TOPIC MODELLING DOKUMEN SKRIPSI MENGGUNAKAN METODE *LATENT SEMANTIC ANALYSIS*

SKRIPSI

UIN SUNAN AMPEL
S U R A B A Y A

Disusun Oleh:

RIFQI HAKIM
NIM: H06216022

PROGRAM STUDI SISTEM INFORMASI
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI SUNAN AMPEL
SURABAYA
2020

PERNYATAAN KEASLIAN

Saya yang bertanda tangan di bawah ini,

NAMA	:	RIFQI HAKIM
NIM	:	H06216022
PROGRAM STUDI	:	Sistem Informasi
ANGKATAN	:	2016

Menyatakan bahwa saya tidak melakukan plagiat dalam penulisan skripsi saya yang berjudul "*TOPIC MODELLING* DOKUMEN SKRIPSI MENGGUNAKAN METODE *LATENT SEMANTIC ANALYSIS*". Apabila suatu saat nanti terbukti saya melakukan tindakan plagiat, maka saya bersedia menerima sanksi yang telah ditetapkan.

Demikian pernyataan keaslian ini saya buat dengan sebenar-benarnya.

Surabaya, 20 Juli 2020

RIFQI HAKIM

NIM. H06216022

LEMBAR PERSETUJUAN PEMBIMBING

JUDUL	:	*TOPIC MODELING* PADA ABSTRAK SKRIPSI MENGGUNAKAN METODE *LATENT SEMANTIC ANALYSIS*
NAMA	:	RIFQI HAKIM
NIM	:	H06216022

Mahasiswa tersebut telah melakukan proses bimbingan dan dinyatakan layak untuk mengikuti Sidang Skripsi.

Surabaya, 20 Juli 2020

Menyetujui,

Dosen Pembimbing 1	Dosen Pembimbing 2
Dwi Rolliawati, MT	Khalid, M. Kom
NIP. 197909272014032001	NIP. 197906092014031002

SITIF

PENGESAHAN TIM PENGUJI SKRIPSI

Skripsi Rifqi Hakim ini telah dipertahankan
di depan tim penguji skripsi
di Surabaya, 22 Juli 2020.

Mengesahkan,
Dewan Penguji

Penguji I

(Dwi Rolliawati, MT)
NIP. 197909272014032001

Penguji II

(Khalid, M.Kom)
NIP. 197906092014031002

Penguji III

(Muhammad Andik Izzuddin, MT)
NIP. 198403072014031001

Penguji IV

(Mohammad Khusnu Milad, M.MT)
NIP. 197901292014031002

Mengetahui,
Plt. Dekan Fakultas Sains dan Teknologi
UIN Sunan Ampel Surabaya

Dr. Evi Fatimatur Rusydiyah, M.Ag.
NIP. 197312272005012003

KEMENTERIAN AGAMA
UNIVERSITAS ISLAM NEGERI SUNAN AMPEL SURABAYA
PERPUSTAKAAN
Jl. Jend. A. Yani 117 Surabaya 60237 Telp. 031-8431972 Fax.031-8413300
E-Mail: perpus@uinsby.ac.id

LEMBAR PERNYATAAN PERSETUJUAN PUBLIKASI
KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademika UIN Sunan Ampel Surabaya, yang bertanda tangan di bawah ini, saya:

Nama : RIFQI HAKIM
NIM : H06216022
Fakultas/Jurusan : SAINS DAN TEKNOLOGI/SISTEM INFORMASI
E-mail address : rifqihakim5889@gmail.com

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Perpustakaan UIN Sunan Ampel Surabaya, Hak Bebas Royalti Non-Eksklusif atas karya ilmiah :
☑ Sekripsi ☐ Tesis ☐ Desertasi ☐ Lain-lain (.....)
yang berjudul :

TOPIC MODELLING DOKUMEN SKRIPSI MENGGUNAKAN METODE LATENT SEMANTIC ANALYSIS

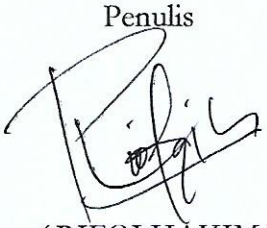
beserta perangkat yang diperlukan (bila ada). Dengan Hak Bebas Royalti Non-Ekslusif ini Perpustakaan UIN Sunan Ampel Surabaya berhak menyimpan, mengalih-media/format-kan, mengelolanya dalam bentuk pangkalan data (database), mendistribusikannya, dan menampilkan/mempublikasikannya di Internet atau media lain secara ***fulltext*** untuk kepentingan akademis tanpa perlu meminta ijin dari saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan atau penerbit yang bersangkutan.

Saya bersedia untuk menanggung secara pribadi, tanpa melibatkan pihak Perpustakaan UIN Sunan Ampel Surabaya, segala bentuk tuntutan hukum yang timbul atas pelanggaran Hak Cipta dalam karya ilmiah saya ini.

Demikian pernyataan ini yang saya buat dengan sebenarnya.

Surabaya, 12 Agustus 2020

Penulis



(RIFQI HAKIM)

ABSTRAK

TOPIC MODELLING DOKUMEN SKRIPSI MENGGUNAKAN METODE *LATENT SEMANTIC ANALYSIS*

Skripsi merupakan penelitian akhir bagi mahasiswa untuk memperoleh gelar sarjana strata-1. Dengan semakin bertambahnya dokumen skripsi, maka akan terbentuk suatu informasi yang dapat diambil dari kumpulan dokumen tersebut. Pada penelitian ini dilakukan penggalian informasi berupa pemodelan topik dan analisis tren topik dari dokumen teks abstrak skripsi Program Studi Sastra Ingris UINSA tahun 2014 sampai 2019. Dari 720 dataset abstraks skripsi dilakukan pemodelan topik dengan metode *Latent Semantic Analysis* yang meliputi *preprocessing* data teks (*case folding*, *tokenizing*, *stemming*, dan *filtering*), pembobotan *term feature selection*, dan perhitungan *Singular Value Decomposition*. Dari pemodelan topik dengan metode LSA dihasilkan 37 topik yang terbagi menjadi dua jenis topik yakni 20 topik linguistik dan 17 topik literatur. Kemudian dengan melakukan analisis penentuan topik disetiap dataset abstrak, diperoleh 7 tren topik untuk masing-masing jenis penelitian. Penelitian didominasi oleh penelitian linguistik tindak tutur yang termasuk dalam bidang sosiolinguistik. Berdasarkan jumlah jenis penelitian yang terbentuk dibandingkan dengan *data real* jenis penelitian Program Studi Sastra Inggris Universitas Islam Negeri Sunan Ampel Surabaya, menghasilkan penelitian linguistik memiliki rata-rata presisi 80 persen dan *recall* 90 persen, sedangkan jumlah penelitian literatur memiliki rata-rata presisi 74 persen dan *recall* 57 persen, serta untuk tingkat akurasi dari analisis jenis penelitian memiliki rata-rata 79 persen.

Kata Kunci: Pemodelan Topik, *Latent Semantic Analysis, Trend Topic*.

ABSTRACT

TOPIC MODELING OF THESIS DOCUMENT USING LATENT SEMANTIC ANALYSIS METHOD

Thesis is the final research for students to obtain a bachelor's degree. With the increasing number of thesis documents, it will form an information that can be taken from the document collection. In this research, information gathering was carried out in the form of topic modeling and topic trend analysis from the abstract text documents of the UINSA English Literature Study Program from 2014 to 2019. From 720 abstraction datasets the thesis was carried out topic modeling using the Latent Semantic Analysis method which includes preprocessing text data (case folding, tokenizing, stemming, and filtering), term feature selection weighting, and Singular Value Decomposition calculations. From the topic modeling using the LSA method, 37 topics were divided into two types of topics namely 20 linguistic topics and 17 literature topics. Then by analyzing the topic determination in each abstract dataset, 7 topic trends are obtained for each type of research. Research is dominated by speech act linguistic research which is included in the field of sociolinguistics. Based on the number of types of research formed compared with real data types of research in English Literature Study Program at Sunan Ampel State Islamic University in Surabaya, resulting in linguistic research has an average precision of 80 percent and recall 90 percent, while the number of literary studies has an average precision of 74 percent and 57 percent recall, and for the accuracy of the analysis of this type of research has an average of 79 percent.

Keywords: Topic Modeling, Latent Semantic Analysis, Trend Topic.

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I
PENDAHULUAN

1.1 Latar Belakang

Dalam dunia pendidikan tinggi, skripsi atau tugas akhir/ skripsi mahasiswa adalah syarat yang digunakan perguruan tinggi Indonesia untuk memperoleh gelar Sarjana Strata-1. Dalam skripsi terdapat abstrak yang berisi penjelasan singkat terhadap keseluruhan isi penelitian, dimana isi dalam skripsi akan dirangkum dalam kalimat-kalimat yang lebih ringkas (Setiawan et als., 2017). Dengan melihat dokumen skripsi yang selalu bertambah setiap tahunnya, maka seharusnya terdapat informasi dari kumpulan dokumen skripsi tersebut. Namun pada umumnya jumlah informasi yang dapat disarikan dari dokumen skripsi selama ini tidak terdapat kajian lebih lanjut, walaupun jumlah dokumen skripsi semakin bertambah (Prilianti & Wijaya, 2014). Sehingga, dengan adanya lulusan Sarjana Strata-1 setiap tahun dan dokumen skripsi yang terus bertambah, perlu untuk menerapkan metode *text mining* dalam penggalian dan pengelolaan informasi pada kumpulan dokumen skripsi tersebut (Hudaya et al., 2018).

Pada penelitian sebelumnya dengan judul "*Topic Modelling* Skripsi Menggunakan Metode *Latent Dirichlet Allocation*" (Alfanzar, 2019) dihasilkan sejumlah topik penelitian dari dokumen skripsi tahun 2014 sampai 2019 pada Program Studi Sastra Inggris UIN Sunan Ampel Surabaya (UINSA). Hal ini yang menjadi landasan dalam penelitian ini menggunakan salah satu teknik *text mining* yakni selain menemukan sejumlah topik penelitian juga untuk mencari informasi jumlah tren topik dalam penelitian skripsi setiap tahun. *Text mining* sendiri, berasal dari metode *data mining* yang dikembangkan untuk klasifikasi topik tugas akhir mahasiswa/ skripsi. Teknik dalam *text mining* seperti yang telah disinggung sebelumnya berhubungan dengan struktur kata dalam dokumen. Sehingga dalam mencari informasi tersebut memerlukan suatu teknik yang dapat membaca, mengambil, dan menampilkan intisari dari kumpulan *text* dalam dokumen menjadi sebuah informasi.

Algoritma pada *text mining* dibuat agar dapat mengidentifikasi data yang bersifat semi terstruktur yakni seperti abstrak, sinopsis, maupun isi dokumen.

Klasifikasi teks pada topik skripsi merupakan suatu proses yang mengelompokan suatu teks dalam skripsi yang mewakili isi keseluruhan skripsi untuk dikelompokkan kedalam suatu kategori tertentu (Somantri et al., 2017). Sehingga salah satu pengolahan dokumen dengan *text mining* melalui pemodelan topik (*topic modeling*) bisa menjadi alternatif solusi untuk mengklasifikasikan topik suatu dokumen, yang mana klasifikasi dilakukan dengan melihat kumpulan kata yang merupakan representasi suatu dokumen (Suhartono, 2015). Sehingga dengan penentuan topik yang telah dilakukan, selanjutnya dapat dilihat tren tiap topik penelitian skripsi yang diangkat mahasiswa.

Usaha untuk pemodelan topik pada dokumen disebut ekstraksi topik. Salah satu metode untuk mengekstrak topik adalah *Latent Semantic Analysis (LSA)* (Farida et al., 2019). Pada penelitian sebelumnya telah dibahas pemodelan topik dengan metode LSA, seperti pada penelitian dengan judul "Algoritma *Latent Semantic Analysis (LSA)* Pada Peringkas Dokumen Otomatis untuk Proses *Clustering* Dokumen" memiliki akurasi *clustering* yang baik pada suatu dokumen dengan persentase *summary document* 40% menghasilkan nilai *f-measure* dengan rata-rata 71.04% (Luthfiarta et al., 2013). Kemudian pada penelitian "Klasifikasi Topik *Multi Label* pada Hadis Shahih Bukhari Menggunakan *K-Nearest Neighbor* dan *Latent Semantic Analysis*" mendapatkan hasil akurasi dengan kombinasi LSA-KNN sebesar 90.28% dengan waktu komputasi sebesar 19 menit 21 detik (Hidayati et al., 2020). Sedangkan pada studi kasus Program Studi Sastra Inggris UIN Sunan Ampel Surabaya, pemodelan topik menggunakan metode *Latent Dirichlet Allocation* yang dibahas dalam penelitian "*Topic Modelling* Skripsi Menggunakan Metode *Latent Dirichlet Allocation*" dan berhasil menghasilkan sejumlah topik penelitian dari dokumen skripsi tahun 2014 sampai 2019 (Alfanzar, 2019).

Dengan melihat penelitian dari studi kasus yang sama yakni pada penelitian (Alfanzar, 2019), maka penelitian ini dimaksudkan untuk mengetahui informasi tren dari topik yang terbentuk pada penelitian skripsi Program Studi Sastra Inggris UINSA setiap tahun antara tahun 2014 sampai 2019. Metode *Latent Semantic Analysis* (LSA) digunakan untuk mengambil kata-kata semantik yang merupakan representasi isi suatu dokumen. Kemudian pemodelan topik dilakukan pada dataset abstrak skripsi Program Studi Sastra Inggris UINSA tahun 2014 sampai dengan

2019, karena abstrak merupakan interpretasi isi dokumen skripsi serta sebagai pengembangan dari penelitian sebelumnya dengan menggunakan metode *LSA* dan penggalian informasi tren topik yang diangkat dalam skripsi setiap tahun.

1.2 Perumusan Masalah

Dengan melihat latar belakang penelitian, maka rumusan masalah penelitian ini yaitu:

1. Bagaimana topik yang dihasilkan melalui proses *topic modelling* dengan metode *Latent Semantic Analysis (LSA)* pada data abstrak skripsi Program Studi Sastra Inggris UINSA tahun 2014 sampai 2019?
2. Bagaimana hasil analisis tren topik tiap tahun pada abstrak skripsi Program Studi Sastra Inggris UINSA dari tahun 2014 sampai 2019?

1.3 Tujuan Penelitian

Dengan melihat rumusan masalah, maka tujuan penelitian ini yaitu:

1. Mengetahui topik apa saja yang terbentuk dari pemodelan topik abstrak skripsi Program Studi Sastra Inggris UINSA dari tahun 2014 sampai 2019 menggunakan metode *Latent Semantic Analysis*.
2. Mengetahui tren topik tiap tahun pada data abstrak skripsi Program Studi Sastra Inggris UIN Sunan Ampel Surabaya dari tahun 2014 sampai 2019.

1.4 Batasan Masalah

Dengan melihat rumusan masalah, maka batasan masalah penelitian ini yaitu:

1. Studi kasus penelitian adalah pada program studi Sastra Inggris UINSA.
2. Data yang digunakan adalah abstrak skripsi dalam penelitian berasal dari *digital library (digilib)* UINSA divisi Program Studi Sastra Inggris tahun 2014 sampai 2019 berjumlah 720 baris data.
3. Metode yang digunakan dalam pemodelan topik adalah *Latent Semantic Analysis (LSA).*

1.5 Manfaat Penelitian

Penelitian ini diharap mampu memberi manfaat, baik dalam hal akademis maupun praktis yaitu:

1.5.1 Manfaat Akademis.

1. Menambah wawasan tentang proses implementasi metode *LSA* dalam menentukan macam topik penelitian di Program Studi Sastra Inggris UINSA.
2. Mengembangkan hasil dari penelitian sebelumnya untuk menemukan informasi tren pada topik penelitian yang dihasilkan

1.5.2 Manfaat Praktis.

1. Membantu pihak Program Studi Sastra Inggris (Prodi Sasing) UINSA dalam mengetahui topik dalam penelitian skripsi 2014-2019.
2. Membantu pihak Prodi Sasing UINSA dalam melihat tren topik skripsi setiap tahun sejak 2014 sampai 2019.

BAB II
KAJIAN PUSTAKA

2.1. Tinjauan Penelitian Terdahulu

Guna memperoleh pemahaman mengenai pemodelan topik, maka dilakukan tinjauan penelitian terkait dengan *topic modelling* pada Program Studi Sastra Inggris UINSA dan penggunaan metode *Latent Semantic Analysis* pada Tabel 2.1 berikut.

Tabel 2.1. Tinjauan Penelitian Terdahulu

Penelitian	Hasil	Korelasi dengan Penelitian
Topic Modelling Skripsi Menggunakan Metode *Latent Dirichlet Allocation*	Hasil klaster topik terbagi dalam 7 klaster topik yang didapatkan dari iterasi ke 1000 pada abstrak skripsi Program Studi Sastra Inggris (UINSA). Dan topik klaster 3 dinilai menjadi topik yang paling sesuai terhadap keseluruhan penelitian	Penelitian melakukan pemodelan topik pada studi kasus yang sama (Program Studi Sastra Inggris UINSA), namun dengan metode yang berbeda.
Klasifikasi Topik *Multi Label* pada Hadis Shahih Bukhari Menggunakan *K-Nearest Neighbor* dan *Latent Semantic Analysis* (Hidayati et al., 2020)	Klasifikasi Hadis Shahih Bukhari ke dalam topik anjuran, larangan, dan informasi terbukti dalam mengatasi kekurangan KNN. Memiliki performansi 0.05% lebih baik dan waktu komputasi 18 menit lebih cepat.	Menggunakan metode yang sama (LSA) untuk pemodelan topik. Namun pada penelitian tidak ditetapkan lebih dulu kategori topiknya.
Deteksi Kemiripan Bagian-bagian Terjemah Al-Qur'an dengan Menggunakan Metode *Latent Semantic Analysis* (Jadhira, Bijaksana, & Wahyudi, 2018)	Menghasilkan tingkat kemiripan dari dua atau lebih halaman yang dipasangkan pada terjemahan Al-Qur'an Bahasa Inggris. Dari pengujian dengan parameter *Rank K* maksimum didapatkan akurasi dan F- measure yaitu 100%.	Pada penelitian menggunakan metode LSA untuk mengecek tingkat kepentingan suatu *term*/ kata berdasarkan kemiripan sebagai bobot untuk pemodelan topik.
Implementasi *Semantic Search* pada *Open Library* menggunakan Metode *Latent Semantic Analysis* (Azharyani & Kusumo, 2019)	Mampu memberikan informasi relevan dengan rata-rata nilai *precision* dan *recall* yang baik. Nilai rata-rata *precision* 57.12% dan nilai rata-rata *recall* 85.08%.	Penelitian berusaha membuat keluaran informasi berupa daftar topik relevan terhadap data masukkan.
Peringkasan Teks Berita Berbahasa Indonesia Menggunakan Metode *Latent Semantic Analysis*	Hasil ringkasan sebesar 50% dari teks dokumen asli dengan nilai *recall* tertinggi yang didapat sebesar 0.71, nilai	Penelitian menggunakan metode untuk meringkas dokumen pada tahap

(LSA) dan Teknik *Steinberger & Jezek* (Saputra, Jerry. Fachrurrozi, 2017)	*precision* tertinggi 0.75, dan nilai *f-measure* tertinggi 0.71	awal untuk mendukung dalam pemodelan topik.

Tabel 2.1 sebagai rujukan penelitian sebelumnya dengan studi kasus serupa, yakni pada penelitian (Alfanzar, 2019) melakukan pemodelan topik (*Topic Modelling*) skripsi menggunakan metode *Latent Dirichlet Allocation*. Dalam penelitian tersebut menghasilkan topik pada skripsi Program Studi Sastra Inggris UINSA meliputi *strategy, maxim, utterance, politeness, advertisement, speech, type, character, positive,* dan *movie*. Sedangkan pada penelitian ini akan mencoba untuk menghasilkan topik dengan metode berbeda serta mencoba untuk melihat tren dari topik penelitian yang terbentuk.

Untuk penelitian yang menggunakan metode *Latent Semantic Analysis* memiliki beberapa kegunaan, seperti klasifikasi topik dokumen, pendeteksi kemiripan semantik (makna) pada tiap halaman dokumen, pencarian dokumen berdasarkan kata semantik (makna), dan peringkas dokumen. Untuk kegunaan klasifikasi, dapat dilihat dalam penelitian (Hidayati et al., 2020), yakni berusaha membangun sistem untuk klasifikasi hadis shahih bukhari terjemahan bahasa Indonesia yang mampu mengenali hadis berdasarkan jenis informasinya yaitu anjuran, larangan, dan informasi. Pada penelitian tersebut juga mengevaluasi sistem dengan *F1*-Score. Dari hasil evaluasi menunjukkan bahwa penambahan fitur *Rule-Based* pada proses ekstraksi dokumen dengan *TF-IDF* menunjukkan nilai yang lebih baik pada pengenalan kata anjuran daripada tanpa penambahan fitur *Rule-Based*. Kemudian pada penggunaan *stemming* dalam pembangunan sistem untuk *preprocessing data* tidak memberikan performansi yang lebih baik daripada tanpa *stemming*, hal ini dikarenakan stemming dapat menghilangkan karateristik dari setiap topik. Korelasi dengan penelitian ini adalah penggunaan metode *Latent Semantic Analysis* digunakan untuk pemodelan topik dengan tidak adanya penetapan kategori topik terlebih dahulu.

Kemudian, untuk deteksi kemiripan kata berdasarkan semantik (makna kata), dapat dilihat pada penelitian (Jadhira et al., 2018) yakni Deteksi Kemiripan Bagian-bagian Terjemah Al-Qur'an dengan Menggunakan Metode *Latent Semantic Analysis* menunjukkan bahwa metode yang digunakan cukup baik dalam mendeteksi kemiripan pasangan halaman yang mengandung kesamaan semantik

yang tinggi, terbukti dengan hasil nilai kemiripan yang mendekati 1. Korelasi dengan penelitian adalah penggunaan metode *Latent Semantic Analysis* untuk mengecek tingkat kepentingan suatu *term*/ kata berdasarkan kemiripan untuk mendukung proses pembobotan dalam pemodelan topik.

Selanjutnya, untuk pencarian dokumen berdasarkan kata semantik, terdapat pada penelitian (Azharyani & Kusumo, 2019) yakni Implementasi *Semantic Search* pada *Open Library* menggunakan Metode *Latent Semantic Analysis*, menunjukkan penggabungan *LSA* dan *weighted tree* didapatkan rata-rata *precision* 57.12% dan rata-rata *recall* 85.08%. Sehingga dapat disimpulkan bahwa sistem pencarian dengan penggabungan *LSA* dan *weighted tree* dapat digunakan untuk mendapatkan informasi yang relevan dalam merepresentasikan hubungan antar kata dan dokumen yang berkaitan, namun rendah dalam menentukan dalam mengukur ketepatan pencarian. Korelasi dengan penelitian adalah berusaha membuat keluaran informasi, namun perbedaannya adalah informasi yang ditampilkan berupa daftar topik yang relevan terhadap data masukkan.

Dan untuk peringkasan dokumen, dapat dilihat pada penelitian (Saputra et al., 2017) dengan judul Peringkasan Teks Berita Berbahasa Indonesia Menggunakan Metode *Latent Semantic Analysis (LSA)* dan Teknik *Steinberger & Jezek*, menghasilkan ringkasan dokumen sebesar 50% dari dokumen asli dengan nilai *recall* tertinggi yang didapat sebesar 0.71, nilai *precision* tertinggi sebesar 0.75, dan nilai *f-measure* tertinggi sebesar 0.71. Korelasi dengan penelitian adalah penggunaan metode *Latent Semantic Analysis* untuk meringkas dokumen, namun bedanya tujuan dari peringkasan adalah untuk mendukung penentuan topik.

Dari penelitian yang pernah dilakukan, metode *Latent Semantic Analysis* memiliki banyak kegunaan dalam pengelolaan atau penggalian informasi suatu dokumen, sehingga dalam penelitian ini mencoba untuk mengembangkan salah satu kegunaan metode *Latent Semantic Analysis* yakni untuk pemodelan topik. Pengembangan yang dilakukan berkaitan dengan pemodelan topik skripsi Program Studi Sastra Inggris untuk melihat kesesuaian dengan penelitian sebelumnya dan lebih lanjut melihat tren topik pada dokumen skripsi tahun 2014 sampai 2019.

2.2. Bahasa *Phyton*

Python dikembangkan pada tahun 1989, oleh Guido van Rossum dan diperkenalkan pada tahun 1991. *Python* sendiri disebut bahasa pemrograman *high level language* (tingkat tinggi) yang dibuat untuk mempermudah *developer* aplikasi dalam pekerjaannya. *Python* dirancang untuk menunjang efisiensi waktu, memudahkan dalam pengembangan program, serta memiliki kompatibilitas dengan berbagai sistem (Qutsiah, Sophan, & Hendrawan, 2016).

Bahasa pemrograman *Python* juga disebut sebagai alat bantu yang dapat digunakan pada proses *text mining* dan tampilan grafis pendukungnya (Petrus, 2019).

2.3. *Text Mining*

Text mining dalam Bahasa Indonesia artinya menambang teks adalah teknik analisis teks yang sumber datanya berasal dari suatu dokumen. *Text Mining* berfungsi untuk mencari intisari berupa kata atau kumpulan kata yang mewakili isi/ informasi dari keseluruhan dokumen. Sehingga dengan adanya intisari dari suatu dokumen, maka dapat dilangsungkan analisis keterkaitan, dan kelas antar dokumen (Hartanto, 2017). Lebih lanjut *text mining* melakukan pengolahan pada data teks yang tidak terstruktur yang merupakan bagian dari keilmuan *information retrieval* (temu balik informasi). *Text mining* memiliki tiga tahapan utama secara berurut yakni *text preprocessing, feature selection,* dan *text analytic* (Priyanto & Ma'arif, 2018). Penjelasan lebih lanjut mengenai tahapan dalam *text mining* adalah sebagai berikut.

2.3.1. Text Preprocessing

Pada tahap *text preprocessing* memiliki fungsi yakni sebagai tahap awal pengolahan teks sebelum diolah lebih lanjut. Data teks yang tidak terstruktur memiliki *noise* seperti angka, karakter khusus, tanda baca atau simbol, dan imbuhan. Tahap ini, data teks dibersihkan/ dinormalkan hingga tersisa bentuk kata dasar saja, yang kemudian dapat dianalisis lebih lanjut (Priyanto & Ma'arif, 2018). Tahapan *text processing* secara umum terdiri dari 4 langkah, yaitu *Case folding, Tokenizing, Stemming,* dan *Filtering* (Hermawan & Ismiati, 2020) yang dijelaskan lebih lanjut sebagai berikut.

1. *Case Folding*

Proses *case folding* dapat dipahami sebagai proses menghilangkan karakter-karakter selain huruf, yakni tanda baca dan angka. Serta mengubah huruf menjadi *lowercase* atau *uppercase* (Rahman, 2017).

2. *Tokenizing*

Tokenizing/ Tokenization merupakan proses yang dimaksudkan untuk memotong kalimat berdasarkan tiap kata yang menyusunnya (Rahman, 2017). Dapat dipahami juga sebagai proses pemotongan satu kalimat menjadi beberapa kata.

3. *Stemming*

Stemming dapat dipahami sebagai proses untuk memotong imbuhan atau mengembalikan suatu kata berimbuhan menjadi kata dasar (Rahman, 2017). Hal ini bertujuan agar setiap kata yang memiliki kata dasar sama dapat dikelompokkan menjadi satu kelompok.

4. *Filtering*

Filtering atau *stopword removal* dapat dipahami sebagai proses penghilangan *stopwords* atau kata-kata yang tidak menggambarkan isi tulisan, sehingga dapat dibuang. Proses ini bertujuan untuk mengurangi jumlah kata yang tidak menggambarkan informasi (Rahman, 2017).

2.3.2. *Feature Selection*

Dalam proses *feature selection* ini bertujuan untuk menemukan kata kunci yang menjadi ciri dari suatu dokumen yang membedakan kata antar dokumen dalam satu korpus. Tahap ini memiliki peran yang penting dalam akurasi *text analytic*. Empat pendekatan yang umum dalam *feature selection* adalah *Document Frequency (DF), Term Frequency (TF), Inverse Document Frequency (IDF)* dan *Term Frequency/Inverse Document Frequency (TF/IDF)* (Priyanto & Ma'arif, 2018). Penjelasan mengenai pendekatan dalam *feature selection* adalah sebagai berikut.

1. *Document Frequency* (DF)

Prinsip dari DF adalah membuang kata yang umum pada suatu dokumen yang ada pada suatu korpus. Sehingga kata yang tersisa dalam suatu dokumen

hanya yang memiliki *overlapping* rendah dibanding kata pada dokumen lain dalam suatu korpus.

2. *Term Frequency* (TF)

Dibandingkan dengan DF, TF tidak melihat kata yang terkandung dalam dokumen lain. Metode TF secara sederhana dipahami sebagai perhitungan jumlah kata dalam suatu dokumen. Kata yang memiliki jumlah kemunculan tinggi akan menjadi ciri dari suatu dokumen dimana kata tersebut berada. TF disimbolkan dengan $tf_{i,j}$(Jumlah kemunculan kata t_j dalam dokumen d_i).

3. *Inverse Document Frequency* (IDF)

IDF memiliki kesamaan dengan TF, yakni menghitung frekuensi kemunculan suatu kata. Namun pada TF hanya menghitung kemunculan kata pada satu dokumen teks, sedangkan IDF menghitung kemunculan kata pada keseluruhan korpus dokumen. IDF diperoleh dengan persamaan sebagai berikut:

$$idf = \log\left(\frac{D}{df_i}\right) \tag{2.1}$$

Keterangan:

D = Jumlah dokumen

Df_i= Jumlah kemunculan kata dalam dokumen

4. *Term Frequency/Inverse Document Frequency* (TF/IDF)

TF/IDF adalah gabungan dari TF dan IDF, yakni dengan mengambil rasio antara nilai TF dan IDF. TF/IDF juga bertujuan untuk mencari bobot suatu kata terhadap dokumen. Persamaan TF/IDF dinotasikan sebagai berikut:

$$W_{i,j} = tf_{i,j} \times idf = tf_{i,j} \times \log\left(\frac{D}{df_i}\right) \tag{2.2}$$

Keterangan:

$W_{i,j}$ = Bobot kata t_j terhadap dokumen d_i

$tf_{i,j}$ = Jumlah kemunculan kata t_j dalam dokumen d_i

D = Jumlah dokumen

df_i = Jumlah kemunculan kata dalam dokumen

2.3.3. *Text Analytic*

Text analytic merupakan tahapan terakhir dari *text mining*. Pada tahap ini data teks yang sudah dibersihkan, diolah lebih lanjut dengan menggunakan berbagai

macam algoritma untuk berbagai kebutuhan analisis. Dua jenis *text analytic* yang paling sering dilakukan adalah *topic modeling* dan *sentiment analysis*. *Topic modeling* sendiri merupakan pendekatan untuk mengelompokkan teks/dokumen teks kedalam beberapa kategori secara otomatis berdasarkan tingkat kesamaan *term*/kata kunci (Priyanto & Ma'arif, 2018). Pada tahap *text analytic* ini yang selanjutnya akan menggunakan teknik *topic modeling* dengan metode *Latent Semantic Analysis*.

2.4. *Topic Modeling*/ Pemodelan Topik

Topic modeling merupakan teknik untuk menyimpulkan topik yang tersembunyi dalam dokumen. *Topic modeling* mewakili setiap dokumen sebagai gabungan dari beberapa topik dan setiap topik merupakan kombinasi gabungan dari beberapa kata, maka *topic modeling* merupakan alat dalam *text mining* untuk mengklasifikasikan dokumen berdasarkan topik nya (M. A. Putra et al., 2019).

Konsep *topic modeling* terdiri dari beberapa entitas yaitu "kata", "dokumen", dan "*corpora*". "Kata" merupakan unit dasar dalam dokumen. "Dokumen" merupakan susunan seluruh kata. Dan "*corpus*" adalah kumpulan dokumen dan *corpora* merupakan bentuk jamak dari *corpus*. Sementara "*topic*" merupakan distribusi dari beberapa kata yang bersifat tetap. Atau secara sederhana, setiap dokumen dalam *corpus* mengandung proporsi kata pembentuk topik sesuai kata-kata yang terkandung di dalamnya (K. B. Putra & Kusumawardani, 2017). Salah satu metode *topic modeling* yang dapat digunakan adalah *Latent Semantic Analysis (LSA).*

2.4.1 Latent Semantic Analysis

Latent Semantic Analysis (LSA) merupakan teori atau metode statistik aljabar yang melakukan ekstraksi struktur kata semantik yang tersembunyi dari kalimat berupa himpunan *term* dari dokumen (Jadhira et al., 2018). *LSA* dipatenkan pada tahun 1988 oleh Karen Lochbaum, Susan Dumais, Scott Deerwester, Richard Harshman, George Furnas, Thomas Landauer, dan Lynn Streeter. Dalam konteks pencarian informasi, metode *LSA* disebut sebagai *Laten Semantic Indexing (LSI)*. *LSA* dapat ditafsirkan sebagai cara yang cepat dan praktis untuk mendapatkan perkiraan *substitutability* kontekstual penggunaan kata-kata dalam segmen teks yang besar yang belum ditentukan makna kesamaan antara kata-kata dan segmen

teks yang mungkin mencerminkan suatu hubungan tertentu. Sebagai metode praktis untuk mengkarakterisasi arti dari kata, *LSA* menghasilkan ukuran hubungan antar kata, dan bagian-bagian yang berkorelasi dengan beberapa fenomena kognitif manusia yang melibatkan asosiasi atau kesamaan semantik. (Ngafifudin, 2017).

Teknik *Latent Semantic Analysis (LSA)* banyak digunakan dalam *Natural Language Processing* (Ngafifudin, 2017). *LSA* didasarkan pada teknik matematika yang diberi nama *Singular Value Decomposition (SVD)* (Hidayati et al., 2020). LSA memiliki metodologi yang dimulai secara berurutan mulai *Preprocessing Data*, *Feature Selection*, dan perhitungan *Singular Value Decomposition (SVD)*.

SVD sendiri adalah salah satu teknik aljabar linear yang digunakan untuk menguraikan (dekomposisi) matriks menjadi tiga buah matriks baru, yaitu matriks orthogonal U, matriks diagonal S, dan Transpose matriks orthogonal V. Rumus *SVD* sendiri dinotasikan sebagai berikut:

$$A_{m\times n} = U_{m\times n}.S_{n\times n}.V^{T}_{n\times n} \tag{2.3}$$

Keterangan:

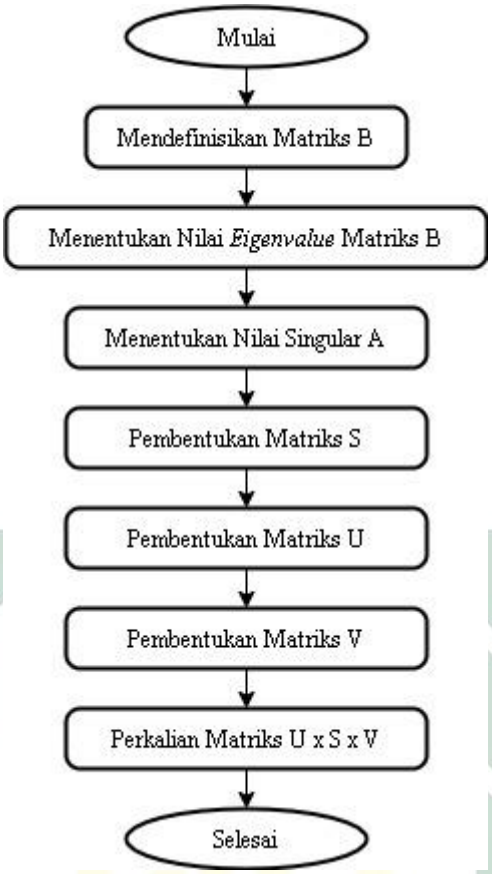
$A_{m\times n}$= matriks A dengan nilai m> n

$U_{m\times n}$= matriks ortogonal berukuran m × n

$S_{n\times n}$= matriks diagonal berukuran n × n, dengan elemen matriks positif atau nol

$V^{T}_{n\times n}$ = matriks ortogonal berukuran n × n, yang merupakan transpose matriks V.

Berdasarkan persamaan *Singular Value Decomposition* pada bagian sebelumnya, perhitungan memerlukan nilai dari matriks U, S, dan V. Sehingga berikut akan dijelaskan tahapan sistematik dalam pencarian nilai matriks U, S, dan V digambarkan pada Gambar 2.1 berikut.



Gambar 2.1. Metodologi *Singular Value Decamposition*

1. Mendefinisikan Mariks B

Pada tahap ini terdapat aturan dalam mendefinisikan matrik B. Berdasarkan matriks $A_{(m\times n)}$ yang diberikan adalah sebagai berikut.

$$\text{Jika } m \leq n \rightarrow B = AA^T \quad (2.4.1)$$

$$\text{Jika } m > n \rightarrow B = A^T A \quad (2.4.2)$$

Keterangan:

B = Matriks B

A = Matriks A

A^T = Transpose Matriks A

Sehingga dengan aturan tersebut menghasilkan matriks bujursangkar dengan dimensi m atau n yang lebih kecil dari matriks A.

2. Menentukan Nilai *Eigenvalue* Matriks B

Pada bagian ini bertujuan untuk menghasilkan nilai λ dari matriks B yang telah diperoleh sebelumnya. Nilai ini disebut nilai *eigenvalue* yang akan digunakan dalam menentukan nilai *eigenvektor* pada normalisasi matrik selanjutnya.

Persamaan karakteristik untuk mendapatkan nilai *eigenvalue* adalah sebagai berikut.

$$|\lambda I - B| = 0 \tag{2.5}$$

Keterangan:

λI = nilai *eigenvalue* sama dengan B

B = nilai matriks B

$|\lambda I - B|$= determinan nilai *eigenvalue* dengan matriks B

Dari Persamaan 2.5 memiliki arti bahwa, terdapat pengurangan nilai λ dengan masing-masing matriks B yang kemudian di determinan kan. Dan hasil determinan memiliki hasil sama dengan nol.

3. Menentukan Nilai *Singular* A (akar kuadrat positif *eigenvalue* matrik B)

Setelah mendapatkan nilai λ, maka untuk mendapatkan nilai singular A adalah sebagai berikut.

$$\sigma_i = \sqrt{\lambda_i} \tag{2.6}$$

Keterangan:

σ_i = nilai singular A

$\sqrt{\lambda_i}$ = akar dari nilai *eigenvalue*

Dari Persamaan 2.6 didapatkan hasil nilai singular yang akan digunakan untuk pembentukan matriks S pada tahap selanjutnya.

4. Pembentukan Matrik S

Dari nilai singular A yang telah didapat selanjutnya membentuk matriks S dengan memperhatikan syarat/ aturan pada matriks A sebagai berikut.

$$\text{Jika } m \leq n\text{, Maka matriks S} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & \sigma_m \end{bmatrix} \tag{2.7.1}$$

$$\text{Jika } m > n\text{, Maka matriks S} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_n \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & 0 \end{bmatrix} \tag{2.7.2}$$

5. Pembentukan Matriks U dan V

Pembentukan kolom matriks U dan V dilakukan dengan normalisasi *eigenvector*. Namun bedanya untuk pembentukan matriks U dilakukan normalisasi

pada matriks AA^T. Sedangkan untuk pembentukan matriks V dilakukan normalisasi pada matriks $A^T A$. Persamaan normalisasi *eigenvector* pada matriks adalah sebagai berikut.

$$(\lambda I - B)x = \begin{bmatrix} \lambda - B & B \\ B & \lambda - B \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad (2.8)$$

Setelah pembentukan matriks U, S, dan V selesai, maka selanjutnya dilakukan perkalian matriks U, S, dan V^T untuk mendapatkan nilai matriks A yang merupakan representasi dari bobot suatu *term*/ kata (Persamaan 2.3). Pada tahap ini suatu kalimat telah diperingkas dan telah dipilih sejumlah term/ kata yang memiliki nilai kepentingan didalam suatu dokumen. Sehingga dengan ini dapat ditentukan topik dari suatu dokumen tersebut untuk kemudian di kalkulasikan pada sebuah tren data.

2.5. Integrasi Keilmuan

Integrasi keilmuan disini bermaksud untuk memaparkan konsep penelitian dalam sudut pandang Islam. Pada bagian ini akan dibahas mengenai topik penelitian yang memiliki hubungan atau korelasi dengan firman Allah Swt dalam kitab suci Al-Qur'an. Topik yang diangkat adalah *topik modeling*/ pemodelan topik yang berkaitan juga dengan pengelompokan/ penggolongan suatu dokumen.

Topic Modeling/ Pemodelan Topik merupakan teknik untuk menyimpulkan topik yang tersembunyi dalam dokumen (M. A. Putra et al., 2019). Hal ini berkaitan dengan penggalian informasi atau makna yang tersembunyi dari suatu berita. Jika pada ilmu keislaman hal ini mirip dengan ilmu tafsir. Tafsir dalam bahasa Arab merupakan bentuk masdar dari kata (*fassara*) yang mana definisi secara bahasa berarti mengungkap dan menampakkan. Kata *tafsir* juga dapat diartikan menerangkan sesuatu yang masih samar atau tidak jelas serta menampakkan atau menyingkap sesuatu yang tertutup (M. Yunus, 2016). Dengan kata lain *topic modeling* dapat dipahami sebagai bentuk penafsiran terhadap dokumen guna menemukan informasi yang tersembunyi atau masih samar didalamnya.

Pentingnya suatu informasi itu di tafsirkan atau digali kembali agar dapat menemukan makna dari informasi itu merupakan bentuk suatu usaha yang dalam agama Islam sering disebut dengan *Tabayyun*. Pengertian *tabayyun* sendiri dalam tafsir Alquran Departement Agama, 2004 pada ayat 6 surat al-Hujurat. Kata *tabayyun* adalah *fiil amr* untuk jamak dari kata kerja *tabayyana*. Masdarnya *at-*

tabayyun, yang artinya adalah mencari kejelasan hakekat sesuatu atau kebenaran suatu fakta dengan teliti, seksama dan hati-hati (Efendi, 2016). Ayat yang melandasi konsep *tabayyun* ini adalah pada surah al-Hujurat ayat 6 sebagai berikut.

يَٰٓأَيُّهَا ٱلَّذِينَ ءَامَنُوٓاْ إِن جَآءَكُمۡ فَاسِقُۢ بِنَبَإٖ فَتَبَيَّنُوٓاْ أَن تُصِيبُواْ قَوۡمَۢا بِجَهَٰلَةٖ فَتُصۡبِحُواْ عَلَىٰ مَا فَعَلۡتُمۡ نَٰدِمِينَ

Artinya:

"Hai orang-orang yang beriman, jika datang kepadamu orang fasik membawa suatu berita, maka periksalah dengan teliti agar kamu tidak menimpakan suatu musibah kepada suatu kaum tanpa mengetahui keadaannya yang menyebabkan kamu menyesal atas perbuatanmu itu (QS. 49:6)".

Dari ayat 6 surat al-Hujarat, kata *fasiq* dan *naba* (berita) dalam bahasa Arab berarti semua bentuk ke*fasiq*an dari berita apa saja (Efendi, 2016). Dengan demikian, arti ayat di atas adalah apabila ada berita atau informasi datang kepadamu, janganlah tergesa untuk menerima dan menyebarkannya, teliti informasi tersebut dan ungkaplah informasi kebenarnnya.

Kemudian lebih lanjut *topic modeling* pada penelitian yang akan dilakukan membahas tentang penentuan topik pada suatu dokumen. Hal ini berkaitan dengan penggalian informasi semantik (tersembunyi) untuk mencari karakteristik dokumen yang kemudian ditetapkan golongan topiknya. Jika melihat dari sudut pandang Islam mengenai penggolongan atau pengelompokkan didasarkan pada Q.S al-Waaqi'ah ayat 7-13 yang berbunyi sebagai berikut.

وَكُنتُمۡ أَزۡوَٰجٗا ثَلَٰثَةٗ (٧) فَأَصۡحَٰبُ ٱلۡمَيۡمَنَةِ مَآ أَصۡحَٰبُ ٱلۡمَيۡمَنَةِ (٨) وَأَصۡحَٰبُ ٱلۡمَشۡـَٔمَةِ مَآ أَصۡحَٰبُ ٱلۡمَشۡـَٔمَةِ (٩) وَٱلسَّٰبِقُونَ ٱلسَّٰبِقُونَ (١٠)

Artinya:

"Dan kamu menjadi tiga golongan. Yaitu golongan kanan, alangkah mulianya golongan kanan itu. Dan golongan kiri, alangkah sengsaranya golongan kiri itu. Dan orang-orang yang paling dahulu (beriman), merekalah yang paling dahulu (masuk surga)".

Dari ayat diatas menggambarkan golongan manusia pada hari kiamat yang menjadi tiga golongan. Dalam proses penggolongan tersebut dapat dipahami bahwa terdapat metode atau perhitungan dari Allah Swt hingga dapat menggolongkan manusia menjadi tiga golongan.

Sehingga dari ayat diatas mengenai *tabayyun* dan pengelompokan diharapkan dapat menjadi landasan integrasi keilmuan dalam penelitian yang akan dilakukan.

BAB III
METODE PENELITIAN

Berdasarkan latar belakang masalah dan tujuan penelitian yang dipaparkan pada bab sebelumnya, maka pada bab ini akan dibahas metodologi yang dilakukan dalam penelitian sebagai berikut.

3.1 Objek Penelitian

Objek dalam penelitian ini adalah pemodelan topik/ *topic modeling* pada dokumen skripsi Program Studi Sastra Inggris UIN Sunan Ampel Surabaya.

3.2 Sumber Data

Penelitian ini menggunakan data abstrak skripsi Program Studi Sastra Inggris UINSA dari tahun 2014 sampai 2019, yang diperoleh dari http://digilib.uinsby.ac.id. Data diambil menggunakan *Web Scrapper* yang merupakan *tool ekstensi Google Chrome*.

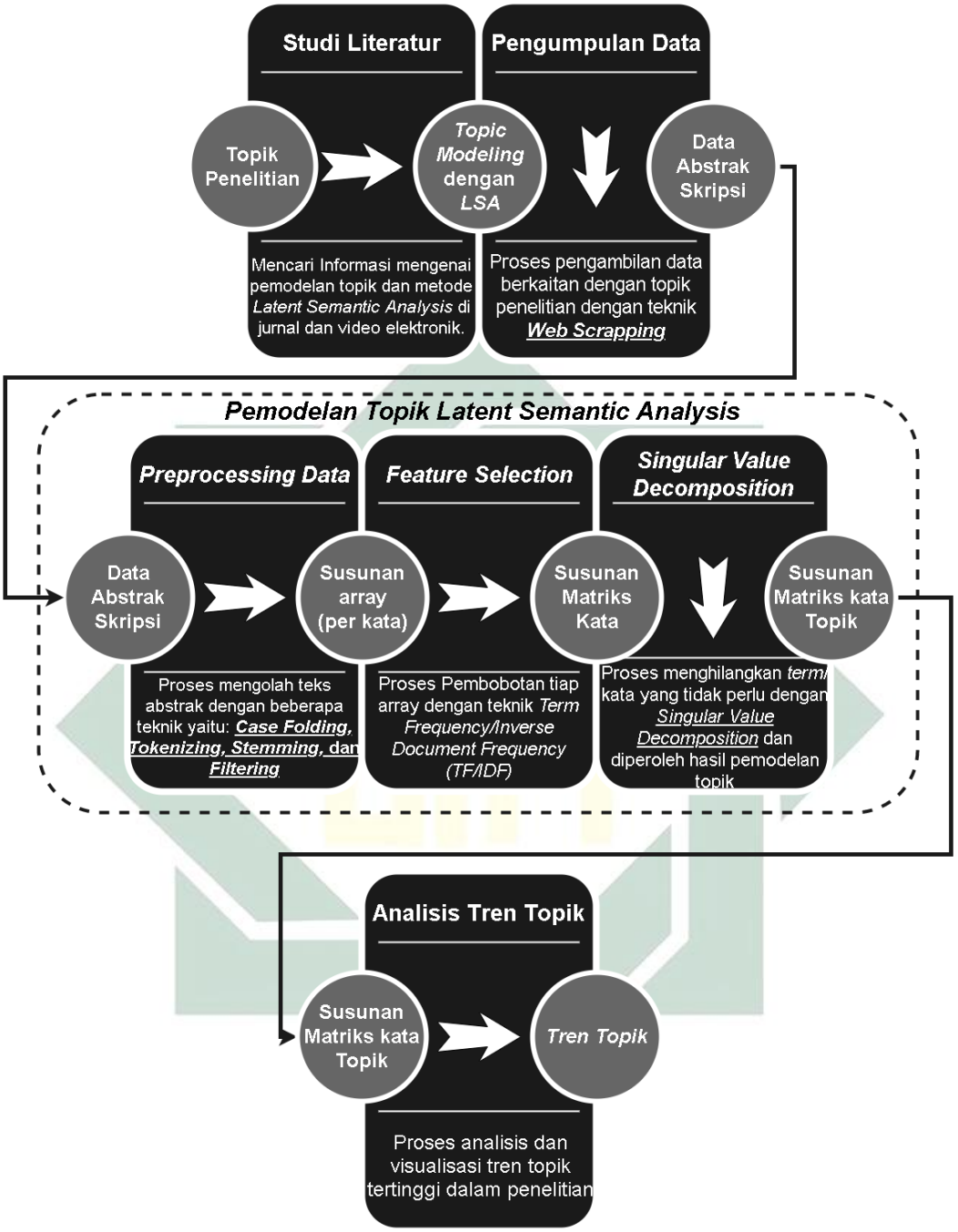
3.3 *Instrument* Penelitian

Beberapa *Instrument* atau alat bantu dalam penelitian ini secara sistematik berdasarkan Gambar 3.1 adalah sebagai berikut.

1. *Web Scrapper*, sebagai salah satu ekstensi *Google Chrome* yang digunakan pada proses pengumpulan data (*crawling data http://digilib.uinsby.ac.id*).
2. *Jupyter Notebook*, sebagai aplikasi *web open-source* untuk pengembangan *software* dengan bahasa pemrograman *python*.
3. *Library* NLTK, Sebagai *library* atau modul untuk proses *preprocessing* data dalam pemrograman *Python*.
4. *Library Scikit-Learn (sklearn) CountVectorizer* dan *TruncatSVD*, sebagai *library* atau modul pemrograman pyhton untuk proses *feature selection* dan dekomposisi matriks (SVD).
5. *Library Matplotlib*, sebagai *library* atau modul pemrograman *python* untuk proses analisis tren topik.

3.4 Desain Penelitian

Desain penelitian digambarkan dalam bentuk diagram alir yang merepresentasikan seluruh langkah dalam penelitian. Diagram alir penelitian, dapat dilihat pada Gambar 3.2 berikut.



Gambar 3.2. Metodologi Penelitian

Berdasarkan Gambar 3.2, terdapat lima tahapan utama dalam penelitian ini yang lebih lanjut dijelaskan secara sistematik sebagai berikut.

3.4.1. Studi Literatur

Tahapan pertama ini berisi proses pencarian informasi yang berhubungan dengan pemodelan topik. Yakni meliputi proses pencarian jurnal yang berkaitan dengan *topic modeling* menggunakan metode *Latent Semantic Analysis*.

3.4.2. Pengumpulan Data

Tahapan kedua pengumpulan data, berisi proses pengumpulan data abstrak skripsi pada website digilib UINSA (http://digilib.uinsby.ac.id). Pengambilan data dilakukan dengan *tool* dalam *google chrome* yaitu *web scrapper*. Dari proses pengumpulan data dengan teknik *web scrapping* didapatkan kumpulan teks abstrak sebagai data penelitian. Prinsip dari *web scrapper* ini adalah membentuk suatu jaring yang memuat urutan halaman dalam suatu website, yang mana jaring tersebut dapat digunakan untuk mengambil *element text* pada suatu *website*. Urutan dari proses pengumpulan data dengan *web scrapper* adalah sebagai berikut.

a. Mengakses *Website* Target
b. Pembuatan *Sitemap web scrapper*
c. Pembuatan *Selector web scrapper*
d. Eksekusi *Selector*/ *Scrape Data*
e. Penyimpanan Data

Pada tahapan selanjutnya, yakni tahap ketiga sampai kelima merupakan bagian utama dalam pemodelan topik. Metode yang digunakan adalah *Latent Semantic Analysis* yang secara bertahap meliputi tahap *preprocessing data*, *feature selection*, dan analisis topik dengan perhitungan *singular value decomposition*.

3.4.3. *Preprocessing* Data

Tahapan ketiga *preprocessing data* atau *text preprocessing*, berisi proses pengolahan awal data teks sebelum diolah lebih lanjut. Data teks tidak terstruktur memiliki *noise* seperti tanda baca, angka, imbuhan, karakter khusus dan lain sebagainya. Dalam tahapan ini, data teks dibersihkan hingga tersisa bentuk dasar saja yang kemudian dapat dianalisis lebih lanjut. Tahapan *text processing* secara umum terdiri dari empat langkah, yaitu *Case folding*, *Tokenizing*, *Stemming*, dan *Filtering*. Pada tahapan ini data yang telah dikumpulkan telah dinormalisasi sehingga data yang tersisa hanyalah kata benda, kata sifat, dan kata kerja dasar.

3.4.4. *Feature Selection*

Tahapan keempat *feature selection*, berisi tahapan pembobotan setiap *term*/ kata yang telah dinormalisasi dari tahap *preprocessing*. Pada tahap ini menggunakan pendekatan pada Persamaan 2.2 tentang perhitungan *Term Frequency/Inverse Document Frequency (TF/IDF)*.

TF/IDF adalah perpaduan dari pendekatan TF dan IDF, tahap ini bertujuan untuk mencari bobot suatu kata terhadap dokumen. Sehingga dengan pemberian bobot ini dapat menyeleksi *term*/ kata yang dapat diproses lebih lanjut. Pada tahap ini menghasilkan bobot setiap kata yang merepresentasikan kepentingan setiap kata dalam dokumen.

3.4.5. Pembentukan Topik (Perhitungan *Singular Value Decomposition*)

Tahap kelima pembentukan topik, berisi proses peringkasan dokumen dari kumpulan kata yang telah diberi bobot pada tahap *feature selection*. Dalam analisis topik dalam *Latent Semantic Analysis*, terdapat proses penyeleksian *term*/ kata lanjutan dengan menyusun dan melakukan perhitungan pada setiap *term*/ kata dalam bentuk matriks. Perhitungan ini bertujuan untuk memberi bobot lanjutan terhadap *term*/ kata sehingga menghasilkan topik yang sesuai. Pada tahap ini menekankan pada Persamaan 2.3 sebagai perhitungan matematis *Singular Value Decomposition (SVD)* untuk pemilihan *term*/ kata sehingga menghasilkan term/ kata yang dianggap sebagai karakteristik suatu dokumen. Sehingga dari tahap ini diharapkan dapat menghasilkan *term*/ kata yang mencerminkan topik suatu dokumen.

3.4.6. Analisis Tren

Tahap keenam analisis tren, berisi proses pemilahan klaster topik sesuai dengan jenis penelitian yang ada pada Program Studi Sastra Inggris UINSA dan proses perhitungan jumlah dokumen yang diklasifikasikan sesuai topik yang terbentuk. Jumlah dari klasifikasi topik tersebut kemudian diambil topik yang memiliki tren paling tinggi dan divisualisasikan hasilnya. Visualisasi dibuat dalam bentuk grafik yang dapat menunjukkan informasi statistik topik yang diangkat dalam penelitian skripsi mahasiswa Program Studi Sastra Inggris UIN Sunan Ampel Surabaya. Grafik yang disajikan diharapkan mampu menginformasikan tren setiap topik yang diangkat dalam penelitian setiap tahun.

3.5 Tempat dan Waktu Penelitian

Penelitian dilaksanakan sejak bulan Maret 2020. Rencana penelitian akan dilaksanakan sampai bulan Mei 2020 karena memperhatikan revisi penelitian dan lain sebagainya. *Timeline* penelitian ini, terdapat pada Tabel 3.2 berikut.

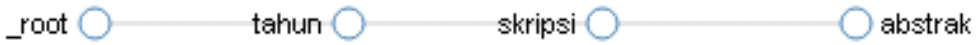
Tabel 3.2. *Timeline* Penelitian

No	Kegiatan	Maret				April				Mei			
		I	II	III	IV	I	II	III	IV	I	II	III	IV
1	Studi Literatur												
2	Pengumpulan Data												
3	*Preprocessing Data*												
4	*Feature Selection*												
5	Pembentukan Topik												
6	Analisis Tren												
7	Penyusunan Laporan												

BAB IV
HASIL DAN PEMBAHASAN

4.1 Pengumpulan Data

Proses dari pengumpulan data berlangsung sekitar 45 menit yang dilakukan dengan eksekusi *selector web scrapper chrome* yang memiliki struktur seperti Gambar 4.3 berikut.



Gambar 4.3. Struktur *Selector* Pengumpulan Data Penelitian

Urutan jaring *selector* melambangkan urutan halaman untuk memperoleh data abstrak skripsi dalam website. Tingkatan halaman untuk memperoleh abstrak skripsi yaitu halaman tahun skripsi, halaman skripsi, dan halaman abstrak skripsi yang berisikan target informasi teks.

Dari proses pengambilan data berhasil dikumpulkan berjumlah 720 dataset abstrak skripsi Program Studi Sastra Inggris UINSA. Dengan rincian data setiap tahunnya dapat dlihat pada Tabel 4.3. berikut.

Table 4.3. Jumlah Dataset Abstrak Pertahun

Tahun	Jumlah Abstrak
2014	71
2015	165
2016	152
2017	120
2018	95
2019	117

Hasil dari pengumpulan data ini adalah sebagai berikut.

1. Jumlah data yang terkumpul berjumlah 720 baris data.
2. Setiap data mengandung informasi berupa tahun skripsi, judul skripsi, dan teks abstrak skripsi.

Sehingga dari tahap pengumpulan data ini telah selesai dan dilanjutkan pada proses selanjutnya, yakni *preprocessing data*.

4.2 *Preprocessing Data*

Pada tahap ini dilakukan beberapa pengolahan pada data yang bertujuan untuk menormalkan dataset, diantaranya dengan menghilangkan karakter spesial dan mengubah semua huruf menjadi huruf kecil (*case folding*), memotong tiap kata dalam teks (*tokenizing*), mengubah kata imbuhan menjadi kata dasar (*stemming*), dan menghilangkan kata yang tidak menggambarkan isi tulisan (*filtering*). Tahap *preprocessing data* menggunakan *jupyter notebook* dengan bahasa *python.* Hasil dari tahap ini secara sistematik adalah sebagai berikut.

4.2.1 *Case Folding*

Tahap ini menghasilkan *dataset* abstrak tanpa karakter seperti angka, tanda baca, dan seluruh teks menjadi huruf kecil. *Pseudocode python* dari tahap *case folding* ini adalah sebagai berikut.

```
Case Folding
Deklarasi:
Doc[abstrak]: datasets kolom abstrak
bad_chars: list of useless characters
Algoritma:
For i=1 to Doc[abstrak] do
   String.lowercase(i)
   For j=1 to bad_chars do
      If(characters(i)==bad_chars(j)) do
         Eliminate characters(i);
      Else do
         Continue;
      End if
   End for
End for
```

Proses *case folding* menggunakan *library str* untuk pengelolaan data string. Prinsip dari *pseudocode* proses *case folding* yang pertama adalah mengubah semua karakter pada dataset abstrak yang tergolong *uppercase* menjadi *lowercase*, dan kedua menghilangkan semua karakter spesial pada dokumen berdasarkan daftar *bad_chars* atau karakter yang tidak diinginkan. *Output* data pada proses *case folding* terdapat pada Tabel 4.4. sebagai berikut.

Tabel 4.4. *Output Case Folding*

Sebelum *Case Folding*	Setelah *Case Folding*
In this study, the author discusses the "swear words" used by characters in a film by Ben Stiller\'s Tropic Thunder. This study consisted of two discussion of the problem, namely: a variety of swear words, and functions as well as	*in this study the author discusses the swear words used by characters in a film by ben stillers tropic thunder this study consisted of two discussion of the problem namely a variety of swear words and functions as well as the reasons*

the reasons for the use of swear words. The author uses descriptive research method because the author presents the results of the analysis in the form of descriptive text. The steps of this research is to categorize words that contain swear words, analyze the functions and reasons use of swear words. The results of this analysis, the authors found many swear words used by a character belonging to a "sex-term". There are four functions used by the characters in the use of swear words that expletive expression, abusive expression, humorous and swear words as an auxiliary function. Based on analysis of the most frequently used functions are functions abusive. And the reasons used in the use of swear words is due to psychological factors and social factors. Psychology is the reason that often appear based on the analysis results.	*for the use of swear words the author uses descriptive research method because the author presents the results of the analysis in the form of descriptive text the steps of this research is to categorize words that contain swear words analyze the functions and reasons use of swear words the results of this analysis the authors found many swear words used by a character belonging to a sexterm there are four functions used by the characters in the use of swear words that expletive expression abusive expression humorous and swear words as an auxiliary function based on analysis of the most frequently used functions are functions abusive and the reasons used in the use of swear words is due to psychological factors and social factors psychology is the reason that often appear based on the analysis results*

Dari hasil *case folding* pada salah satu abstrak diperoleh hasil seperti Tabel 4.5, yang mana karakter selain huruf telah dihilangkan serta mengubah seluruh teks menjadi *lowercase*.

4.2.2 *Tokenizing*

Selanjutnya pada tahap Tokenizing menghasilkan potongan-potongan kata untuk setiap dataset abstrak. *Pseudocode python* untuk proses *tokenizing* adalah sebagai berikut.

```
Tokenizing
Deklarasi:
Doc[abstrak]: datasets abstrak after case folding
Tokenize: function RegexpTokenizer for tokenizing from
nltk.tokenize library
Algoritma:
For i=1 to Doc[abstrak] do
   If(characters(i)==space) do
      Tokenize(i);
   Else do
      Continue;
   End if
End for
```

Pada proses *tokenizing* ini menggunakan *library python RegexpTokenizer*. Prinsip dari *pseudocode Tokenizing* adalah memisahkan setiap kata dalam dataset

pada hasil *output case folding* dengan fungsi *regxtoken*. Dan *Output* dari proses *tokenizing* terdapat pada Tabel 4.5. berikut.

Tabel 4.5. *Output Tokenizing*

Sebelum *Tokenizing*	Setelah *Tokenizing*
in this study the author discusses the swear words used by characters in a film by ben stillers tropic thunder this study consisted of two discussion of the problem namely a variety of swear words and functions as well as the reasons for the use of swear words the author uses descriptive research method because the author presents the results of the analysis in the form of descriptive text the steps of this research is to categorize words that contain swear words analyze the functions and reasons use of swear words the results of this analysis the authors found many swear words used by a character belonging to a sexterm there are four functions used by the characters in the use of swear words that expletive expression abusive expression humorous and swear words as an auxiliary function based on analysis of the most frequently used functions are functions abusive and the reasons used in the use of swear words is due to psychological factors and social factors psychology is the reason that often appear based on the analysis results	*'in','this','study','the','author', 'discusses','the','swear','words', 'used','by','characters','in','a', 'film','by','ben','stillers','tropic', 'thunder','this','study','consisted', 'of','two','discussion','of','the', 'problem','namely','a','variety', 'of','swear','words','and','functions', 'as','well','as','the','reasons', 'for','the','use','of','swear','words', 'the','author','uses','descriptive', 'research','method','because','the', 'author','presents','the','results', 'of','the','analysis','in','the', 'form','of','descriptive','text', 'the','steps','of','this','research', 'is','to','categorize','words','that', 'contain','swear','words','analyze', 'the','functions','and','reasons', 'use','of','swear','words','the', 'results','of','this','analysis', 'the','authors','found','many','swear', 'words','used','by','a','character', 'belonging','to','a','sexterm','there', 'are','four','functions','used','by', 'the','characters','in','the','use', 'of','swear','words','that','expletive', 'expression','abusive','expression', 'humorous','and','swear','words','as', 'an','auxiliary','function','based', 'on','analysis','of','the','most', 'frequently','used','functions','are', 'functions','abusive','and','the', 'reasons','used','in','the','use','of', 'swear','words','is','due','to', 'psychological','factors','and', 'social','factors','psychology','is', 'the','reason','that','often', 'appear','based','on','the', 'analysis','results'*

Dari hasil *Tokenizing* pada salah satu dataset abstrak diperoleh hasil seperti Tabel 4.6, yang mana teks keseluruhan dalam dataset telah dipisahkan berdasarkan kata/ *term*. Pada tahap ini diharapkan dapat membantu proses selanjutnya yakni

filtering dan *stemming*, karena tahap selanjutnya membutuhkan susunan *term* yang terpisah dalam beberapa string.

4.2.3 *Filtering*

Pada tahap selanjutnya yakni *filtering* menghasilkan potongan-potongan kata untuk setiap dataset abstrak yang telah dibuang berbagai kata yang berasal dari *stopwords*/ kata dalam bahasa Inggris yang tidak disertakan dalam penggalian informasi lebih lanjut. Proses *filtering* dilakukan dengan menghilangkan kata yang terindikasi dalam daftar *stopwords* bahasa inggris dan daftar kata tambahan dari pihak sastra inggris UINSA. Daftar stopwords bahasa inggris yang digunakan diambil dari salah satu *library NLTK python*, lebih tepatnya *library NLTK.corpus* dilanjutkan dengan *import stopwords.words('english')*. Daftar kata yang hendak dibuang dalam *dataset* adalah sebagai berikut.

```
'only','for',"hasn't",'my','have','themselves',"haven't",'too',
"mustn't",'when',"don't","wasn't",'she','yourselves',"wouldn't",
"mightn't",'doing','then','over','no','so','more','ma','there',
'be','o','ve','won','while','with','your','itself','them',
'very','a',"couldn't","needn't","should've","you'd","isn't",
'again','y','these','between','of','but','below','down',
'myself',"you're",'to','those','d','most','as',"that'll",'our',
'theirs','not','will','don','under','off','during','needn',
'after',"aren't",'on','were','i','s','m','was','hadn','is',
'mightn','shan','do','all','from',"it's",'we','ain','where',
'whom','in','nor','ours','further','t','his','ourselves','can',
"doesn't",'ll','and','other','if','him','hers','has','few',
'couldn','being','here','up','just',"won't",'re','into',
'should','what','their','out','now',"hadn't",'didn','because',
'at','which','herself','or','by','are',"didn't","shan't",'hasn',
'some',"weren't",'am',"wouldn",'weren','this','same','mustn',
'they','any','above','its','own',"you'll",'yourself','against',
'through','how','you','why','doesn','yours','been','me',
"shouldn't",'had','her',"she's",'that','than','who','aren',
'does','about','shouldn','haven','it','each','the','an','such',
'he','did','wasn','himself','once','isn','both','before',
'having','until',"you've"
```

Daftar kata stopwords yang berasal dari *library NLTK.corpus* pada python berjumlah 139 kata. Kemudian daftar *stopwords* ditambah dengan daftar kata tambahan yang tidak disertakan dalam proses pemodelan topik berdasarkan keterangan pihak sastra Inggris UINSA pada penelitian (Alfanzar, 2019). Daftar kata tersebut adalah sebagai berikut.

```
'thesis','researcher','researched','researchers','analysis',
'analyzes','study','research','article','analyze','analyzing',
'analyzed','meaning','language','writer','method','found',
'theory','result','describe','function'
```

Daftar kata tambahan yang tidak digunakan dalam pemodelan topik berjumlah 21, sehingga total dari daftar *stopwords* berjumlah 150 kata. Dengan proses *filtering* ini, maka seluruh kata *stopwords* dan daftar kata tambahan akan dihilangkan dari dataset abstrak. *Pseudocode python* pada tahap *filtering* ini adalah sebagai berikut.

```
Filtering
Deklarasi:
Doc[abstrak]: datasets abstrak after tokenizing
Stopwords: list of stopwords from library python
Useless_words: list of stopwords from Sastra Inggris UINSA
Remove: function str.replace python for replace words
Algoritma:
For i=1 to Doc[abstrak] do
   For j=1 to Stopwords do
      If(words(i)==stopwords(j)) do
         Remove words(i);
      Else do
         Continue;
      End if
   End for
   For h=i to useless_words do
      If(words(i)==useless_words(j)) do
         Remove words(i);
      Else do
         Continue;
      End if
   End for
End for
```

Prinsip dari *pseudocode filtering* adalah menyeleksi *dataset* untuk menghilangkan kata-kata yang sesuai dengan daftar *stopwords* dan *useless_words*. Sehingga pemodelan topik dapat disusun dengan kata-kata yang dianggap lebih menggambarkan topik informasi. Pada tahap selanjutnya akan dicoba untuk mengubah setiap kata yang berimbuhan menjadi kata dasar. Sehingga menjadikan setiap *term* dengan imbuhan yang berbeda menjadi satu bentuk dasarnya.

4.2.4 *Stemming*

Selanjutnya pada tahap *Stemming* akan menghilangkan imbuhan pada setiap kata sehingga menghasilkan abstrak yang tersusun atas kata/ *term* dasar. *Pseudocode python* untuk proses *stemming* adalah sebagai berikut.

```
Stemming
Deskripsi:
Doc[abstrak]: datasets abstrak after filtering
Adverbs: s, ing, ed, etc
Algoritma:
For i=1 to Doc[abstrak] do
   For j=1 to adverbs do
      If(adverbs(j)contain in words(i)) do
```

```
          Remove adverb
       End if
    End for
End for
```

Pada proses *stemming* menggunakan *library python PorterStemmer*. Prinsip pada *pseudocode stemming* adalah mengubah seluruh kata dalam dataset menjadi bentuk dasar dengan fungsi *word_stemmer* untuk melakukan *looping* pada seluruh *dataset*. Hasil *Output* dari proses *stemming* terdapat pada Tabel 4.6. berikut.

Tabel 4.6. *Output Stemming*

Sebelum *Stemming*	Setelah *Stemming*
'author','discusses','swear','words',	author discuss swear word
'used','characters','film','ben',	use charact film ben
'stillers','tropic','thunder',	stiller tropic thunder
'consisted','two','discussion',	consist two discuss
'problem','namely','variety','swear',	problem name varieti
'words','functions','well','reasons',	swear word function well
'use','swear','words','author','uses',	reason use swear word
'descriptive','author','presents',	author use descript
'results','form','descriptive','text',	author present result
'steps','categorize','words','contain',	form descript text step
'swear','words','functions','reasons',	categor word contain
'use','swear','words','results',	swear word function
'authors','many','swear','words',	reason use swear word
'used','character','belonging',	result author mani swear
'sexterm','four','functions',	word use charact belong
'used','characters','use','swear',	sexterm four function use
'words','expletive','expression',	charact use swear word
'abusive','expression','humorous',	explet express abus
'swear','words','auxiliary','based',	express humor swear word
'frequently','used','functions',	auxiliari base frequent
'functions','abusive','reasons',	use function function
'used','use','swear','words','due',	abus reason use use swear
'psychological','factors','social',	word due psycholog factor
'factors','psychology','reason',	social factor psycholog
'often', 'appear','based','results'	reason often appear base result

Dengan hasil dari *stemming* ini maka struktur *dataset* yang digunakan telah dinormalkan dari *noise* dan diharapkan dapat membantu proses pemodelan topik menjadi lebih baik. Sehingga pada proses *feature selection* dapat dilakukan pembobotan pada dataset yang telah bersih dari kata-kata yang tidak diperlukan.

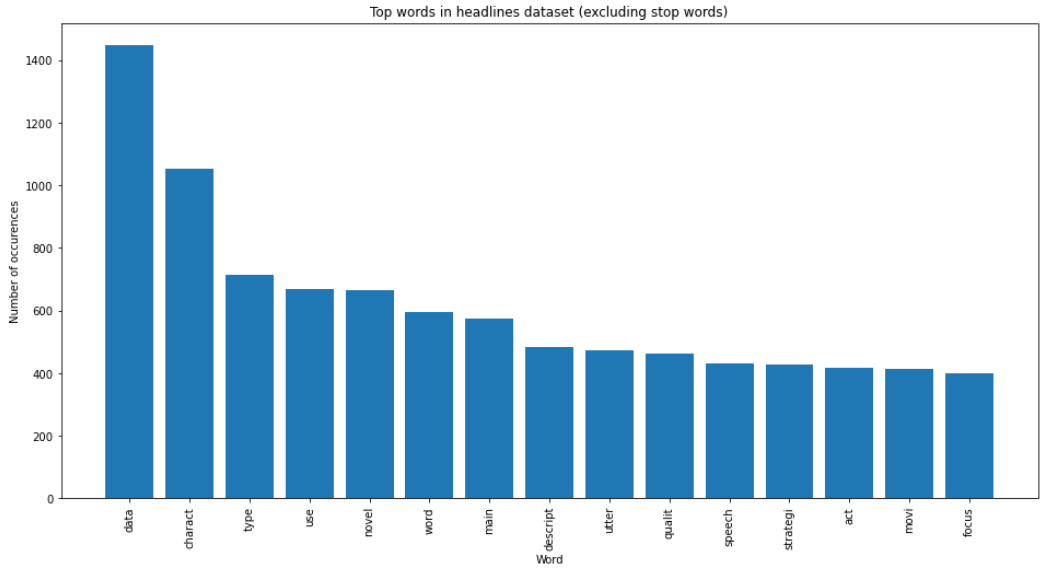
4.3 *Feature Selection*

Pada tahap *Feature Selection* dilakukan pemberian bobot dan mengubah *term*/kata menjadi nilai matriks yang akan membantu pada tahap perhitungan *Singular Value Decomposition.* Tahap *feature selection* pada penelitian ini dilakukan dengan menggunakan *library python CountVectorizer* untuk menghitung tingkat kepentingan setiap kata pada seluruh dokumen dan menghasilkan bobot dari

setiap kata yang telah dihitung tingkat kepentingannya. *Pseudocode python* untuk proses *Feature Selection* adalah sebagai berikut.

```
Feature Selection
Deskripsi:
Doc[abstrak]: datasets abstrak after stemming
Row[abstrak]: total documents (row of datasets)
Algoritma:
For i=1 to Doc[abstrak] do
   Count_words = total words in Doc[abstrak]
   For j=1 to Doc[abstrak](i) do
      Count_words(j) = total words in Doc[abstrak](i)
      Weight_words =
      count_words(j) x log(row[abstrak]/count_words)
   End for
End for
```

Prinsip dari *pseudocode feature selection* adalah memberikan bobot pada setiap kata dalam dataset sesuai dengan perhitungan *feature selection* TF-IDF pada Persamaan 2.2. Hasil bobot kata kemudian diurutkan dari tertinggi sampai terendah. Hasil dari *feature selection* divisualisasikan dalam 15 kata populer atau kata dengan kemunculan paling banyak pada Gambar 4.4 berikut.



Gambar 4.4. Kata dengan Frekuensi Tertinggi

Dari Gambar 4.4 menunjukkan 15 kata dengan kemunculan (*number of occurrences*) terbanyak adalah *data*, *charact*, *type*, *use*, *novel*, *word*, *main*, *descript*, *utter*, *quality*, *speech*, *strategy*, *act*, *movie*, dan *focus*. Selanjutnya dari bobot yang terbentuk dihitung kembali menggunakan perhitungan *Singular Value*

Decomposition (SVD) yang akan membentuk sejumlah topik dari keseluruhan dokumen.

4.4 Pembentukan Topik

Pada tahap pembentukan topik, dihasilkan beberapa topik yang terbentuk dari keseluruhan dokumen. Tahap ini merupakan tahap akhir dari pemodelan topik dengan membuat susunan kata berdasarkan tingkat kepentingan yang memiliki makna semantik yang terkandung dalam dokumen. Kata-kata yang terbentuk membentuk suatu kalimat yang menjadi topik dokumen. Tahap akhir pemodelan topik ini dilakukan dengan *library python TruncatedSVD* atau perhitungan matematika *Singular Value Decomposition*. *Psuedocode python* untuk proses analisis topik adalah sebagai berikut.

```
Analysis Topic (Singular Value Decomposition)
```

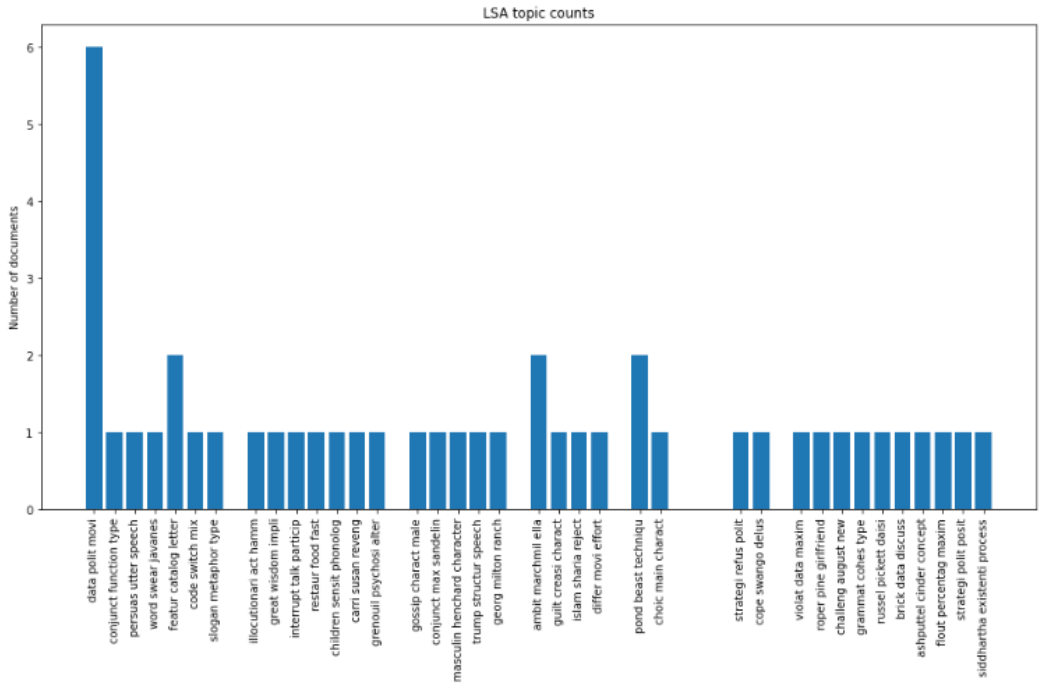
```
Deskripsi:
Doc[abstrak]: datasets abstrak after feature selection
Matriks_B[]: Array(doc[abstrak])x transpose.Array(doc[abstrak])
Eigenvalue_B[]: determinane of matriks_B
Singluar[]: √eigenvalue_B
Matriks_A[]: replace all value doc[abstrak] to 0
Matriks_D: transpose.Array(doc[abstrak]) x Array(doc[abstrak])
Eigenvalue_D: determinane of Matriks_D
```

```
Algoritma:
#find Matriks S
For i=1 to ordo_m_Matriks_A[] do
   For j=1 to ordo_n_Matriks_A[] do
      For k=1 to singular[] do
         If(i=j) do
            Matriks_S[i][j] =
            Replace.Matriks_A[i][j] with singular[k]
         Else do
            Matriks_S[i][j] = 0;
         End if
      End for
   End for
End for
#find matriks U
For i=1 to row_matriks_B[] do
   For j=1 to column_matriks_B[] do
      If(column_matriks_B is not 0)do
         Matriks_U[i][j] =
         eigenvalue_B[i][j] - singularColumn_eigenvalue_B(j)
      Else if(column_matriks_B is 0) do
         Matriks_U[i][j] = 0;
      End if
   End for
   If(total row count of Matriks_U[i] = 0)do
      Matriks_U[i][i] = 1;
   Else do
      Matriks_U[i][i] = 0;
   End if
End for
```

```
#find matriks V
For i=1 to row_matriks_D[] do
   For j=1 to column_matriks_D[] do
      If(column_matriks_D is not 0)do
         Matriks_V[i][j] =
         eigenvalue_D[i][j] - singularColumn_eigenvalue_D(j)
      Else if(column_matriks_D is 0) do
         Matriks_V[i][j] = 0;
      End if
   End for
   If(total row count of Matriks_V[i] = 0)do
      Matriks_V[i][i] = 1/√2;
   Else do
      Matriks_V[i][i] = 0;
   End if
End for
SVD = Matriks_U x Matriks_S x transpose.Matriks_V
```

Prinsip pada *pseudocode Analysis Topic (Singular Value Decomposition)* adalah melakukan seleksi kata lanjutan dengan perhitungan *Singular Value Decomposition* (Persamaan 2.3). Hasil dari seleksi lanjutan akan mendapatkan daftar kata yang bernilai dalam ordo dari hasil perkalian Matriks_U, Matriks_S, dan Matriks_V^t. Sehingga untuk hasil akhir setiap kata disusun berdasarkan urutan terbesar sampai terkecil dan dipecah dalam 3 kata sebagai topik yang terbentuk.

Perhitungan SVD menghasilkan sejumlah topik dengan makna semantik yang tersusun atas 3 kata yang berjumlah 37 topik dari 720 dataset. Hasil *running script python* menghasilkan topik dan frekuensi topik pada Gambar 4.5 berikut.



Gambar 4.5. Topik dan Frekuensi Topik pada *Dataset*.

Secara rinci daftar topik yang terbentuk dapat dilihat pada Tabel 4.7 berikut.

Tabel 4.7. Daftar Topik yang Terbentuk

Topik ke	Nama Topik	Frekuensi
1	*data polit movi*	6
2	*conjunct function type*	1
3	*persuas utter speech*	1
4	*word swear javanes*	1
5	*featur catalog letter*	2
6	*code switch mix*	1
7	*slogan metaphor type*	1
8	*illocutionari act hamm*	1
9	*great wisdom impli*	1
10	*interrupt talk particip*	1
11	*restaur food fast*	1
12	*children sensit phonolog*	1
13	*carri susan reveng*	1
14	*grenouil psychosi alter*	1
15	*gossip charact male*	1
16	*conjunct max sandelin*	1
17	*masculin henchard character*	1
18	*trump structur speech*	1
19	*georg milton ranch*	1
20	*ambit marchmil ella*	2
21	*guilt creasi charact*	1
22	*islam sharia reject*	1
23	*differ movi effort*	1
24	*pond beast techniqu*	2
25	*choic main charact*	1
26	*strategi refus polit*	1
27	*cope swango delus*	1
28	*violat data maxim*	1
29	*roper pine girlfriend*	1
30	*challeng august new*	1
31	*grammat cohes type*	1
32	*russel pickett daisi*	1
33	*brick data discuss*	1
34	*ashputtel cinder concept*	1
35	*flout percentag maxim*	1
36	*strategi polit posit*	1
37	*siddhartha existenti process*	1
Total		45

Pada hasil frekuensi topik hanya 45 dokumen yang teridentifikasi pada topik yang terbentuk. Yang artinya untuk frekuensi yang dihasilkan pada tiap topik tidak dapat digunakan untuk melihat hasil tren topik. Sehingga diperlukan proses analisis tren topik lanjutan dengan melihat frekuensi topik pada tiap dataset abstrak berdasarkan kandungan kata topik dalam setiap dataset.

4.5 Analisis Tren

Pada tahap ini dilakukan analisis terhadap sejumlah topik yang telah terbentuk dari tahap sebelumnya. Analisis pada tahap ini bertujuan untuk menemukan sejumlah topik yang memiliki tren tertinggi, artinya sejumlah topik tersebut adalah yang paling banyak diangkat dalam penelitian skripsi oleh mahasiswa program studi sastra Inggris UINSA. Langkah-langkah dalam tahapan ini yakni sebagai berikut.

4.5.1 Menentukan Jenis Penelitian dalam Topik yang Terbentuk

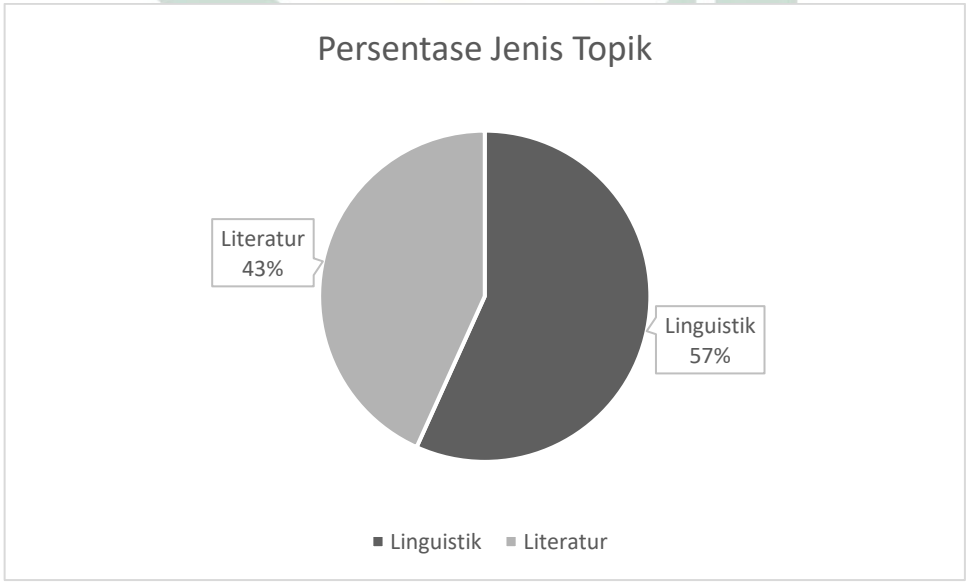
Berdasarkan topik yang terbentuk pada Tabel 4.8, dilakukan analisis jenis topik dengan menentukan setiap topik termasuk jenis penelitian linguistik atau literatur. Penentuan jenis topik dikonsultasikan dengan pihak Sastra Inggris seperti pemahaman kata-kata yang termasuk bidang linguistik dan literatur. Hasil dari penentuan jenis topik terdapat pada Tabel 4.8 berikut.

Tabel 4.8. Jenis Penelitian pada Topik yang Terbentuk

Topik	Kategori
conjunct function type	Linguistik
persuas utter speech	Linguistik
word swear javanes	Linguistik
featur catalog letter	Linguistik
code switch mix	Linguistik
slogan metaphor type	Linguistik
illocutionari act hamm	Linguistik
great wisdom impli	Linguistik
interrupt talk particip	Linguistik
children sensit phonolog	Linguistik
conjunct max sandelin	Linguistik
trump structur speech	Linguistik
pond beast techniqu	Linguistik
strategi refus polit	Linguistik
cope swango delus	Linguistik
violat data maxim	Linguistik
grammat cohes type	Linguistik

flout percentag maxim	Linguistik
strategi polit posit	Linguistik
brick data discuss	Linguistik
data polit movi	Linguistik
restaur food fast	Literatur
carri susan reveng	Literatur
grenouil psychosi alter	Literatur
gossip charact male	Literatur
masculin henchard character	Literatur
georg milton ranch	Literatur
ambit marchmil ella	Literatur
guilt creasi charact	Literatur
islam sharia reject	Literatur
differ movi effort	Literatur
choic main charact	Literatur
roper pine girlfriend	Literatur
challeng august new	Literatur
russel pickett daisi	Literatur
ashputtel cinder concept	Literatur
siddhartha existenti process	Literatur

Dari Tabel 4.8, Penentuan jenis topik didapatkan hasil bahwa dari 37 topik, terdapat 21 topik linguistik dan 16 topik literatur. Dengan persentase jenis topik yang terbentuk dapat dilihat pada Gambar 4.6 berikut.



Gambar 4.6. Persentase Perbandingan Jenis Topik

Dari Gambar 4.6, didapatkan persentase jenis topik yakni 57 persen berjenis linguistik dan 43 persen berjenis literatur.

4.5.2 Menentukan Topik pada Setiap Baris Abstrak Skripsi

Setelah setiap topik ditentukan klasternya (linguistik atau literatur), maka selanjutnya adalah menentukan topik pada setiap dokumen skripsi yang mana dalam penelitian ini menggunakan abstrak skripsi. Pada penentuan topik dokumen dilakukan dengan menghitung kemunculan kata setiap topik (3 kata/ *term*) dalam setiap dataset abstrak. Sehingga dengan nilai kemunculan kata suatu topik yang terhitung paling banyak maka suatu dataset abstrak dipahami termasuk dalam topik yang terhitung paling banyak tersebut. *Pseudocode* pada *microsoft excel* untuk penentuan topik setiap abstrak adalah sebagai berikut.

```
Klasifikasi topik abstrak
Deskripsi:
X_abstrak: dataset abstract
X_topik: 38 list of topics
Algoritma:
#count value linguistik each abstrak
For i=1 to x_abstrak do
   For j=1 to x_topic do
      If(x_abstrak(i) contain x_topic(j))do
         X_topic(j)++
      Else do
         Continue;
      End if
   End for
End for
#classification abstract
For i=1 to x_abstrak do
   X_topic[i] = max value of x_topic[i]
End for
```

Prinsip *pseudocode* klasifikasi topik abstrak adalah melakukan pembacaan setiap kata dalam *dataset* abstrak untuk digolongkan pada topik tertentu. Setiap abstrak dibaca dan ditentukan lebih condong pada suatu topik tertentu berdasarkan kemunculan katanya. Adapun hasil dari penentuan topik tersebut adalah sebagai berikut. Tabel 4.9.

Tabel 4.9. Jumlah Topik Penelitian pada Abstrak Skripsi Sastra Inggris 2014-2019

Topik	Jumlah
conjunct function type	57
persuas utter speech	35
word swear javanes	51
featur catalog letter	16
code switch mix	20
slogan metaphor type	25
illocutionari act hamm	160
great wisdom impli	13

interrupt talk particip	19
children sensit phonolog	10
conjunct max sandelin	4
trump structur speech	22
pond beast techniqu	10
strategi refus polit	11
cope swango delus	3
violat data maxim	65
grammat cohes type	32
flout percentag maxim	12
strategi polit posit	34
brick data discuss	83
data polit movi	90
restaur food fast	3
carri susan reveng	6
grenouil psychosi alter	3
gossip charact male	55
masculin henchard character	38
georg milton ranch	7
ambit marchmil ella	9
guilt creasi charact	41
islam sharia reject	5
differ movi effort	19
choic main charact	171
roper pine girlfriend	4
challeng august new	24
russel pickett daisi	3
ashputtel cinder concept	9
siddhartha existenti process	16

Dari hasil Tabel 4.9 yang diperoleh, setiap topik diberi nilai berdasarkan jumlah kemunculan tiap kata topik dalam 720 dataset abstrak. Hasil identifikasi topik dalam setiap abstrak menghasilkan jumlah topik yang beragam, mulai dari 1 topik sampai 5 topik dalam setiap dataset abstrak.

4.5.3 Menentukan Jenis Penelitian pada Setiap Baris Abstrak Skripsi

Setelah menentukan topik, maka selanjutnya menentukan jenis penelitian pada setiap dataset abstrak. Hal ini untuk menganalisis suatu dataset abstrak dengan topik lebih dari satu yang seharusnya hanya memiliki satu jenis penelitian (linguistik atau literatur). Penentuan jenis penelitian ditentukan dari jenis topik yang dimiliki setiap dataset abstrak. Penentuan salah satu jenis penelitian (linguistik atau literatur) untuk setiap dataset dilakukan dengan melihat jenis topik yang paling

banyak. Hal ini dilakukan agar suatu dataset abstrak dapat diketahui jenis penelitiannya walaupun memiliki topik yang lebih dari satu dengan jenis topik yang berbeda. *Pseudocode* untuk menentukan jenis penelitian setiap dataset abstrak adalah sebagai berikut.

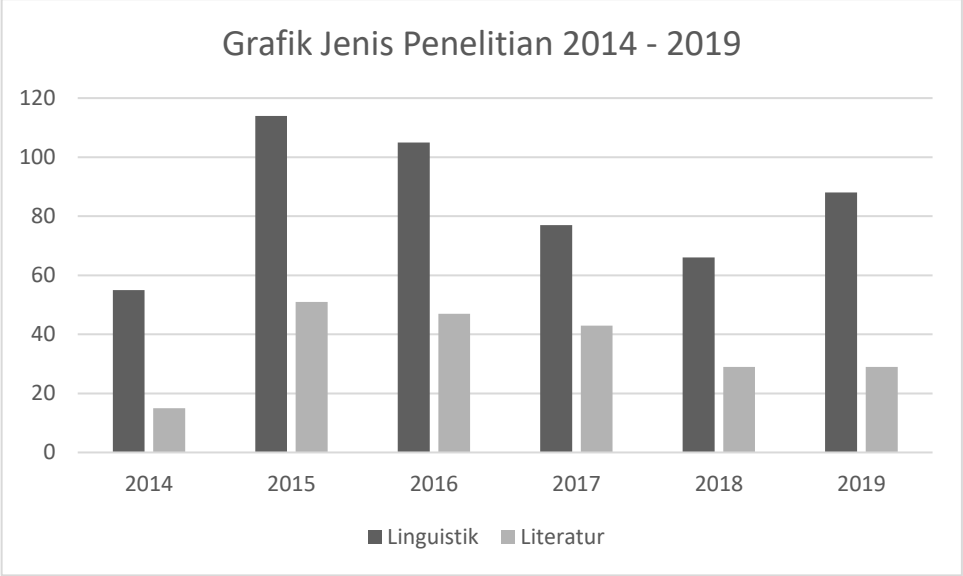
```
Klasifikasi jenis penelitian abstrak
Deskripsi:
X_linguistik: range cell that contain linguistik topic
X_literatur: range cell that contain literatur topic
X_topic_abstrak: abstract topics from klasifikasi topik abstrak
X_clasification_jenis: data clasification jenis penelitian each
abstract
Algoritma:
#count linguistik
For i=1 to x_topic_abstrak do
   For j=1 to x_linguistik do
      If(x_topic_abstrak(i)==x_linguistik(j))do
         Linguistik[i]++;
      Else do
         Continue;
      End if
   End for
End for
#count literatur
For i=1 to x_topic_abstrak do
   For j=1 to x_literatur do
      If(x_topic_abstrak(i)==x_literatur(j))do
         Literatur[i]++;
      Else do
         Continue;
      End if
   End for
End for
#compare linguistik and literatur
For i=1 to x_topic_abstrak do
   If(linguistik[i]>literatur[i])do
      X_classification_jenis[i] = linguistik;
   Else do
      X_classification_jenis[i] = literatur;
End for
```

Pseudocode klasifikasi jenis penelitian abstrak menunjukkan proses pembacaan setiap topik yang telah ditentukan pada setiap *dataset* abstrak. Pembacaan topik pada setiap dataset abstrak dilakukan untuk menentukan jenis penelitian (linguistik atau literatur) pada setiap dataset abstrak. Topik yang telah diklasifikasikan jenisnya pada Tabel 4.8 digunakan sebagai acuan untuk penentuan jenis penelitian disetiap dataset. Berdasarkan jumlah data penelitian pada Tabel 4.3, maka jenis penelitian setiap tahun dapat dilihat pada Tabel 4.10 berikut.

Tabel 4.10. Jumlah Jenis Penelitian Setiap Tahun

Tahun	Dataset	Linguistik	Literatur
2014	71	55	16
2015	165	114	51
2016	152	105	47
2017	120	77	43
2018	95	66	29
2019	117	88	29

Dari data jenis penelitian setiap tahun dibagi menjadi dua jenis penelitian yakni linguistik dan literatur dari tahun 2014 sampai 2019. Kemudian dari data yang telah diperoleh tersebut divisualisasikan dalam grafik yang dapat dilihat pada Gambar 4.7 berikut.



Gambar 4.7. Grafik Jenis Penelitian Setiap Tahun

Model grafik yang dihasilkan pada data jenis penelitian skripsi sastra inggris UINSA mengalami fluktuasi untuk masing-masing jenis penelitian, dengan data tiap tahun untuk jenis penelitian linguistik selalu lebih banyak daripada literatur.

4.5.4 Menghitung Jumlah Tren Topik

Setelah mengetahui data statistik untuk jenis penelitian setiap tahun maka selanjutnya akan dicoba untuk menampilkan tren topik dari masing-masing jenis penelitian setiap tahun. Untuk menampilkan tren topik terlebih dahulu dihitung topik dan diambil beberapa topik yang paling sering muncul. *Pseudocode* untuk menentukan tren topik adalah sebagai berikut.

```
Menghitung Tren Topik
Deskripsi:
List_year: 2014 - 2019
Algoritma:
#Tren linguistik
For j=1 to list_year do
   Tren_topics_linguistik =
   get 7 largest linguistik topic in list_year(j)
End for
#tren literatur
For j=1 to list_year do
   Tren_topics_literatur =
   get 7 largest literatur topic in list_year(j)
End for
```

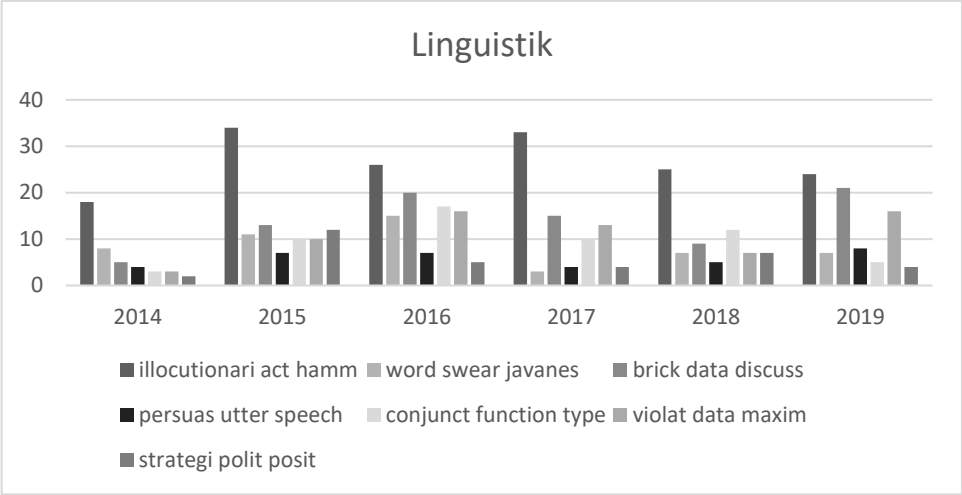
Pada *pseudocode* menghitung tren topik, diambil 7 topik tertinggi pada setiap jenis penelitian. Pengambilan topik tertinggi dilakukan berdasarkan tahun. Sehingga menghasilkan perhitungan jumlah topik untuk jenis penelitian linguistik adalah sebagai berikut. Tabel 4.11

Tabel 4.11. Jumlah 7 Tren Topik Linguistik pada Dokumen Setiap Tahun

Topik linguistik	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**
illocutionari act hamm	18	34	26	33	25	24
word swear javanes	8	11	15	3	7	7
brick data discuss	5	13	20	15	9	21
persuas utter speech	4	7	7	4	5	8
conjunct function type	3	10	17	10	12	5
violat data maxim	3	10	16	13	7	16
strategi polit posit	2	12	5	4	7	4

Dari 20 topik linguistik, diambil 7 topik linguistik tertinggi setiap tahun yakni *illocutionari act hamm, word swear javanes, brick data discuss, persuas utter speech, conjunct function type, violat data maxim,* dan *strategi polit posit*. 7 topik yang diambil karena dianggap sebagai tren topik berdasarkan jumlah tertinggi yang teridentifikasi pada dataset abstrak setiap tahun.

Dari hasil jumlah topik Tabel 4.11, kemudian divisualisasikan dalam bentuk grafik. Visualisasi topik mengambil 7 topik tertinggi untuk jenis penelitian linguistik. Hasil dari visualisasi tersebut sebagai berikut. Gambar 4.8.



Gambar 4.8. Tren 7 Topik Linguistik Tertinggi Setiap Tahun

Dari hasil visualisasi dapat dilihat bahwa dalam penelitian linguistik, topik *illocutionari act hamm* selalu menempati peringkat tertinggi dalam penelitian setiap tahun. Topik *illocutionari act hamm* menjadi yang tertinggi dengan rata-rata jumlah penelitian berkisar 27 penelitian tiap tahun.

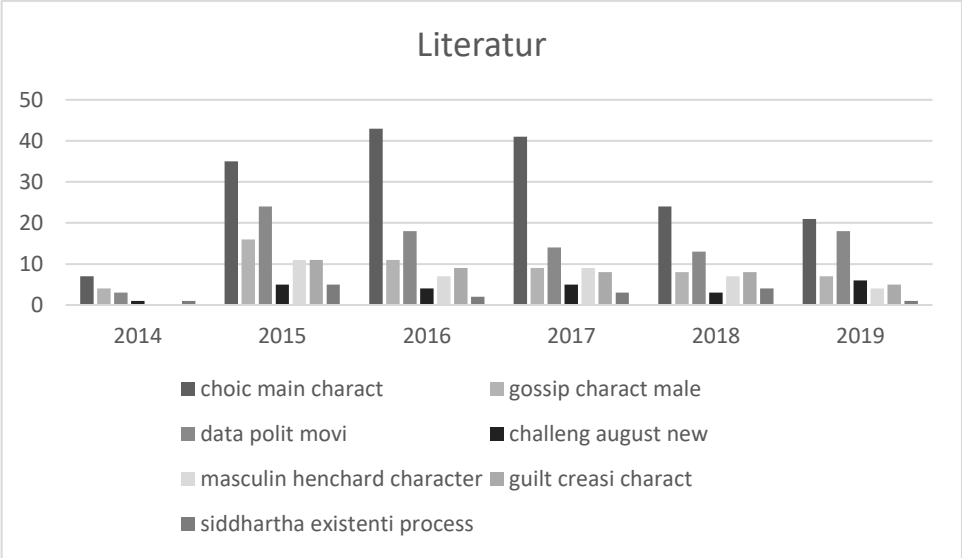
Sedangkan pada penelitian literatur, menghasilkan data Tabel 4.12 berikut.

Tabel 4.12. Jumlah 7 Tren Topik Literatur pada Dokumen Setiap Tahun

Literatur	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**
choic main charact	7	35	43	41	24	21
gossip charact male	4	16	11	9	8	7
differ movi effort	2	4	3	2	2	6
challeng august new	1	5	4	5	3	6
masculin henchard character		11	7	9	7	4
guilt creasi charact		11	9	8	8	5
siddhartha existenti process	1	5	2	3	4	1

Pada penelitian literatur, 7 topik yang memiliki jumlah tertinggi teridentifikasi pada dataset abstrak setiap tahun adalah *choic main charact, gossip charact male, differ movi effort, challeng august new, masculin henchard character, guilt creasi charact, dan siddhartha existenti process*.

Dengan data penelitian literatur Tabel 4.12, maka selanjutnya dapat divisualisasikan dalam bentuk grafik untuk melihat pola setiap topiknya. Hasil dari visualisasi topik literatur adalah. Gambar 4.9.



Gambar 4.9. Grafik Top 7 Topik Literatur Setiap Tahun

Dari hasil visualisasi topik literatur memiliki urutan yang beragam. Dengan jumlah rata-rata topik terbanyak dimiliki oleh topik *choice main character*. Rata-rata jumlah penelitian setiap tahun pada topik *choice main character* adalah 28 penelitian setiap tahun.

4.6 Pengujian Hasil

Pada bagian ini dimaksudkan untuk mengetahui persentase presisi, *recall*, dan akurasi dari jenis penelitian setiap tahun pada Tabel 4.10. Pengujian dilakukan dengan membandingkan hasil data Tabel 4.10 dengan *data real* jenis penelitian setiap tahun. *Data real* diperoleh dari Kepala Program Studi (Kaprodi) Sastra Inggris UINSA. Perbandingan antara jumlah jenis penelitian hasil analisis dengan *data real* setiap tahun adalah sebagai berikut. Tabel 4.13.

Tabel 4.13. Perbandingan Hasil analisis dan *Data Real* Jenis Penelitian

Tahun	Dataset	Hasil Analisis		*Data Real*	
		Linguistik	Literatur	Linguistik	Literatur
2014	71	54	16	55	15
2015	165	99	66	114	51
2016	152	96	56	105	47
2017	120	72	48	77	43
2018	95	57	38	66	29
2019	117	82	35	88	29

Dari data Tabel 4.14, menunjukkan data untuk jenis penelitian linguistik lebih banyak daripada penelitian literatur, baik hasil analisis maupun pada *data real*. Namun untuk jumlah yang dihasilkan tidak mirip, sehingga perlu untuk

melakukan perhitungan tingkat kesalahan antara data hasil analisis dengan *data real* pada Tabel 4.13 untuk mengetahui tingkat akurasinya. Dari hasil perbandingan hasil analisis dengan *data real* didapatkan nilai *true positif, true negatif, false positif,* dan *false negatif* sebagai berikut. Tabel 4.14.

Tabel 4.14. Nilai *True-False Positive dan True-False Negative*

Tahun	Linguistik				Literatur			
	TP	*TN*	*FP*	*FN*	*TP*	*TN*	*FP*	*FN*
2014	42	4	12	12	4	42	12	12
2015	74	23	25	43	23	74	43	25
2016	74	38	22	18	38	74	18	22
2017	53	39	19	9	39	53	9	19
2018	44	26	13	12	26	44	12	13
2019	64	25	18	10	25	64	10	18

Sehingga dari nilai pada Tabel 4.14, maka tingkat akurasinya terdapat pada Tabel 4.15 berikut.

Tabel 4.15. Persentase Kesalahan Jenis Penelitian

Tahun	Linguistik		Literatur		*Accuracy (%)*
	Presisi (%)	*Recall (%)*	*Presisi (%)*	*Recall (%)*	
2014	0,782	0,796	0,267	0,250	0,671
2015	0,772	0,752	0,431	0,458	0,667
2016	0,848	0,967	0,936	0,733	0,875
2017	0,779	0,968	0,953	0,707	0,842
2018	0,848	1,000	1,000	0,744	0,895
2019	0,795	0,946	0,862	0,581	0,812
Rata-rata	0,804	0,905	0,742	0,579	0,794

Dari Tabel 4.15 pada hasil linguistik menunjukkan tingkat presisi, dan *recall* yang cukup baik dengan rata-rata presisi 80 persen, dan *recall* 90 persen. Hal ini menunjukkan bahwa hasil analisis linguistik memiliki tingkat ketepatan, dan keberhasilan menemukan informasi yang baik. Sedangkan pada hasil literatur menunjukkan tingkat presisi yang cukup baik, namun untuk *recall* tidak begitu baik dengan rata-rata presisi 74 persen, dan *recall* 57 persen. Hal ini menunjukkan bahwa hasil analisis literatur memiliki tingkat ketepatan yang cukup baik namun keberhasilan menemukan informasi yang kurang baik. Sedangkan untuk tingkat akurasi memiliki hasil yang cukup baik yakni 79 persen, yang menunjukkan kedekatan hasil prediksi dan *data real* yang cukup baik.

4.7 Pembahasan

Hasil pembentukan topik menunjukkan 21 topik sesuai dengan *cluster* 7 topik pada iterasi ke 1000 dari penelitian Alfanzar (Alfanzar, 2019). Jika dicocokan hasil *cluster* topik dengan kandungan kata topik yang terbentuk pada penelitian ini sebagai berikut. Tabel 4.16.

Tabel 4.16. Klaster Hasil Pembentukan Topik

Hasil Pembentukan Topik Penelitian Ini	**Cluster Topik** (Alfanzar, 2019)
conjunct max sandelin	*Topik 1: story, cohesion, grammatical, type, sentence, implicature, short, conjunction, movie, qualitative.*
grammat cohes type	
differ movi effort	
great wisdom impli	
word swear javanes	*Topik 2 maxim, feature, woman, movie, utterance, process, words, flout, character, qualitative.*
flout percentag maxim	
choic main charact	
violat data maxim	
conjunct function type	*Topik 3: style, speech, movie, utterance, type, advertisement, illocutionary, character, function, qualitative.*
illocutionari act hamm	
data polit movi	
persuas utter speech	*Topik 4: speech, advertisement, words, affix, identity, error, toward, qualitative, three, feature.*
featur catalog letter	
slogan metaphor type	*Topik 5: strategy, deixis, politeness, type, positive, person, character, utterance, conversation, social*
strategi refus polit	
strategi polit posit	
trump structur speech	*Topik 6: student, english, sign, school, vocabulary, trump, language, donald, surabaya, learning.*
gossip charact male	*Topik 7: novel, focus, criticism, problem, characterization, experience, analyze, conflict, struggle, become.*
masculin henchard character	
guilt creasi charact	
code switch mix	
interrupt talk particip	
children sensit phonolog	
pond beast techniqu	
cope swango delus	
brick data discuss	
restaur food fast	
carri susan reveng	
grenouil psychosi alter	
georg milton ranch	
ambit marchmil ella	

islam sharia reject	
roper pine girlfriend	
challeng august new	
russel pickett daisi	
ashputtel cinder concept	
siddhartha existenti process	

Dari hasil pencocokan Tabel 4.16 terdapat 17 topik yang masih belum dapat dikategorikan pada klaster topik (Alfanzar, 2019). Hal ini dikarenakan pemodelan topik pada penelitian sebelumnya dibatasi pembentukan topik hanya 7 topik serta jumlah dataset yang digunakan hanya berjumlah 584 dataset, sedangkan pada penelitian ini tidak ditentukan jumlah pembentukan topiknya sehingga menghasilkan jumlah topik maksimal (37 topik) yang dapat terbentuk dari 720 dataset. Sehingga untuk hasil topik yang ditemukan penelitian ini lebih mendominasi pada jumlah topik yang dihasilkan, karena jumlah dataset yang lebih banyak.

Jika melihat topik tertinggi yang terbentuk pada jenis linguistik yang disesuaikan dengan pemahaman (Fauziah, 2018), maka dihasilkan *scope* pada Tabel 4.17 berikut.

Tabel 4.17. *Scope* Topik Terpopuler

Topik	**Makna**	**Scope**
Illocutionary act hamm	Aksi tindak tutur/ ilokusi	Sosiolinguistik

Jika merujuk topik tertinggi pada hasil tren pada Gambar 4.8 menunjukkan topik *illocutionary act hamm* sebagai yang tertinggi dengan makna semantik topik yang tergolong pada bidang sosiolinguistik. Bidang topik penelitian sosiolinguistik merupakan satu kajian untuk mempelajari bahasa dalam konteks sosiokultural (Ahasa, 1986), kemudian pada penelitian (Yendra, S. S., 2016) yang menerapkan sosiolinguistik untuk memahami sosiokultural masyarakat Minangkabau. Maka hal ini sejalan dengan hasil penelitian (Lei & Liu, 2019) yang menyebutkan dari 42 *Social Science Citation Index (SSCI)-indexed journals of applied linguistics*, sebagian besar topik yang sering dibahas tetap populer selama 12 tahun, beberapa (terutama masalah sosiokultural/ fungsional/ identitas) mengalami peningkatan minat yang signifikan. Sehingga hal ini menunjukkan bidang topik tentang sosiolinguistik yang merupakan kajian dalam sosiokultural dan sering digunakan dalam penelitian mahasiswa strata-1 Sastra Inggris UINSA untuk mengetahui

sosiokultural masyarakat dalam penelitian 5 tahun terakhir ini, sesuai dengan tren bidang topik secara umum.

BAB V
PENUTUP

5.1 Kesimpulan

Topic modeling pada 720 dataset abstrak skripsi Program Studi Sastra Inggris UINSA menggunakan metode *Latent Semantic Analysis* (LSA), menghasilkan kesimpulan sebagai berikut.

1. Dalam kurun waktu 5 tahun dihasilkan total 37 topik skripsi Sastra Inggris, yang terbagi menjadi 2 jenis yaitu 21 topik berjenis linguistik yang mempunyai tingkat presisi, dan *recall* yang baik dengan rata-rata presisi 80 persen, dan *recall* 90 persen. Serta 16 topik lainnya berjenis literatur yang memiliki tingkat presisi cukup baik dan *recall* yang kurang baik dengan rata-rata presisi 74 persen, dan *recall* 57 persen. Serta untuk tingkat akurasi kedua jenis topik yang menunjukkan kedekatan hasil prediksi dan *data real* memiliki hasil yang cukup baik yakni 79 persen.
2. Selanjutnya dalam proses analisis tren topik menghasilkan 7 topik tertinggi untuk setiap jenis penelitian. Topik *illocutionary act hamm* (Sosiolinguistik) sebagai topik dengan rata-rata paling tinggi sebesar 27 penelitian setiap tahun pada jenis linguistik dan topik *choice main character* (literatur) sebagai topik dengan rata-rata paling tinggi sebesar 28 penelitian tiap tahun pada jenis literatur.

5.2 Saran

Dari hasil yang telah diperoleh, maka beberapa saran yang dapat dilakukan untuk penelitian selanjutnya adalah:

Saran untuk memperbaiki hasil presisi, dan *recall* yang kurang baik pada jenis literatur adalah dengan menyeleksi topik literatur atau linguistik yang memiliki makna abstrak, seperti *restaur food fast*, *brick data discuss*, dan *islam sharia reject*. Sehingga penggolongan topik pada abstrak dapat mengubah tingkat akurasi. Secara garis besar untuk mengubah tingkat akurasi ada pada tahap analisis tren topik. Yang mana perlu ada langkah selain menggolongkan abstrak berdasarkan kandungan kata topik.

DAFTAR PUSTAKA

Ahasa, D. A. N. P. E. B. (1986). *HUBUNGAN VARIASI BAHASA DENGAN KELOMPOK SOSIAL*. 12–13. Retrieved from https://media.neliti.com/media/publications/78731-ID-hubungan-variasi-bahasa-dengan-kelompok.pdf

Alfanzar, A. I. (2019). *Topic modelling skripsi menggunakan metode latent dirichlet allocation*. Universitas Islam Negeri Sunan Ampel.

Azharyani, I., & Kusumo, D. S. (2019). *Implementasi Semantic Search pada Open Library menggunakan Metode Latent Semantic Analysis (Studi Kasus : Open Library Universitas Telkom)*. *6*(2), 8987–8998.

Efendi, E. (2016). TABAYYUN DALAM JURNALISTIK. *Jurnal KOMUNIKA ISLAMIKA*, *3*(3). https://doi.org/10.1017/CBO9781107415324.004

Farida, I. N., Kom, M., Swanjaya, D., & Kom, M. (2019). *ARTIKEL IMPLEMENTASI METODE LATENT DIRICHLET ALLOCATION UNTUK MENENTUKAN TOPIK TEKS BERITA Oleh : DIMAS ARYANTO SAPUTRO Dibimbing oleh : UNIVERSITAS NUSANTARA PGRI KEDIRI TAHUN 2019 SURATPERNYATAAN ARTIKEL SKRIPSI TAHUN 2019*.

Fauziah, S. (2018). Kesantunan Sebagai Kajian Sosiolinguistik. *Journal of Chemical Information and Modeling*, *53*(9), 1689–1699. https://doi.org/10.1017/CBO9781107415324.004

Hartanto. (2017). Text Mining Dan Sentimen Analisis Twitter Pada Gerakan Lgbt. *Intuisi : Jurnal Psikologi Ilmiah*, *9*(1), 18–25.

Hermawan, L., & Ismiati, M. B. (2020). *Pembelajaran Text Preprocessing berbasis Simulator Untuk Mata Kuliah Information Retrieval*. *17*(2), 188–199.

Hidayati, D. C., Al Faraby, S., & Adiwijaya, A. (2020). Klasifikasi Topik Multi Label pada Hadis Shahih Bukhari Menggunakan K-Nearest Neighbor dan Latent Semantic Analysis. *JURIKOM (Jurnal Riset Komputer)*, *7*(1), 140. https://doi.org/10.30865/jurikom.v7i1.2013

Hudaya, C. S., Fakhrurroja, H., & Alamsyah, A. (2018). Analisis Persepsi

Konsumen Terhadap Brand Go-Jek Pada Media Sosial Twitter Menggunakan Metode Sentiment Analysis Dan Topic Modelling. *Jurnal Mitra Manajemen*, *2*(4), 273–285. Retrieved from http://e-jurnalmitramanajemen.com/index.php/jmm/article/view/125/69

Jadhira, A. A., Bijaksana, M. A., & Wahyudi, B. A. (2018). Deteksi Kemiripan Bagian-bagian Terjemah Al-Qur'an dengan Menggunakan Metode Latent Semantic Analysis. *EProceedings of Engineering*, *5*(3), 7649–7657. Retrieved from https://libraryeproceeding.telkomuniversity.ac.id/index.php/engineering/article/view/7286

Lei, L., & Liu, D. (2019). Research Trends in Applied Linguistics from 2005 to 2016: A Bibliometric Analysis and Its Implications. *Applied Linguistics*, *40(3)*, 540–561. https://doi.org/doi:10.1093/applin/amy003

Luthfiarta, A., Zeniarja, J., & Salam, A. (2013). Algoritma Latent Semantic Analysis (LSA) Pada Peringkas Dokumen Otomatis Untuk Proses Clustering Dokumen. *Seminar Nasional Teknologi Informasi & Komunikasi Terapan 2013 (SEMANTIK 2013)*, *2013*(November), 13–18.

M. Yunus, B. (2016). Tafsir Tarbawī. *Al-Bayan: Jurnal Studi Ilmu Al- Qur'an Dan Tafsir*, *1*(1), 1–7. https://doi.org/10.15575/al-bayan.v1i1.1670

Ngafifudin, D. (2017). KONSEP DAN PRINSIP SERTA ALGORITMA DALAM LATENT SEMENTIC INDEXING. Retrieved July 23, 2020, from http://hirupmotekar.com/2017/05/30/dwi-ngafifudin-konsep-dan-prinsip-serta-algoritma-dalam-latent-sementic-indexing/

Petrus, J. (2019). MODEL GRAFIS JEJARING PENULIS KARYA ILMIAH. *Prosiding Seminar Nasional Pakar Ke 2*, *1*, 1–6.

Prilianti, K. R., & Wijaya, H. (2014). *Aplikasi Text Mining untuk Automasi Penentuan Tren Topik Skripsi dengan Metode K-Means Clustering*. *2*(1), 1–6.

Priyanto, A., & Ma'arif, M. R. (2018). Implementasi Web Scrapping dan Text Mining untuk Akuisisi dan Kategorisasi Informasi dari Internet (Studi Kasus: Tutorial Hidroponik). *Indonesian Journal of Information Systems*, *1*(1), 25–

33. https://doi.org/10.24002/ijis.v1i1.1664

Putra, K. B., & Kusumawardani, R. P. (2017). Analisis Topik Informasi Publik Media Sosial di Surabaya Menggunakan Pemodelan Latent Dirichlet Allocation (LDA). *Jurnal Teknik ITS*, *6*(2), 4–9. https://doi.org/10.12962/j23373539.v6i2.23205

Putra, M. A., Sari, P. K., Prodi, S., Bisnis, M., Informatika, T., Ekonomi, F., & Telkom, U. (2019). *ANALISIS BRAND AWARENES S PRODUK OTOMOTIF PADA MASA RILIS MENGGUNAKAN DYNAMIC NETWORK ANALYSIS (STUDI KASUS PADA ALL NEW NISSAN LIVINA) BRAND AWARENESS ANALYSIS OF AUTOMOTIVE PRODUCTS DURING THE RELEASE USING DYNAMIC NETWORK ANALYSIS (CASE STUDY ON ALL*. *6*(2), 1926–1933.

Qutsiah, S. A., Sophan, M. K., & Hendrawan, Y. F. (2016). APLIKASI PEMBELAJARAN MATEMATIKA DASAR BANGUN DATAR MENGGUNAKAN PYTHON PADA PERANGKAT BERGERAK. *SCAN (Jurnal Teknologi Informasi Dan Komunikasi)*, *XI*, 13–22. Retrieved from http://www.ejournal.upnjatim.ac.id/index.php/scan/article/view/868/716

Rahman, A. (2017). Online News Classification Using Multinomial Naive Bayes. *Itsmart*, *6*(1), 32–38. https://doi.org/10.1177/1096348015584441

Saputra, Jerry. Fachrurrozi, M. Y. (2017). Peringkasan Teks Berita Berbahasa Indonesia Menggunakan Metode Latent Semantic Analysis (LSA) dan Teknik Steinberger & Jezek. *Computer Science and ICT*, *3*(1), 215–219.

Setiawan, A., Darmanta, J., Tinaliah, & Yoannita. (2017). Peringkasan dokumen berita Bahasa Indonesia menggunakan metode Cross Latent Semantic Analysis. *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, *3*(2), 94. https://doi.org/10.26594/register.v3i2.1161

Somantri, O., Wiyono, S., & Dairoh, D. (2017). Optimalisasi Support Vektor Machine (Svm) Untuk Klasifikasi Tema Tugas Akhir Berbasis K-Means. *Telematika*, *13*(2), 59. https://doi.org/10.31315/telematika.v13i2.1722

Suhartono, D. (2015). *Probabilistic Latent Semantic Analysis (PLSA) untuk*

Klasifikasi Dokumen Teks Berbahasa Indonesia. Retrieved from http://arxiv.org/abs/1512.00576

Yendra, S. S., M. H. (2016). PENERAPAN SOSIOLINGUISTIK DALAM MEMAHAMI SOSIOKULTURAL MINANGKABAU UNTUK PENDIDIKAN KARAKTER; CIME'EH DAN INSYA ALLAH ORANG MINANGKABAU. *JURNAL IPTEKS TERAPAN Research of Applied Science and Education*, *10*, 71–80. https://doi.org/http://dx.doi.org/10.22216/jit.2016.v10i1.466