# BAB I PENDAHULUAN

## Latar Belakang

Dewasa ini, kebutuhan manusia akan informasi semakin tinggi. Perkembangan teknologi dan ilmu pengetahuan menyebabkan ketersediaan informasi meningkat, terutama informasi berupa data non-linguistik atau data numerik. Hal tersebut bertolak belakang dengan kebutuhan user akan informasi yang mudah dipahami dan dimengerti dengan cepat, sehingga waktu yang digunakan *user* untuk memahami informasi yang berbentuk data numerik relatif lebih lama dibandingkan dengan informasi berupa teks atau berita. Sehingga para peneliti dan pengembang berlomba-lomba untuk mengembangkan aplikasi-aplikasi yang mampu menghasilkan informasi dalam bentuk text dengan *input* data *non-linguistik* atau data numerik. Salah satunya yaitu aplikasi atau sistem *Data-to-text* (D2T) yang diperkenalkan oleh Reiter (2011).

Sistem D2T ini dapat menerima berbagai masukan berupa data non linguistik mulai dari data numerik, *event logs,*  maupun data yang dihasilkan dari sensor sekalipun. Dengan kemampuannya yang dapat menerjemahkan data numerik kedalam data berbentuk tekstual secara otomatis, membuat sistem D2T ini menjadi salah satu bagian dari sistem *Natural Language Generation* (NLG) dimana D2T ini dapat menerjemahkan data ke dalam teks dengan mengasumsikan bahwa data yang digunakan pada dasarnya benar dan akurat (Gkatzia et al., 2017). Karena arsitekturnya yang mirip dengan arsitektur sistem NLG, membuat sistem D2T ini sangat erat dengan proses linguistik dan juga proses analisis data, maka Reiter (2011) memaparkan bahwa setidaknya ada empat langkah dalam tahapan pembangunan sistem D2T, yaitu: (*signal analysis, data interpretation, document planning, microplanning* dan *realisation*).

Sudah banyak implementasi sistem D2T yang menjadi solusi dalam menyediakan informasi tekstual dalam beberapa bidang. Contohnya pada bidang peramalan cuaca, yaitu aplikasi *Data-to-text Weather Prediction* (DWP) yang dapat menghasilkan ringkasan cuaca selama satu bulan dan prediksi cuaca dalam bentuk berita dari masukan berupa data klimatologi dan data kualiats udara selama satu tahun (Putra et al., 2017). Lalu, ada *Forecast Generator* (FOG) yang diperkenalkan oleh (Kittredge & Driedger, 1994), aplikasi tersebut dapat mengkonversi peta cuaca menjadi ramalan dalam bentuk kalimat dengan pengolahan bahasa alami. Selain itu, ada *SumTime-Mousam* yang diperkenalkan oleh (Sripada & Reiter, 2003), aplikasi ini dapat menghasilkan ramalan cuaca laut tekstual untuk rig minyak lepas pantai. Contoh lainnya yaitu pada bidang kesehatan, yaitu *BABYTALK family System* yang diperkenalkan oleh (Portet et al., 2009), aplikasi ini mampu membuat sebuah ringkasan peristiwa yang terjadi selama 45 menit dari sinyal psikologis kontinyu dan diskrit, seperti pengaturan peratalatan dan pemberian obat dalam bentuk kalimat. Selain itu Hunter et al., (2011) memperkenalkan sistem yang dapat menghasilkan ringkasan dari pergantian keperawatan yang berasal dari pencatatan pasien elektronik di Neonatal Intensive Care Unit (NICU)*.* Pada biadang ekonomi, terdapat *Knowledge-Based Report Generator* yang diperkenalkan oleh (Kukich, 1983) yang dapat mengkonversi data stok produk (non-linguistik) menjadi laporan stok pada suatu pasar. Ada juga sistem D2T yang digunakan untuk menerima masukan berupa data *streaming* (Abidin et al., 2018). Mengingat luasnya penerapan sistem D2T tersebut, membuat sistem ini menjadi salah satu pilihan utama yang bisa digunakan untuk menerjemahkan data *non-linguistik* agar lebih mudah dipahami oleh masyarakat luas, tanpa menghilangkan makna asli yang terkandung pada data tersebut.

Setidaknya ada dua masalah utama yang harus diperhatikan dalam pembangunan sebuah *corpus* pada sistem D2T. Pertama, beragamnya jenis informasi yang disimpan dalam sebuah *corpus*, sehingga aspek sumber daya menjadi salah satu faktor yang penting dan harus diperhitungkan. Kedua, beragamnya jenis user, sehingga corpus yang dibangun harus bisa mencakup berbagai kebutuhan dari setiap user (Soehn et al., 2007). Maka pada penelitian ini penulis akan mengembangkan sebuah sistem D2T yang dapat menerima input berupa data *unspecific* atau data yang tidak terikat pada suatu bidang apapun, baik data tersebut memiliki identitas berupa informasi header, kategori, ataupun tidak, sehingga corpus dan keluaran yang dihasilkan bersifat umum atau *unspecific* sesuai dengan informasi yang terkandung dalam data tersebut. Karena seringkali pembangunan sistem D2T ini hanya terdapat untuk satu bidang spesifik seperti yang sudah penulis jelaskan pada bagian sebelumnya. Hal inilah yang menjadi latar belakang pengembangan sistem D2T pada penelitian ini, sehingga sistem D2T yang dibangun diharapkan mampu menerima masukan data *non-linguistik* dalam bidang apapun, tidak terbatas pada suatu bidang spesifik, dan dapat menerima masukan baik data tersebut memiliki informasi berupa *header* atau tidak.

Data yang digunakan dalam penelitian ini meliputi data nilai tukar mata uang asing terhadap rupiah (kurs), data klimatologi, data kualitas udara, dan data partikel udara Kota Beijing. Untuk mengolah data yang bersifat eksak tesebut secara manual tentu saja dibutuhkan sumber daya manusia dan waktu yang cukup untuk menganalisa dan menarik informasi dari data tersebut, sehingga dengan adanya D2T ini sumber daya tersebut bias kita minimalisir. Karena pada dasarnya di era *Big Data* ini ketersediaan data semakin meningkat, mudah diakses, variatif, dan juga dinamis. Namun jika tidak didampingi dengan sebuah sistem yang dapat mengelola data tersebut sehingga informasi yang diperoleh mudah dipahami maka akan dirasa sangat sulit jika kita harus menganalisis data tersebut secara manual. Maka tidak heran sistem D2T ini bisa menjadi suatu solusi yang dapat mengurangi sumber daya, dan mempermudah dalam penyampaian serta analisis suatu data khusunya pada data yang bersifat *unspecific.*

Untuk membangun sistem D2T yang dapat menerima masukan data *unspecific*, maka digunakan beberapa penerapan *Machine Learning* seperti *Gradient Descent*, lalu diterapkan beberapa fitur untuk analisis data seperti *Statistical Tools, Time-Series Analysis, Exponential Smoothing, Linear Model*, *Knuth Morris Pratt* (KMP), *Pearson Correlation,* dan metode lainnya seperti *Fuzzy Membership Function* dan *Crisp Membership Function* untuk menginterpretasikan data masukan. Selain itu, untuk mengefisienkan *Development Time*, penulis menggunakan beberapa *package* yang tersedia dalam R.

## Rumusan Masalah

Berdasarkan latar belakang masalah yang ada, maka permasalah dalam skripsi ini dirumuskan sebagai berikut:

1. Bagaimana pengembangan model dari sistem *Data-to-Text* untuk membangkitkan berita pada data *unspecific* dengan menggunakan pendekatan *Time Series* dan *Machine Learning*?
2. Bagaimana proses implementasi sistem *Data-to-text* untuk data *unspecific* dalam R?
3. Bagaimana eksperimen dan hasil eksperimen dari sistem *Data-to-text* yang dikembangkan?

## Tujuan Penelitian

Berdasarkan rumusan masalah di atas, maka tujuan penelitian ini adalah sebagai berikut:

1. Mengembangkan model sistem *Data-to-text* untuk membangkitkan berita pada data *unspecific* dengan menggunakan pendekatan *Time Series* dan *Machine Learning.*
2. Mengimplementasi model *Data-to-text* untukdata *unspecific* menggunakan bahasa pemrograman R .
3. Menganalisis kualitas sistem dengan melakukan eksperimen, dan pembahasan hasil eksperimen

## Manfaat Penelitian

Manfaat yang dapat diperoleh dari penelitian ini adalah:

1. Diharapkan dapat menambahkan pengetahuan tentang sistem *Data-to-text, Natural Language Processing, Time-Series* serta penerapannya dalam membangkitkan bahasa alami untuk mendeskripsikan data *unspecific*.
2. Dapat menjadi salah satu alternatif dan pelengkap dalam menyampaikan hasil analisis data secara otomatis oleh sistem *Data-totext.*
3. Dapat menjadi salah satu referensi dalam pembangunan sistem *Data-to-text* yang memanfaatkan bahasa pemrograman R beserta fiturnya seperti *packages.*

## Batasan Masalah

Dalam penelitian ini, permasalahan dibatasi hal-hal berikut inil:

1. Pembangunan sisten *Data-to-text* denganpendekatan *Time Series* ini hanya didasarkan pada data *numerical* dan *categorical* yang berbentuk eksak.
2. Pembangunan sisten *Data-to-text* untuk data *unspecific* ini menggunakan bahasa pemrograman R, HTML, dan JavaScript.

## Sistematika Penulisan

Sistematika penulisan skripsi ini diuraikan menjadi lima bab, yaitu:

**BAB I PENDAHULUAN**

BAB I teridiri dari latar belakang, rumusan masalah, batasan masalah, tujuan penelitian yang akan dilakukan, dan sistematikan penulisan.

**BAB II TINJAUAN PUSTAKA**

BAB II terdiri dari beberapa kajian singkat tentang teori-teori dan konsep yang dibutuhkan dalam penelitian. Terdiri dari pembahasan mengenai *Natural Language Processing, Natural Language Generation, Data-to-text, Machine Learning, Time-series, String Matching, R Programming,* dan lainnya*.*

**BAB III METODOLOGI PENELITIAN**

BAB III terdiri dari langkah-langkah yang akan dilakukan dalam penelitian. Terdiri dari desain penelitian dari sistem D2T yang akan dikembangkan, metode pengembangan perangkat lunak D2T, serta alat dan bahan yang digunakan dalam penelitian D2T.

**BAB IV HASIL PENELITIAN DAN PEMBAHASAN**

BAB IV berisi pengumpulan data yang terdiri dari data kurs, data klimatologi, data kualitas udara, dan data partikel udara. Pengembangan model sistem D2T yang terdiri dari *Unspecific Data Handling, Signal Analysis, Data Interpretation, Document Planning,* dan *Microplanning*. Selain itu, pada bab ini dijelaskan mengenai pengembangan sistem D2T menggunakan *Linear Sequential Model*, desain eksperimen, dan hasil dan analisa hasil eksperimen yang mengukur tingkat *Readability, Computation Time,* analisis *Representative Text* dengan plot grafis, serta perbandingan dengan penlitian terkait.

**BAB V KESIMPULAN DAN SARAN**

BAB V berisi kesimpulan yang didapat selama penelitian. Selain itu, pada bab ini dipaparkan saran-saran dalam meningkatkan kualitas dan kuantitas hasil penelitian sistem D2T.

**LAMPIRAN**

Berisi dokumen-dokumen yang menunjang keabsahan penelitian.