenario Preprocessing Model #4 LDA

4.2.2 *Hyperparameter Tuning* Model

LDA adalah model "tas-of-words" yang berarti urutan kata-kata tidak masalah. LDA adalah model generatif dimana dokumen diinterpretasikan kata demi kata dengan memilih campuran topik θ Dirichlet (α). Didalam model LDA juga akan dilakukan testing untuk mendapatkan *hyperparameter* terbaik. Ada beberapa poin yang perlu diketahui tentang *hyperparamter* ini yaitu parameter *chunksize, pasess, iterations* dan *eval_every*.

Chunksize adalah jumlah dokumen yang akan diproses secara bersamaan dalam satu iterasi. Nilai chunksize yang lebih besar dapat meningkatkan kecepatan proses training, tetapi juga dapat menyebabkan memory error jika memori tidak cukup. *Passes* adalah jumlah iterasi yang dilakukan saat proses training. Nilai passes yang lebih besar dapat meningkatkan keakuratan model, tetapi juga dapat meningkatkan waktu training. *Iterations* adalah jumlah iterasi yang dilakukan saat

proses training untuk setiap chunksize. Nilai iterations yang lebih besar dapat meningkatkan keakuratan model, tetapi juga dapat meningkatkan waktu training. *Eval_every* adalah jumlah iterasi yang dilakukan sebelum model di-evaluasi. Nilai eval_every yang lebih kecil dapat meningkatkan waktu training, tetapi juga dapat menyebabkan model tidak di-evaluasi secara teratur selama proses training. Semua hyperparameter tersebut dapat mempengaruhi hasil model LDA. Nilai yang tepat untuk setiap hyperparameter tergantung pada dataset yang akan diolah oleh model. Oleh karena itu, hyperparameter terbaik harus dioptimalkan secara terpisah untuk setiap dataset dengan menggunakan teknik-teknik seperti grid search atau random search.

Hyperparameter yang akan digunakan dalam model yang bersifat tetap adalah *chunksize* sebesar 500, *passes* sebesar 20, *iterations* sebesar 400 dan *eval_every* sebesar 1. Penggunaan *hyperparameter* ini merupakan hasil percobaan yang dilakukan sebelum eksperimen inti dilaksanakan agar model yang didapatkan merupakan model yang stabil dalam menemukan topik.

4.2.3 Hasil Evaluasi

Pada model pertama dengan penerapan stemmer truncating menggunakan algoritma sastra dan statistik menggunakan N-gram, menghasilkan hasil sebagai berikut. Tabel dibawah menunjukkan bahwa model dengan parameter *num_topics* 12 tertinggi dibandingkan model dengan parameter lainnya, terlihat bahwa model dengan parameter *num_topics* 13 tiba-tiba mengalami penurunan skor koherensi.

Dalam mengukur kinerja dari model LDA akan digunakan ukuran *coherence score*. *Coherence score* adalah ukuran yang digunakan untuk mengukur seberapa baik sebuah model topic modeling (seperti LDA) dapat mengelompokkan dokumen atau teks menjadi topik yang terkait satu sama lain. *Coherence score* berguna untuk mengevaluasi kualitas model topic modeling dan membantu dalam proses pemilihan model terbaik.

Salah satu metode yang sering digunakan untuk menghitung *coherence score* pada model LDA adalah dengan menggunakan metode "*c_v*" (*coherence via vector*) atau "umum". Metode ini menghitung *coherence score* dengan cara menghitung rata-rata nilai kata yang terkait dengan setiap topik dalam model. Nilai kata terkait diukur dengan menggunakan salah satu dari beberapa ukuran seperti pointwise mutual information (PMI) atau normalized discounted cumulative gain (nDCG). Untuk menghitung *coherence score c_v* pada model LDA, langkah-langkahnya adalah sebagai berikut:

1. Terapkan preprocessing pada dokumen yang akan dianalisis, seperti menghilangkan kata yang tidak bermakna (*stop words*) dan mengubah kata menjadi bentuk dasar (*stemming*).
2. Buat matriks dokumen-kata (*document-word matrix*) dari dokumen yang telah diolah. Matriks ini menyatakan frekuensi kemunculan setiap kata dalam setiap dokumen.
3. Terapkan model LDA pada matriks dokumen-kata untuk mengelompokkan dokumen ke dalam beberapa topik.

4. Hitung nilai kata terkait untuk setiap topik dengan menggunakan salah satu ukuran seperti PMI atau nDCG.
5. Hitung rata-rata nilai kata terkait untuk setiap topik. Rata-rata nilai ini akan menjadi *coherence score c_v* untuk model LDA tersebut.

Semakin tinggi nilai *coherence score c_v*, semakin baik kualitas model LDA dalam mengelompokkan dokumen ke dalam topik yang terkait satu sama lain. Sebaliknya, semakin rendah nilai *coherence score c_v*, semakin buruk kualitas model LDA dalam mengelompokkan dokumen ke dalam topik yang terkait. Sedangkan rincian *Coherence Score c_v* dapat dilihat pada tabel di bawah ini.

Tabel 4.1 Hasil Model 1 Berdasarkan *num_topics* dan *Coherence Score*

num_topic	***Coherence Score***
5	0.27133171516309085
6	0.2971736788233644
7	0.29240999859208355
8	0.26112189162789196
9	0.2613736247613906
10	0.27865901817194094
11	0.29152524844140554
12	**0.30063692761114086**
13	0.29540634305924396

Pada model kedua dengan penerapan *truncating stemmer* menggunakan algoritma sastrawi dan tanpa *stemmer statistik* menggunakan N-gram menghasilkan hasil sebagai berikut, tampak pada tabel dibawah menunjukkan sekali lagi bahwa *num_topics* terbaik ada di nomor 12 sedangkan nomor 13 mengalami penurunan drastis. Adapun rincian model *Coherence Score* 2 dapat dilihat pada tabel di bawah ini.

Tabel 4.2 Hasil Model 2 Berdasarkan *num_topics* dan *Coherence Score*

num_topic	*Coherence Score*
5	0.2707055338461587
6	0.2675425046160588
7	0.2934398993267315
8	0.27595091223242224
9	0.26867870976956343
10	0.26935085311052265
11	0.2914966299699736
12	**0.29045220547061973**
13	0.2616608200409875

Pada model ketiga tanpa menerapkan *stemmer truncating* dan menerapkan statistik stemmer menggunakan N-gram menghasilkan hasil sebagai berikut. Pada table dibawah, num_topik 13 merupakan yang memiliki skor koherensi tertinggi dibandingkan num_topik lainnya namun belum melampaui skor koherensi model 1 dengan num_topik 12. Adapun rincian hasil model 3 dapat dilihat pada tabel di bawah.

Tabel 4.3 Hasil Model 3 Berdasarkan *num_topics* dan *Coherence Score*

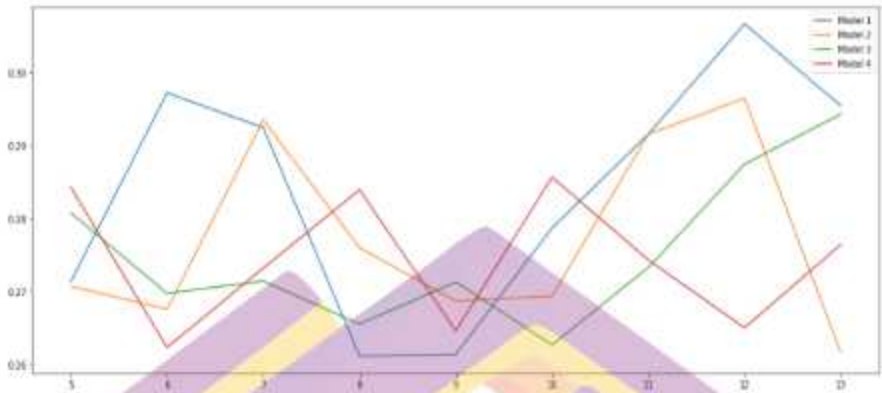
num_topic	*Coherence Score*
5	0.2807391633129487
6	0.2697187116111068
7	0.2714141435689421
8	0.26551254616333464
9	0.27127137525113465
10	0.26268148630169136
11	0.2732334644825967
12	0.2874133128546404
13	**0.29426375154662393**

Model keempat tanpa menerapkan stemmer truncating dan tanpa menerapkan statistik stemmer menggunakan N-gram menghasilkan hasil sebagai berikut. Tabel dibawah menunjukkan hasil skor koherensi untuk setiap *num_topic*, namun hasil di atas menunjukkan ketidakseimbangan model yang dihasilkan sehingga model tersebut dapat dikatakan sebagai model prematur. Adapun rincian hasil model 4 dapat dilihat pada tabel di bawah ini.

Tabel 4.4 Hasil Model 4 Berdasarkan *num_topics* dan *Coherence Score*

num_topic	*Coherence Score*
5	0.28425625289831985
6	0.2623154155964427
7	0.27318274466551923
8	0.2838990576791418
9	0.26459909991573677
10	**0.28562531741521590**
11	0.2743142933006359
12	0.2650150569553048
13	0.27640386606676076

Sedangkan perbandingan hasil dari hasil model jika di komparasikan adalah sebagai berikut, didalam gambar bisa dilihat bahwa pada model pertama memiliki *coherence score* tertinggi pada *num_topics* ke 12.



Gambar 4.6. Hasil Komparasi Model 1,2,3 dan 4 Berdasarkan *num_topics* dan *Coherence Score*

Sedangkan tabel perbandingan hasil keempat model di atas dapat dilihat pada tabel berikut.

Tabel 4.5 Perbandingan Hasil dari 4 Model Berdasarkan *num_topics*, *Coherence Score*, Nilai Total dan Rata-rata *Cohorence Score*

num_topics	*COHERENCE SCORE*			
	Model 1	**Model 2**	**Model 3**	**Model 4**
5	0.271331715163090 85	0.2707055338461587	0.2807391633129487	0.28425625289831985
6	0.2971736788233644	0.2675425046160538	0.2697187116111068	0.2623154155964427
7	0.2924099985920835 5	0.2934398993267315	0.271414143568942 1	0.27318274466551923
8	0.2611218916278919 6	0.2759509122324222 4	0.2655125461633346 4	0.2838990576791418
9	0.2613736247613906	0.2686787097695634 3	0.2712713752511346 5	0.2645990999157367 7
10	0.2786590181719409 4	0.2693508531105226 5	0.2626814863016913 6	0.2856253174152159 6

Tabel 4.5 (Lanjutan) Perbandingan Hasil dari 4 Model Berdasarkan *num_topics*, *Coherence Score*, Nilai Total dan Rata-rata *Cohorence Score*

num_topics	*COHERENCE SCORE*			
	Model 1	**Model 2**	**Model 3**	**Model 4**
11	0.29152524844140554	0.291496629969973 6	0.2732334644825967	0.274314293300635 9
12	**0.30663692761114086**	0.296452205470619 73	0.2874133128546404	0.265015056955304 8
13	0.29540634305924396	0.261666820040987 5	0.29426375154662393	0.276403866066767 6
SUM	**2.555638446**	2.495278068	2.476247955	2.469611104
AVERAGE	**0.283959827**	0.277253119	0.275138662	0.274401234

Dari hasil Tabel 5, model terbaik adalah model pertama yang menerapkan truncating stemmer menggunakan algoritma sastra dan menerapkan stemmer statistik menggunakan N-gram dengan jumlah num_topics terbaik berada di nomor 12 dan jika dilihat dari jumlah nilai koherensi, model pertama memiliki jumlah terbanyak dibandingkan dengan model lainnya dan memiliki rata-rata nilai koherensi terbesar yang dimiliki oleh model pertama dibandingkan dengan model lainnya.

4.3 Pembuatan Model VSM Untuk Membuat Dataset Baru

Pada alur penelitian 1 akan melibatkan pengggunaan bersilang pada tahap preprocessing tambahan antara library sastrawi yang mewakili *stemmer truncating* dan library N-gram yang mewakili *stemmer statistical* sebelum dimasukkan pada model 1, 2, 3, dan 4. Penggunaan library sastrawi dalam Bahasa Indonesia menemukan kata dasar pada sebuah kata yang memiliki imbuhan afiks, infiks,

sufiks serta konfiks. Imbuhan affiks adalah imbuhan yang terletak didepan kata seperti kata "ber-sambung" adapun contoh lain tentang penggunaan afiks pada kata berbahasa indoensia seperti di-, ter-, pen-, per-, se-, me-, dan ke-. Sedangkan untuk infiks contoh katanya adalah "t-em-ali" yang kata dasarnya tali sedangkan infiksnya adalah -em- contoh lain dari imbuhan ini adalah -el, -er, -e, dan -in. Sedangkan untuk sufiks adalah imbuhan yang terletak pada akhiran kata dasar contoh untuk jenis ini adalah "sambung-kan" disini -kan merupakan imbuhan dan sambung merupakan kata dasar tersebut. Adapun konfiks adalah penggunaan kata dasar beserta imbuhan/afiks depan dan imbuhan belakang/sufiks yang digunakan dalam satu kata atau bisa dikatakan sebagai perpaduan 2 imbuhan dan 1 kata dasar.

N-gram yang mewakili *stemmer statistical* merupakan model yang digunakan untuk memprediksi kata berikutnya berdasarkan kecocokan penyandingan kata sebeumnya atau N-1. Makanya konsep N-gram sendiri disebut sebagai model probabilistic dikarenakan dapat melakukan pemodelan kata dan membuat keterhubungan antara kata yang saling bersandingan baik dalam posisi belakang kata atau depan kata maka model ini akan menyendingkan kata tersebut lalu memisahkannya dengan *underscore*. Misalkan dalam corpus akan kita temukan nama kota yang selalu diikuti dengan nama kotanya misalkan kota makassar maka model N-gram akan menyandingkan dua kata itu menjadi kota_makassar dengan tanda *underscore*. Sedangkan dalam penelitian ini akan digunakan N = 2 dan N = 3 pada pemodelan N-gram dimana yang dimaksud tersebut adalah menangkap keterhubungan antara kata dengan jumlah kata 2 dan maksimal 3. Sedangkan detail

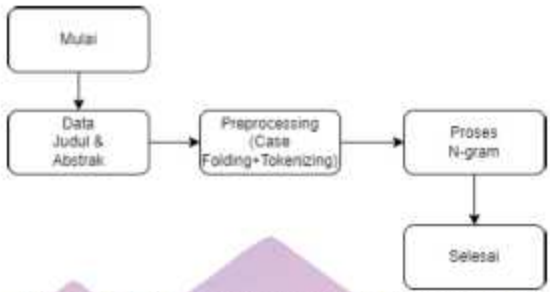
skenarionya penggunaan libarary sastrawi dan N-gram dapat kita lihat dalam penjelasan dibawah ini.

4.3.1 Pembagian *Preprocessing* Data 4 Skenario

Pada Preprocessing yang dilakukan pada data skenario pertama akan diterapkan proses stemmer sastrawi dan proses N-gram sebelum dilanjutkan pada proses pembuatan model, seperti yang tampak pada gambar dibawah ini.

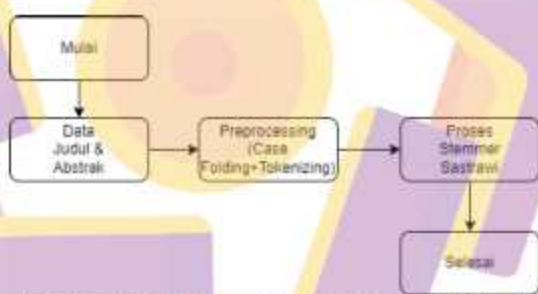
Gambar 4.7. Skenario Preprocessing Model #1 VSM

Pada Preprocessing yang dilakukan pada data skenario kedua akan diterapkan proses stemmer sastrawi namun tanpa proses stemmer N-gram sebelum dilanjutkan pada proses pembuatan model, seperti yang tampak pada gambar dibawah ini.



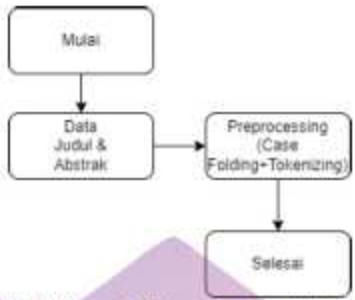
Gambar 4.8. Skenario Preprocessing Model #2 VSM

Pada Preprocessing yang dilakukan pada data skenario ketiga akan diterapkan proses stemmer Sastrawi tapi tanpa proses stemmer N-gram sebelum dilanjutkan pada proses pembuatan model, seperti yang tampak pada gambar dibawah ini.



Gambar 4.9. Skenario Preprocessing Model #3 VSM

Pada Preprocessing yang dilakukan pada data skenario keempat tidak dilakukan sama sekali proses stemmer sastrawi dan proses N-gram sebelum dilanjutkan pada proses pembuatan model, seperti yang tampak pada gambar dibawah ini.



Gambar 4.10. Skenario Preprocessing Model #4 VSM

4.3.2 Split Data Training Dan Testing

Teknik split data traning dan testing yang digunakan pada percobaan ini adalah sebesar 50 persen dari dataset untuk data traning sedangkan untuk data testing yang digunakan adalah sebesar 16 persen dari dataset.

4.3.3 Hasil Evaluasi

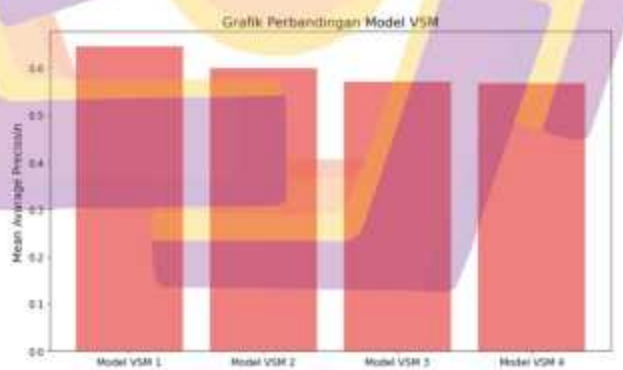
Pada alur penelitian ini akan digunakan *mean average precision* (mAP) untuk mengukur kinerja model VSM. MAP adalah salah satu metrik yang digunakan untuk mengevaluasi kinerja model VSM. VSM adalah salah satu teknik pemodelan yang digunakan untuk mengklasifikasikan dokumen berdasarkan isinya. Model VSM mengkonversi dokumen menjadi vektor spasi yang menggambarkan fitur-fitur yang terkandung dalam dokumen tersebut.

Untuk mengevaluasi kinerja model VSM menggunakan MAP, pertama-tama terlebih dahulu harus ditentukan label yang tepat untuk setiap dokumen.

Kemudian, setiap dokumen diuji dengan model VSM dan dihasilkan sejumlah prediksi label yang mungkin. Setiap prediksi label kemudian dibandingkan dengan label yang sebenarnya, dan dihitung berapa banyak label yang benar dan salah yang dihasilkan oleh model VSM.

Untuk setiap dokumen, *average precision* dihitung dengan mengalikan nilai *precision* pada setiap langkah dengan *recall* pada langkah yang sama, kemudian menjumlahkan semua hasil tersebut dan membagi dengan jumlah total langkah. *Average precision* dihitung untuk setiap dokumen, kemudian dijumlahkan dan dibagi dengan jumlah total dokumen untuk menghasilkan MAP.

Nilai MAP yang tinggi menunjukkan bahwa model VSM mampu memberikan prediksi label yang benar pada tingkat keseluruhan yang tinggi. Sebaliknya, nilai MAP yang rendah menunjukkan bahwa model VSM kurang efektif dalam memberikan prediksi label yang benar.



Gambar 4.11. Gambar Hasil Perbandingan keempat model VSM

Pada grafik diatas dan table yang dibawah menunjukkan bahwa penerapan model VSM dengan bantuan *stemmer truncating* mengguakan library sastrawi dan *stemmer statistical* menggunakan n-gram mempengaruhi kualitas *mean average precision* (mAP) pada model yang jika semakin tinggi maka semakin baik pula model yang dihasilkan berdasarkan hasil evaluasi.

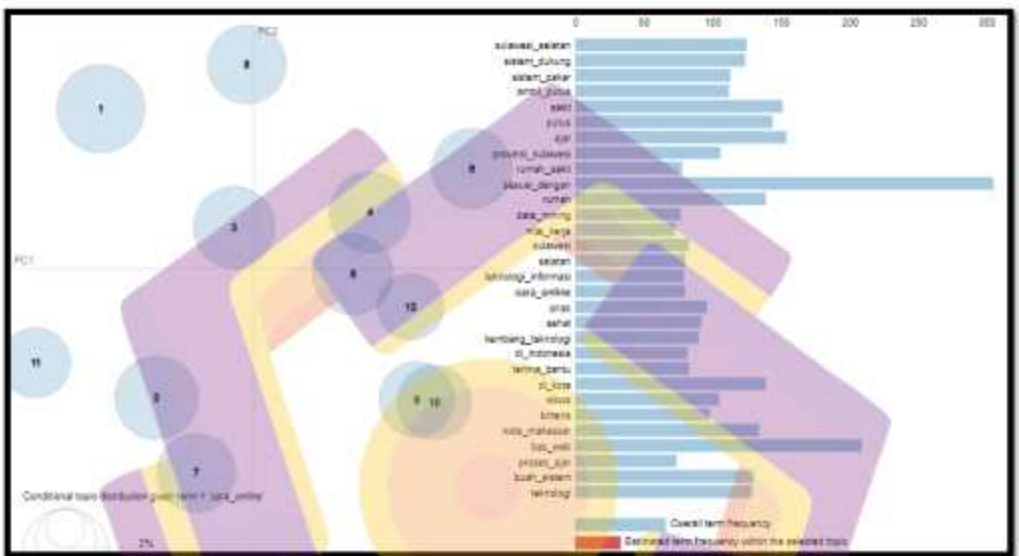
Tabel 4.6 Perbandingan hasil model berdasarkan *Coherence score*

Model	mAP
VSM + Sastrawi + N-gram	0.6438269428983715
VSM + Sastrawi	0.598394166633515
VSM + N-gram	0.5690989314917886
VSM	0.5655649174245564

4.4 Hasil Pembuatan Dataset Baru Berdasarkan Model LDA dan VSM Terbaik

Pada hasil alur penelitian ini akan di load ulang model LDA dan juga hasil yang didapatkan berupa *latent topics* dengan jumlah 12 topik yang merupakan hasil num_topics tertinggi dengan skema scenario yang pertama yaitu memberlakukan penggunaan *stemmer truncating* dalam hal ini adalah library sastrawi dan *stemmer statistical* menggunakan n-gram. Setelah itu akan di load model VSM untuk melakukan proses *indexing* terhadap *latent topics* tadi yang dianggap sebagai *query* untuk disandingkan dalam vector yang dibuat bersamaan dengan dataset awal. Maka setelah proses perhitungan menggunakan VSM (*Cosine Similarity*) maka

akan tersisa judul-judul yang tidak terindeks. Setelah itu data set yang terindeks akan digunakan untuk membuat dataset baru.



Gambar 4.12. Penyebaran Klusterisasi Laten topik

Tabel 4.7 Hasil dari 12 topik yang dihasilkan oleh model terbaik LDA

Topik Ke-	Isi Topik	Interpretasi
1	provinsi_sulawesi stmik_dipanegara ajar sesuai_dengan provinsi dosen sulawesi dipanegara stmik dalam_bentuk	Maksud topik adalah banyaknya mahasiswa menggunakan objek pada

Tabel 4.7 (Lanjutan) Hasil dari 12 topik yang dihasilkan oleh model tebaik LDA

Topik Ke-	Isi Topik	Interpretasi
		Instansi Provinsi Sulawesi seletan, STMIK DIpanegaga
2	di_indonesia na ve_bayes covid indonesia bayes virus kumpul_data dapat_bantu sehat data_mining	Maksud topik adalah banyaknya mahasiswa penggunaan metode naïve bayes yang kadang disandingkan dengan Analisa data covid untuk keperluan data mining.
3	sistem_pakar sakit pakar orang mendiagnosa_sakit diagnosa_sakit gejala diagnosa masalah_sebut tangan	Maksud topik adalah banyaknya mahasiswa dalam membuat model sistem pakar untuk mendiagnosa sebuah penyakit berdasarkan gejala yang muncul

Tabel 4.7 (Lanjutan) Hasil dari 12 topik yang dihasilkan oleh model tebaik LDA

Topik Ke-	Isi Topik	Interpretasi
4	bas_mobile layan rumah_sakit sehat pasien mobile jasa di_bidang dapat_bantu obat	Maksud topik adalah banyakna mahasiswa yang membahas tentang pembuatan aplikasi berbasis mobile untuk pelayanan Kesehatan pada rumah sakit/apotik
5	barang perangkat_lunak internet_of alat deteksi sensor judul sedia diri_dari	Maksud Topik adalah banyaknay mahasiswa yang menggunakan teknologi IoT dalam mendeteksi menggugnakan sensor
6	algoritma beli data_mining jual motor rupa_salah dapat_bantu tanam toko beri_informasi	Maksud topik adalah mahasiswa menggunakan sistem informasi penjualan dengan memanfaatkan data mining

Tabel 4.7 (Lanjutan) Hasil dari 12 topik yang dihasilkan oleh model tebaik LDA

Topik Ke-	Isi Topik	Interpretasi
7	ambil_putus terima_bantu terima dapat_bantu sistem_dukung kriteria calon knearest_neighbor pilih rumah	Maksud topik adalah banyaknya mahasiswa yang mengambil judul terkait sistem pengambilan keputusan dan kadang lebih bnayk disadingkan dengan metode kNN
8	kota_makassar nilai_kerja karyawan kota data_mining layak sesuai_dengan program berkas usaha	Maksud topik adalah banyaknya mahasiswa menggunakan data minig dalam mengimplementasikan penilaian kinerja karyawan pada suatu instansi
9	di_kota teknologi_informasi hal_sebut media digital	Maksud topik adalah banyaknya mahasiswa membuat sistm informasi

Tabel 4.7 (Lanjutan) Hasil dari 12 topik yang dihasilkan oleh model tebaik LDA

Topik Ke-	Isi Topik	Interpretasi
	Universitas dipa rumah kembang teknologi layan dokumen	Untuk pelayanan pada kampus terutama dari universitas dipa makassar
10	siswa ajar anak sekolah guru oleh_karena proses_ajar cara_online absensi buat_buah	Maksud topik adalah banyaknya mahasiswa yang membahas terkait absensi online pada sekolah yang melibatkan guru dan siswa ketika kelas online berlangsung
11	sistem_dukung nilai_kerja karyawan dukung kriteria di_kota lokasi sensor tempat usaha	Maksud topik adalah banyaknya mahasiswa yang membahas terkait sistem pendukung keputusan namun dalam kasus ini dihubungkan dengan topik

Tabel 4.7 (Lanjutan) Hasil dari 12 topik yang dihasilkan oleh model tebaik LDA

Topik Ke-	Isi Topik	Interpretasi
12	ajar buat_buah dapat_bantu bagai_media oleh_karena proses_ajar teknologi_informasi anak game didik	Maksud topik ini adalah banyaknya mahasiswa mengangkat judul terkati masalah teknologi informasi untuk media pengajaran pada anak berbasis game

Sedangkan ke-12 hasil tersebut akan dimasukkan sebagai *query* didalam model VSM untuk dihitung similaritas masing-masing *query* sehingga akan ada judul yang terindeks dan ada juga tidak terindeks sama sekali, maka dari itu akan

dibuatkan dataset baru dari hasil yang terindeks saja. Sedangkan untuk detailnya hasil yang ditemukan peneliti adalah sebagai berikut.

Tabel 4.8 Hasil perbandingan dataset awal dan dataset terbaru berdasarkan hasil *indexing* model VSM dari *latent topics* LDA

Topik Ke-	Isi Topik	Dokumen Terindeks Dengan Nilai Similaritas > 20% / 0.2 Format (Id Dokumen -> Nilai Similaritas)
1	provinsi_sulawesi stmik_dipanegara ajar sesuai_dengan provinsi dosen sulawesi dipanegara stmik dalam_bentuk	405 -> 0.364511884583410051 03 -> 0.4154786286255201 131 -> 0.27976748154620623 114 -> 0.5197991819642086 258 -> 0.2848092757932537 319 -> 0.4505202619423689
2	di_indonesia na ve_bayes covid indonesia bayes virus kumpul_data dapat_bantu sehat data_mining	26 -> 0.27684755685912954 360 -> 0.2880126326923307 138 -> 0.29686728699257864 100 -> 0.21567517969765132 352 -> 0.39375533581651895 25 -> 0.39209885760922186 29 -> 0.22612799280241924
3	sistem_pakar sakit pakar orang mendiagnosa_sakit diagnosa_sakit gejala diagnosa masalah_sebut tangan	333 -> 0.24983775985808726 328 -> 0.33802391422801326 242 -> 0.4333607779841579

Tabel 4.8 (Lanjutan) Hasil perbandingan dataset awal dan dataset terbaru berdasarkan hasil *indexing* model VSM dari *latent topics* LDA

Topik Ke-	Isi Topik	Dokumen Terindeks Dengan Nilai Similaritas > 20% / 0.2 Format (Id Dokumen -> Nilai Similaritas)
3	sistem_pakar sakit pakar orang mendiagnosa_sakit diagnosa_sakit gejala diagnosa masalah_sebut tangan	333 -> 0.24983775985808726 328 -> 0.33802391422801326 242 -> 0.4333607779841579 105 -> 0.34652347530481664 265 -> 0.2522740957289497
4	bas_mobile layan rumah_sakit sehat pasien mobile jasa di_bidang dapat_bantu obat	318 -> 0.3862192228863237 72 -> 0.22170886030597084 83 -> 0.20561115752038564 165 -> 0.20237750601586738 25 -> 0.3089257224707318
5	barang perangkat_lunak internet_of alat deteksi sensor judul sedia diri_dari	212 -> 0.224198817746251O3

Tabel 4.8 (Lanjutan) Hasil perbandingan dataset awal dan dataset terbaru berdasarkan hasil *indexing* model VSM dari *latent topics* LDA

Topik Ke-	Isi Topik	Dokumen Terindeks Dengan Nilai Similaritas > 20% / 0.2 Format (Id Dokumen -> Nilai Similaritas)
6	algoritma beli data_mining jual motor rupa_salah dapat_bantu tanam toko beri_informasi	175 -> 0.2296596814422748 89 -> 0.22487347600047997 200 -> 0.29224177979865923 231 -> 0.2590716933131802 60 -> 0.20654920427984702 286 -> 0.2867099245941459
7	ambil_putus terima_bantu terima dapat_bantu sistem_dukung kriteria calon knearest_neighbor pilih rumah	87 -> 0.21458609170920254 29 -> 0.2009043230023346 33 -> 0.20224515422044884 282 -> 0.22265625080716572 352 -> 0.38273116215541847 103 -> 0.2863330044206244 217 -> 0.23793253980787066 407 -> 0.23820819579430574

Tabel 4.8 (Lanjutan) Hasil perbandingan dataset awal dan dataset terbaru berdasarkan hasil *indexing* model VSM dari *latent topics* LDA

Topik Ke-	Isi Topik	Dokumen Terindeks Dengan Nilai Similaritas > 20% / 0.2 Format (Id Dokumen -> Nilai Similaritas)
8	kota_makassar nilai_kerja karyawan kota data_mining layak sesuai_dengan program berkas usaha	175 -> 0.3333884463432839 292 -> 0.2041925132428356 152 -> 0.24332135198836352 79 -> 0.21406051413294122 155 -> 0.35093751150901387 424 -> 0.21748104878983177 169 -> 0.24913275754477915 216 -> 0.3025782867925436
9	di_kota teknologi_informasi hal_sebut media digital universitas_dipa rumah kembang_teknologi layan dokumen	193 -> 0.20704872479776223 257 -> 0.28876893436558565 95 -> 0.24625918367917565
10	siswa ajar anak sekolah guru oleh_karena proses_ajar cara_online absensi buat_buah	90 -> 0.3703646369696927 344 -> 0.22119583812073224 267 -> 0.21302005229994916 377 -> 0.2696412343611373

Tabel 4.8 (Lanjutan) Hasil perbandingan dataset awal dan dataset terbaru berdasarkan hasil *indexing* model VSM dari *latent topics* LDA

Topik Ke-	**Isi Topik**	**Dokumen Terindeks Dengan Nilai Similaritas > 20% / 0.2 Format (Id Dokumen -> Nilai Similaritas)**
11	sistem_dukung nilai_kerja karyawan dukung kriteria di_kota lokasi sensor tempat usaha	76 -> 0.2904187197885561 315 -> 0.365042291816731 236 -> 0.2480790927159667 3 99 -> 0.4326818629165389 299 -> 0.20097122552661562
12	ajar buat_buah dapat_bantu bagai_media oleh_karena proses_ajar teknologi_informasi anak game didik	334 -> 0.21782948435098368 306 -> 0.23590955964811755 374 -> 0.28428815226114923 112 -> 0.3125861887515339 272 -> 0.26494784928954723
Total Dataset Awal 430 Data		**Tersisa 58 Data yang Akan Dibuat Dataset Baru**

Maka selanjutnya peneliti melakukan proses pembuatan dataset tersebut berdasarkan id yang terlah terindeks yang melebihi angka 0.2 atau 20 persen. Standar pengambilan 30 persen diambil dari standar plagiarism yang ditetapkan pada tiap kampus untuk syarat penulisan ilmiah dapat diterima.

4.5 Hasil Model VSM Baru Yang Berinteraksi Dengan Mahasiswa

Pada hasil alur penelitian ini akan kita gunakan dataset baru yang sebelumnya menjadi output dari alur penelitian ketiga setelah itu kita membuat model VSM baru dengan alur pembuatan seperti sebelumnya namun memilih penggunaan preprocessing yang melibatkan *stemmer truncating* menggunakan library sastrawi dan *stemmer statistical* menggunakan n-gram karena scenario inilah yang bisa menghasilkan mAP tertinggi pada model VSM. Selanjutnya akan dilakukan split data traning dan testing lalu memeasukkan pada *word embedding*, dan mahasiswa selanjutnya memasukkan kata kunci / *query* dan akan dihitung hasil similaritasnya menggunakan VSM (*Cosine similarity*) jika dalam pedetekesian similaritasnya melebihi 20 persen maka akan dimunculkan pesan "Judul Ditolak" dan jika tidak melebih maka akan memunculkan pesan "Judul Dapat Ditindak Lanjuti".

Peneliti melakukan pengujian model VSM baru dengan ukuran pembenaran berdasarkan penilaian pribadi peneliti. Peneliti mencoba mengilustrasikan diri sebagai mahasiswa yang melakukan pengajuan judul sebagai *query* dan memberikan laporan hasil yang didapatkan pada model serta penilaian peneliti terhadap hasil yang didapatkan. Sehingga hal tersebut dapat digambarkan pada tabel berikut ini.

Tabel 4.9 Hasil pengujian model VSM baru berdasarkan dataset baru.

Query	Hasil	Penilaian
Rancang Bangun Game 2d Shooter Pltformer Menggunakan Metode Finite State Machine	Judul Dapat Ditindak Lanjuti	**Tepat**. Karena mulai dari Topik dan Metode yang digunakan Jarang Dibahas.
Implementasi Aplikasi Mobile Water Drinking Assistant	Judul Dapat Ditindak Lanjuti	**Tepat**. Karena mulai dari Topik dan Metode yang digunakan Jarang Dibahas.
Rancang Bangun Sistem Parkir Otomatis Pada Kampus Menggunakan Minimum Sistem Arduino Dengan Website	Judul Dapat Ditindak Lanjuti	**Tepat**. Karena mulai dari Topik dan Metode yang digunakan Jarang Dibahas.
Perbandingan Algoritma Naive Bayes dan C. 45 Dalam Klasifikasi Data Mining	Judul Ditolak Detail Similaritas: 156 -> 0.24864250946441657 36 -> 0.24237300376690152	**Tepat**. Karena algoritma yang digunakan sering diajukan serta topik tentang data mining sudah sering disebutkan

Tabel 4.9 (Lanjutan) Hasil pengujian model VSM baru berdasarkan dataset baru.

Query	Hasil	Penilaian
Implementasi Data Mining Pemilihan Pelanggan Potensial Menggunakan Algoritma K Means	Judul Ditolak Detail Similaritas: 253 -> 0.24237300376690152 31 -> 0.22316040165025441	**Tepat**. Karena algoritma yang digunakan sering diajukan serta topik tentang data mining sudah sering disebutkan
Sistem Pendukung Keputusan Penentuan Dosen Pembimbing Skripsi Menggunakan Metode Profile Matching	Judul Ditolak Detail Similaritas: 275 -> 0.48759270373795427	**Tepat**. Karena algoritma yang digunakan sering diajukan serta topik tentang sistem pendukung keputusan sudah sering disebutkan
Sistem Pakar Diagnosa Penyakit Pencernaan Pada Manusia Menggunakan Metode Certainty Factor Berbasis Web	Judul Ditolak Detail Similaritas:	**Tepat.** Karena sistem pakar sudah sering diteliti apalagi memiliki implementasi dalam bentuk web
Sistem Pendukung Keputusan Penentuan Dosen Pembimbing Skripsi Menggunakan Metode Profile Matching	Judul Ditolak Detail Similaritas: 275 -> 0.48759270373795427	**Tepat.** Karena sistem pendukung keputusan sering diajukan dan diteliti serta algoritma/metode yang diajukan

Tabel 4.9 (Lanjutan) Hasil pengujian model VSM baru berdasarkan dataset baru.

Query	Hasil	Penilaian
PERANCANGAN SISTEM INFORMASI PENILAIAN KINERJA KARYAWAN	Judul Ditolak Detail Similaritas: 207 -> 0.23034529052681363 307 -> 0.22116907689112686 27 -> 0.44882204568797557	**Tepat.** Karena pembahan topik terkait penilaian kinerja karyawan sudah sering diangkat.
MODEL ALGORITMA K-NEAREST NEIGHBOR (K-NN) UNTUK PREDIKSI KELULUSAN MAHASISWA	Judul Ditolak Detail Similaritas:	**Tepat.** Karena algorima KNN sering diangkat dan juga topik terkait prediksi kelulusan mahasiswa

Dari hasil diatas model telah bekerja dengan baik dalam menilai apakah pengajuan judul terdeteksi sebagai topik yang sering dibahas dalam judul yang sering diangkat dalam penelitian mahasiswa jurusan Teknik informatika universitas Dipa Makassar.

4.6 Perbandingan Hasil Pada Penelitian Sebelumnya

Tabel 4.10 Perbandingan Hasil Pada Penelitian Sebelumnya.

Peneliti	Tahun	Dataset	Metode	Evaluasi Metode	Num_topics yang digunakan
Lau dkk	2014	1000 Artikel New York Times	LDA, pLSi, CTM	PMI, NPMI, LCP Word Intrusion	50, 100, 150
Aletras and Stevenson	2014	47,229 Artikel Wikipedia	LDA, CTM	PMI, ESA Human Judge	50,100, 200,600, 800
Lau and Baldwin	2016	100,000 artikel dari berbagai media	LDA	NPMI Word intrusion	300
Koltcov	2018	20 Portal berita dengan	VLDA, GLDA	Jaccard Index	2-120

Tabel 4.10 (Lanjutan) Perbandingan Hasil Pada Penelitian Sebelumnya.

Peneliti	Tahun	Dataset	Metode	Evaluasi Metode	Num_topics yang digunakan
		15,404 artikel berita			
Koltcov dkk	2019	Dataset 20 ortal berita rusia dengan jumlah data 699,746	LDA	Shannon entropy, Log-likelihood and perplexity	0-601
Koltcov and Ignatenko	2020	Dataset 20 portal berita perancis	VLDA, GLDA	KI, divergence	2-120

Berbagai peneliti menawarkan berbagai macam metode dan macam-macam model evaluasi namun dalam penelitian mereka menemukan semakin bnayak dataset yang digunakan model akan membentuk interpretasi topik semantik yang

lebih baik. Karena pada penelitian ini menggunakan dataset yang tergolong sedikit akhirnya menghasilkan *coherence score* yang tidak menyentuh standar penilaian evaluasi model yang ideal yang dimana membutuhkan angka *coherence score* sebesar 0.4-0.7 sedangkan dalam penelitian ini hanya mendapatkan *coherence score* sebesar 0.30663692761114086.

BAB V

PENUTUP

5.1. Kesimpulan

Berdasarkan hasil penelitian ini ada 3 poin yang perlu disimpulkan dari hasil penelitian.

1. Model LDA terbaik untuk kasus berhasil didapatkan ditandai dengan mendapatkan *coherence score* tertinggi dengan melibatkan scenario penerapan *stemmer truncating* menggunakan library sastrawi dan juga penerapan *stemmer statistical* menggunakan n-gram. Sehingga dari model tersebut didapatkan 12 *latent topic* yang merupakan topik-topik inti penelitian dikalangan mahasiswa.
2. Model VSM terbaik yang menggunakan skenario yang sama dengan model LDA terbaik yaitu melibatkan penerapan *stemmer truncating* menggunakan library sastrawi dan juga penerapan *stemmer statistical* menggunakan n-gram, sehingga model VSM terbaik dapat menemukan judul dan abstrak yang dikategorisasi sebagai pengambilan topik yang monoton berdasarkan hasil analisa yang dikeluarkan oleh LDA (*latent topics*).
3. Berdasarkan hasil pada kesimpulan kedua akan didapatkan dataset baru yang terindeks oleh model VSM terbaik, dataset dapat dikurangi untuk mereduksi dimensi vector dan juga tidak menggangu pembobotan pada model VSM baru dalam melakukan filter pengajuan judul mahasiswa. Dari hasil tersebut didapatkanlah dataset baru yang dapat diproses oleh model

VSM baru. Sehingga model VSM baru dapat melakukan filtering apakah judul yang diajukan oleh mahasiswa yang termasuk monoton atau tidak dengan melihat nilai similaritasnya jika diatas 20 % maka akan dianjurkan untuk mencari topik penelitian lain. Sedangkan jika lebih kecil dari 20 % akan diberikan langkah berupa judul dapat ditindak lanjuti.

5.2. Saran

Saran dalam penelitian ini adalah model LDA dan VSM harusnya dapat digabungkan dalam proses otomatisasi dalam menemukan secara periodik berdasarkan semester berjalan. Karena data tren penelitian tiap periode berbeda maka topik-topik tren yang baru harus ditangkap oleh model LDA dan secara langsung diproses oleh model VSM untuk dimasukkan pada sistem rekomendasi penolakan pengajuan judul penelitian skripsi mahasiswa, yang dimana akan berinteraksi langusng dengan mahasiswa.

DAFTAR PUSTAKA

PUSTAKA BUKU

Arikunto, Suharsimi. 2010. Prosedur Penelitian Suatu pendekatan Praktek. Jakarta: Rineka Cipta.

Budiharto, W, 2016, Knowledge and Information Retrieval, Deeppublish, Indonesia.

Bhargav Srinivasa-Desikan, 2018, Natural Language Processing and Computational Linguistics: A Practical Guide to Text Analysis with Python, Gensim, SpaCy, and Keras, Packt Publishing

Connoly, Thomas; Begg, Carolyn; Strachan, A, (2003). *Database Systems : A Practical Approach to Design, Implementation and Management* (3rd editio), Addison Wesley

Russell, S. J., & Norvig, P, 2010. *Artificial Intelligence: A Modern Approach* (Prentice H), Prentice Hall

Steyvers M & Griffiths T, 2007, handbook of semantic analysis, University of Colorado Institute of Cognitive Science Series) 1st Edition

Sugiyono. (2017). Metode Penelitian dan Pengembangan. Bandung: Alfabeta

Sugiyono. (2016) Metode Penelitian Kuantitatif, Kualitatif dan R&D. Bandung: Alfabeta.

PUSTAKA MAJALAH, JURNAL ILMIAH ATAU PROSIDING

Alif Iffan Alfanzar, Khalid & Indri Sudanawati Rozas, 2020, Topic Modelling Skripsi Menggunakan Metode Latentdiriclhet Allocation, Jurnal Sistem Informasi (JSil) vol. 7 No.1

Bucher S, Clarke C.L & Cormack G. V, 2010, Information retrieval – implementing and evaluation serach Engines, MIT Press

Bagus Wicaksono Arianto & Gangga Anuraga, 2020, Pemodelan Topik Pengguna Twitter Mengenai Aplikasi "Ruangguru", Jurnal Imu Dasar, Vol. 21 No. 2 hal 149 – 154

Baris Ozyurt & M. Ali Akacayol, 2021, A new topic modeling based approach for aspect extraction in aspect based sentiment analysis: SS-LDA, Expert Systems with Applications vol.168

Eka Sabna, 2020, Penerapan Text Mining Untuk Pengelompokan Penelitian Dosen, Jurnal Ilmu Komputer (JIK) Vol. 2 Hal 161-164

Graham L Giller, 2012, The Statistical Properties of Random Bitstreams and the Sampling Distribution of Cosine Similarity, SSRN

Heidarian, A. and Dinneen, M. J., 2016, A Hybrid Geometric Approach for Measuring Similarity Level among Documents and Document Clustering, Proceedings - 2016 IEEE 2nd International Conference on Big Data Computing Service and Applications, BigDataService 2016, 1994, pp. 142–151.

I Made Artah Agasta, 2018, Pengaruh Stemmer Bahasa Indonesia Terhadap Peforma Analisis Sentimen Terjemahan Ulasan Film, Jurnal Teknokompak, Vol. 12, No. 1

I. Vayansky and S.A.P. Kumar, 2020, A Review of Topic Modeling Methods, Elsevier Ltd

Kabil Boukhari & Mohamed Nazih Omri, 2020, DL-VSM based document indexing approach for information retrieval, Journal of Ambient Intelligence and Humanized Computing

M.Geetha & N.Vimala, 2020, Survey On Semantic Information Retrieval Techniques In Bigdata, International journal of scientific & technology research, vol 9

Rani, R., Lobiyal, D.K, 2021, An extractive text summarization approach using tagged-LDA based topic modeling, SpingerLink, Multimed Tools Appl 80, 3275–3305

P. Mahalakshmi & N. Sabiyath Fathima, 2020, An Art of Review on Conceptual based Information Retrieval, Webology Vol. 18

Venkat N. Gudivada, Dhana L. Rao & Amogh R, 2018, Information Retrieval: Concepts, Models, and Systems, vol 38 page 331-401

PUSTAKA ELEKTRONIK

SAS Insights. 1 april 2022. Big Data What It Is and Why It Matters. Retrieved, https://www.sas.com/en_us/insights/big-data/what-is-big-data.html

Suhartono, D, 1 april 2022. Natural Language Processing. Retrieved, https://socs.binus.ac.id/2013/06/22/natural-language-processing

Wildcard, 23 Desember 2003. Wildchar Query, https://nlp.stanford.edu/IR-book/html/htmledition/wildcard-queries-1.html