BAB II

LANDASAN TEORI DAN TINJAUAN PUSTAKA

2.1 Tinjauan Pustaka

Penelitian tentang pengelompokan dokumen teks dengan metode *K-Means Clustering* dan *Naïve Bayes Classifier* bukan pertama kalinya dilakukan, terdapat beberapa penelitian sebelumnya yang telah menyelesaikan studi kasus dengan berbagai macam metode.

Penelitian yang berjudul Analisis Perbandingan Metode *Clustering Single Linkage* dan *K-Menas* pada Pengelompokkan Tugas Akhir Teknik Informatika oleh Febrian Rizky Ramadhan (2017) disimpulkan bahwa pengujian dengan menggunakan metode *cluster analysis of variance* menunjukkan bahwa hasil cluster dari metode *single linkage* merupakan cluster yang ideal daripada metode *k-means*. Hal itu dikarenakan nilai *variance* metode *single linkage* lebih kecil daripada nilai *variance* metode *k-means*. Sedangkan Pengujian dengan menggunakan metode *silhouette coefficient* menunjukkan bahwa hasil *cluster* dari metode single linkage memiliki nilai *silhouette* yang lebih baik daripada metode *k-means*. Hal itu dikarenakan pada pembentukan 3, 5, dan 7 cluster, nilai silhouette dari metode *single linkage* lebih mendekati nilai 1 atau sudah berada pada *cluster* yang tepat.

Oman Somantri, Slamet Wiyono, dan Dairoh (2017) melakukan penelitian dengan judul Optimalisasi *Support Vektor Machine* (SVM) untuk Klasifikasi Tema Tugas Akhir Berbasis *K-Means*. Setelah dilakukan eksperimen terdapat perbedaan antara model dari SVM dibandingkan dengan model SVM + K-Means, dimana

tingkat akurasi sebelumnya 85,38% menjadi 86,21% . Sehingga Model SVM dan *K-Means* dapat digunakan oleh para pengambil kebijakan dalam menglasifikasikan kategori tugas akhir sebagai pendukung keputusan dalam penentuan tema tersebut. *K-Means* menjadi model optimalisasi untuk dapat meningkatkan tingkat akurasi model SVM dalam mengklasifikasikan kategori tema tugas akhir.

Ria Melita, dkk (2018) melakukan penelitian tentang Penerapan Metode *Term Frequency Inverse Document Frequency* (TF-IDF) dan *Cosine Similarity* pada Sistem Temu Kembali Informasi untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus : Syarah Umdatil Ahkam). Hasil penelitian ini menjelaskan bahwa metode term *frequency inverse document frequency* (tf-idf) dan *cosine similarity* telah berhasil diterapkan dalam sistem dengan baik dimana sistem dapat memberikan output berupa dokuemn yang relevan yaitu syarah hadits sesuai dengan *query* yang di-inputkan, dengan melalui 3 tahapan teks *preprocessing* yaitu *tokenizing*, *stopword removal* atau *filtering*, dan *steamming*. Hasil pengujian stemming model Nazief Adriani yang telah dilakukan menunjukkan hasil akurasi sebesar 90,93% yang menyatakan bahwa stemming model ini memiliki tingkat akurasi yang tinggi. Kemudian, pengujian sistem yang dilakukan dengan menggunakan confusion matrix dalam penelitian ini diadapat nilai *precision* 100% *recall* 88,7%, , *accuracy* 88,73% dan *error rate* 11,27%. Sehingga sistem dapat dikatakan baik, dikarenakan sistem baik adalah sistem yang memiliki nilai *recall* dan *precision* tinggi serta tingkat akurasi yang tinggi pula.

Penelitian yang berjudul Implementasi *K-Means Clustering* pada Terjemahan Al-Qur'an Berdasarkan Keterkaitan Topik oleh Ahmad Salam Wahid Faizin (2018)

diperoleh kesimpulan bahwa penelitian menggunakan algoritme *K-Means* untuk melakukan clustering terjemahan ayat-ayat Al-Qur'an mempunyai akurasi 43%. Hasil tersebut dikarenakan pada beberapa ayat Al-Qur'an terdapat lebih dari satu topik, dan algoritme *K-Means* hanya melakukan clustering berdasrkan satu *topic* saja.

Penelitian yang dilakukan oleh Kitami Akromunisa (2019) dengan judul Klasifikasi Dokumen Tugas Akhir (Skripsi) Menggunakan *K-Nearst Neighbor*. Hasil penelitian ini dapat disimpulkan bahwa klasifikasi menggunakan metode k-nearest neighbor bisa digunakan untuk mengklasifikasi dokumen tugas akhir berdasarkan data intisari bahasa Indonesia , abstrak bahasa inggris dan judul skripsi, dengan hasil akurasi yang lebih besar tanpa melalui proses stemming. Pada pembagian data yaitu *Spli into train test sets* dan *KFold cross validation* menghasilkan akurasi yang lebih besar menggunakan *Split into train test sets.* Untuk *Split into train test sets* dengan rasio perbandingan 9:1 menghasilkan akurasi lebih besar dibandingkan dengan rasio perbandingan 6:4, 7:3, 8:2. Untuk pembagian data menggunakan *kfold cross validation* akurasi terbesar didapatkan k maksimal yaitu 10 dengan nilai akurasi sebesar 86,4% berdasarkan data uji judul skripsi.

Klasifikasi dokumen karya akhir mahasiswa menggunakan *Naïve Bayes Classifier* (NBC) berdasarkan abstrak karya akhir di Jurusan Teknik Elektro Universitas Negeri Jakarta oleh Nur Indah Pratiwi & Widodo (2017) menghasilkan penelitian dengan menggunakan 100 dokumen abstrak, 90 dokumen sebagai data train dan 10 dokumen sebagai data test. Data diambil dari skripsi mahasiswa

Jurusan Teknik Elektro Universitas Negeri Jakarta dari 14 Maret 2014 sampai dengan 27 Maret 2014. Setelah melakukan proses pengembangan perangkat lunak, dihasilkan sebuah sistem klasifikasi yang bernama Sistem Klasifikasi Dokumen Skripsi. Sistem di implementasi menggunakan PHP dan MySQL, dan diuji menggunakan *K-Fold Cross Validation* (10 Fold). Berdasarkan pada hasil uji Sistem didapatkan hasil tingkat akurasi sebesar 81%. Oleh karena itu, dapat disimpulkan bahwa Sistem Klasifikasi Dokumen Abstrak Karya Akhir Menggunakan Algoritma *Naïve Bayes* di Jurusan Teknik Elektro telah berhasil dikembangkan.

Adapun hasil perbandingan tinjauan pustaka disajikan dalam Tabel 2.1, 2.2, dan 2.3.

Tabel 2. 1 Perbandingan Tinjauan Pustaka

No	Judul	Peneliti	Metode	Hasil
1.	Analisis Perbandingan Metode *Clustering Single Linkage* dan *K-Means* pada Pengelompokkan Tugas Akhir Teknik Informatika	Febrian Rizky Ramadhan (2017)	*Clustering Single Linkage* dan *K-Means*	Hasil penelitian ini dapat disimpulkan pengujian dengan menggunakan metode *cluster analysis of variance* menunjukkan bahwa hasil cluster dari metode *single linkage* merupakan cluster yang ideal daripada metode *k-means*. Hal itu dikarenakan nilai *variance* metode *single linkage* lebih kecil daripada nilai *variance* metode *k-means*. Sedangkan Pengujian dengan menggunakan metode *silhouette coefficient* menunjukkan bahwa hasil *cluster* dari metode single linkage memiliki nilai *silhouette* yang lebih baik daripada metode *k-means*. Hal itu dikarenakan pada pembentukan 3, 5, dan 7 cluster, nilai silhouette dari metode *single linkage* lebih mendekati nilai 1 atau sudah berada pada *cluster* yang tepat.
2.	Optimalisasi *Support Vektor Machine* (SVM) untuk Klasifikasi Tema Tugas Akhir Berbasis *K-Means*	Oman Soemantri, Slamet Wiyono dan Dairoh (2016	*Support Vektor Machine* (SVM) dan K-Means	Setelah dilakukan eksperimen terdapat perbedaan antara model dari SVM dibandingkan dengan model SVM + *K-Means*, dimana tingkat akurasi sebelumnya 85,38% menjadi 86,21%. Sehingga Model SVM dan *K-Means* dapat digunakan oleh para pengambil kebijakan dalam menglasifikasikan kategori tugas akhir sebagai pendukung keputusan dalam penentuan tema tersebut. *K-Means* menjadi model optimalisasi untuk dapat meningkatkan tingkat akurasi model SVM dalam mengklasifikasikan kategori tema tugas akhir.

Tabel 2. 2 Perbandingan Tinjauan pustaka (Lanjutan)

No	Judul	Peneliti	Metode	Hasil
3.	Penerapan Metode *Term Frequency Inverse Document Frequency* (TF-IDF) dan *Cosine Similarity* pada Sistem Temu Kembali Informasi untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus : Syarah Umdatil Ahkam)	Ria Melita, dkk (2018	*Term Frequency Inverse Document Frequency* (TF-IDF) dan *Cosine Similarity*	Metode term *frequency inverse document frequency* (tf-idf) dan *cosine similarity* telah berhasil diterapkan dalam sistem dengan baik dimana sistem dapat memberikan output berupa dokuemn yang relevan yaitu syarah hadits sesuai dengan *query* yang di-inputkan, dengan melalui 3 tahapan *teks preprocessing* yaitu *tokenizing*, *stopword removal* atau *filtering*, dan *steaming*. Hasil pengujian stemming model Nazief Adriani yang telah dilakukan menunjukkan hasil akurasi sebesar 90,93% yang menyatakan bahwa stemming model ini memiliki tingkat akurasi yang tinggi. Kemudian, pengujian sistem yang dilakukan dengan menggunakan *confusion matrix* dalam penelitian ini diadapat nilai *precision* 100% *recall* 88,7%, , *accuracy* 88,73% dan *error rate* 11,27%.
4.	Implementasi *K-Means Clustering* pada Terjemahan Al-Qur'an Berdasarkan Keterkaitan Topik	Ahmad Salam Wahid Faizin (2018)	*K-Means Clustering*	Penelitian menggunakan algoritme *K-Means* untuk melakukan clustering terjemahan ayat-ayat Al-Qur'an mempunyai akurasi 43%. Hasil tersebut dikarenakan pada beberapa ayat Al-Qur'an terdapat lebih dari satu topik, dan algoritme *K-Means* hanya melakukan clustering berdaarkan satu topic saja.

Tabel 2. 3 Perbandingan Tinjauan pustaka (Lanjutan)

No	Judul	Peneliti	Metode	Hasil
5.	Klasifikasi Dokumen Tugas Akhir (Skripsi) Menggunakan *K-Nearst Neighbor*	Kitami Akromunnisa (2019)	K-Nearst Neighbor	Pada pembagian data yaitu *Spli into train test sets* dan *KFold cross validation* menghasilkan akurasi yang lebih besar menggunakan *Split into train test sets.* Untuk *Split into train test sets* dengan rasio perbandingan 9:1 menghasilkan akurasi lebih besar dibandingkan dengan rasio perbandingan 6:4, 7:3, 8:2. Untuk pembagian data menggunakan *kfold cross validation* akurasi terbesar didapatkan k maksimal yaitu 10 dengan nilai akurasi sebesar 86,4% berdasarkan data uji judul skripsi.
6.	Klasifikasi Dokumen Karya Akhir Mahasiswa Menggunakan *Naïve Bayes Classifier* (NBC) Berdasarkan Abstrak Karya Akhir Di Jurusan Teknik Elektro Universitas Negeri Jakarta	Nur Indah Pratiwi & Widodo (2017)	*Naïve Bayes Classifier*	Hasil penelitian ini menggunakan 100 dokumen abstrak, 90 dokumen sebagai data train dan 10 dokumen sebagai data test. Data diambil dari skripsi mahasiswa Jurusan Teknik Elektro Universitas Negeri Jakarta dari 14 Maret 2014 sampai dengan 27 Maret 2014. Setelah melakukan proses pengembangan perangkat lunak, dihasilkan sebuah sistem klasifikasi yang bernama Sistem Klasifikasi Dokumen Skripsi. Sistem di implementasi menggunakan PHP dan MySQL, dan diuji menggunakan K-Fold Cross Validation (10 Fold). Berdasarkan pada hasil uji Sistem didapatkan hasil tingkat akurasi sebesar 81%. Oleh karena itu, dapat disimpulkan bahwa Sistem Klasifikasi Dokumen Abstrak Karya Akhir Menggunakan Algoritma Naïve Bayes di Jurusan Teknik Elektro telah berhasil dikembangkan.

2.2 Landasan Teori

2.2.1 *Text Mining*

Text mining adalah cabang dari bidang data *mining*. *Text mining* memiliki kekhasan tersendiri karena menggunakan teks sebagai datanya. *Text mining* biasa digunakan untuk memperoleh informasi yang tidak diketahui sebelumnya, dari sumber tertulis, secara otomatis. Informasi yang digali ini akan menjadi suatu fakta baru yang dapat diteliti lebih lanjut. *Text mining* berbeda dengan web *search*. Web *search* bertujuan untuk menemukan kembali informasi yang telah ditulis oleh seseorang atau telah ada sebelumnya, sedangkan *text mining* bertujuan untuk menggali informasi yang belum diketahui sebelumnya dari sebuah sumber tertulis (Hearst, 1999)

2.2.2 *Text Preprocessing*

Preprocessing adalah proses normalisasi teks sehingga informasi yang dimuat merupakan bagian yang padat dan ringkas namun tetap merepresentasikan informasi yang termuat didalamnya. Dalam tahap ini, terdapat beberapa proses yatu :

1. *Cleansing*

   Cleansing adalah proses membersihkan dokumen dari komponen-komponen seperti tanda baca.

2. *Casefolding*

   Case folding merupakan tahapan yang mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf ‘a’ sampai dengan ‘z’ yang diterima. Karakter selain huruf dihilangkan dan dianggap *delimiter* (pembatas) (Triawati, 2009).

3. *Tokenizing*

   Tahap *tokenizing / parsing* adalah tahap pemotongan string input berdasarkan tiap kata yang menyusunnya (Triawati, 2009)

4. *Stopwords Removal*

   Stopwords removal adalah proses penghilangan kata yang tidak penting pada deskripsi melalui pengecekan kata-kata hasil *parsing* deskripsi apakah dalam daftar kata yang tidak penting atau tidak.

2.2.3 *Term Frequency-Inverse Document Frequency (TF-IDF)*

Metode *TF-IDF* (merupakan suatu cara untuk memberikan bobot hubunngan suatu kata (token) terhadap suatu dokumen. Metode ini menggunakan dua konsep dalam perhitungan bobot yaitu, frekuensi kemunculan sebuah kata di dalam sebuah dokumen dan *inverse* dari frekuensi dokumen yang mengandung kata tersebut. Frekuensi dokumen yang mengandung kata tersebut menunjukkan seberapa umum kata tersebut. Bobot hubungan antara sebuah kata dan sebuah dokumen akan tinggi apabila frekuensi kata tersebut tinggi didalam dokumen dan

frekuensi keseluruhan dokumen yang mengandung kata tersebut yang rendah pada kumpulan dokumen (*database)* (Robertson, 2004).

Rumus umum untuk pembobotan TF-IDF yang ditulis oleh Intan dan Defeng (2006) :

$$w_{td} = tf_{td} * idf \quad (2.1)$$

$$w_{td} = tf_{td} * log\ (\frac{N}{df_t}) \quad (2.2)$$

Keterangan :

w_{td} = bobot kata/token t_t terhadap dokuemen d_d

tf_{td} = jumlah kemunculan kata/token t_t dalam dokumen d_d

N = jumlah semua dokumen dalam *database*

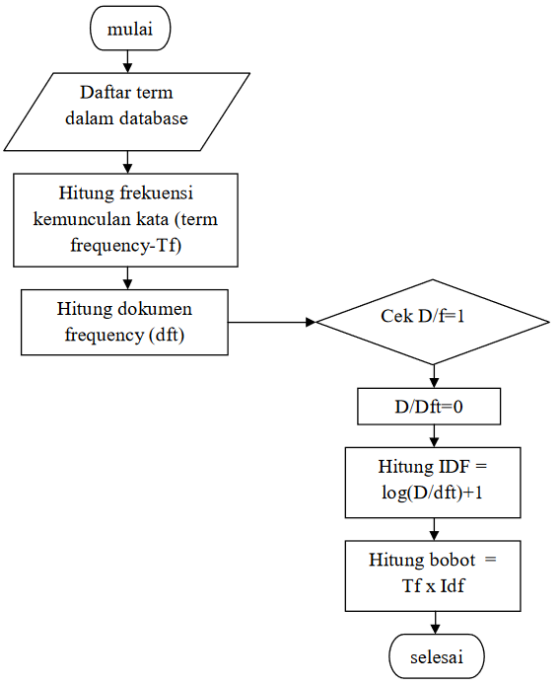
df_t = jumlah dokumen yang mengandung kata/token t_t

Berdasarkan rumus (2.2) , berapapun besarnya nilai tf_{td}, apabila N = df_t dimana sebuah kata/token muncul di semua dokumen, maka akan didapatkan hasil 0 (nol) untuk perhitungan idf , sehingga perhitungan bobotnya diubah menjadi sebagai berikut :

$$w_{td} = tf_{td} * (log\ (\frac{N}{df_t}) + 1) \quad (2.3)$$

Rumus (2.3) dapat dinormalisasikan dengan rumus (2.4) dengan tujuan menstandarisasi nilai bobot (w_{td}) kedalam interval 0 s.d. 1, rumusnya adalah sebagai berikut :

$$w_{td} = \frac{tf_{td} * (log\ (\frac{N}{df_t}) + 1)}{\sqrt{\sum_{k=1}^{t} (tf_{td})^2 \times [(log\ (\frac{N}{df_t}) + 1)]^2}} \quad (2.4)$$



Gambar 2. 1 Diagram alir pembobotan TF-IDF (Safitri, 2013)

2.2.4 *Cosine Similarity*

Cosine similarity atau kemiripan kosinus adalah ukuran jarak yang digunakan untuk data yang berupa vektor. Vektor dokumen adalah sebuah vektor yang menyatakan frekuensi kemunculan istilah dalam dokumen tersebut. Secara matematis *cosine* diformulasikan sebagai berikut (Suyanto, 2017) :

$$sim(x, y) = \frac{x.y^t}{\|x\|\|y\|} \quad (2.5)$$

di mana y^t adalah *transpose* dari vektor y, dan $\|x\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_p^2}$ adalah jarak dari vektor $x = x_1^2 + x_2^2 + \cdots + x_p^2$. Secara konsep formula ini menyatakan kosinus sudut antara x dan vektor y. Dengan demikian, formula ini dapat menhasilkan nilai dalam rentang [0,0;1,0], jika vektor x dan vektor y tidak memiliki kemiripan sama sekali, maka kedua vektor tersebut akan membentuk 90 derajat (orthogonal atau tegak lurus) sehingga nilai kosinus sudutnya adalah 0, artinya $sim(x, y)$ = 0. Sebaliknya, jika kedua vektor sama persis, maka keduanya akan memebentuk sudut 0 derajat (berimpit) sehingga $sim(x, y) = 1$

2.2.5 Klasterisasi

Klasterisasi adalah salah satu teknik *unsupervised learning* dimana kita tidak perlu melatih metode tersebut atau dengan kata lain, tidak ada fase *learning.* Masuk dalam pendekatan *unsupervised learning* adalah metode-metode yang tidak membutuhkan label atau *output* dari setiap titik data yang kita investigasi. Sebaliknya *supervised learning* adalah metode yang memerlukan *training* (melatih) dan *testing* (menguji). Masuk dalam kategori ini adalah regresi, *Artificial neural network* (ANN), analisis diskriminan (LDA) dan *support vector machine* (Santosa and Umam, 2018)

2.2.6 Klasifikasi

Teknik klasifikasi yaitu bagaimana mempelajari sekumpulan data sehingga dihasilkan aturan yang bisa mengklasifikasikan atau mengenali data-data baru yang belum pernah dipelajari. Klasifikasi dapat didefinisikan

sebagai proses untuk menyatakan suatu objek data sebagai salah satu kategori (kelas) yang telah didefinisakan sebelummnya (Zaki and Meira, 2013).

Model klasifikasi dapat dibangun berdasarkan pengetahuan seorang pakar (ahli). Namun mengingat himpunan data yang sangat besar, model klasifikasi lebih sering dibangun menggunakan teknik pembelajaran dalam bidang *machine learning.* Proses pembelajaran secara otomatis terhadap suatu data mampu menghasilkan model klasifikasi (fungsi target) yang memetakan objek data x (input) ke salah satu kelas y yang telah didefinisikan sebelummnya. Jadi, proses pebelajaran memelukan masukan (*input*) berupa himpunan data latih (training set) yang berlabel (memiliki atribut kelas) dan mengeluarkan *output* yang berupa sebuah model kalsifikasi (Suyanto, 2017).

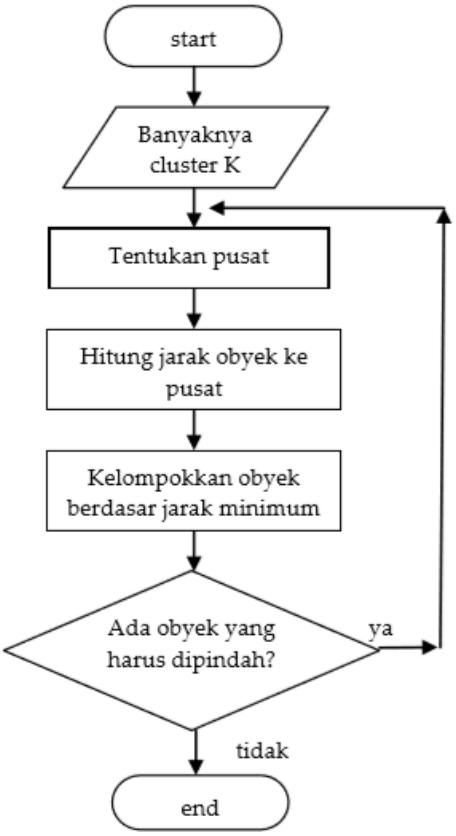
2.2.7 *K-Means Clustering*

Algoritma *k-means* bekerja dengan cara membagi data kedalam *k* buah *cluster* yang telah ditentukan. Pada Wu et al. (2012) dinyatakan bahwa riset tentang *k-means* telah dilakukan banyak penelitian dari beragam disiplin ilmu sejak tahun 1950-an, diantaranya adalah Lloyd (1957, 1982), Forgey (1965), Friedman dan Rubin (1967) serta macQueen (1967) (Han & Kamber 2011).

K-means bertujuan untuk meminimasi *Sum of Squared Error* (SSE) antara objek data dengan sejumlah k *centroid.* Langkah-langkah

pengkalsteran menggunkan algoritma *k-means* adalah sebagai berikut (Suyanto, 2018):

1. Dari himpunan data yang akan diklasterisasi, dipilih sejumlah k objek data secara acak sebagai titik pusat (*centroid*) awal.
2. Setiap objek yang bukan *centroid* dimasukkan ke klaster terdekat berdasarkan suatu ukuran jarak.
3. Setiap *centroid* diperbaharui berdasarkan rata-rata dari objek yang ada di dalam setiap klaster.
4. Lakukan iterasi untuk langkah kedua dan ketiga sampai *centroid* konvergen dan stabil, di mana semua centroid yang dihasilakan pada iterasi saat ini sama persis (atau berbeda tipis dengan toleransi tertentu yang diinginkan user) dengan semua *centroid* yang dihasilkan pada iterasi sebelumnya dan SEE stabil tidak mengalami peubahan yang berarti. Pada Gambar 2.2 ditunjukkan diagram alir pada algoritma *K-Means Clustering.*



Gambar 2. 2 Diagram Alir Algoritma *K-Means* (Andayani, 2007)

2.2.8 *Naïve Bayes Classifier*

Metode Naïve Bayes classifier merupakan metoda klasifikasi yang berdasar kepada teorema *bayes*, sebuah teorema yang terkenal di dalam bidang ilmu probabilitas. Selain itu, metode ini turut didukung oleh ilmu statistika khususnya dalam penggunaan data petunjuk untuk mendukung keputusan pengklasifikasian. Metode ini sangat luas dipakai dalam berbagai bidang, khususnya dalam proses klasifikasi dokumen (Pratiwi and Widodo, 2018).

Dasar dari Naïve Bayes yang dipakai dalam pemrograman adalah rumus Bayes :

P (A|B) = (P(B|A) * P(A))/P(B) (2.6)

Peluang kejadian A sebagai B ditentukan dari peluang B saat A, peluang A, dan peluang B. Pada pengaplikasiannya nanti rumus ini berubah menjadi : P(Ci|D) = (P(D|Ci)*P(Ci)) / P(D) (2.7)

Naïve Bayes Classifier atau bisa disebut sebagai merupakan model penyederhanaan dari Metode *Bayes* yang cocok dalam pengklasifikasian teks atau dokumen. Persamaannya adalah:

$$V_{MAP} = arg\ max\ P(V_j \mid a_1, a_2\ , \ldots\ldots a_n) \quad (2.8)$$

Menurut persamaan (2.8), maka persamaan (2.9) dapat ditulis :

$$V_{MAP} = \underset{v_j \in V}{\arg max}\ \frac{P(a_1,a_2\ ,\ldots\ldots a_n|v_j)P(v_j)}{P(a_1,a_2\ ,\ldots\ldots a_n)} \quad (2.9)$$

$P(a_1, a_2\ , \ldots\ldots a_n)$ konstan, sehingga dapat dihilangkan menjadi :

$$V_{MAP} = \underset{v_j \in V}{\arg max}\ P(a_1, a_2\ , \ldots\ldots a_n|v_j)P(v_j) \quad (2.10)$$

Karena $P(a_1, a_2\ , \ldots\ldots a_n|v_j)$ sulit dihitung, maka akan diasumsikan bahwa kata pada dokumen tidak mempunyai keterkaitan.

$$V_{MAP} = \underset{v_j \in V}{\arg max}\ P(v_j) \prod_i P(a_i \mid v_j\ ) \quad (2.11)$$

Keterangan :

$$P(v_j) = \frac{|docs_j|}{|Contoh|} \quad (2.12)$$

$$P(w_k|v_j) = \frac{|n_k+1|}{n+|kosakata|} \quad (2.13)$$

Dimana untuk :

$P(v_j)$: probabilitas setiap dokumen terhadap sekumpulan dokumen

$P(w_k|v_j)$: probabilitas kemunjulan kata w_k pada suatu dokumen dengan kategori *class* v_j.

$|docs_j|$: frekuensi dokumen pada setiap kaktegori.

$|Contoh|$: jumlah dokumen yang ada

n_k : frekuensi kata pada dokumen test

Pada persamaan (2.13) terdapat 1 penambahan pembilang, hal ini dilakukan untuk mengantisipasi jika terdapat suatu kata pada dokuemn uji yang tidak ada pada setiap dokumen data *training*.

2.2.9 *Silhoutte Cofficient*

Silhouette Coefficient digunakan untuk melihat kualitas dan kekuatan *cluster*, seberapa baik suatu objek ditempatkan dalam suatu *cluster*. Metode ini merupakan gabungan dari metode *cohesion* dan *separation.* Tahapan perhitungan *Silhouette Coefficient* (Handoyo et al., 2014) :

1. Hitung rata-rata jarak dari suatu dokumen misalkan i dengan semua dokumen lain yang berada dalam satu *cluster*

$$a(i) = \frac{1}{|A|-1} \sum j \in_{A, j \neq i} \; d(i,j) \quad (2.14)$$

dengan j adalah dokumen lain dalam satu *cluster A* dan $d(i,j)$ adalah jarak antara dokumen i dan j.

2. Hitung rata-rata jarak dari dokumen i tersebut dengan semua dokumen di *cluster* lain, dan diambil nilai terkecilnya.

$$a(i) = \frac{1}{|A|-1} \sum j \in C \; d(i,j) \quad (2.15)$$

dengan $d(i, C)$ adalah jarak rata-rata dokumen dengan semua objek pada *cluster* lain C dimana A $\neq C$.

$$b(i) = \min C \; \neq \; A \; d(i, C) \quad (2.16)$$

3. Nilai *Silhouette Coefficient* nya adalah :

$$s(i) = \frac{b(i)-1(a)}{\max(a(i).b(i))} \quad (2.17)$$

Nilai hasil *silhouette coefficient* terletak pada kisaran nilai 1 hingga -1. Semakin nilai *silhouette coefficient* mendekati nilai 1, maka semakin baik pengelompokan data dalam satu cluster. Sebaliknya jika nilai *silhouette coefficient* mendekati nilai -1, maka semakin buruk pengelompokan data di dalam satu cluster (Pramesti et al., 2017)

Untuk mengukur kualitas klaster dalam suatu klasterisai kita dapat menghitung rata-rata *silhouette coefficient* dari semua objek dalam klaster tersbut. Sementara itu, untuk mengukur kualitas suatu klasterisasi, kita dapat menggunakan rata-rata *silhouette coefficient* dari semua objek dalam himpunan data (Suyanto, 2017).

2.2.10 *Confusion Matrix*

Evaluasi dapat digunakan menggunakan suatu ukuran tertentu tujuh diantaranya diilustrasikan pada Tabel 2.4 (Han, Kamber and Pei, 2011).

Tabel 2. 4 Ukuran Evaluasi Model Klasifikasi

No	Ukuran	Rumus
1	*Accuracy* atau tingkat pengenalan	$\frac{TP + TN}{P + N}$
2	*Error rate* atau tingkat kesalahan atau kekeliruan klasifikasi	$\frac{FP + FN}{P + N}$
3	*Recall* atau *sensitivity* atau true *positive rate*	$\frac{TP}{P}$
4	*Specificity* atau *true negative rate*	$\frac{TN}{N}$
5	*Precision*	$\frac{TP}{TP + FP}$
6	F atau F_1 atau F-*score* atau rata-rata harmonik dari *precision* dan *recall*	$\frac{2 \times precision \times recall}{precission + recall}$
7	F_β dimana β adalah sebuah bilangan rill nonegatif	$\frac{(1 + \beta^2) \times precision \times recall}{\beta^2 \times precission + recall}$

Dimana FP adalah *true positives*, TN adalah *true negative*, FP adalah *false positives*, P adalah jumlah sample positif, dan N adalah jumlah sampel negatif. Istilah tersebut dapat diilustrasikan dengan menggunakan *confusion matrix* pada Gambar 2.3. *Confusion matrix* sangat berguna untuk

menganalisis kualitas model klasifikasi dalam mengenali *tuple-tuple* dari kelas yang ada (Suyanto, 2018).

	Ya	Tidak	Jumlah
Ya	TP	FN	P
Tidak	FP	TN	N
Jumlah	P	N	P + N

Gambar 2. 3 *Confusion Matrix*

2.2.11 *Latent Dirichlet Allocation* (LDA)

Latent Dirichlet Allocation (LDA) merupakan metode topik *modelling* dan topik analisis yang paling popular saat ini. LDA muncul sebagai salah satu metode yang dipilih melakukan analisis pada dokumen yang berukuran sangat besar. LDA dapat digunakan untuk meringkas, melakukan klasterisasi, menghubungkan maupun memproses data yang sangat besar karena LDA menghasilkan daftar topik yang diberi bobot untuk masing-masing dokumen (Putra dan Kusumawardani, 2017).

LDA merupakan model probabilistic generative dari kumpulan tulisan yang disebut *corpus.* Ide dasar yang diusulkan metode LDA adalah setiap dokumen dipresentasikan sebagai campuran acak atas topik yang tersembunyi, yang mana setiap topik memiliki karakter yang ditentukan berdasarkan distribusi kata-kata yang terdapat di dalamnya (Blei *et al.*, 2003).

2.2.12 Python

Python merupakan bahsa pemrograman yang berorientasi obyek dinamis, dapat digunakan untuk bermacam-macam pengembangan

perangkat lunak. Python menyediakan dukungan yang untuk integrasi dengan bahasa pemrograman lain dan alat-alat bantu lainnya. Python hadir dengan pustaka-pustaka standar yang dapat diperluas serta dapat dipelajari hanya dalam beberapa hari. Sudah banyak programmer Python yang menyatakan bahwa mereka mendapat produktivitas yang lebih tinggi. Python didistribusikan dibawah lisensi *Open Source* yang disetujui oleh OSI (*Open Source Intiatives*), sehingga Python bebas digunakan, gratis digunakan untuk produk-produk komersil. Beberapa keunggulan Python apabila dibandingkan bahasa pemrograman lain adalah (Santoso, 2016) :

1. Syntaxnya sangat bersih dan mudah dibaca.
2. Kemampuan melakukan pengecekan syntax yang kuat.
3. Berorientasi obyek secara intuitif.
4. Kode-kode prosedur dinyatakan pada ekspresi natural.
5. Modularitas yang penuh, mendukung hirarki paket.
6. Penanganan error berdasarkan ekspasi.
7. Tipe-tipe data dinamis berada pada tingkat sangat tinggi.
8. Library standar dapat diperluas dan modul dari pihak ketiga dapat dibuat secara virtual untuk setiap kebutuhan.
9. Ekstensi dan modul-modul dapat secara mudah ditulis dalam C. C++ (atau java untuk Jython atau NET. untuk IronPython).
10. Dapat dimasukkan kedalam aplikasi sebagai antar muka skrip.

2.2.13 Abstrak

Abstrak merupakan sintesa keseluruhan dari isi dokumen karya ilmiah. Abstrak mengandung konsep, pernyataan masalah, pendekatan, dan kesimpulan yang dirangkai sedemikian rupa saling berkaitan dan memiliki makna utuh sehingga menggambarkan keseluruhan isi tulisan. Abstrak walaupun sulit untuk ditulis di awal, tetapi dengan memahami keseluruhan isi tulisan, akan mudah ditulisakan di akhir penulisan, walaupun penempatannya selalu di awal secara manuskrip (Nasution, 2017).

2.2.14 ACM (*Association for Computing Machinary*)

Association for Computing Machinery (Asosiasi untuk Permesinan Komputer),adalah sebuah serikat ilmiah dan pendidikan komputer pertama di dunia yang didirikan pada tahun 1947. Anggota ACM sekitar 78.000 terdiri dari para profesional dan para pelajar yang tertarik akan komputer.

Dalam dekade sejak 1960-an, ACM, bersama dengan masyarakat komputasi profesional dan ilmiah terkemuka, telah berusaha untuk menyesuaikan rekomendasi kurikulum dengan lanskap teknologi komputer yang berubah dengan cepat. Ketika bidang komputasi terus berkembang, dan disiplin ilmu terkait komputasi yang baru muncul, laporan kurikulum yang ada akan diperbarui, dan laporan tambahan untuk disiplin komputasi baru akan disusun.

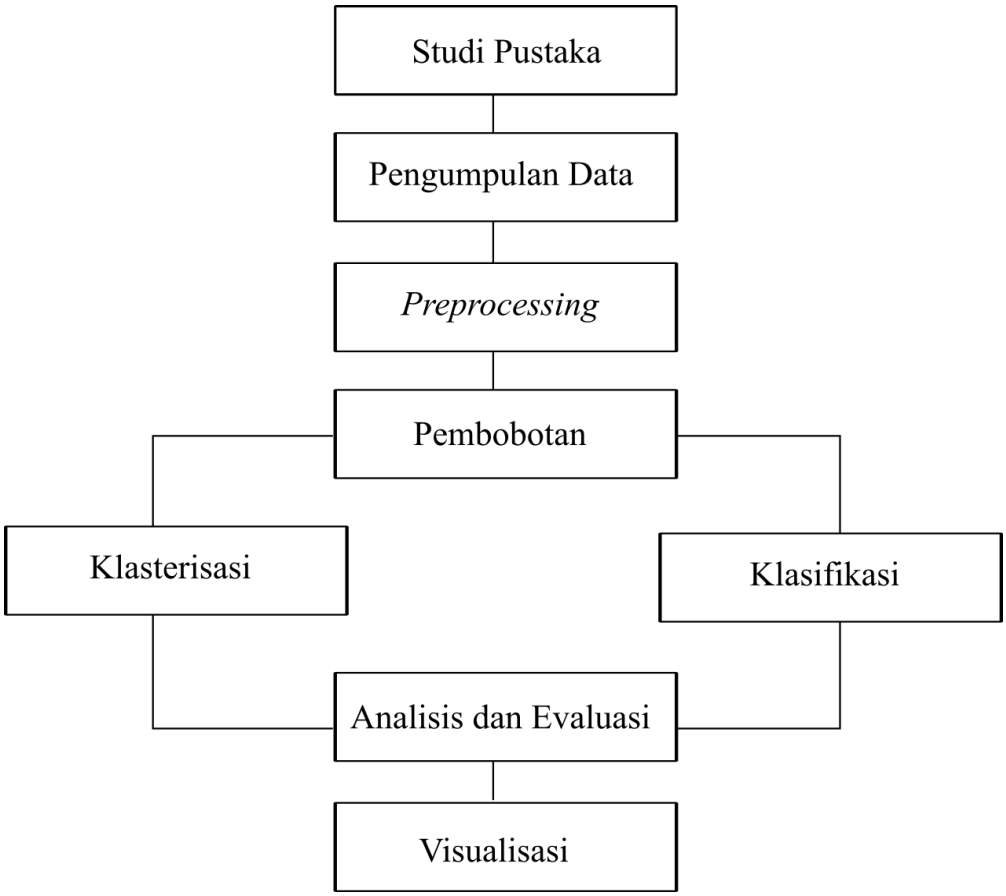
ACM (*Association for Computing Machinery*) *Computing Curricula* 2005 memberikan pedoman kurikulum sarjana untuk lima sub-disiplin komputasi yang ditentukan yaitu :

1. *Computer Engineering* (CE)
2. *Computer Science* (CS)
3. *Information System* (IS)
4. *Information Technology* (IT)
5. *Software Engineering* (SE)

BAB III

METODE PENELITIAN

Dalam penelitian ini, ada bebeapa tahapan yang dilakukan. Tahapan-tahapan tesebut disajikan pada Gambar 3.1.



Gambar 3. 1 Tahapan Penelitian

3.1 Studi Pustaka

Tahap studi pustaka dilakukan dengan cara membaca dan memahami tentang *text mining, clustering , k-means, python dan* teori-teori yang relevan dengan penelitian. Literatur-literatur yang dipelajari didapat dari buku, skripsi, tesis, karya ilmiah, dan internet. Studi pustaka yang dilakukan peneliti selain untuk memahami teori-teori terkait juga untuk mengkaji hasil-hasil penelitian sebelumnya sehingga terhindar dari penelitian yang sama.

3.2 Pengumpulan Data

Pengumpulan data berasal dari intisari dokumen yang ada di dalam dokumen tugas akhir S1 Teknik Informatika UIN Sunan Kalijaga Yogyakarta pada laman *Digital Library* UIN Sunan Kalijaga Yogyakarta dari tahun 2010 sampai dengan 2018 sebanyak 493 data.

3.3 *Preprocessing*

Tahap awal dalam pengolahan data adalah dengan melakukan tahap *preprocessing*. Tahapan ini diperlukan untuk merubah text yang belum terstruktur menjadi text yang terstruktur. Sehingga akan ditemukan kata-kata yang dapat mewakili dari suatu teks tersebut. Pada penelitian ini tahapan *preprocessing* meliputi pembersihan teks atau *cleansing* berupa tanda baca. Kemudian *casefolding* yaitu merubah semua huruf menjadi huruf kecil , Setelah itu *tokenizing* yaitu memotong text menjadi kata-kata dan terakhir *stopword removal* yaitu menghilangkan kata yang sering muncul.

Pada penelitian ini tidak dilakukan proses *stemming* karena pada penelitian yang dilakukan oleh Kitami Akromunisa (2019) telah dilakuan perbandingan tahapan *preprocessing* dengan dan tanpa proses *stemming. Hasilnya stemming* pada bahasa Indonesia menurunkan akurasi terhadap hasil klasifikasi. Saran yang diberikan adalah sebaiknya penelitian yang menggunakan klasifikasi text bahasa Indonesia tidak menggunakan proses *stemming.* Penelitian yang dilakukan oleh Yudi Wibisono dan Masayu Leylia Khodra (2005) juga mejelaskan mengenai eksperimennya tentang *clustering* berita bahasa Indonesia mengenai penggunaan pembobotan tf-idf tanpa *steamming* menghasilkan kualitas *cluster* terbaik.

3.4 Pembobotan

Penelitian ini menggunakan pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF) yang bertujuan untuk mengukur nilai *Term Frequency (TF)* dan *Inverse Document Frequency* (IDF) pada setiap kata di setiap dokumen.

3.5 Klasterisasi

Data yang telah dilakukan tahap *preprocessing* dan pembobotan TF-IDF yang menghasilkan keluaran vektor dokumen maka selajutnya akan masuk pada tahap klasterisasi berdasarkan ukuran jarak yaitu *Euclidean distance*. Metode klasterisasi yang digunakan pada penelitian ini adalah metode *k-means clustering*.

3.6 Klasifikasi

Tahap klasifikasi ini dilakukan dengan melakukan pelabelan secara manual terlebih dahulu. Model klasifikasi yang digunakan adalah kalsifikasi *Naïve Bayes Calsifier*.

3.7 Analisis dan Evaluasi

Analisis dilakukan untuk mengetahui hasil model klasterisasi dengan metode *K-Means Clustering* dan klasifikasi dengan metode *Naïve Bayes Classifier* . Kemudian untuk mengetahui kelayakan atau kualitas hasil klasterisasi digunakan evaluasi *silhouette coefficient* dan klasifikasi dengan evaluasi *confusion matrix.*

3.8 Visualisasi

Analisis data relatif sulit dilakukan hanya dengan mengandalkan nilai-nilai tendensi sentral maupun sebaran data. Bahkan grafik statistik juga tidak cukup dengan menjelaskan karakteristik data. Oleh karena itu, diperlukan teknik khusus untuk memvisualisasikan data. Visualisasi data memiliki dua kegunaan, yaitu : memudahkan analisis bagi perancang dan memudahkan pembacaan laporan bagi pengguna (*user*). Bahkan saat ini visualisasi data sudah menjadi sebuah seni indah dan bernilai tinggi (Suyanto, 2017). Dalam penelitian ini visualisasi data yang diguanakan adalah *topic modelling* dan *wordcloud* yaitu untuk mengetahui kata-kata yang paling sering dibicarakan pada dokumen.

3.9 Kebutuhan Sistem

Penelitian ini membutuhkan perangkat keras dan perangkat lunak, sebagai berikut :

a. Perangkat Lunak

Perangkat lunak adalah aplikasi atau program yang diguanakan untuk pembuatan sistem. Perangkat lunak yang dibutuhkan adalah sebagai berikut :

1. Windows 10
2. Python 3.7
3. Anaconda

b. Perangkat Keras

Perangkat keras adalah *device* yang digunakan untuk menunjang pembuatan sistem. Perangkat keras yang digunakan yaitu laptop dengan spesifikasi sebagai berikut:

1. SSD 512 GB
2. RAM 8 GB
3. Processor Intel Core i7

BAB IV

HASIL DAN PEMBAHASAN

4.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah data abstrak bahasa Indonesia atau intisari tugas akhir Teknik Informatika UIN Sunan Kalijaga Yogyakarta dari tahun 2010 sampai dengan 2018 sebanyak 493. Data tersebut diambil dari laman *Digital Library* UIN Sunan Kalijaga berbentuk dokumen.

Dokumen tersebut kemudian diseleksi dengan mengambil abstrak bahasa Indonesia atau intisari tanpa menggunakan kata kunci. Intisari tersebut kemudian dipindah ke dalam *file* excel dan dirubah kedalam *file* csv. Contoh data intisari disajikan dalam Tabel 4.1 dan 4.2.

Tabel 4. 1 Contoh Intisari Tugas Akhir

No	Intisari
1.	Akademi Angkatan Udara Yogyakarta merupakan salah satu lembaga pendidikan militer bagi para Karbol yang bertujuan untuk membentuk seorang Perwira serta mampu mengembangkan pribadi sebagai kader pemimpin bangsa atau TNI AU. Karbol merupakan sebutan bagi Taruna di AAU. Perancangan sistem informasi penilaian untuk mengetahui perkembangan nilai non akademis Karbol dalam bidang kepribadian perlu dilakukan. Pengembangan sistem dalam penelitian ini menggunakan metode model spiral. Metode tersebut terdiri dari tahap komunikasi pelanggan, perencanaan, analisi resiko, perekayasaan, konstruksi dan peluncuran, evaluasi pelanggan. Model spiral adalah salah satu bentuk evolusi yang dimiliki oleh model prototyping dan digabungkan dengan model waterfall. Penelitian ini menghasilkan Sistem Informasi dan monitoring perkembangan nilai Karbol dalam bidang kepribadian yang dapat digunakan sebagai paparan dan bahan sidang dewan akademis Pengasuh Karbol di lingkungan Akademi Angkatan Udara.
2	Akhir-akhir ini pengolahan citra digital di banyak negara maju menjadi bidang yang digeluti oleh banyak peneliti karena menarik untuk

Tabel 4. 2 Contoh Intisari Tugas Akhir (Lanjutan)

No	Intisari
	diterapkan pada berbagai kegiatan, baik kegiatan analisis maupun produksi. Salah satu cabang dalam dari citra digital adalah pengenalan pola. Penelitian ini menggunakan Tesseract sebagai alat untuk mengenali pola dari huruf Hiragana.Penelitian ini dilakukan untuk mengetahui seberapa besar Tesseract mampu mengenali sebuah teks jepang dan juga teks tulisan tangan. Penelitian ini menggunakan 1 citra sebagai data latih yang berisi 74 huruf hiragana yang diproses melalui sebuah pelatihan dan menghasilkan data pelatihan untuk masing-masing huruf. Pada penelitian mempunyai beberapa kriteria pengujian berdasarkan ukuran huruf dan juga resolusi untuk mencari hasil terbaik dalam pengenalan pola Sistem ini mampu mengenali 74 Huruf Hiragana dengan memakai Tesseract Engine. Sistem pengenalan pola ini juga mampu melakukan pelatihan data menggunakan Tesseract Engine. Sistem juga dapat mengenali citra dengan prosentase keberhasilan terbaik 98,24 % dengan resolusi gambar 200dpi dan ukuran huruf 18 Sistem ini juga bisa mengenali citra tulisan tangan dengan prosentase keberhasilan terbaik 90 % dengan resolusi gambar 200dpi.
3	Akhir-akhir ini, banyak muncul perangkat lunak permainan (game) komputer yang menyediakan fasilitas untuk dapat bermain dalam suatu jaringan komputer. Fasilitas ini memungkinkan permainan dapat dimainkan oleh beberapa orang sekaligus dengan menggunakan beberapa buah komputer yang terhubung dalam Local Area Network (LAN). Ular Tangga adalah salah satu jenis permainan papan untuk anak-anak yang dimainkan oleh 2 orang atau lebih. Antar pemain akan berusaha menjadi yang pertama sampai di kotak 100 (finish). Oleh karena itu, peneliti ingin merancang perangkat lunak permainan Ular Tangga yang dapat dimainkan multiplayer dalam suatu jaringan komputer. Perangkat lunak yang dikembangkan menggunakan Microsoft Visual Basic 6.0 sebagai bahasa pemrograman, kontrol winsock pada Visual Basic sebagai jembatan komunikasi antar komputer, dan CorelDRAW X4 sebagai desain gambar. Tahap penelitian meliputi analisis kebutuhan, perancangan, implementasi, dan pengujian. Strategi pengujian yang digunakan adalah pengujian alpha dan pengujian beta. Berdasarkan penelitian yang dilakukan diperoleh kesimpulan bahwa telah berhasil dirancang dan diimplementasikan perangkat lunak dari permainan Ular Tangga yang dapat dimainkan multiplayer dengan 2, 3, dan 4 pemain di jaringan komputer, yakni jaringan LAN (Local Area Network). Perangkat lunak ini juga dapat berjalan dengan baik di sistem operasi Windows 7 dan Windows XP.

4.2 Pelabelan Manual

Intisari tugas akhir yang diperoleh dari *Digital Library* UIN Sunan Kalijaga akan masuk pada tahap pelabelan secara manual. Pelabelan secara manual ini digunakan pada proses klasifikasi dokumen. Model klasifikasi merupakan proses pembelajaran yang memerlukan data latih berlabel. Data intisari akan dikelompokan ke dalam 5 kelompok yaitu :

1. *Computer Engineering* (CE)
2. *Computer Science* (CS)
3. *Information System* (IS)
4. *Information Technology* (IT)
5. *Software Engineering* (SE)

Kelompok tersebut didasarkan pada ACM (*Association for Computing Machinery*) *Computing Curricula* 2005. Pada Tabel 4.3, 4.4 dan 4.5 menunjukan contoh intisari dengan pelabelan secara manual.

Tabel 4. 3 Contoh Intisari Pelabelan Manual

No	Intisari	Label
1.	Para petani cabai sekarang mengembangkan sebuah metode bertanam di media yang kecil yaitu polybag. Sistem bercocok tanam ini membutuhkan lahan yang tidak terlalu besar. Merawat tanaman cabai itu mudah – mudah susah terutama saat penyiraman tanaman. Cara penyiraman tanaman cabai itu harus tepat yaitu tepat waktu, suhu, dan kelembaban tanah. Hal ini jarang sekali diperhatikan oleh petani. Padahal dengan cara penyiraman yang tepat hasil panen akan melimpah dan mengurangi kematian pada tanaman cabai. Kelembaban tanah yang terlalu tinggi membuat tanaman cabai akan membusuk dan kelembaban tanah yang rendah akan membuat tanaman cabai layu dan kering. Prototipe penyiraman otomatis pada tanaman cabai berbasis mikrokontroller ATMega16 ini dibuat untuk memudahkan petani cabai pada media polybag. Sehingga mengurangi dampak kematian pada tanaman cabai dan dapat meningkatkan hasil panen tanaman cabai. Prototipe sistem pernyiraman tanaman otomatis ini mengaplikasikan sensor DHT11 (kelembaban) untuk mendeteksi kelembaban tanah dan sensor LM35 (suhu) untuk mendeteksi suhu disekitar tanaman cabai. Prototipe sistem dilengkapi dengan pompa air untuk penyiraman tanaman. Prototipe sitem akan mengirimkan aktifitas yang dilakukan oleh sistem kepada user yang akan dikirmkan melalui layanan SMS.	CE
2.	Khat merupakan salah satu cabang kesenian Islam yang bersumber dari tulisan Arab. Ada beragam jenis khat yang dapat dinikmati keindahan bentuknya. Namun kurangnya pengetahuan orang awam mengenai jenis-jenis khat dan bentuk khat yang terkadang rumit membuat sebagian orang hanya bisa menikmati keindahan khat dari bentuknya tanpa dapat membaca atau mengetahui kalimat apayang terkandung didalam khat tersebut. Salah satu cara yang dapat digunakan untuk membantu orang awam dalam mengenali jenis khat adalah dengan computer vision. Didalam computer vision terdapat tahapan yang berperan penting dalam mengenali karakter huruf. Tahapan tersebut ialah segmentasi citra dengan operasi deteksi tepi. Oleh karena itu penelitian ini mengkaji lebih lanjut mengenai pengunaan operator deteksi tepi dalam proses segmentasi citra khat Arab. Operator yang digunakan ialah operator sobel dan canny. Objek pada penelitian ini adalah citra kaligrafi hasil tulisan tangan yang telah melalui proses digitalisasi menggunakan scanner merk EPSON StylusTX101 dan kemudian disimpan dalam format jpg. Citra khat berjumlah 40 citrayang terdiri dari 8 jenis khat dan masing-masing jenis khat berjumlah 5 buah citra. Tahapan yang dilakukan pada penelitian antara lain preprocessing , processing dan analisis. Sedangkan untuk pengujiannya dilakukan sebanyak dua kali dengan parameter pengujian timing tun, SNR, dan pengamatan visual. Pengujian pertama dilakukan preprocessing sedangkan pada pengujian kedua tidak dilakukan preprocessing. Hal ini dilakukan untuk menguji kinerja operator ketika menggunakan preprocessing dan ketika tanpa preprocessing.	CS

Tabel 4. 4 Contoh Intisari Pelabelan Manual (Lanjutan)

No	Intisari	Label
	Hasil penelitian menunjukkan bahwa pada pengujian timing run sobel non preprocessing memiliki waktu pemrosesan tercepat dengan rata-rata nilai 0.4689 detik/citra. Sobel non preprocessing juga lebih unggul dalam pengujian visual dengan nilai persentase sebesar 65.25%. Sedangkan pada pengujian SNR, canny preprocessing memiliki rata-rata nilai tertinggi yaitu 8.4901 dB. Maka dapat disimpulkan bahwa operator sobel lebih baik pada pengujian timing run dan pengujian visual. Sedangkan canny lebih baik pada pengujian SNR.	
3.	Informasi di dalam kehidupan sehari-hari sudah menjadi kebutuhan pokok. Kebutuhan informasi menjadi masalah ketika kebutuhan tersebut tidak dapat dirumuskan dengan baik sehingga tidak mewakili kebutuhan itu sendiri. Koperasi Pegawai Republik Indonesia merupakan organisasi dengan data manajerial yang kompleks. Data manajerial yang kompleks serta kebutuhan informasi yang cepat menyebabkan Koperasi Pegawai Republik Indonesia membutuhkan sebuah sistem informasi manajemen yang digunakan untuk mengolah data manajerial secara cepat dan akurat. Penelitian ini bertujuan untuk mengembangkan sebuah sistem informasi manajemen untuk Koperasi Pegawai Republik Indonesia dengan Studi Kasus di Koperasi Pegawai Republik Indonesia Garuda Ngrampal. Alur penelitian yang digunakan adalah identifikasi masalah, analisis masalah, analisis kebutuhan,perancangan, implementasi aplikasi, sertapengujian. Pemodelan sistem menggunakan Unified Modeling Language (UML). Bahasa Pemrograman yang digunakan adalah Delphi dengan database PostgreSQL. Penelitian ini menghasilkan sebuah sistem informasi manajemen Koperasi Pegawai Republik Indonesia yang mencakup pengolahan unit usaha simpan pinjam, pengolahan data keanggotaan serta laporannya. Hasil pengujian sistem dengan 8 orang responden, terdiri dari 3 orang pengurus, 2 orang pengawas dan 3 orang anggota Koperasi Pegawai Republik Indonesia Garuda menunjukan bahwa sistem sudah berjalan dengan baik dan memiliki tampilan yang nyaman serta mudah digunakan.	IS
4.	Jejaring sosial memberikan pengaruh luar biasa dalam kehidupan manusia saat ini. Hubungan antara manusia dengan manusia yang lain semakin mudah dilakukan. Informasi tentang seseorang pun dapat dengan mudah diperoleh dengan adanya media jejaring sosial. Hal ini tidak terlepas dari perkembangan Teknologi internet dan komputer yang semakin pesat. Membaca adalah sebuah kebutuhan bagi masyarakat. Bahkan saat ini membaca seakan menjadi gaya hidup baru di masyarakat. Hal ini dapat dilihat dari banyaknya pengunjung toko-toko buku dan pameran-pameran buku yang diadakan secara periodik. Namun, besarnya antusias masyarakat tidak sebanding dengan kemampuan masyarakat untuk membeli buku, terutama bagi pelajar ataupun mahasiswa. Lendabook berupaya mengadopsi beberapa teknologi yang sudah ada untuk memecahka permasalahan di atas.	IT

Tabel 4. 5 Contoh Intisari Pelabelan Manual (Lanjutan)

No	Intisari	Label
	Dengan menggunakan konsep jejaring sosial dan memanfaatkan teknologi internet. Lendabook merupakan situs jejaring sosial yang dibangun dengan tujuan untuk memudahkan masyarakat dalam berbagi dengan cara saling pinjam meminjamkan buku.	
5.	Salah satu kegiatan transaksi akademik yang ada diperguruan tinggi lebih banyak berhubungan dengan administrasi persuratan. UIN Sunan Kalijaga Yogyakarta memiliki sistem informasi surat terpusat yang dapat memudahkanproses pencarian surat dan pengawasan peredaran surat. Untuk mengetahui berapa tingkat kehandalan dalam sistem informasi surat yang dikembangkan oleh UIN Sunan Kalijaga, peneliti melakukan pengujian terhadap sistem tersebut. Penelitian ini di fokuskan pada pengujian sistem yang disusun berdasarkan prosedur pengujian Reliability (kehandalan) dengan menggunakan metode McCall. Pengujian fakto rreliability memiliki tujuan untuk mengetahui seberapa jauh sebuah program atau aplikasi dapat diharapkan melakukan fungsinya dengan baik sehingga mampu memenuhi kebutuhan pengguna. Berdasarkan pengujian diketahui bahwa hasil presentase faktor Reliability pada Sistem Informasi Surat UIN Sunan Kalijaga dengan menggunakan teori McCalls Quality Factors sebesar 85,11%. Nilai tersebut adalah hasil penjumlahan dari keseluruhan presentase subfaktor (Accuracy, Consistency, Error Tolerance, Modularity, Simplicity) yang kemudian dibagi 5. Dari hasil penelitian dapat disimpulkan bahwa Sistem Informasi Surat UIN Sunan KalijagaYogyakarta mempunyai nilai reliabilitas yang baik.	SE

4.3 Preprocessing

Preprocessing ini dilakukan untuk membersihkan data intisari bahasa Indonesia. Berikut ini adalah tahapan dari *preprocessing.*

4.2.1 *Cleansing*

Tahap *clenansing* bertujuan untuk membersihkan intisari terhadap tanda baca . Tabel 4.6 dan 4.7 menunjukan contoh hasil *cleansing* intisari tugas akhir.

Tabel 4. 6 Contoh Hasil *Celansing* Intisari Tugas Akhir

No	Intisari
1	Akademi Angkatan Udara Yogyakarta merupakan salah satu lembaga pendidikan militer bagi para Karbol yang bertujuan untuk membentuk seorang Perwira serta mampu mengembangkan pribadi sebagai kader pemimpin bangsa atau TNI AU Karbol merupakan sebutan bagi Taruna di AAU Perancangan sistem informasi penilaian untuk mengetahui perkembangan nilai non akademis Karbol dalam bidang kepribadian perlu dilakukan Pengembangan sistem dalam penelitian ini menggunakan metode model spiral Metode tersebut terdiri dari tahap komunikasi pelanggan perencanaan analisi resiko perekayasaan konstruksi dan peluncuran evaluasi pelanggan Model spiral adalah salah satu bentuk evolusi yang dimiliki oleh model prototyping dan digabungkan dengan model waterfall Penelitian ini menghasilkan Sistem Informasi dan monitoring perkembangan nilai Karbol dalam bidang kepribadian yang dapat digunakan sebagai paparan dan bahan sidang dewan akademis Pengasuh Karbol di lingkungan Akademi Angkatan Udara
2.	Akhirakhir ini pengolahan citra digital di banyak negara maju menjadi bidang yang digeluti oleh banyak peneliti karena menarik untuk diterapkan pada berbagai kegiatan baik kegiatan analisis maupun produksi Salah satu cabang dalam dari citra digital adalah pengenalan pola Penelitian ini menggunakan Tesseract sebagai alat untuk mengenali pola dari huruf HiraganaPenelitian ini dilakukan

Tabel 4. 7 Contoh Hasil *Celansing* Intisari Tugas Akhir (Lanjutan)

No	Intisari
	untuk mengetahui seberapa besar Tesseract mampu mengenali sebuah teks jepang dan juga teks tulisan tangan Penelitian ini menggunakan 1 citra sebagai data latih yang berisi 74 huruf hiragana yang diproses melalui sebuah pelatihan dan menghasilkan data pelatihan untuk masingmasing huruf Pada penelitian mempunyai beberapa kriteria pengujian berdasarkan ukuran huruf dan juga resolusi untuk mencari hasil terbaik dalam pengenalan pola Sistem ini mampu mengenali 74 Huruf Hiragana dengan memakai Tesseract Engine Sistem pengenalan pola ini juga mampu melakukan pelatihan data menggunakan Tesseract Engine Sistem juga dapat mengenali citra dengan prosentase keberhasilan terbaik 9824 dengan resolusi gambar 200dpi dan ukuran huruf 18 Sistem ini juga bisa mengenali citra tulisan tangan dengan prosentase keberhasilan terbaik 90 dengan resolusi gambar 200dpi
3.	Akhirakhir ini banyak muncul perangkat lunak permainan game komputer yang menyediakan fasilitas untuk dapat bermain dalam suatu jaringan komputer Fasilitas ini memungkinkan permainan dapat dimainkan oleh beberapa orang sekaligus dengan menggunakan beberapa buah komputer yang terhubung dalam Local Area Network LAN Ular Tangga adalah salah satu jenis permainan papan untuk anakanak yang dimainkan oleh 2 orang atau lebih Antar pemain akan berusaha menjadi yang pertama sampai di kotak 100 finish Oleh karena itu peneliti ingin merancang perangkat lunak permainan Ular Tangga yang dapat dimainkan multiplayer dalam suatu jaringan komputer Perangkat lunak yang dikembangkan menggunakan Microsoft Visual Basic 60 sebagai bahasa pemrograman kontrol winsock pada Visual Basic sebagai jembatan komunikasi antar komputer dan CorelDRAW X4 sebagai desain gambar Tahap penelitian meliputi analisis kebutuhan perancangan implementasi dan pengujian Strategi pengujian yang digunakan adalah pengujian alpha dan pengujian beta Berdasarkan penelitian yang dilakukan diperoleh kesimpulan bahwa telah berhasil dirancang dan diimplementasikan perangkat lunak dari permainan Ular Tangga yang dapat dimainkan multiplayer dengan 2 3 dan 4 pemain di jaringan komputer yakni jaringan LAN Local Area Network Perangkat lunak ini juga dapat berjalan dengan baik di sistem operasi Windows 7 dan Windows XP

Berikut *source code* yang digunakan dalam tahan *cleansing*:

```
import string
import re
cleantext=[]
for i in intisari:
    clean=re.sub("["+string.punctuation+"]", "",i)
    cleantext.append(clean)
```

4.2.2 *Casefolding*

Casefolding merupakan tahap untuk merubah huruf pada intisari tugas akhir menjadi huruf kecil. Tahap ini diperlukan karena besar kecilnya huruf akan berdampak pada perhitungan tf-idf. Tabel 4.8 dan 4.9 menunjukan contoh hasil proses *casefolding* pada intisari tugas akhir.

Tabel 4. 8 Contoh Hasil *Casefolding* Intisari Tugas Akhir

No	Intisari
1.	akademi angkatan udara yogyakarta merupakan salah satu lembaga pendidikan militer bagi para karbol yang bertujuan untuk membentuk seorang perwira serta mampu mengembangkan pribadi sebagai kader pemimpin bangsa atau tni au karbol merupakan sebutan bagi taruna di aau perancangan sistem informasi penilaian untuk mengetahui perkembangan nilai non akademis karbol dalam bidang kepribadian perlu dilakukan pengembangan sistem dalam penelitian ini menggunakan metode model spiral metode tersebut terdiri dari tahap komunikasi pelanggan perencanaan analisi resiko perekayasaan konstruksi dan peluncuran evaluasi pelanggan model spiral adalah salah satu bentuk evolusi yang dimiliki oleh model prototyping dan digabungkan dengan model waterfall penelitian ini menghasilkan sistem informasi dan monitoring perkembangan nilai karbol dalam bidang kepribadian yang dapat digunakan sebagai paparan dan bahan sidang dewan akademis pengasuh karbol di lingkungan akademi angkatan udara
2.	akhirakhir ini pengolahan citra digital di banyak negara maju menjadi bidang yang digeluti oleh banyak peneliti karena menarik untuk diterapkan pada berbagai kegiatan baik kegiatan analisis maupun produksi salah satu cabang dalam dari citra digital adalah

Tabel 4. 9 Contoh Hasil *Casefolding* Intisari Tugas Akhir (Lanjutan)

No	Intisari
	pengenalan pola penelitian ini menggunakan tesseract sebagai alat untuk mengenali pola dari huruf hiragana penelitian ini dilakukan untuk mengetahui seberapa besar tesseract mampu mengenali sebuah teks jepang dan juga teks tulisan tangan penelitian ini menggunakan 1 citra sebagai data latih yang berisi 74 huruf hiragana yang diproses melalui sebuah pelatihan dan menghasilkan data pelatihan untuk masingmasing huruf pada penelitian mempunyai beberapa kriteria pengujian berdasarkan ukuran huruf dan juga resolusi untuk mencari hasil terbaik dalam pengenalan pola sistem ini mampu mengenali 74 huruf hiragana dengan memakai tesseract engine sistem pengenalan pola ini juga mampu melakukan pelatihan data menggunakan tesseract engine sistem juga dapat mengenali citra dengan prosentase keberhasilan terbaik 9824 dengan resolusi gambar 200dpi dan ukuran huruf 18 sistem ini juga bisa mengenali citra tulisan tangan dengan prosentase keberhasilan terbaik 90 dengan resolusi gambar 200dpi
3.	akhirakhir ini banyak muncul perangkat lunak permainan game komputer yang menyediakan fasilitas untuk dapat bermain dalam suatu jaringan komputer fasilitas ini memungkinkan permainan dapat dimainkan oleh beberapa orang sekaligus dengan menggunakan beberapa buah komputer yang terhubung dalam local area network lan ular tangga adalah salah satu jenis permainan papan untuk anakanak yang dimainkan oleh 2 orang atau lebih antar pemain akan berusaha menjadi yang pertama sampai di kotak 100 finish oleh karena itu peneliti ingin merancang perangkat lunak permainan ular tangga yang dapat dimainkan multiplayer dalam suatu jaringan komputer perangkat lunak yang dikembangkan menggunakan microsoft visual basic 60 sebagai bahasa pemrograman kontrol winsock pada visual basic sebagai jembatan komunikasi antar komputer dan coreldraw x4 sebagai desain gambar tahap penelitian meliputi analisis kebutuhan perancangan implementasi dan pengujian strategi pengujian yang digunakan adalah pengujian alpha dan pengujian beta berdasarkan penelitian yang dilakukan diperoleh kesimpulan bahwa telah berhasil dirancang dan diimplementasikan perangkat lunak dari permainan ular tangga yang dapat dimainkan multiplayer dengan 2 3 dan 4 pemain di jaringan komputer yakni jaringan lan local area network perangkat lunak ini juga dapat berjalan dengan baik di sistem operasi windows 7 dan windows xp

Berikut *source code* yang digunakan dalam tahap *casefolding*:

```
loweri=[]
for i in cleantext:
    loweri.append(i.lower())
```

4.2.3 *Tokenizing*

Tokenizing atau tokenisasi adalah tahap untuk membagi text, kalimat atau paragraf intisari tugas akhir menjadi kata-kata. Tabel 4.10, 4.11 dan 4.12 menunjukan contoh hasil penerapan proses *tokenizing* pada intisari tugas akhir.

Tabel 4. 10 Contoh Hasil *Tokenizing* Intisari Tugas Akhir

No	Intisari
1.	'akademi', 'angkatan', 'udara', 'yogyakarta', 'merupakan', 'salah', 'satu', 'lembaga', 'pendidikan', 'militer', 'bagi', 'para', 'karbol', 'yang', 'bertujuan', 'untuk', 'membentuk', 'seorang', 'perwira', 'serta', 'mampu', 'mengembangkan', 'pribadi', 'sebagai', 'kader', 'pemimpin', 'bangsa', 'atau', 'tni', 'au', 'karbol', 'merupakan', 'sebutan', 'bagi', 'taruna', 'di', 'aau', 'perancangan', 'sistem', 'informasi', 'penilaian', 'untuk', 'mengetahui', 'perkembangan', 'nilai', 'non', 'akademis', 'karbol', 'dalam', 'bidang', 'kepribadian', 'perlu', 'dilakukan', 'pengembangan', 'sistem', 'dalam', 'penelitian', 'ini', 'menggunakan', 'metode', 'model', 'spiral', 'metode', 'tersebut', 'terdiri', 'dari', 'tahap', 'komunikasi', 'pelanggan', 'perencanaan', 'analisi', 'resiko', 'perekayasaan', 'konstruksi', 'dan', 'peluncuran', 'evaluasi', 'pelanggan', 'model', 'spiral', 'adalah', 'salah', 'satu', 'bentuk', 'evolusi', 'yang', 'dimiliki', 'oleh', 'model', 'prototyping', 'dan', 'digabungkan', 'dengan', 'model', 'waterfall', 'penelitian', 'ini', 'menghasilkan', 'sistem', 'informasi', 'dan', 'monitoring', 'perkembangan', 'nilai', 'karbol', 'dalam', 'bidang', 'kepribadian', 'yang', 'dapat', 'digunakan', 'sebagai', 'paparan', 'dan', 'bahan', 'sidang', 'dewan', 'akademis', 'pengasuh', 'karbol', 'di', 'lingkungan', 'akademi', 'angkatan', 'udara'
2.	'akhirakhir', 'ini', 'pengolahan', 'citra', 'digital', 'di', 'banyak', 'negara', 'maju', 'menjadi', 'bidang', 'yang', 'digeluti', 'oleh', 'banyak', 'peneliti', 'karena', 'menarik', 'untuk', 'diterapkan', 'pada', 'berbagai', 'kegiatan', 'baik', 'kegiatan', 'analisis', 'maupun', 'produksi', 'salah', 'satu',

Tabel 4. 11 Contoh Hasil *Tokenizing* Intisari Tugas Akhir (Lanjutan)

No	Intisari
	'cabang', 'dalam', 'dari', 'citra', 'digital', 'adalah', 'pengenalan', 'pola', 'penelitian', 'ini', 'menggunakan', 'tesseract', 'sebagai', 'alat', 'untuk', 'mengenali', 'pola', 'dari', 'huruf', 'hiraganapenelitian', 'ini', 'dilakukan', 'untuk', 'mengetahui', 'seberapa', 'besar', 'tesseract', 'mampu', 'mengenali', 'sebuah', 'teks', 'jepang', 'dan', 'juga', 'teks', 'tulisan', 'tangan', 'penelitian', 'ini', 'menggunakan', '1', 'citra', 'sebagai', 'data', 'latih', 'yang', 'berisi', '74', 'huruf', 'hiragana', 'yang', 'diproses', 'melalui', 'sebuah', 'pelatihan', 'dan', 'menghasilkan', 'data', 'pelatihan', 'untuk', 'masingmasing', 'huruf', 'pada', 'penelitian', 'mempunyai', 'beberapa', 'kriteria', 'pengujian', 'berdasarkan', 'ukuran', 'huruf', 'dan', 'juga', 'resolusi', 'untuk', 'mencari', 'hasil', 'terbaik', 'dalam', 'pengenalan', 'pola', 'sistem', 'ini', 'mampu', 'mengenali', '74', 'huruf', 'hiragana', 'dengan', 'memakai', 'tesseract', 'engine', 'sistem', 'pengenalan', 'pola', 'ini', 'juga', 'mampu', 'melakukan', 'pelatihan', 'data', 'menggunakan', 'tesseract', 'engine', 'sistem', 'juga', 'dapat', 'mengenali', 'citra', 'dengan', 'prosentase', 'keberhasilan', 'terbaik', '9824', 'dengan', 'resolusi', 'gambar', '200dpi', 'dan', 'ukuran', 'huruf', '18', 'sistem', 'ini', 'juga', 'bisa', 'mengenali', 'citra', 'tulisan', 'tangan', 'dengan', 'prosentase', 'keberhasilan', 'terbaik', '90', 'dengan', 'resolusi', 'gambar', '200dpi'
3.	'akhirakhir', 'ini', 'banyak', 'muncul', 'perangkat', 'lunak', 'permainan', 'game', 'komputer', 'yang', 'menyediakan', 'fasilitas', 'untuk', 'dapat', 'bermain', 'dalam', 'suatu', 'jaringan', 'komputer', 'fasilitas', 'ini', 'memungkinkan', 'permainan', 'dapat', 'dimainkan', 'oleh', 'beberapa', 'orang', 'sekaligus', 'dengan', 'menggunakan', 'beberapa', 'buah', 'komputer', 'yang', 'terhubung', 'dalam', 'local', 'area', 'network', 'lan', 'ular', 'tangga', 'adalah', 'salah', 'satu', 'jenis', 'permainan', 'papan', 'untuk', 'anakanak', 'yang', 'dimainkan', 'oleh', '2', 'orang', 'atau', 'lebih', 'antar', 'pemain', 'akan', 'berusaha', 'menjadi', 'yang', 'pertama', 'sampai', 'di', 'kotak', '100', 'finish', 'oleh', 'karena', 'itu', 'peneliti', 'ingin', 'merancang', 'perangkat', 'lunak', 'permainan', 'ular', 'tangga', 'yang', 'dapat', 'dimainkan', 'multiplayer', 'dalam', 'suatu', 'jaringan', 'komputer', 'perangkat', 'lunak', 'yang', 'dikembangkan', 'menggunakan', 'microsoft', 'visual', 'basic', '60', 'sebagai', 'bahasa', 'pemrograman', 'kontrol', 'winsock', 'pada', 'visual', 'basic', 'sebagai', 'jembatan', 'komunikasi', 'antar', 'komputer', 'dan', 'coreldraw', 'x4', 'sebagai', 'desain', 'gambar', 'tahap', 'penelitian', 'meliputi', 'analisis', 'kebutuhan', 'perancangan', 'implementasi', 'dan', 'pengujian', 'strategi', 'pengujian', 'yang', 'digunakan', 'adalah', 'pengujian', 'alpha', 'dan', 'pengujian', 'beta', 'berdasarkan', 'penelitian', 'yang', 'dilakukan', 'diperoleh', 'kesimpulan', 'bahwa', 'telah', 'berhasil', 'dirancang', 'dan', 'diimplementasikan', 'perangkat', 'lunak', 'dari',

Tabel 4. 12 Contoh Hasil *Tokenizing* Intisari Tugas Akhir (Lanjutan)

No	Intisari
	'permainan', 'ular', 'tangga', 'yang', 'dapat', 'dimainkan', 'multiplayer', 'dengan', '2', '3', 'dan', '4', 'pemain', 'di', 'jaringan', 'komputer', 'yakni', 'jaringan', 'lan', 'local', 'area', 'network', 'perangkat', 'lunak', 'ini', 'juga', 'dapat', 'berjalan', 'dengan', 'baik', 'di', 'sistem', 'operasi', 'windows', '7', 'dan', 'windows', 'xp'

Berikut *source code* yang digunakan dalam tahap *tokenizing*:

```
split=[]
for i in loweri :
    split.append(i.split())
```

4.2.1 *Stopword Removal / Filtering*

Stopword Removal merupakan proses menghilangkan kata-kata yang sering muncul atau yang dianggap tidak memiliki makna. Tabel 4.13 dan 4.14 adalah contoh penerapan proses *stopword removal.*

Tabel 4. 13 Contoh Hasil *Stopword Removal* Intisari Tugas Akhir

No	Intisari
1.	akademi angkatan udara yogyakarta merupakan salahsatu lembaga pendidikan militer para karbol bertujuan membentuk seorang perwira mampu mengembangkan pribadi kader pemimpin bangsa tni au karbol merupakan sebutan taruna aau perancangan sistem informasi penilaian mengetahui perkembangan nilai non akademis karbol bidang kepribadian perlu dilakukan pengembangan sistem penelitian menggunakan metode model spiral metode tersebut terdiri tahap komunikasi pelanggan perencanaan analisi resiko perekayasaan konstruksi peluncuran evaluasi pelanggan model spiral salah satu bentuk evolusi dimiliki model prototyping digabungkan model waterfall penelitian menghasilkan sistem informasi monitoring perkembangan nilai karbol bidang kepribadian dapat digunakan paparan bahan sidang

Tabel 4. 14 Contoh Hasil *Stopword Removal* Intisari Tugas Akhir (Lanjutan)

No	Intisari
	dewan akademis pengasuh karbol lingkungan akademi angkatan udara
2.	akhirakhir pengolahan citra digital banyak negara maju menjadi bidang digeluti banyak peneliti menarik diterapkan berbagai kegiatan baik kegiatan analisis maupun produksi salah satu cabang citra digital pengenalan pola penelitian menggunakan tesseract alat mengenali pola dari huruf hiraganapenelitian dilakukan mengetahui seberapa besar tesseract mampu mengenali sebuah teks jepang teks tulisan tangan penelitian menggunakan 1 citra data latih berisi 74 huruf hiragana diproses melalui sebuah pelatihan menghasilkan data pelatihan masingmasing huruf penelitian mempunyai beberapa kriteria pengujian berdasarkan ukuran huruf juga resolusi mencari hasil terbaik pengenalan pola sistem mampu mengenali 74 huruf hiragana memakai tesseract engine sistem pengenalan pola juga mampu melakukan pelatihan data menggunakan tesseract engine sistem juga dapat mengenali citra prosentase keberhasilan terbaik 9824 resolusi gambar 200dpi ukuran huruf 18 sistem juga mengenali citra tulisan tangan prosentase keberhasilan terbaik 90 resolusi gambar 200dpi
3.	akhirakhir banyak muncul perangkat lunak permainan game komputer menyediakan fasilitas bermain suatu jaringan komputer fasilitas memungkinkan permainan dimainkan beberapa orang sekaligus menggunakan beberapa buah komputer terhubung local area network lan ular tangga salah satu jenis permainan papan anakanak dimainkan 2 orang lebih antar pemain berusaha menjadi pertama kotak 100 finish karena peneliti merancang perangkat lunak permainan ular tangga dapat dimainkan multiplayer suatu jaringan komputer perangkat lunak dikembangkan menggunakan microsoft visual basic 60 bahasa pemrograman kontrol winsock visual basic jembatan komunikasi antar komputer coreldraw x4 desain gambar tahap penelitian meliputi analisis kebutuhan perancangan implementasi pengujian strategi pengujian digunakan pengujian alpha pengujian beta berdasarkan penelitian dilakukan diperoleh kesimpulan telah berhasil dirancang diimplementasikan perangkat lunak permainan ular tangga dapat dimainkan multiplayer 2 3 4 pemain jaringan komputer jaringan lan local area network perangkat lunak juga dapat berjalan baik di sistem operasi windows 7 windows xp

Berikut *source code* yang digunakan dalam tahap *stopword removal*:

```
From Sastrawi.StopWordRemover.StopWordRemoverFactory import
StopWordRemoverFactory
factory=StopWordRemoverFactory()
txtrmvstop=[]
for i in loweri :
    stopword=factory.create_stop_word_remover()
    txtrmvstop.append(stopword.remove(i))
```

4.4 Pembobotan TF-IDF

Untuk menjadikan kata di dalam setiap dokumen menjadi sebuah vektor maka perlu adanya pembobotan. Pembobotan yang digunakan adalah TF-IDF. TF atau Term Frequency menyatakan jumlah berapa banyak keberadaan suatu term/kata pada satu dokumen kemudian DF (*Document Frequency*) menyatakan banyak dokumen yang mengandung kata/term sedangkan IDF atau *Inverse Document Frequency* menyatakan seberapa penting kata/term tersebut. Untuk rumus TF-IDF telah dijelaskan pada landasan teori 2.2.3.

Contoh intisari yang dihitung manual dengan menggunakan pembobotan TF-IDF. Contoh intisari akan dilakuan pembobotan TF-IDF dapat dilihat pada Tabel 4.15.

Tabel 4. 15 Contoh Kalimat di dalam Intisari untuk Perhitungan TF-IDF

No	Intisari
1,	yogyakarta merupakan kota tujuan wisata indonesia mempunyai keberagaman kuliner selalu berkembang berkesinambungan banyak tempat kuliner yogyakarta membuat tempat kuliner tersebut tentu diketahui orang banyak terutama masyarakat tinggal yogyakarta
2.	yogyakarta merupakan salah satu pusat wisata budaya banyak wisatawan baik wisatawan negeri maupun mancanegara datang kota yogyakarta
3.	yogyakarta salah satu tujuan wisata indonesia memiliki jumlah hotel terus bertambah

Contoh perhitungan manual TF-IDF dapat dilihat pada Tabel 4.16 dan 4.17.

Tabel 4. 16 Contoh Hasil Perhitungan TF-IDF

Term	TF				IDF	TF * IDF		
	D1	D2	D3	DF		D1	D2	D3
yogyakarta	3	2	1	6	0,301029996	0,903089987	0,602059991	0,301029996
merupakan	1	1	0	2	0,176091259	0,176091259	0,176091259	0
kota	1	1	0	2	0,176091259	0,176091259	0,176091259	0
tujuan	1	0	1	2	0,176091259	0,176091259	0	0,176091259
wisata	1	0	1	2	0,176091259	0,176091259	0	0,176091259
indonesia	1	0	1	2	0,176091259	0,176091259	0	0,176091259
mempunyai	1	0	0	1	0,477121255	0,477121255	0	0
keberagaman	1	0	0	1	0,477121255	0,477121255	0	0
kuliner	3	0	0	3	0	0	0	0
selalu	1	0	0	1	0,477121255	0,477121255	0	0
berkembang	1	0	0	1	0,477121255	0,477121255	0	0
berkesinambungan	1	0	0	1	0,477121255	0,477121255	0	0
banyak	2	0	0	2	0,176091259	0,352182518	0	0
tempat	2	0	0	2	0,176091259	0,352182518	0	0
membuat	1	0	0	1	0,477121255	0,477121255	0	0
tersebut	1	0	0	1	0,477121255	0,477121255	0	0
tentu	1	0	0	1	0,477121255	0,477121255	0	0
diketahui	1	0	0	1	0,477121255	0,477121255	0	0
orang	1	0	0	1	0,477121255	0,477121255	0	0
terutama	1	0	0	1	0,477121255	0,477121255	0	0
masyarakat	1	0	0	1	0,477121255	0,477121255	0	0
tinggal	1	0	0	1	0,477121255	0,477121255	0	0

Tabel 4. 17 Contoh Hasil Perhitungan TF-IDF (Lanjutan)

Term	TF				IDF	TF * IDF		
	D1	D2	D3	DF		D1	D2	D3
satu	0	1	0	1	0,477121255	0	0,477121255	0
pusat	0	1	0	1	0,477121255	0	0,477121255	0
budaya	0	1	0	1	0,477121255	0	0,477121255	0
wisatawan	0	1	0	1	0,477121255	0	0,477121255	0
baik	0	1	0	1	0,477121255	0	0,477121255	0
negeri	0	1	0	1	0,477121255	0	0,477121255	0
maupun	0	1	0	1	0,477121255	0	0,477121255	0
mancanegara	0	1	0	1	0,477121255	0	0,477121255	0
datang	0	1	0	1	0,477121255	0	0,477121255	0
memiliki	0	0	1	1	0,477121255	0	0	0,477121255
jumlah	0	0	1	1	0,477121255	0	0	0,477121255
hotel	0	0	1	1	0,477121255	0	0	0,477121255
terus	0	0	1	1	0,477121255	0	0	0,477121255
bertambah	0	0	1	1	0,477121255	0	0	0,477121255
salah	0	1	0	1	0,477121255	0	0,477121255	0

Berikut *source code* perhitungan TF-IDF dengan menggunakan Python :

```
from sklearn.feature_extraction.text import
TfidfVectorizer
vectorizer = TfidfVectorizer()
tfidf = vectorizer.fit_transform(documents)
pd.DataFrame(tfidf.toarray(),
columns=vectorizer.get_feature_names())
```

4.5 Klasterisasi Metode *K-Means Clustering*

Setelah dilakuakan perhitungan bobot dan jarak maka akan masuk ke dalam proses klasterisasi dengan menggunakan *K-Means Clustering* . *Library* yang digunakan pada *python* adalah *sklearn*. Data yang digunakan pada model klasterisasi hanya mengguanakan data intisari tanpa label. . Dalam proses klasterisasi ini dilakukan eksperiman dengan menentukan jumlah *k* (*klaster*) yaitu 2,4,5,6, dan 7. Eksperimen ini dilakukan untuk mengetahui jumlah klaster terbaik berdasarkan jarak *Cosine Similarity.*

Berikut *source code* untuk menghitung *cosine similarity*:

```
from sklearn.metrics.pairwise import linear_kernel
cosine_similarities=linear_kernel(tfidf[0:1], tfidf).flatten()
cosine_similarities
```

Berikut *source code* untuk klasterisasi menggunakan *K-Menas Clustering*:

```
from sklearn.cluster import KMeans
number_of_clusters= n #n adalah jumlah klaster yang diinginkan
km = KMeans n_clusters=number_of_clusters, random_state=42)
km.fit(dist)
km.fit
print("Top terms per cluster:")
order_centroids = km.cluster_centers_.argsort()[:, ::-1]
terms = vectorizer.get_feature_names()
for i in range(number_of_clusters):
    top_ten_words = [terms[ind] for ind in order_centroids[i, :10]]
    print("Cluster {}: {}".format(i, ' '.join(top_ten_words)))
```

4.6 Klasifikasi Metode *Naïve Bayes Classifier*

Model klasifikasi *Naïve Bayes* dilakukan setelah melalui proses pelabelan manual, *preprocessing* dan pembobotan tf-idf. Sebelum masuk pada tahap klasifikasi maka ditentukan terlebih dahulu pembagian terhadap data uji dan data latih. Pada eksperimen yang dilakukan pembagian data latih dan data uji menggunakan dengan perbandingan 7 : 3 , 8 : 2 dan 9 : 1. Hal ini bertujuan untuk mengetahui ukuran performasi atau kulaitas model klasifikasi *Naïve Bayes* berdasarkan pembagian data tersebut.

Berikut *source code* klasifikasi dengan menggunakan metode *Naïve Bayes Clasifier:*

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.model_selection import train_test_split

Pintisari =dataset['Pintisari'].values
Plabel =dataset['Plabel'].values
#pembobotan tf-idf
vectorizer = TfidfVectorizer(smooth_idf=False, norm='l1')
X = vectorizer.fit_transform(Pintisari).toarray()
#test_split dapat dirubah sesuai keinginan
x_train, x_test, y_train, y_test = train_test_split(Pintisari,
Plabel, test_size = 0.2, random_state=42)
vect = CountVectorizer()
vect.fit(x_train)
x_train_dtm = vect.transform(x_train)
x_test_dtm = vect.transform(x_test)
      nb = MultinomialNB()
      nb.fit(x_train_dtm,y_train)
y_pred_class = nb.predict(x_test_dtm)
```

4.7 Analisis dan Evaluasi

4.7.1 Hasil Evaluasi *Silhouette Coefficient*

Klasterisasi yang telah dilakukan menggunakan *K-Means Clustering* kemudian di evaluasi dengan menggunakan metode *Silhouette Coefficient.* Metode tersebut bertujuan untuk mengukur kualitas klasterisasi dengan mencari rata-rata *Silhouette Coefficient* dari semua objek dalam himpunan data.

Berikut *source code* untuk menghitung rata-rata *silhouette coefficient:*

```
from sklearn.metrics import silhouette_score
silhouette_avg = silhouette_score(dist,km.labels_)
```

Ekperimen yang telah dilakukan dengan meggunkaan jumlah *k* 2,3,4,5,6 dan 7 menghasilkan rata-rata *silhouette* sebagai berikut :

k = 2 menghasilkan rata-rata *silhouette sebesar* 0.0594565792967913

k = 3 menghasilkan rata-rata *silhouette sebesar* 0.0663662495824928

k = 4 menghasilkan rata-rata *silhouette sebesar* 0.0539393366402383

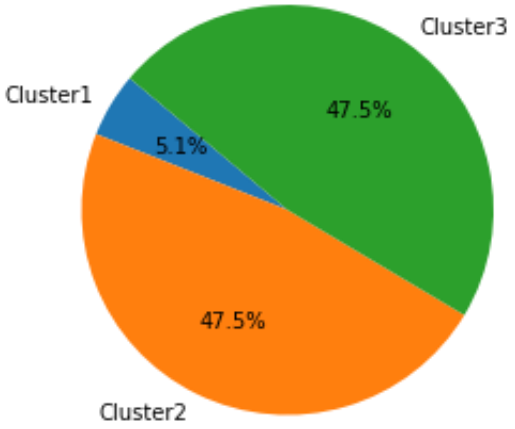
k = 5 menghasilkan rata-rata *silhouette sebesar* 0.0588686833775039

k = 6 menghasilkan rata-rata *silhouette sebesar* 0.0487474557823854

k = 7 menghasilkan rata-rata *silhouette sebesar* 0.0619909881428413

Berdasarkan eksperimen yang telah dilakukan menujukan bahwa jumlah k = 3 paling mendekati nilai 1 atau bisa dikatakan bahwa pengelompokan dengan menggunakan jumlah klaster tersebut sangat padat

dan baik. Hasil prosentase yang dihasilkan dengan k = 3 disajikan dalam bentuk *piechart* dapat dilihat pada Gambar 4.1.



Gambar 4. 1 Prosentase dengan Jumlah k=3

Berdasarkan Gambar 4.1 prosentase *cluster* 1 sebesar 5,1%, *cluster* 2 sebesar 47,5% dan *cluster* 3 sebesar 47,5%. Dari ketiga *cluster* tersebut, *cluster* 2 dan *cluster* 3 mengasilkan prosentase yang sama.

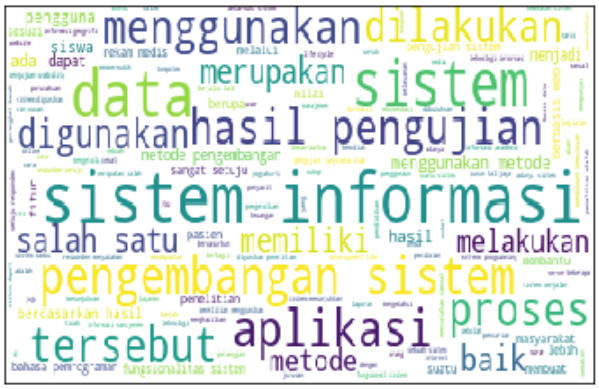
Berikut *source code* untuk menampilakan prosentase jumlah k=3 dalam bentuk *piechart* :

```
import matplotlib.pyplot as plt
def PieChart(score,labels):
    fig1=plt.figure(); fig1.add_subplot(111)

plt.pie(score,labels=labels,autopct='%1.1f%%',startangle=1
40)
    plt.axis('equal');plt.show()
    return None

for i in kelas:
    score_se=[len([True for t in kelas if t==0]),len([True
for t in kelas if t==2]), len([True for t in kelas if
t==2])]
    Label_se=['Cluster1','Cluster2','Cluster3']

PieChart(score_se,Label_se)
```

Kemudian untuk mengetahui gambaran/ isi dari setiap *cluster*, maka dilakukan visualisasi data. Visualisasi tersebut menggunakan visualisasi *worcloud*. *Wordcloud* pada setiap *cluster* dapat dilihat pada Gambar 4.2, 4.3 dan 4.4.

Gambar 4. 2 *Wordcloud Cluster* 1 dengan k=3

Berdasarkan visualisasi *wordcloud cluster* 1 pada Gambar 4.2, 5 kata/term teratas yang sering muncul pada setiap dokumen adalah kata perangkat, lunak, faktor, sistem dan pengujian. Sehingga dapat diberikan kesimpulkan bahwa isi dari *cluster* tersebut adalah mengenai pengujian perangkat lunak.



Gambar 4. 3 *Wordcloud Cluster* 2 dengan k=3

Wordcloud cluster 2 pada pada Gambar 4.3 memiliki 5 kata/term teratas yaitu sistem, informasi, pengembangan, aplikasi dan pengujian. Sehingga dapat diketahui bahwa *cluster* 2 berisi dokumen-dokumen tentang pengembangan sistem informasi.

Gambar 4. 4 *Wordcloud Cluster* 3 dengan k=3

Gambar 4.4 menggambarkan *wordcloud* dari cluster 3 dengan 5 kata teratas adalah sistem, data, aplikasi, metode, dan nilai. Pada *cluster* ini dapat diberikan kesimpulan bahwa rata-rata dokumen tersebut berisikan mengenai hal-hal yang berkaitan dengan data, aplikasi dan sistem.

Selain menggunakan visualisasi *wordcloud* untuk mengetahui isi dari setiap *cluster*, penelitian ini juga menggunakan *topic modelling*. *Topic Modelling* bertujuan untuk menentukan topik pada setiap *cluster* secara otomatis. Pada Tabel 4.18 dan 4.19 menunjukan hasil *topic modelling* dengan jumlah *cluster* 3.

Tabel 4. 18 Hasil *Topic Modelling* dengan k= 3

Cluster	Topik
1	Topik 1: sistem informasi dapat aplikasi medis rekam penelitian berbasis Topik 2: sistem pengujian hasil setuju responden berdasarkan sangat menyatakan Topik 3: sistem metode pengembangan informasi pengujian menggunakan penelitian digunakan Topik 4: sistem informasi pengujian satu merupakan dilakukan penelitian salah Topik 5: sistem informasi data pengembangan metode aplikasi menggunakan penelitian
2	Topik 1: metode menggunakan penelitian sistem informasi digunakan data jaringan Topik 2: sistem hasil data penelitian metode menggunakan dilakukan nilai Topik 3: sistem menggunakan hasil metode algoritma data jaringan aplikasi Topik 4: sistem data menggunakan informasi metode penelitian jaringan hasil Topik 5: hasil proses berdasarkan sistem dilakukan menggunakan data dapat

Tabel 4. 19 Hasil *Topic Modelling* dengan k=3 (Lanjutan)

Cluster	Topik
3	Topic 1: sistem informasi pengujian kinerja penelitian hasil perangkat lunak Topic 2: yogyakarta uin sunan kalijaga sistem pengujian karakteristik dilakukan Topic 3: pengujian sistem kualitas penelitian game teori dilakukan baik Topic 4: faktor sebesar sistem correctness usability reliability pengujian efficiency Topik 5: sistem kualitas berdasarkan memiliki baik lunak perangkat penelitian

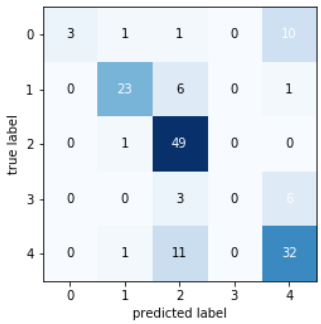
Berikut *souce code* pada *topic modelling* :

```
from sklearn.decomposition import
LatentDirichletAllocation as LDA

n_topics = 5
lda = LDA(n_components=n_topics, learning_method='batch',
random_state=0).fit(tf)

def print_Topics(model, feature_names, Top_Topics,
n_top_words):
    for topic_idx, topic in
enumerate(model.components_[:Top_Topics]):
        print("Topic #%d:" %(topic_idx+1))
        print(" ".join([feature_names[i]
                        for i in topic.argsort()[:-
n_top_words - 1:-1]]))

Top_Words=8
print('Printing top {0} Topics, with top {1}
Words:'.format(n_topics, Top_Words))
print_Topics(lda, tf_terms, n_topics, Top_Words)
```

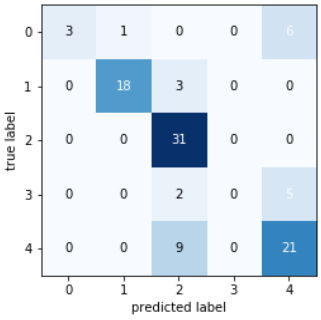
4.7.2 Hasil Evaluasi *Confusion Matrix*

Evaluasi klasifikasi *Naïve Bayes* dilakukan dengan menggunakan *Confusion Matrix.* Evaluasi ini bertujuan untuk mengetahui tingkat *accuracy*, *precision* dan *recall* suatu klasifikasi.

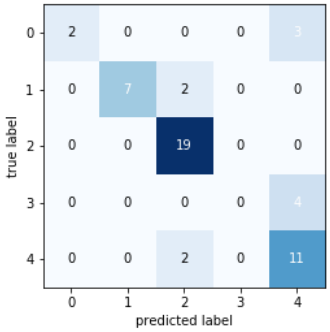
Pembagian data antara data latih dan data uji dilakukan eksperimen dengan perbandingan 7:3 , 8 : 2 , dan 9 : 1. *Confuntion Matrix* dengan perbandingan tersebut dapat dilihat pada Gambar 4.5, 4 .6 dan 4.7.



Gambar 4. 5 *Confusion Matrix* dengan Perbandingan Data 7:3



Gambar 4. 6 *Confusion Matrix* dengan Perbandingan Data 8:2



Gambar 4. 7 *Confusion Matrix* dengan Perbandingan Data 9:1

Berikut *source code* untuk membentuk *confusion matrix* :

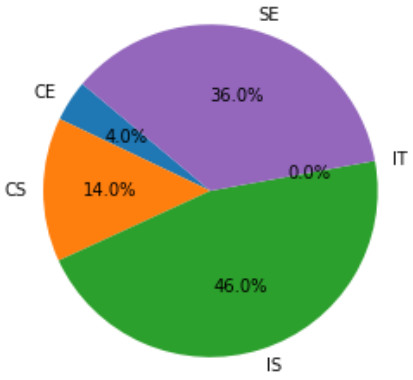
```
from sklearn import metrics
import matplotlib.pyplot as plt
conf = confusion_matrix(y_test, y_pred_class)
fig, ax = plot_confusion_matrix(conf_mat = conf)
plt.show()
```

Setelah terbentuk *confusion matrix* pada data dengan perbandingan 7:3, 8:2 dan 9:1 tahap selanjutnya yaitu mencari prosentase *accuracy, precision* dan *recall* dari klasifikasi. Prosentase tersebut dapat dilihat pada Tabel 4.20.

Tabel 4. 20 Prosentase *Accuracy, Precission* dan *Recall* pada Setiap Pembagian Data

Data	*Accuracy*	*Precission*	*Recall*
7 : 3	72,29%	71 %	72%
8 : 2	73,73%	72%	74%
9 : 1	78%	75%	78%

Untuk melihat prosentase prediksi dengan *accuracy* tertinggi pada semua kategori yaitu CE sebesar 4 %, CS sebesar 14%, IS sebesar 46%, TI sebesar 0% dan SE sebesar 36% disajikan dalam Gambar 4.8.



Gambar 4. 8 Prosentase Prediksi dengan *Accuracy* Tertinggi

Kemudian untuk melihat kata yang sering muncul pada semua dokumen intisari tugas akhir S1 Teknik Informatika UIN Sunan Kalijaga Yogyakarta dapat dilihat dengan menggunakan visualisasi *wordcloud* pada Gambar 4.9

Gambar 4. 9 *Wordcloud* Intisari Bahasa Indonesia

Visualisasi *wordcloud* pada Gambar 4.9 menunjukan kata teratas adalah aplikasi, data, sistem, dan menggunakan