SKRIPSI

PERBANDINGAN KLASIFIKASI TEKS DENGAN PEMODELAN VECTOR SPACE MODEL DAN LATENT SEMANTIC INDEXING

Sterenlie

NPM: 2017730014

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN
2022

UNDERGRADUATE THESIS

COMPARISON OF TEXT CLASSIFICATION WITH VECTOR SPACE MODELING AND LATENT SEMANTIC INDEXING

Sterenlie

NPM: 2017730014

DEPARTMENT OF INFORMATICS
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES
PARAHYANGAN CATHOLIC UNIVERSITY
2022

LEMBAR PENGESAHAN

PERBANDINGAN KLASIFIKASI TEKS DENGAN PEMODELAN VECTOR SPACE MODEL DAN LATENT SEMANTIC INDEXING

Sterenlie

NPM: 2017730014

Bandung, 19 Januari 2022

Menyetujui,

Pembimbing

Digitally signed by Luciana Abednego

Luciana Abednego, M.T.

Ketua Tim Penguji

Digitally signed by Veronica Sri Moertini

Dr. Veronica Sri Moertini

Anggota Tim Penguji

Digitally signed by Mariskha Tri Adithia

Mariskha Tri Adithia, P.D.Eng

Mengetahui,

Ketua Program Studi

Digitally signed by Mariskha Tri Adithia

Mariskha Tri Adithia, P.D.Eng

PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa skripsi dengan judul:

PERBANDINGAN KLASIFIKASI TEKS DENGAN PEMODELAN VECTOR SPACE MODEL DAN LATENT SEMANTIC INDEXING

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,
Tanggal 19 Januari 2022

Sterenlie
NPM: 2017730014

ABSTRAK

Klasifikasi adalah proses menemukan model atau fungsi yang mendeskripsikan dan membedakan kategori. Dokumen perlu diklasifikasi agar tersusun secara sistematis dalam kategori yang ditentukan. Metode penting untuk memberi kategori kepada sekian banyak dokumen yang senantiasa mengalami banyak perkembangan dari waktu ke waktu yaitu klasifikasi teks. Ada banyak metode atau teknik klasifikasi untuk dokumen, masing-masing teknik klasifikasi memiliki kekurangan dan kelebihan. Penelitian ini mencari algoritma untuk melakukan klasifikasi yang paling akurat untuk dokumen teks dengan membandingkan pengelompokan dokumen-dokumen untuk menemukan metode algoritma pengklasifikasian yang paling baik. Dokumen yang digunakan adalah dataset berita, dengan banyaknya jumlah dokumen yang tersebar menimbulkan kesulitan dalam mengelompokkan dokumen tersebut berdasarkan kategorinya. Oleh karena itu penulis melakukan penelitian untuk mengklasifikasikan dokumen berita ke dalam 5 kategori (Teknologi, Kesehatan, Politik, Ekonomi dan Olahraga).

Pada penelitian ini kumpulan dokumen akan dibersihkan dengan *text preprocessing*. Setelah dibersihkan melalui *text preprocessing*, dokumen akan dimodelkan dengan menggunakan *Vector Space Model* yang akan menjadi sebuah vektor. Vektor terdiri dari banyaknya kata dari seluruh dokumen yang ada agar terbentuk matriks *document-term* dan terbentuk fitur dari teks *Term Frequency-Inverse Document Frequency* (TF-IDF). Matriks *document-term* mempunyai dimensi yang cukup besar. Maka untuk mengurangi dimensi matriks vektor dan merepresentasikan dokumen ke dalam konsep akan digunakan *Latent Semantic indexing* dengan metode *Singular Value Decomposition* (SVD). Hasil dari fitur TF-IDF dan *Latent Semantic Indexing* nantinya akan digunakan untuk melakukan klasifikasi. Klasifikasi dokumen akan dilakukan dengan metode *Naive Bayes*, *Support Vector Machine* (SVM), dan *K-Nearest Neighbor* (KNN). Hasil klasifikasi dokumen dari beberapa metode tersebut akan dibandingkan dengan parameter *F-Measure* dan *Accuracy*. Eksperimen yang dilakukan yaitu membandingkan algoritma klasifikasi yang paling akurat dan mencari nilai k yang paling terbaik pada metode KNN dan SVD.

Hasil pengujian pada penelitian ini menunjukkan bahwa algoritma *Naive Bayes* dengan fitur TF-IDF, algoritma *Support Vector Machine* dengan fitur TF-IDF dan algoritma *Support Vector Machine* dengan metode LSI dapat mengklasifikasi teks dengan tingkat keberhasilan 90% lebih. Metode LSI bekerja baik ketika digunakan pada algoritma KNN karena didapatkan penaikkan nilai *F-Measure* dan *Accuracy*. Hal ini menunjukkan bahwa algoritma TFIDF-Naive Bayes, TFIDF-SVM dan LSI-SVM dalam mengklasifikasikan dataset berita yang digunakan cukup optimal dan metode LSI sangat berpengaruh terhadap metode KNN.

Kata-kata kunci: Klasifikasi teks, *Vector Space Model*, *Latent Semantic Indexing*, *Naive Bayes*, *Support Vector Machine*, *K-Nearest Neighbor*

ABSTRACT

Classification is the process of finding a model or function that describes and distinguishes categories. Documents need to be classified so that they are arranged systematically in the specified categories. An important method for categorizing only one of many documents that has undergone many developments over time is text classification. There are many classification methods or techniques for documents, each classification technique has advantages and disadvantages. This study looks for ways to perform the most accurate classification for text documents by comparing the grouping of documents to find the best classification method. The document used is a collection of news data, with the large number of scattered documents causing difficulties in grouping these documents by category. Therefore, the authors conducted a study to classify news documents into 5 categories (Technology, Health, Politics, Economics and Sports).

In this study, a collection of documents will be cleaned by text preprocessing. After cleaning through text preprocessing, the document will be modeled using the Vector Space Model which will become a vector. The vector consists of many words from all existing documents to form a document-term matrix and form features from the Term Frequency-Inverse Document Frequency (TF-IDF) text. The term-document matrix has quite large dimensions. So to reduce vector dimensions and represent documents into concepts, Latent Semantic indexing will be used with the Singular Value Decomposition (SVD) method. The results from the TF-IDF feature and Latent Semantic Indexing will later be used to perform classification. Document classification will be carried out using the Naive Bayes method, Support Vector Machine (SVM), and K-Nearest Neighbor (KNN). The results of document classification from these methods will be compared with the F-Measure and Accuracy parameters. The experiment carried out is to compare the most accurate algorithms and find the best k value in the KNN and SVD methods.

The test results in this study indicate that the Naive Bayes algorithm with the TF-IDF feature, the Support Vector Machine algorithm with the TF-IDF feature and the Support Vector Machine algorithm with the LSI method can classify text with a success rate of 90% more. The LSI method works well when used in the KNN algorithm because it increases the F-Measure and Accuracy values. This shows that the TFIDF-Naive Bayes, TFIDF-SVM and LSI-SVM algorithms in classifying news datasets used are quite optimal and the LSI method is very influential on the KNN method.

Keywords: *Text classification*, *Vector Space Model*, *Latent Semantic Indexing*, *Naive Bayes*, *Support Vector Machine*, *K-Nearest Neighbor*

KATA PENGANTAR

Puji syukur kepada Tuhan Yesus Allah Yang Maha Esa atas segala berkat, rahmat dan karunia-Nya yang berlimpah sehingga penulis dapat menyelesaikan penyusunan skripsi yang berjudul "Perbandingan Klasifikasi Teks dengan Pemodelan Vector Space Model dan Latent Semantic Indexing". Selama proses penyusunan skripsi ini, penulis menghadapi banyak kendala dan berbagai masalah. Penulis menyadari bahwa banyaknya bantuan dan dukungan yang diberikan kepada penulis dari berbagai pihak, baik langsung maupun tidak langsung. Oleh karena itu, pada kesempatan ini penulis ingin mengucapkan terima kasih yang sebesar-besarnya kepada:

1. Orang tua dan Keluarga yang selalu memberikan dukungan kepada penulis setiap saat baik berupa doa atau dukungan mental serta materiil.
2. Ibu Natalia, M.Si. selaku dosen pembimbing yang telah membimbing dengan sangat sabar kepada penulis dan memberikan dukungan maupun bantuan kepada penulis dalam proses penyusunan skripsi ini.
3. Ibu Dr. Veronica Sri Moertini dan Ibu Mariskha Tri Adithia, P.D.Eng. selaku dosen penguji yang telah memberikan kritik, saran, serta masukan yang membangun sehingga penelitian ini menjadi lebih baik.
4. Seluruh dosen Teknik Informatika UNPAR yang telah memberikan ilmu dari awal kuliah sampai penulis dapat menyelesaikan penyusunan skripsi ini.
5. Teman-teman 'GRBLG' yang sejak awal telah setia bersama berjuang mengarungi masa-masa sulit dan pahit di Teknik Informatika UNPAR khususnya Reynard Rafferty Susilo, Dio Antares, David Christopher Sentosa, Reinalta Sugianto, Rio Aurelio Sumantri, Fritz Humphrey Silalahi, Leonard Wang, dan Juan Nandrissa Redemptino.
6. Teman Spesial. Nicky Sahnia Putri yang selalu ada untuk menemani, mendukung, menghibur, bercanda tawa, memberi semangat dan motivasi kepada penulis dalam penyusunan skripsi ini.
7. Teman-teman PMV Bhumi Pharsjia yang menemani dan mendukung penulis dalam menyusun skripsi ini.
8. Teman-teman SMA khususnya Shelly Eka Putri Budiman yang selalu memberikan arahan dan semangat untuk mengingatkan penulis mengerjakan skripsi.
9. Teman-teman Teknik Informatika UNPAR angkatan 2017 yang telah berbagi ilmu, memberikan dukungan, dan membantu penulis dalam menyelesaikan skripsi ini.
10. Pihak-pihak lain yang tidak dapat penulis sebutkan satu per satu yang sudah membantu dalam penulisan skripsi ini.

Penulis menyadari bahwa skripsi ini masih jauh dari kata sempurna karena keterbatasan pengetahuan dan kemampuan penulis. Oleh karena itu, penulis memohon maaf jika terdapat kekurangan pada skripsi ini. Penulis juga mengharapkan kritik dan saran yang membangun untuk menyempurnakan skripsi ini. Semoga skripsi ini dapat bermanfaat bagi pembaca atau pihak yang meneruskan penelitian ini.

Bandung, Januari 2022

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR KODE PROGRAM

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Teknologi informasi merupakan salah satu hal yang tidak dapat dipisahkan dari keberadaan manusia. Dengan adanya teknologi internet mempermudah bagi siapapun untuk mendapatkan informasi yang diinginkan. Informasi tersebut dapat diakses melalui halaman *website* contohnya detik.com, liputan6.com, tribunews.com, sindonews.com dan kompas.com. *Website* kompas.com menjadi media tertua dan merupakan satu-satunya media *online* yang termasuk dalam generasi pertama di antara kelima media tersebut. Hal ini menunjukkan bahwa kompas.com memiliki kemampuan adaptasi yang baik dalam menghadapi perkembangan zaman [4]. Salah satu cara untuk mengorganisasikan informasi dalam jumlah banyak dan dapat dipahami oleh para pencari informasi atau pembaca adalah dengan melakukan klasifikasi dokumen berdasarkan kategorinya. Dengan banyaknya informasi yang didapatkan dari *website*, pembaca membutuhkan waktu yang tidak singkat untuk mendapatkan informasi yang sesuai dengan keinginan pembaca. Untuk mempermudah pembaca untuk mencari informasi sesuai kebutuhan dan tujuan yang ingin dicapai, dibutuhkan suatu metode yang dapat mengklasifikasikan dokumen secara otomatis sesuai dengan isi dalam artikel yang ada pada *website* tersebut. Ada banyak metode atau teknik klasifikasi untuk dokumen, masing-masing teknik klasifikasi memiliki kekurangan dan kelebihan. Oleh karena itu, penelitian ini mencari algoritma untuk melakukan klasifikasi yang paling akurat untuk dokumen teks dengan membandingkan pengelompokan dokumen-dokumen berita untuk menemukan metode algoritma pengklasifikasian yang paling akurat.

Klasifikasi merupakan suatu metode untuk memprediksi kategori dari suatu *item* atau data. Klasifikasi adalah proses menemukan model atau fungsi yang cocok untuk mendeskripsikan dan membedakan sebuah kategori dengan kategori lainnya. Data yang diklasifikasi merupakan informasi yang dapat berupa teks, gambar, suara, video dan obyek multimedia lainnya. Informasi dalam bentuk teks merupakan fokus utama dalam penelitian ini. Secara umum, klasifikasi teks telah dianggap sebagai metode penting untuk mengelola dan memproses sekian banyak dokumen digital yang senantiasa mengalami banyak perkembangan dari waktu ke waktu [5].

Klasifikasi memerlukan informasi atau dokumen sebagai suatu dataset yang dibagi menjadi data *training* dan data *testing*. Klasifikasi pada penelitian ini membutuhkan data *training* yang sudah memiliki kategori awal. Berita merupakan salah satu *dataset* yang mudah digunakan karena berita mempunyai kategori sesungguhnya, kategori berita yang dimaksudkan seperti berita politik, olahraga, ekonomi, kesehatan, kebudayaan, otomotif dan lain-lain [6]. Berita yang sudah mempunyai kategori dapat dilakukan *supervised learning*. Berita yang akan digunakan sebagai data *training* dan data *testing* didapat dengan berbagai cara salah satunya dengan menggunakan *web scraping*. Berita yang digunakan tersebut perlu dilakukan *preprocessing text*. Tahapan *preprocessing text* diantaranya adalah *tokenizing*, *case folding*, *stopwords*, dan *stemming*. Setelah dilakukan *preprocessing text*, dokumen-dokumen akan dimodelkan dengan *Vector Space Model* (VSM).

Vector Space Model (VSM) adalah model yang dapat digunakan untuk melihat tingkat kedekatan atau kesamaan (similarity) dokumen dengan cara pembobotan *term*. *Vector Space Model* (VSM) merupakan pemodelan dengan merepresentasikan dokumen menjadi sebuah vektor [7]. *Vector Space*

Model (VSM) memiliki beberapa kekurangan seperti kata yang terdaftar sangat banyak sehingga menyebabkan dimensi vektor yang besar dan tidak memperhatikan kemiripan antara dua kata. Untuk mengurangi dimensi vektor dapat dengan menggunakan *Latent Semantic Indexing* (LSI) dan merepresentasikan dokumen ke dalam kategori yang akan meningkatkan deteksi dokumen yang relevan [8]. *Latent Semantic Indexing* (LSI) berfungsi untuk mendapatkan suatu pemodelan yang efektif untuk merepresentasikan dengan baik hubungan antara dokumen dan kategori yang dicari dengan memecah dokumen term ke dalam beberapa matriks.

Setelah dokumen direpresentasikan dalam bentuk dimensi vektor yang sederhana dengan menggunakan *Latent Semantic Indexing* (LSI), selanjutnya dilakukan klasifikasi mengelompokkan dalam masing-masing kategori. Kemudian teks akan dikategorikan atau diklasifikasikan dengan algoritma-algoritma klasifikasi. Berbagai macam algoritma klasifikasi yang banyak digunakan dalam melakukan klasifikasi berupa teks diantaranya adalah *Naive Bayes*, *Support Vector Machine* (SVM) dan *K-Nearest Neighbor* (KNN) dengan menggunakan *Latent Semantic Indexing* (LSI) lalu dibandingkan keakuratannya dengan algoritma-algoritma tersebut.

Metode pengklasifikasian *Naive Bayes* pertama kali dikemukakan oleh ilmuwan inggris Thomas Bayes. Ide dari metode *Naive Bayes* yang dikenal sebagai Teorema Bayes yaitu dengan menggunakan metode probabilitas dan statistik yang dapat memprediksi probabilitas atau peluang keanggotaan kategori suatu data yang akan masuk ke dalam kategori tertentu [9]. *Support Vector Machine* (SVM) merupakan metode klasifikasi untuk data linear dan nonlinear. *Support Vector Machine* (SVM) merupakan suatu teknik untuk menemukan *hyperplane* sebagai pemisah yang bisa membagi data menjadi 2 kategori yang berbeda [10]. *Support Vector Machine* (SVM) memiliki kelebihan diantaranya adalah dalam menentukan jarak menggunakan *support vector* sehingga proses komputasi menjadi cepat. *Support vector* adalah objek data terluar yang paling dekat dengan *hyperplane*. Objek yang disebut *support vector* paling sulit diklasifikasikan, karena posisi yang hampir tumpang tindih dengan kategori lain. *Support vector* inilah yang diperhitungkan untuk menemukan *hyperplane* yang paling optimal oleh *Support Vector Machine* (SVM).

Metode *K-Nearest Neighbor* (KNN) adalah proses untuk mengelompokkan data ke dalam kategori-kategori yang telah ditentukan sebelumnya berdasarkan jarak terdekat atau tingkat kemiripan data tersebut [11]. *Nearest Neighbor* adalah suatu pemodelan untuk mencari kasus dengan menghitung kedekatan antara kasus baru dengan kasus lama, yaitu berdasarkan pada pencocokan bobot dari sejumlah fitur yang ada [12]. Prinsip kerja *K-Nearest Neighbor* (KNN) yaitu untuk melakukan klasifikasi suatu data berdasarkan data *train* yang diambil dari k tetangga terdekatnya.

Banyak penelitian yang telah dilakukan dalam mengklasifikasikan berita menggunakan klasifikasi teks, dalam penelitian yang dilakukan Septian dan tim [13] serta Asyarie dan partnernya [14] dalam mengklasifikasikan berita menggunakan metode klasifikasi *Naive Bayes* menunjukkan bahwa *Naive Bayes* merupakan metode yang sederhana dengan akurasi yang tinggi, namun kekurangan dari metode *Naive Bayes* yaitu menimbulkan masalah performasi jika memproses data yang besar. Penelitian yang dilakukan [15] menyatakan bahwa metode *Support Vector Machine* (SVM) memiliki kelebihan yaitu dapat diketahui memiliki tingkat akurasi yang baik dan tidak dipengaruhi oleh besar kecilnya data uji. Sementara itu, algoritma *K-Nearest Neighbor* (KNN) menghasilkan performa yang baik dengan nilai sederhana dan efisien [16]. Oleh karena itu, untuk menentukan metode atau algoritma dari ketiga tersebut yang paling akurat dan efisien untuk digunakan, penelitian ini bertujuan untuk melakukan perbandingan hasil akurasi pada algoritma *Naive Bayes*, *Support Vector Machine* (SVM) dan *K-Nearest Neighbor* (KNN) terhadap fitur TF-IDF dan fitur dari *Latent Semantic Indexing* (LSI).

Dalam penelitian ini akan dibangun sebuah perangkat lunak untuk melakukan klasifikasi dokumen berita dengan *Naive Bayes*, *Support Vector Machine*, dan *K-Nearest Neighbor* dengan menggunakan *Latent Semantic Indexing*. Data yang akan digunakan pada penelitian ini berupa berita dari hasil *web scraping* (.csv). Bahasa yang digunakan dalam pembuatan perangkat lunak ini adalah bahasa pemrograman Python dan C# untuk *User Interface* (UI). Penelitian ini menggunakan 3 metode yang digunakan untuk klasifikasi diantaranya adalah *Naive Bayes*, *Support Vector Machine*

(SVM) dan *K-Nearest Neighbor* (KNN). Pengujian yang dilakukan akan bersifat eksperimental dan fungsional, diantaranya mencari nilai k yang paling terbaik pada metode *K-Nearest Neighbor* (KNN) dan *Singular Value Decomposition* (SVD) dengan menggunakan metrik evaluasi *F-Measure* dan *Accuracy*.

1.2 Rumusan Masalah

Berdasarkan latar belakang diatas, berikut rumusan masalah dari penelitian ini:

1. Bagaimana pembuatan fitur dari teks menggunakan *Vector Space Model* (VSM) ?
2. Bagaimana pembuatan fitur dari teks dengan melakukan pendekatan *Latent Semantic Indexing* (LSI) untuk mereduksi dimensi pada *Vector Space Model* (VSM) ?
3. Bagaimana klasifikasi dokumen menggunakan algoritma *Naïve Bayes*, *Support Vector Machine*, dan *K-Nearest Neighbor* ?
4. Bagaimana melakukan perbandingan evaluasi performa antara algoritma *Naïve Bayes*, *Support Vector Machine*, dan *K-Nearest Neighbor*?

1.3 Tujuan

Berdasarkan rumusan masalah diatas, tujuan dari penelitian adalah sebagai berikut:

1. Mengimplementasikan *Vector Space Model* (VSM) untuk pembuatan fitur dari teks.
2. Mempelajari reduksi dimensi pada *Vector Space Model* (VSM) sehingga mendapatkan representasi dokumen yang dimensinya lebih kecil dengan *Latent Semantic Indexing* (LSI).
3. Mengaplikasikan algoritma *Naïve Bayes*, *Support Vector Machine*, dan *K-Nearest Neighbor* untuk klasifikasi dokumen.
4. Menerapkan perbandingan evaluasi performa antara algoritma *Naïve Bayes*, *Support Vector Machine*, dan *K-Nearest Neighbor*.

1.4 Batasan Masalah

Adapun batasan masalah yang digunakan agar penelitian dapat fokus adalah sebagai berikut:

1. Penelitian hanya membahas dokumen berita dari situs berita *online* yaitu kompas.
2. Kategori berita yang digunakan adalah Teknologi, Kesehatan, Politik, Ekonomi dan Olahraga.

1.5 Metodologi

Berikut langkah-langkah yang dilakukan dalam pembuatan skripsi:

1. Melakukan observasi pada portal online untuk mencari sekumpulan dataset.
2. Melakukan studi literatur mengenai *Vector Space Model* (VSM).
3. Melakukan studi literatur mengenai *Latent Semantic Indexing* (LSI).
4. Melakukan studi literatur mengenai *Naive Bayes Classifier*.
5. Melakukan studi literatur mengenai *Support Vector Machine* (SVM).
6. Melakukan studi literatur mengenai *K-Nearest Neighbor* (KNN).
7. Melakukan pencarian dataset.
8. Melakukan *text pre-processing* pada dataset atau dokumen.
   (a) *Case Folding*
   (b) *Tokenizing*
   (c) *Filtering*
   (d) *Stopword Removing*
   (e) *Stemming*
9. Merancang struktur data dan analisa yang cocok untuk menyelesaikan masalah.

10. Melakukan eksperimen untuk mencari nilai k yang paling terbaik pada metode *K-Nearest Neighbor* (KNN) dan *Singular Value Decomposition* (SVD).
11. Melakukan perbandingan hasil akurasi dan *f-measure* dari algoritma *Naïve Bayes*, *Support Vector Machine* (SVM), dan *K-Nearest Neighbor* (KNN).
12. Melakukan rancangan desain perangkat lunak.
13. Mengimplementasikan hasil perancangan perangkat lunak.
14. Melakukan pengujian fungsional dan pengujian eksperimental terhadap perangkat lunak yang dibuat.
15. Membuat kesimpulan terhadap penelitian yang sudah dilakukan.
16. Menulis dokumen skripsi.

1.6 Sistematika Pembahasan

Sistematika pembahasan merupakan uraian tentang susunan penulisan itu sendiri yang dibuat secara teratur dan terperinci sehingga dapat memberikan gambaran secara menyeluruh. Adapun sistematika pembahasan pada dokumen skripsi ini terbagi menjadi lima bab, yaitu:

1. Bab 1. Pendahuluan
   Bab ini membahas tentang latar belakang, rumusan masalah, tujuan, batasan masalah, metodologi, dan sistematika pembahasan.
2. Bab 2. Landasan Teori
   Bab ini membahas tentang teori-teori atau pemikiran yang digunakan untuk mendukung pembuatan skripsi mengenai klasifikasi, algoritma klasifikasi, evaluasi klasifikasi, klasifikasi teks, *text preprocessing*, *Vector Space Model*, *Singular Value Decomposition*, *Latent Semantic Indexing*, *web scraping*, dan *library* python.
3. Bab 3. Analisis
   Bab ini membahas tentang analisis masalah yang telah dideskripsikan pada Bab 1 yaitu analisis masalah, metode pengujian, analisa algoritma, analisis dataset, analisis *library*, dan *flowchart*.
4. Bab 4. Eksperimen
   Bab ini membahas tentang lingkungan eksperimen dalam membandingkan algoritma klasifikasi dan mencari nilai k yang paling terbaik pada metode *K-Nearest Neighbor* (KNN) dan *Singular Value Decomposition* (SVD).
5. Bab 5. Perancangan Model dan Implementasi
   Bab ini membahas tentang perancangan antarmuka yang ingin dibangun dan implementasi antarmuka , berisi tentang pengujian fungsional terhadap perangkat lunak yang telah dibuat, diagram kelas, penjelasan kelas dan diagram aktivitas.
6. Bab 6. Kesimpulan dan Saran
   Bab ini membahas tentang kesimpulan akhir dari hasil penelitian yang telah dilakukan dan saran-saran untuk melakukan pengembangan di penelitian lebih lanjut.