PROPOSAL PROYEK PEMROSESAN BAHASA ALAMI

Peringkasan Teks Berita Bahasa Inggris dengan Strategi Ekstraktif menggunakan Algoritma K-Nearest Neighbour

Disusun oleh:

12S17005 Kiky Purnamasari Napitupulu

12S17006 Tripheni Simanjuntak

12S17023 Jessycha Royanti Tampubolon

11S4037 - PEMROSESAN BAHASA ALAMI

PROGRAM STUDI SARJANA SISTEM INFORMASI
FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO
INSTITUT TEKNOLOGI DEL

NOVEMBER 2020

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

1. PENDAHULUAN

Bab ini berisi penjelasan mengenai latar belakang pengerjaan proyek, tujuan yang ingin dicapai dalam proyek, manfaat yang diperoleh dalam pengerjaan proyek dan ruang lingkup pengerjaan proyek.

1.1 Latar Belakang

Perkembangan teknologi informasi dan komunikasi berdampak menciptakan ledakan informasi seperti bertambahnya jumlah data teks berita dengan cepat dan dalam jumlah besar. Hal ini menyebabkan semua informasi berita dituntut untuk bisa diakses dengan cepat dan tidak butuh banyak waktu untuk dibaca. Teknologi peringkasan teks adalah solusi untuk membantu permasalahan tersebut. Peringkasan teks mengacu pada tugas mengompresi sejumlah besar data teks atau artikel teks yang panjang menjadi bentuk yang lebih ringkas dengan proses pemilihan informasi penting untuk memudahkan pembaca dalam memahami teks [1]. Dengan banyaknya informasi tekstual seperti artikel berita, peringkasan teks penting untuk akses data teks, dimana dapat membantu pembaca melihat konten atau poin utama dalam data teks tanpa harus membaca semua teks. Ringkasan teks dapat memberikan gambaran yang lebih baik kepada pembaca tentang informasi apa yang dikandung dokumen sebelum memutuskan untuk membacanya secara keseluruhan

Terdapat dua strategi dalam peringkasan teks yaitu strategi *extractive summarization* dan strategi *abstractive summarization*. Peringkasan teks dengan strategi *extractive summarization* merupakan strategi peringkasan yang menerapkan fitur linguistik dan statistik dalam membangun kalimat sehingga tidak melakukan perubahan kata pada dokumen hasil ringkasan [2]. Sedangkan, peringkasan teks dengan strategi *abstractive summarization* merupakan strategi peringkasan yang lebih alami dan memiliki komputasi yang lebih sulit karena menerapkan proses *paraphrase* pada seluruh isi dokumen hasil ringkasan [3]. Berdasarkan kedua strategi ini, strategi *extractive* akan menghasilkan ringkasan teks yang lebih kaku dibandingkan dengan strategi peringkasan *abstractive* karena hasil ringkasan pada strategi *extractive* diperoleh berdasarkan frekuensi kemunculan kata pada dokumen asli. Metode yang dapat diterapkan dalam strategi peringkasan *extractive* adalah seperti *Term Frequency-Inverse Document Frequency (TF-IDF) method* yang memanfaatkan model *bag of words*, *cluster based method*, *graph theoretic approach* yang direpresentasikan dalam *undirected graph*, LSA *Method*, *Machine Learning approach* dengan K Nearest Neighbour dan Naive Bayes, serta metode peringkasan menggunakan *neural network* [4]. Strategi peringkasan *abstractive* dapat dikelompokkan kedalam dua metode yaitu *structure base methods* dan *semantic based methods* [5].

Peringkasan ekstraktif lebih sederhana dan merupakan praktik umum pada penelitian peringkasan teks otomatis saat ini. Salah satu metode yang dapat diterapkan dalam strategi peringkasan *extractive* yaitu *Machine Learning approach*, salah satu pendekatan

yang saat ini paling sering digunakan oleh banyak studi di bidang *text summarization.* Untuk peringkasan teks secara ekstraktif, *Machine Learning approach* menganggap peringkasan teks sebagai sebuah *classification problem* yaitu dengan memprediksi apakah sebuah kalimat layak untuk dijadikan sebagai ringkasan atau tidak. *Machine Learning approach* memiliki performa yang baik dan dapat menghasilkan ringkasan yang baik dengan data latih yang terbatas. Salah satu algoritma pada metode M*achine Learning approach* adalah *K-Nearest Neighbour* (KNN). Algoritma *K-Nearest Neighbour* akan mengklasifikasi kalimat (data latih) ke dalam dua kategori, kalimat penting dan kalimat tidak penting. Kalimat yang dipilih menjadi ringkasan berdasarkan kedekatan jarak antara data uji ke data latih dengan mengacu pada nilai hasil ekstraksi fitur. Posisi kalimat hasil ringkasan akan sama urutannya dengan kalimat asli dari dokumen [6].

1.2 Tujuan

Tujuan dari proyek ini yaitu untuk membangun suatu sistem peringkasan teks yang akan menghasilkan ringkasan teks bahasa Inggris menggunakan strategi ekstraktif dengan algoritma K-Nearest Neighbour (KNN)

1.3 Manfaat

Dengan memanfaatkan algoritma K-Nearest Neighbour (KNN) dalam menghasilkan sistem peringkasan teks diharapkan dapat memudahkan pengguna dalam mendapatkan informasi dalam berita BBC News.

1.4 Ruang Lingkup

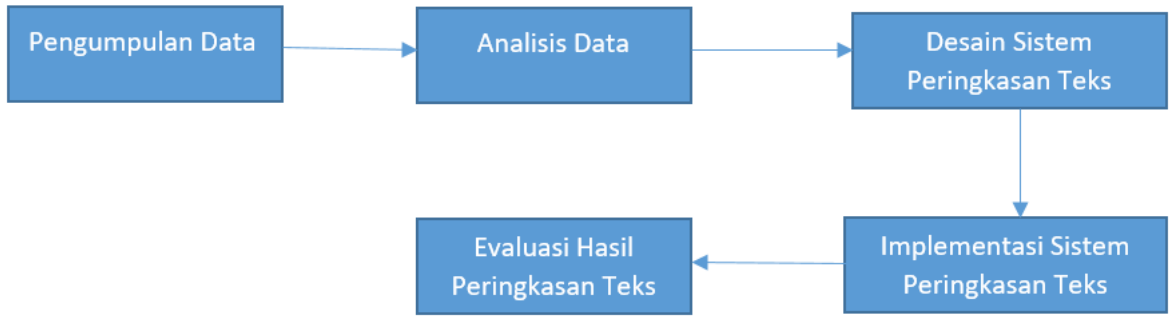
Ruang lingkup dari proyek ini dibatasi dengan data awal berupa dataset bahasa Inggris yaitu *BBC News Summary,* dimana dataset ini akan digunakan sebagai data untuk peringkasan teks bahasa Inggris dengan strategi ekstraktif menggunakan algoritma K-Nearest Neighbour (KNN) pada sistem yang akan dibangun

2. ISI

Bab ini berisi penjelasan meliputi tahapan pemrosesan bahasa alami yang akan diterapkan dalam bentuk diagram alir disertai penjabaran yang terdiri dari pengumpulan data, analisis data, desain sistem peringkasan teks, implementasi sistem peringkasan teks dan evaluasi hasil peringkasan teks.

2.1 Tahapan Pemrosesan Bahasa Alami pada Peringkasan Teks

Sub bab ini menjelaskan tentang tahapan pemrosesan bahasa alami yang akan diterapkan untuk membangun sistem peringkasan teks. Secara umum, langkah-langkah aktivitas yang dilakukan adalah sebagai berikut.



Gambar 1. Diagram Tahapan Pemrosesan Bahasa Alami pada Peringkasan Teks

1. Pengumpulan Data

Data yang akan diringkas adalah data yang dapat diperoleh secara *open source* yaitu *BBC News Summary*. Data ini dipublikasikan oleh *kaggle* dan dapat diakses pada *link* berikut: *https://www.kaggle.com/pariza/bbc-news-summary/data*

2. Analisis Data

Data *BBC New Summary* adalah kumpulan data untuk *extractive text summarization* yang memiliki 417 artikel berita sejak tahun 2004 hingga 2005 yang berasal dari BBC. Dalam data, terdapat kumpulan pasangan artikel dan ringkasannya masing-masing dalam bahasa Inggris dengan 5 kategori yaitu bisnis, politik, olahraga, teknologi dan hiburan dengan format file txt. Baris pertama dari teks artikel adalah judulnya masing-masing.

3. Desain Sistem Peringkasan Teks

Pada tahap desain, akan dilakukan perancangan untuk membangun sebuah model peringkasan teks dengan menggunakan metode ekstraktif serta merancang pembangunan sebuah sistem untuk peringkasan teks.

4. Implementasi Sistem Peringkasan Teks

Pada tahap implementasi akan dilakukan peringkasan teks dengan membangun model menggunakan metode ekstraktif dengan menggunakan algoritma K-Nearest Neighbor (KNN) yang selanjutnya hasil dari *text summarization* akan dianalisis kualitas data nya dengan menggunakan metode evaluasi ROUGE-L.

5. Evaluasi Hasil Peringkasan Teks

Pada tahap ini dilakukan evaluasi dan analisis terhadap hasil ringkasan yang telah diperoleh dengan algoritma K-Nearest Neighbor (KNN). Tahap ini bertujuan untuk memastikan bahwa proyek ini dapat menghasilkan hasil ringkasan yang baik dan sesuai dengan makna dokumen aslinya. Adapun pendekatan yang akan digunakan untuk evaluasi hasil ringkasan tersebut adalah metode ROUGE-L. ROUGE-L merupakan matrik evaluasi yang menghitung nilai kesamaan struktur dokumen ringkasan dengan ringkasan pembanding secara statistik serta akan menghasilkan skor yang menunjukkan nilai perbandingan antara hasil ringkasan dengan ringkasan referensi. Nilai atau skor ROUGE ini juga akan dijadikan sebagai kriteria pembanding yang didasarkan pada skor ROUGE ringkasan referensi.

3. RENCANA KERJA

Bab ini berisi penjelasan mengenai jadwal kegiatan proyek dalam bentuk *Gantt Chart* dan pembagian tugas masing - masing anggota kelompok dalam pengerjaan proyek.

3.1 Jadwal Kegiatan

Berikut adalah jadwal kegiatan pelaksanaan aktivitas pemrosesan bahasa alami pada peringkasan teks yang ditampilkan dalam bentuk *Gantt Chart.*

Tabel 1 *Gantt Chart* Jadwal Kegiatan

No	Aktivitas	Waktu (minggu)			
		1	2	3	4
1	Mengumpulkan data				
2	Melakukan analisis data				
3	Membuat desain sistem peringkasan teks				
4	Melakukan implementasi sistem peringkasan teks				
5	Melakukan evaluasi hasil peringkasan teks				

3.2 Pembagian Tugas

Berikut adalah tabel pembagian tugas setiap anggota kelompok.

Tabel 2 *Team Member and Roles*

Member	*Role*	*Task*
Kiky Purnamasari Napitupulu	*System Analyst*	Berperan dalam perencanaan, pengkoordinasian, pengerjaan serta menganalisis hasil dari yang sudah dikerjakan

	Programmer	Berperan untuk mengimplementasikan code untuk membangun sebuah sistem serta melakukan pengujian terhadap sistem yang sudah dibangun
Tripheni Simanjuntak	*System Analyst*	Berperan dalam perencanaan, pengkoordinasian, pengerjaan serta menganalisis hasil dari yang sudah dikerjakan
	Programmer	Berperan untuk mengimplementasikan code untuk membangun sebuah sistem serta melakukan pengujian terhadap sistem yang sudah dibangun
Jessycha Royanti Tampubolon	*System Analyst*	Berperan dalam perencanaan, pengkoordinasian, pengerjaan serta menganalisis hasil dari yang sudah dikerjakan
	Programmer	Berperan untuk mengimplementasikan code untuk membangun sebuah sistem serta melakukan pengujian terhadap sistem yang sudah dibangun

DAFTAR PUSTAKA

[1] C. Zhai and S. Massung, "A Practical Introduction to Information Retrieval and Text Mining Kindle Edition," *ACM/Association for Computing Machinery,* 2016.

[2] J. M. Kumar and G. R, "Extractive Text Summarization Using Sentence Ranking," *2019 International Conference on Data Science and Communication (IconDSC),* 2019.

[3] R. Adelia, S. Suyanto and U. N. Wisestya, "Indonesian Abstractive Text Summarization Using Bidirectional Gated Reccurent Unit," *Procedia Computer Science,* vol. 157, pp. 581-588, 2019.

[4] V. Gupta and G. Lehal, "A Survey of Text Summarization Extractive Techniques," *Journal of Emerging Technologies in Web Intelligence,* vol. 2, no. 3, 2010.

[5] S. Yeasmin, P. B. Tumpa, A. M. Nitu, M. P. Uddin4, E. Ali5 and M. I. Afjal, "Study on Abstractive Text Summarization Techniques," *American Journal of Engineering Research (AJER),* vol. 6, pp. 254-255, 2017.

[6] R. Indrianto, M. A. Fauzi and L. Mufliklah, "ARTIKEL BERITA KESEHATAN MENGGUNAKAN K-NEAREST NEIGHBOR BERBASIS FITUR STATISTIK," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer,* 2017.