BAB I
PENDAHULUAN

1.1 Latar Belakang

Informasi dalam bentuk teks memiliki jumlah yang sangat banyak dan tersebar di berbagai sumber baik media cetak seperti surat kabar, majalah, maupun media elektonik seperti situs web, *e-mail*, *chatting* atau media sosial. Teks merupakan paparan bahasa baik dalam bentuk lisan maupun tulisan, yang memiliki makna, memiliki kegunaan untuk umum, bersifat praktis, serta berhubungan dengan dunia nyata [1]. Teks dapat terdiri dari satu kata, beberapa kata, satu kalimat, hingga beberapa susunan kalimat. Suatu informasi dari teks dapat diambil dan dianalisis ke dalam suatu bidang penelitian yang meliputi kategori teks, pencarian topik secara spesifik, analisis sentimen atau analisis emosi, hingga *spam filtering*. Spesifikasi *text mining* secara umum adalah untuk mengetahui topik atau cakupan dari permasalahan yang disediakan dalam teks [2].

Setiap layanan yang ada dalam website tidak dapat terlepas dari data teks (document). Data teks sangat penting dalam sebuah layanan pada website karena memberi komunikasi berupa informasi baik dari penyedia layanan kepada pengguna (pengunjung) atau dari pengguna (penggunjung) kepada penyedia layanan tersebut. Setiap penyedia layanan berbasis web tentu tidak luput dari pertanyaan yang diajukan oleh pengguna layanan tersebut, mulai dari pertanyaan teknis hingga pertanyaan yang bersifat umum. Tidak terkecuali Rumah 2 in 1.

Rumah2 in 1 adalah sebuah penyedia layanan *ecommerce property* terbaru di Indonesia yang memiliki keunggulan dengan sistem afiliasi dan berbagai sistem cerdas yang sedang dikembangkan. Perkembangan teknologi terkini yang sangat pesat seperti big data, sistem pakar, dan kecerdasan buatan harus diterapkan guna dapat mengolah berbagai data dan memenuhi kebutuhan pengguna. Untuk mendukung dalam memenuhi kebutuhan pengguna, salah satu yang cocok untuk diterapkan dalam Rumah 2 in 1 adalah pengelompokkan teks atau kategorisasi teks.

Salah satu fitur yang disediakan oleh Rumah 2 in 1 adalah live chat dan helpdesk. Lewat dua fitur tersebut pengguna dapat memberi masukan dan mengajukan pertanyaan. Sampai sejauh ini ada lebih dari 50 pertanyaan yang telah diajukan oleh pengguna lewat layanan tersebut. Namun pertanyaan tersebut belum terstruktur atau belum dikelompokkan. Hal itu menimbulkan ketidakefisien admin dalam mendistribusikan dan membalas pertanyaan. Hal tersebut dapat diperbaiki dengan memanajemen perntanyaan tersebut kedalam kelompok atau kategori. Untuk itu perlu suatu model yang dapat menilai kemiripan teks dan mengkategorikannya.

Clustering dapat diartikan sebagai sebuah teknik yang mengkasifikasikan instansi atau data kedalam kelas-kelas dengan mengghitung jarak anatara instansi-instansi. Instansi yang memiliki jarak terdekat dengan instansi lainnya dikelomokkan kedalah grup yang sama, sebaliknya instansi yang memiliki jarak jauh terhadap instansi lainnya dikelompokkan ke dalam grup lainnya [3]. *Clustering* juga dapat diaplikasikan untuk mengelompokkan atau mengorganisasi teks atau dokumen ke dalam *cluster*. Document *clustering* sudah dipelajari secara intensif

karena dapat diaplikasikan pada area yang bervariasi seperti web mining, search engines, dan information retrieval [4].

Salah satu metode yang dapat digunakan dalam analisis *cluster* adalah metode *Hierarchial Clustering* dengan mengelompokkan data secara hirarkis. Kelebihan metode ini adalah biaya efisiensi yang lebih rendah dan metode ini juga dapat mengelompokkan data tanpa mendefinisikan jumlah *cluster* terlebih dahulu, sehingga *output* metode ini dapat memberi saran jumlah *cluster*. Salah satu algoritma yang populer dan luas digunakan adalah Agglomerative Hierarchial *Cluster*ing [5].

Berdasarkan data-data di atas, penulis tertarik meneliti *Document Clustering* yang akan diaplikasikan pada data pertanyaan pengguna layanan Rumah 2 in 1. Penelitian ini dapat bermanfaat dapat mengorganisir data teks yang ada pada Rumah 2 in 1, khususnya data pertanyaan pengguna sehingga memudahkan dalam menganalisis kebutuhan pengguna dan mempermudah pembuatan FAQ. Meskipun penelitian ini menggunakan data pertanyaan pengguna layanan pada Rumah 2 in 1, namun penelitian ini dapat aplikasikan ke dalam data teks lain, sehingga penelitian ini dapat bermanfaat bagi penelitian selanjutnya dalam bidang *text processing*.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan di atas, Rumusan Masalah pada penelitian ini adalah:

1. Bagaimana membuat suatu model yang dapat mengelompokkan dan mengkategorikan data pertanyaan pengguna pada Rumah 2 in 1 secara otomatis?

2. Bagaimana mengevaluasi model *cluster* dengan menggunakan parameter Korelasi *Cophenetic*?

1.3 Batasan Masalah

Supaya penelitian yang dilakukan lebih terarah dan mencapai sasaran, maka diperlukan suatu pembatasan dalam penelitianatau ruang lingkup kajian sebagai berikut:

1. Data yang digunakan adalah data pertanyaan pengguna layanan Rumah 2 in 1 dengan jumlah data sebanyak 50 pertanyaan.
2. Sistem yang diimplementasikan berbasis *web*.
3. Algoritma Stemmer yang digunakan adalah Porter Stemmer untuk Bahasa Indonesia.
4. Pembotan pada teks atau dokumen menggunakan metode *Tf-Idf*.
5. Menggunakan ngram range 1-2 pada *Tf-Idf*.
6. Menggunakan metode *Cosine Similarity* untuk mengukur kemiripan antar dokumen.
7. Membandingkan 3 model pada algortima Agglomerative Hierarchial *Cluster*ing yaitu Single Link, Complete Link, dan Ward.
8. Mengevaluasi hasil model *cluster*ing menggunakan parameter korelasi *cophenetic*.

1.4 Maksud dan Tujuan Penelitian

Sebagai sebuah penelitian, Tujuan yang akan dicapai dalam penelitian ini adalah sebagai berikut:

1. Membuat suatu model yang dapat mengkategorikan suatu teks atau dokumen berbahasa Indonesia dengam studi kasus pertanyaan pengguna layanan Rumah 2 in1 .
2. Mengukur nilai korelasi *cophenetic* untuk mempertimbangkan seberapa baik model *cluster* yang dihasilkan.

1.5 Metode Penelitian

Metode penlitian yang digunakan dalam penelitian ini adalah metode pengumpulan data, metode pengembangan sistem dan metode pengujian.

1.5.1 Metode Pengumpulan Data

Metode yang digunakan dalam pengumpulan data adalah metode sejarah/wawancara, metode studi kasus, dan metode studi pustaka.

1.5.1.1 Metode Sejarah/Wawancara

Metode wawancara terhadap obyek diperlukan untuk memperoleh info tentang strategi bisnis yang sedang dijalankan oleh Rumah 2 in 1, kemudian didapat kebutuhan-kebutuhan yang diperlukan dalam menunjang kebutuhan bisnis tersebut. Metode ini juga digunakan untuk meperoleh data primer berupa dokumen pertanyaan pengguna pada Rumah 2 In 1. Dokumen pertanyaan ini yang nanti berfungsi sebagai acuan dalam penelitian ini.

1.5.1.2 Metode Studi Kasus

Metode ini digunakan untuk menganalisis kebutuhan obyek penelitian Rumah 2 in 1 sebagai layanam *ecommerce property*. Studi kasus mengacu pada situs Rumah 2 in 1 yaitu www.rumah2in1.com. Dari studi kasus yang ada peneliti dapat memahami kebutuhan dalam pengelolaan data teks atau dokumen sebagai strategi dalam memahami kebutuhan konsumen dan penunjang strategi bisnis. Melalui studi kasus ini, penelitian ini bertujuan untuk membantu mendorong obyek memperbaiki kualitas pelayanan yang ditawarkan.

1.5.1.3 Metode Studi Pustaka

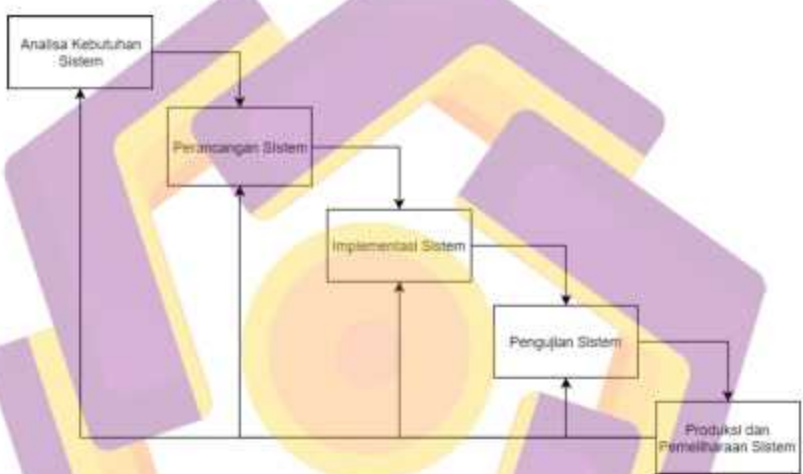
Metode studi pustaka digunakan untuk memperkaya informasi dari mulai perancangan hingga hasil dari sistem. Studi pustaka yang dipakai adalah jurnal internasional yang dipublish serta buku-buku yang bersumber dari perpustakaan dan yang tersedia secara gratis di internet.

1.5.2 Metode Analisis Data

Untuk metode mencari kata dasar atau *Stemming*, peneliti menggunakan metode *Porter Stemmer* untuk Bahasa Indonesia. Pada bagian *Information Retrieval* peneliti menggunakan metode *term frequency–inverse document frequency* (Tf/Idf). Sedangkan untuk metode *clustering*, peneliti membandingkan 3 model yang ada pada Algoritma Agglomerative Hierarchial *Clustering* yaitu Single Link, Complete Link, dan Ward.

1.5.3 Metode Pengembangan Sistem

Peneliti melakukan tahap pengembagan sistem secara prosedural dengan pendekatan secara sistematis. Sehingga pada tahap pengembangan sistem, peneliti menggunakan metode *Waterfall*. Untuk alur pengembangan sistem dapat digambarkan sebagai berikut:



Gambar 1.1 Alur pengembangan sistem menggunakan metode Waterfall

1.5.4 Metode Pengujian

Peneliti menggunakan 2 macam pendekatan dalam pengujian yaitu *Black Box testing* dan *White Box testing*:

1. Black Box Testing (higher levels of testing)

   Untuk menguji secara umum yang dicapai oleh sistem dalam hal ini adalah sistem dapat mengasilkan suatu produk yang dapat melakukan

analisis sentimen secara tepat. Pengujian ini akan mengarah pada *Acceptance Testing*.

2. White Box Tesing (lower levels of testing)

White Box Testing adalah pengujian struktur internal sistem pada level programmer. Pada kategori pengujian ini, peneliti menerapakn *unit testing* untuk menguji beberapa fungsi yang ada dalam program telah sesuai dengan fungsionalnya.

1.6 Sistematika Penulisan

Supaya dokumentasi penelitian ini sistematis, berikut adalah sistematika penulisan skripsi ini:

BAB I : PENDAHULUAN

1.1 Latar Belakang

Menjelaskan latar belakang atau alasan peneliti melakukan penelitian ini.

1.2 Rumusan Masalah

Merumuskan masalah-masalah yang akan dipecahkan hingga penelitian ini berakhir.

1.3 Batasan Masalah

Mengarahkan penelitian supaya penelitian lebih terarah dan tidak keluar dari cakupan.

1.4 Maksud dan Tujuan Penellitian: Tujuan yang dicapai dalam penelitian.

1.5 Metode Penilitian: Metode penelitian yang digunakan supaya penelitian lebih terstruktur.

1.6 Sistematika Penulisan: Susunan dokumentasi tulisan dalam penelitian supaya sistematis.

BAB II: LANDASAN TEORI

2.1. Tinjauan Pustaka

Berisi tetang perbandingan penelitian ini dengan penelitian lain sejenis atau memiliki kemiripan dari beberapa jurnal 5 tahun terakhir.

2.2. Dasar Teori

Berisi tentang dasar teori yang digunakan dalam merancang dan membangun sistem ini. Dasar diambil dari beberapa sumber ilmiah yang dapat dipertanggungjawabkan seperi buku dan jurnal ilmiah. Teori yang akan dibahas di antaranya adalah teori tentang Data, Informasi, Teks, *Data Mining*, *Text Mining*, *Information Retrieval*, *Noisy text*, *Text Classification*, *Stemming*, *Porter Stemmer* Bahasa Indonesia, Clusteing, *Entity Relationship Diagram* (EFD), *Flowchart*, *Data Flow Diagram* (DFD), Algoritma Agglomerative Hierarchial *Clustering*, Evaluasi model menggunakan Rand Index dan Normalized mutual information.

BAB III: ANALISIS DAN PERANCANGAN

Memuat analisis dan perancangan sistem yang dibuat, mulai dari analisis kebutuhan fungsional dan non-fungsional, rancangan *Entity Relationship Diagram*, rancangan proses model seperti *Flowchart* dan *Data Flow Diagram*, struktur atau arsitektur text mining dan analisis sentimen, rancangan *testing* model, rancangan *testing* sistem.

BAB IV: IMPLEMENTASI DAN PEMBAHASAN

Memuat tentang implementasi sistem dengan mengacu pada ANALISIS dan PERANCANGAN yang telah dibuat serta pembahasan masing-masing komponen yang membentuk sistem.

BAB V: PENUTUP

Memuat tentang kesimpulan dari proses dan hasil penelitian, serta saran dari pihak kedua terhadap sistem yang telah dibuat.