Pengembangan Sistem *Data-To-Text* (D2T) untuk Membangkitkan Berita pada Data *Unspecific*

Muhammad Ridwan[*], Lala Septem Riza[#], Enjun Junaeti[#]
Departemen Pendidikan Ilmu Komputer
Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam
Universitas Pendidikan Indonesia
Bandung, Indonesia
just.muhammadridwan@student.upi.edu, lala.s.riza@upi.edu, enjun@upi.edu

Sistem *Data-to-Text* menjadi salah satu pilihan untuk menerjemahkan data *non-linguistik* kedalam bentuk tekstual. Namun seiring dengan perkembangan teknologi, beragamnya bidang dari suatu data dan beragamnya pengguna menjadi salah satu fokus yang harus diperhatikan dalam pengembangan sistem *Data-to-Text.* Penelitian ini bertujuan untuk mengembangkan sistem *Data-to-Text* dengan masukan berupa data *unspecific*, sebagai solusi agar sistem *Data-to-Text* dapat menerima masukan berupa data dari bidang atau domain apapun, baik data tersebut memiliki identitas berupa informasi *header*, tipe data, *rule* ataupun tidak. Maka digunakan pendekatan *Fuzzy Rule* untuk menginterpretasikan data *unspecific* tersebut. Selain itu digunakan beberapa algoritma *Machine Learning* seperti *Gradient Descent,* dan analisis lainnya seperti *Exponential Smoothing, Knuth-Morris-Pratt, Statistical tools* dan *Pearson Correlation Coefficient.* Sistem yang dikembangkan dapat menghasilkan informasi berupa ringkasan data, informasia data terkini dan informasi prediksi. Pengembangan sistem dilakukan dalam bahasa pemrograman R dengan memanfaatkan beberapa *packages* yang tersedia. Eksperimen dilakukan dengan mengukur tingkat *Readibility* dari berita yang dibangkitkan, *Computation Time*, dan membandingkan hasil dengan penelitian terkait. Hasil eksperimen menunjukan bahwa informasi yang dihasilkan terbukti merepresentasikan data yang diberikan dan dapat dipahami oleh tingkat siswa pada tingkat sekolah dasar sekalipun, serta waktu komputasi cukup baik. Sistem ini mampu menhasilkan informasi berdasarkan data meteorologi, data klimatologi, data keuangan, dan data *time series* lainnya.

Kata Kunci— *Data-to-Text; Natural Language Generation; Machine Learning; General purpose; General Corpus; Fuzzy Rule-based; Crisp Rule-based; Time-series Analysis; Exponential Smoothing; Linear Model; Gradient Descent; Kunth-morris-pratt; Pearson Correlation Coefficient*

I. PENDAHULUAN

Seiring perkembangan teknologi saat ini, ketersediaan informasi kian meningkat, terutama informasi berupa data *non-linguistik* atau data *numerik*. Data *non-linguistik* dapat disajikan kedalam bentuk teks atau *linguistik* untuk mempermudah dalam penarikan sebuah informasi [1]. Sehingga dikembangkan sistem *Data-to-Text* (D2T) yang mampu menghasilkan informasi dalam bentuk tekstual dengan masukan berupa data *non-linguistik* atau *data numerik* [2]. Data yang digunakan bisa didapatkan dari berbagai sumber data seperti hasil rekaman sebuah sensor, *event logs*, maupun sumber data lainnya yang dicatat secara berkala.

Data-to-Text (D2T) merupakan bagian dari sistem *Natural Language Generation* (NLG) dimana D2T menerjemahkan data ke dalam bentuk teks dengan mengasumsikan bahwa data yang digunakan pada dasarnya benar dan akurat [3]. Sebuah sistem NLG setidaknya terbagi dalam empat bagian utama yaitu (*macroplanning, microplanning, linguistic realization* dan *presentation*), dimana setiap bagian memiliki sub bagian sendiri, seperti pada *macroplanning* terdapat sub bagian *content planning*, *text planning*, dan *Rhetorical structure theory* (RST) dan pada *microplanning* terdapat *lexicalization* [4].

Arsitektur D2T hampir serupa dengan NLG yang terbagi kedalam empat tahapan utama (*signal analysis, data interpretation, document planning, microplanning and realisation*) [2]. D2T merupakan salah satu solusi yang bisa digunakan untuk menerjemahkan data *non-linguistik* kepada masyarakat tanpa menghilangkan makna yang terdapat didalam data tersebut, denga tujuan agar informasi berupa teks yang disajikan lebih mudah dipahami dibandingkan dengan data *non-linguistik.*

Pada penelitian ini, akan dibahas mengenai pengaplikasian D2T untuk membangkitkan berita berdasarkan data *unspecific* dengan jangka waktu jam, harian, mingguan, bulanan atau bahkan tahunan namun ditabatasi hanya untuk data yang berjenis eksak dan *time series*. Data *unspecific* yang dimaksud adalah data yang tidak terikat pada suatu domain apapun, baik data tersebut memiliki identitas berupa informasi *header*, kategori ataupun tidak, sehingga *corpus* dan keluaran yang dihasilkan bersifat se-*unspecific* mungkin.

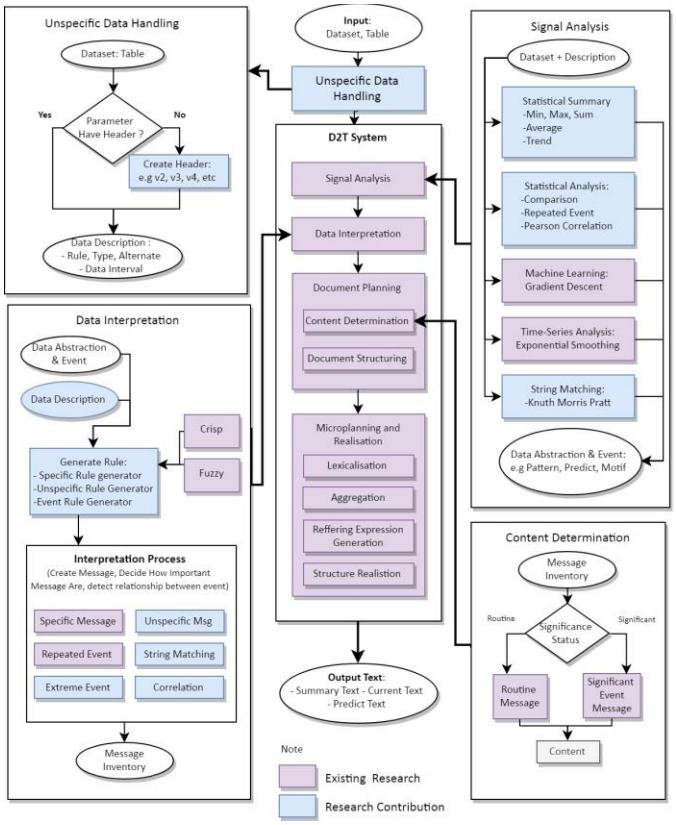
Setidaknya ada dua masalah utama yang harus diperhatikan dalam pembangunan sebuah *corpus.* Pertama, beragamnya jenis informasi yang disimpan dalam sebuah *corpus*, sehingga aspek sumber daya menjadi salah satu faktor yang penting dan harus diperhitungkan. Kedua, beragamnya jenis *user*, sehingga *corpus* yang dibangun harus bisa mencakup berbagai kebutuhan dari setiap *user* [5].

Dalam penelitian ini, akan dibangun sebuah sistem D2T yang tidak terikat pada suatu bidang apapun dan dapat menerima masukan berupa data *unspecific* lalu menghasilkan keluaran se-*unspecific* mungkin, agar sistem yang dibangun diharapkan dapat menjadi solusi dari masalah yang sudah disebutkan sebelumnya. Untuk mencapai hal tersebut, pengembangan sistem D2T ini menggunakan bantuan *Machine Learning* seperti *Linear Regession, Knuth-Morris-Pratt* (KMP), *Pearson Correlation, dan Statistical Tools* untuk pengolahan, analisis, dan prediksi data. Selain itu juga, penulis menggunakan memanfaatkan beberapa *packages* yang tersedia dalam R untuk mengefisienkan *Development Time*. Berbagai penelitian sudah dilakukan untuk membangun sebuah sistem *Data-to-Text* untuk setiap bidang spesifik tertentu, misalnya pada bidang klimatologi terdapat sistem *Forecast Generator* (FOG) yang dapat mengkonversi peta cuaca menjadi ramalan dalam bentuk kalimat dengan pengolahan bahasa alami [6], lalu terdapat sisem *Data-to-text Weather Prediction* (DWP) yang mampu

menghasilkan ringkasan berita klimatologis dan cuaca selama satu bulan serta memberikan informasi prediksi untuk satu hari berikutnya [7], selain itu ada *SumTime-Mousam*, aplikasi ini dapat menghasilkan ramalan cuaca laut tekstual untuk rig minyak lepas pantai [8]. Pada bidang kesehatan, sistem *BabyTalk* diperkenalkan dengan tujuan menghasilkan ringkasan teks dari data neonatal selama 45 menit yang nantinya akan digunakan sebagai bahan pendukung keputusan presentasi modalitas yang terjadi saat itu [9], sistem *BT-Nurse* meringkas kejadian selama *shift* keperawatan berlangsung, berdasarkan hasil rekaman medis elektronik pasien [10]. Pada bidang ekonomi, *Knowledge-Based Report Generator* mampu menghasilkan laporan stok berdasarkan data stok produk (*non-linguistik*) suatu pasar [11].

II. Model Sistem Data-to-Text untuk Data Unspecific

Model *Data-to-Text* terbagi kedalam empat bagian utama yaitu *signal analysis*, *data interpretation*, *document planning*, *microplanning and realisation* [2]. Pada penelitian dilakukan pengembangan model Data-to-Text untuk data *unspecific*, sehingga berfokus pada tahapan *Signal Analysis* dan *Data Interpretation*, lalu terdapat tambahan tambahan yaitu *Unspecific Data Handler* seperti pada "Gambar 1". Model ini menggambarkan ringkasan tahap pengembangan sistem.



Gambar 1 Model *Data-to-Text* untuk Data *Unspecific*

A. *Unspecific Data Handler*.

Pada proses ini, dilakukan pengecekan dataset, apakah terdapat header pada dataset, jika tidak maka nama header menjadi *default* (DateTime, v2, v3, v4, dst). Lalu, proses selanjutnya adalah pembacaan *file datadescription.csv* pada folder *Config*, *Data description* ini terdiri dari nama parameter, tipe parameter (*numerical* / *categorical*), *rule* (*crisp* / *fuzzy*), dan *alternate* yang akan berfungsi untuk me*replace* nama parameter pada tampilan akhir teks. Seperti yang ditampilkan pada tabel I.

Tabel I *Data Description*

	ColName	Type	Rule	Alternate
1	CloudCoverage	numeric	crisp	Cloud Coverage
2	Temperature	numeric	fuzzy	*NA*
3	WindSpeed	numeric	crisp	Wind Speed
4	WindDirection	numeric	crisp	Wind Direction
5	Rainfall	numeric	fuzzy	*NA*

B. *Signal Analysis*

Pada proses ini dilakukan penerapan *Statistical tools* seperti *Min*, *Max*, *Mean*, dan *Machine Learning*. *Linear Regression* digunakan untuk menentukan *trend* dari suatu parameter [12], *Knuth-Morris-Pratt* digunakan untuk mencari pola data yang sama dengan pola data pada minggu ini [13], lalu *Pearson Correlation Coefficient* digunakan untuk melihat keterkaitan antar setiap parameter, sedangkan *Time Series Analysis* dan *Exponential Smoothing* digunakan dalam penentuan prediksi untuk data selanjutnya [14]. Proses ini dilakukan untuk mencari pola diskret dari suatu data, yang nantinya akan diproses lebih lanjut pada tahapan selanjutnya, untuk lebih lengkapnya digambarkan pada "Gambar. 2".

	ColName	MaxDate	MaxValue	MaxIndex	MinDate	MinValue	MinIndex	SumValue	Average	Trend
1	V2	09/01/2015 00:00	14.657	177	05/01/2002 00:00	8.279	17	2174.527	10.3548904761905	+
2	V3	06/01/2018 00:00	13.03707	210	06/01/2002 00:00	6.913	18	2086.51976	9.93580838095238	+
3	V4	09/01/2015 00:00	22.2083	177	05/01/2001 00:00	12.87269	5	3517.31851	16.7491357619048	+
4	V5	09/01/2015 00:00	15.08156	177	05/01/2001 00:00	5.62349	5	2040.76856	9.71794552380952	+
5	V6	06/01/2018 00:00	10.52963	210	06/01/2002 00:00	4.716	18	1512.70414	7.20335304761905	+
6	V7	05/01/2009 00:00	3.948	101	05/01/2010 00:00	1.178	113	603.21457	2.87245033333333	+
7	V8	11/01/2016 00:00	3.03798	191	05/01/2002 00:00	1.062	17	280.79225	1.33710595238095	+
8	V9	06/01/2014 00:00	11.265	162	09/01/2001 00:00	4.897	9	1759.46607	8.37840985714286	+
9	V10	12/01/2013 00:00	11.443	156	12/01/2001 00:00	5.672	12	1858.65248	8.85072609523809	+

Gambar. 2. Hasil Proses *Signal Analysis* untuk *Statistical Summary*

Selain itu, dalam proses ini terdapat proses penelusuran *Extreme Event*, pertama-tama dicari selisih antara data ke i dan i+i, lalu dijumlahkan selisihnya sesuai dengan apakah selisih tersebut bernilai negatif (*decreasing*) atau positif (*increasing*), lalu jumlah selisih negatif dan positif yang paling tinggi, disimpan kedalam bentuk *list*, informasi yang disimpan berupa: jumlah kenaikan/penurunan, indeks awal dan akhir dari kenaikan/penurunan, dan lamanya kenaikan/penurunan berlangsung.

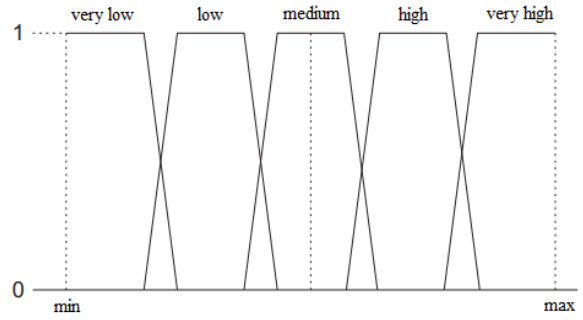
Sedangkan untuk suatu data dikategorikan *repeated event* jika jumlah baris data nilai yang sama secara berturut-turut melebihi n baris data, user dapat menentukan nilai n dengan mengubah value dalam folder *config*. Namun jika user tidak mendefinisikan, maka digunakan nilai default, yaitu: untuk data dengan jangka waktu jam, batas minimum yang digunakan yaitu 6 jam. Untuk data dengan jangka waktu harian, batas minimum yang digunakan yaitu satu minggu atau 7 hari. Untuk data dengan jangka waktu bulanan, digunakan batas minimum yaitu satu caturwulan atau setara dengan 4 bulan, dan untuk data dengan jangka waktu tahunan, digunakan batas minimum selama satu kuartal atau 3 bulan.

Semua proses ini diterapkan pada data yang bertipe *numerikal*, untuk data yang bertipe *catagorical* diterapkan proses khusus yaitu *Motif Discovery* dimana sistem akan mencari apakah terdapat pola data yang sama dengan pola data ke (n-i) hingga ke n, dimana n adalah indeks data untuk hari ini, sedangkan i merupakan limit parameter seperti mingguan,

caturwulan, kuartal seperti yang digunakan pada proses *Repeated Event.*

C. *Data Interpretation*

Karena berita yang akan dibangkitkan berupa berita *unspecific*, tahap ini terbagi menjadi dua proses utama yaitu, interpretasi *Rule-based* menggunakan himpunan *Crisp* dan interpretasi menggunakan himpunan *Fuzzy*. Namun untuk data *unspecific* yang beum mempunyai *Rule* sebagai acuan interpretasi, maka ada proses khusus bernama *Unspecific Rule Generator*, dimana data yang bertipe unspecific (yang tidak terdapat dan tidak didefinisikan dalam *Data Description*) akan menggunakan konsep *Rule-based* dengan menggunakan himpunan *fuzzy*. Data unspecific akan di interpretasi berdasarkan *corpus* yang terdapat pada file "UnspecificAdjective.csv", dengan nilai keanggotaan yang merupakan modifikasi *Fuzzy membership function* pada *Unspecific Trend Generator* seperti yang bisa dilihat di Gambar 3, dimana untuk nilai keanggotaan pada setiap parameter sangat bergantung dari rentang nilai maksimum dan nilai minimum yang diperoleh dari tahapan *Signal Analysis* [15].



Gambar. 3. *Fuzzy membership function for Unspecific Parameter*

Selain untuk data *unspecific*, terdapat interpretasi data yang sudah difenisikan pada *Data Description*, diantaranya *AirQuality* [16], *WindSpeed* [17], *WindDirection* [18], dan *CloudCoverage* [19] menggunakan *Crisp membership function* sedangkan untuk *Temperature* [20], *Rainfall* [21], dan parameter-parameter yang belum teridentifikasi (*unspecific*) maka akan dilakukan interpretasi data dengan *Fuzzy membership function*, selain itu pengguna dapat mengkostumisasi proses ini dengan mengubah nilai ataupun menambah parameter baru untuk diinterpretasi. Pengguna cukup menambahkan atau mengubah parameter pada file MainConfig.csv yang terdapat dalam folder *Config* dan memasukan file fungsi keanggotan pada folder *Corpus* dengan format nama file [nama_parameter]Adjective.csv.

D. *Document Planning*

Pada proses ini dilakukan pemilihan konten (*Content Determination*) dan pembentukan struktur teks (*Document Structuring*) [22]. Untuk proses pemilihan konten, dilakukan dengan membagi konten kedalam dua kelompok, yaitu *Routine Message* dan *Significant Event Message* [7], sedangkankan *Document Structuring* dilaukan dengan cara membuat skema berdasarkan *Target Text* yang dibuat [7]. Pada tahapan ini pengguna dapat mengkostumisasi apa saja yang akan ditampilkan pada teks. Untuk konten ringkasan dan deskripsi data terkini konten dipilih dengan menglompokan kedalam dua kelompok sebelumnya, sedangkan pada informasi prediksi hanya menggunakan kelompok *Routine Message*.

E. *Microplanning and Realisation*

Pada tahap ini setidaknya ada empat hal yang perlu dilakukan yaitu, *Lexicalisation, Aggregation, Referring Expression Generation* dan *Structure Realisation*. Pada tahap *Lexicalisation* dilakukan proses representasi anatara perubahan data, misalnya "*turn progressively to*", "*decrease to*", "*keep stable at*", dan lain sebagainya [6]. Pada tahap *Aggregation* dilakukan ketika akan menghubungkan beberapa pesan menjadi satu kesatuan dengan menggunakan *Simple Conjunction Referring to Contrast Value* [22]. *Referring Expression Generation* dilakukan dengan cara membangkitkan secara *random* berdasarkan *corpus* yang di buat [6]. *Structure Realisation* dilakukan penerapan dengan menyusun semua konten kedalam struktur yang telah ditentukan [2], proses selanjutnya adalah mengganti nama parameter dengan nilai *alternate* yang terdapat pada *Data Description.*

III. Desain Eksperimen

Eksperimen dilakukan dengan cara membandingkan hasil dengan penelitian sebelumnya serta membangkitkan berita dengan 12 *test-case* [23] seperti pada tabel II yang kemudian setiap hasil eksperimen dilakukan pengukuran pada empat aspek yaitu, *Readibility* dan *Computation Time* [23][24], serta perbandingan keluaran sistem *unspecific* dibandingkan dengan sistem pada suatu bidang yang sudah spesifik, juga validasi *representative text*. Pengujian *Readibility* dilakukan dengan menggunakan aplikasi *Readibility Analyzer* pada situs datayze dan readabilityformulas, untuk *Computation Time* dilakukan dengan menggunakan fungsi system.time() dalam R, sedangkan validasi *representative text* dilakukan dengan membandingkan informasi dengan visualisasi data.

Tabel II *Test-Case Eksperimen*

Kode Dataset	Dataset	Detail dan Sumber
CE_NH	1 Jan 2002 – 1 Okt 2018 (bulanan)	Situs web Bank Indonesia (https://www.bi.go.id/) kurs rupiah terhadap valuta asing dengan rentang bulanan, selama 1 Januari 2002 hingga 1 Oktober 2018 tanpa menggunakan header
CE_WH	1 Jan 2002 – 1 Okt 2018 (bulanan)	Situs web Bank Indonesia (https://www.bi.go.id/) kurs rupiah terhadap valuta asing dengan rentang bulanan, selama 1 Januari 2002 hingga 1 Oktober 2018 tanpa menggunakan header
CE_WHM	1 Jan 2002 – 1 Okt 2018 (bulanan)	Situs web Bank Indonesia (https://www.bi.go.id/) kurs rupiah terhadap valuta asing dengan rentang bulanan, selama 1 Januari 2002 hingga 1 Oktober 2018 tanpa menggunakan header
CL_WH	1 Januari 2016 - 31 Desember 2017 (harian)	Data klimatologi pada situs web www.MeteoGalicia.gal, selama satu tahun pada periode 2016-2017 dengan menggunakan header
CL_WHM	1 Januari 2016 - 31 Desember 2017 (harian)	Data klimatologi pada situs web www.MeteoGalicia.gal, selama satu tahun pada periode 2016-2017 dengan kustomisasi *Corpus*, dengan menggunakan header
AQ_NH	1 Januari 2016 - 31 Desember 2017 (harian)	Data kualitas udara pada situs web www.MeteoGalicia.gal, selama satu tahun pada periode 2016-2017 tanpa menggunakan header
AQ_WH	1 Januari 2016 - 31	Data kualitas udara pada situs web www.MeteoGalicia.gal, selama satu

Kode Dataset	Dataset	Detail dan Sumber
	Desember 2017 (harian)	tahun pada periode 2016-2017 dengan menggunakan header
AQ_WHM	1 Januari 2016 - 31 Desember 2017 (harian)	Data kualitas udara pada situs web www.MeteoGalicia.gal, selama satu tahun pada periode 2016-2017 dengan kustomisasi *Corpus*, dengan menggunakan header
BPM_NH	1 Januari 2016 - 31 Desember 2017 (per jam)	Data partikel udara pada situs web http://archive.ics.uci.edu/ml/ datasets/Beijing+PM2.5+Data, selama dua tahun pada periode 2010-2011 tanpa menggunakan header
BPM_WH	1 Januari 2016 - 31 Desember 2017 (per jam)	Data partikel udara pada situs web http://archive.ics.uci.edu/ml/ datasets/Beijing+PM2.5+Data, selama dua tahun pada periode 2010-2011 dengan menggunakan header
BPM_WHM	1 Januari 2016 - 31 Desember 2017 (per jam)	Data partikel udara pada situs web http://archive.ics.uci.edu/ml/ datasets/Beijing+PM2.5+Data, selama dua tahun pada periode 2010-2011 dengan kustomisasi *corpus*, dengan menggunakan header

IV. HASIL DAN PEMBAHASAN

A. Perbandingan *Output* Sistem dengan penelitian terkait

Untuk mempermudah perbandingan *output* dengan penelitian terkait seperti DWP [7], penelitian Ramos [23], dan lainnya, peneliti menggunakan data pada penelitian DWP [7], yaitu data klimatologi dari stasiun MeteoGalicia. Perbandingan *output* dapat dilihat pada tabel III.

Tabel III Perbandingan Output Sistem

Penelitian	Ouput
CL_NH Ou*tput*	According to the daily Climatology data between 07/06/2016 00:00 to 07/06/2017 00:00, with parameters: V2, V3, V4, V5, and V6, it showed that V3 trend is decreased and V6 trend is constant but the rest is increased. V3, and V5 parameters are higher than last week's data, but the rest parameters are lower than last week's data. There were some repeating value more than 36 days: V6 stayed constant at very low during 6 Jul - 18 Aug 2016. V2 fluctuated sharply (increased 90.9 points and decreased 87.6 points), V4 fluctuated dramatically (increased 20.05 points and decreased 27.03 points), V5 fluctuated extremely (increased 270 points and decreased 270 points), and V6 fluctuated extremely (increased 67.2 points and decreased 64.1 points). V6 appears to have a highest impact to all variable with moderate relationship in average. Today data show that: V2 in medium condition. V3 in high condition. V4 in low condition. V5 in very high condition. V6 in very low condition. V6 reached their lowest value on this day. V5 reached their highest value on this day. According to the prediction result, it's forecasted that V2 will normaly move to medium. V3 will stay stable at high. V4 will stay stable at low. V5 will begin to turn to very high. V6 will stay stable at very low.
CL_WH Ou*tput*	Regarding to the daily Climatology data between 07/06/2016 00:00 to 07/06/2017 00:00, with parameters: Cloud Coverage, Temperature, Wind Speed, Wind Direction, and Rainfall, it can be seen that Temperature trend is decreased and Rainfall trend is constant but the rest is increased. Temperature, and Wind Direction parameters are higher than last week's data, but the rest parameters are lower than last week's data. There were some repeating value more than 36 days: Rainfall stayed constant at no rain during 6 Jul - 18 Aug 2016. Cloud Coverage fluctuated rapidly (increased 90.9 points and decreased 87.6 points), Wind Speed fluctuated significantly (increased 20.05 points and decreased 27.03 points), Wind Direction fluctuated rapidly (increased 270 points and decreased 270 points), and Rainfall fluctuated sharply (increased 67.2 points and decreased 64.1 points). Rainfall appears to have a highest impact to all variable with moderate relationship in average. Today data illustrate that: Cloud Coverage in mostly cloudy condition. Temperature in warm condition. Wind Speed in light Breeze condition. Wind Direction in North West condition. Rainfall in no rain condition. Rainfall reached their lowest value on this day. Wind Direction reached their highest value on this day. Based on prediction result, it's estimated that tomorrow sky will be light rain although it's covered by partly cloudy sky. Followed by temperature which decreased to warm. Cloud Coverage will change progressively to partly cloudy. Temperature will still stable at warm. Wind Speed will stay constant at light Breeze. Wind Direction will moderately turn to West. Rainfall will start to change to light rain.
[7] *Output*	Regarding to the prediction result, tomorrow sky state will be light rain although its covered by partly cloudy sky. Followed by temperature which decreased to warm. According to the air quality state, it will start to change to good. According to the monthly summary result, this month was cooler and wetter than average. With average number of rain days, accordingly the total rain so far is well below the average. There was rain on everyday for 7 days from 02nd to 08th and intense rain was dropped in 06th. The wind for the month was light breeze in average. Average air quality was admissible. Average temperature was increased but 05 th was the coldest day of the month with 13.3 celcius degree temperature.
[23] *Output*	With respect to the air quality state, it will be variable although is expected to improve to good, favored by the wind during the coming days
[25] *Output*	Winds northwest 15 diminishingto light monday afternoon. Cloudy with occasional light snow. Fog patches. Visibilities 2 to 5 nm in snow. Belle isle. Northeast gulf northeast coast. Gale warning in belle isle and northeast gulf issued. Gale warning in northeast coast continued. Freezing spray warning continued. Winds southwest 15 to 20 knots increasing to west gales 35
[26] *Outptut*	-Light rian showers are likely -Sunny intervals with rain being possible – less likely than not. -Sunny with rain being unlikely

Berdasarkan pada tabel III jumlah konten tentu semakin banyak, namun secara tekstual aplikasi ini tidak sebaik *output* DWP [7] pada penjelasannya, dikarenakan konsep aplikasi ini dibangun untuk data *unspecific* sehingga mampu membangkitkan berita berdasarkan data apapun selama data tersebut mengikuti format data inputan, sedangkan pada penelitian DWP data inputan harus sama dengan yang ada pada penelitian (parameter). Pada penelitian DWP terdapat dua data inputan yaitu klimatologi dan kualitas udara, sehingga konten yang muncul terdapat dua bagian. Sedangkan pada Ramos, teks yang dibangun hanya untuk kualitas udara dan kecepatan angin saja [23]. Pada penelitian Goldberg terdapat pesan mengenai angin dan salju secara terpisah-pisah [25] sedangkan dalam penelitian Gkatzia hanya disampaikan terkait keadaan langit

[26]. Hal ini menunjukan bahwa D2T yang telah dibangun sebelumnya hanya untuk data yang spesifik, tidak *unspecific*, sehingga ada kemungkinan sistem sebelumnya tidak berjalan jika diberikan data *time series* yang lain.

B. Hasil Eksperimen

Pada aspek *Readibility* dilakukan penilaian berdasarkan *Flesch Reading Ease Score* yang didapatkan menggunakan *tools Readibility Analyzer* pada situs www.datayze.com dan situs www.Readabilityformulas.com, sehingga didapatkan hasil pada tabel IV.

Tabel IV Hasil Pengukuran Aspek *Readibility*

Kode Dataset	*Flesch Reading Ease Score* (Readabilityformulas)	*Flesch Reading Ease Score* (Datayze)
CE_WH	82.01	81
CE_NH	82.02	79.8
CE_WHM	81.71	80.8
CL_WH	68.93	70.2
CL_NH	47.46	52.6
CL_WHM	59.98	63.3
AQ_WH	70.76	69.4
AQ_NH	74.96	72.4
AQ_NHM	70.78	69.8
BPM_WH	77.95	75.6
BPM _NH	81.92	79.6
BPM _NHM	70.78	71.8
Rata-rata	72.43	72.19
Rata-rata Keseluruhan	72.31	

Hasil *Computation Time* didapatkan dengan menjalankan fungsi system.time() dalam bahasa R, seperti system.time(source("D2T_Main.R")), sehingga mendapatkan hasil seperti pada tabel V.

Tabel V Hasil Pengukuran *Computation Time*

Kode Dataset	*Running Time (s)*
CE_WH	4.62
CE_NH	3.72
CE_WHM	3.58
CL_WH	4.36
CL_NH	4.24
CL_WHM	3.37
AQ_WH	5.19
AQ_NH	5.91
AQ_NHM	5.11
BPM_WH	45.02
BPM _NH	44.64
BPM _NHM	45.25
Rata-rata	14.58

V. KESIMPULAN

Pengembangan sistem *Data-to-Text* untuk data *unspecific* dengan menggunakan *Mchine Learning* sangat bermanfaat, dimana sistem dapat bekerja tanpa dengan inputan berupa tabel data numerik berjenis apapun. Hal ini menjadi salah satu keunggulan, karena sistem masukan apapun baik data tersebut memiliki informasi berupa header, tipe data, *rule*, ataupun tidak, sistem akan tetap dapat digunakan.

Kesimpulan dari keseluruhan hasil eksperimen yang dilakukan, keluaran dari sistem terbukti merepresentasikan data yang diberikan. Penelitian ini memeroleh nilai rata-rata keseluruhan 72.31 pada aspek Readability yang artinya keluaran dari sistem ini tergolong dalam kategori *plain english* yang berarti dapat dipahami oleh anak usia remaja sekalipun. Sehingga hal ini menjawab masalah pada latar belakang dimana sistem ini mampu menghasilkan teks keluaran yang mudah dipahami untuk berbagai input data *unspecific*. Sedangkan pada aspek Computation Time diperoleh rata-rata waktu komputasi 2-5 detik untuk data berukuran dibawah 1 ribu baris. Namun untuk data berukuran lebih dari 18 ribu data proses terjadi lebih lama dengan durasi sekitar 45 detik.

Untuk penelitian berikutnya dapat dilakukan pengembangan *Corpus* agar jenis keluaran yang dihasilkan menjadi lebih variatif, terutama pada bagian *Content Determination* diharapkan dapat menggunakan algoritma *Reinforcement Learning* untuk pemilihan konten pada teks keluaran. Perbaikan *UI*/*UX* sehingga pengguna dapat lebih leluasa dalam mengkostumisasi aplikasi.

DAFTAR PUSTAKA

[1] P. Grestl, "Linking linguistic and non-linguistic information," Data & Knowledge Engineering.,vol. 8, pp. 205-222, 1992.

[2] E. Reiter, "An Architecture for Data-to-Text Systems," Comput. Intell., vol. 27, no. 1, pp. 23–40, 2011.

[3] D. Gkatzia, O. Lemon, and V. Rieser, "Data-to-Text Generation Improves Decision-Making Under Uncertainty," *IEEE Comput. Intell. Mag.*, vol. 12, no. 3, pp. 10–17, 2017.

[4] J. Bateman and M. Zock, "Natural Language Generation," Oxford Handb. Comput. Linguist., no. December 2017, pp. 1–21, 2012.

[5] J. Soehn et al, "Requirements of a User-Friendly, Unspecific-Purpose Corpus Query Interface", Proceeding of the LREC 2008 Workshop..Sustainability of Language Resources and Tools for Natural Language Processing, vol. 1, pp.27-32, 2008.

[6] R. I. Kittredge and N. Driedger, "Using Natural-Language Processing to Produce Weather Forecasts," IEEE Expert. Syst. their Appl., vol. 9, no. 2, pp. 45–53, 1994.

[7] B. Putra, L. S. Riza, and Y. Wihardi, "Pengembangan Sistem Data-to-Text untuk Membangkitkan Berita Cuaca dengan Pendekatan Time-Series dalam R," 2017.

[8] J. Hunter *et al.*, "Bt-Nurse: Computer generation of natural language shift summaries from complex heterogeneous medical data," *J. Am. Med. Informatics Assoc.*, vol. 18, no. 5, pp. 621–624, 2011.

[9] F. Portetv *et al*, "Automatic Generation of Textual Summaries from Neonatal Intensive Care Data", pp. 1-45, 2009

[10] J. Hunter *et al.*, "Bt-Nurse: Computer generation of natural language shift summaries from complex heterogeneous medical data," *J. Am. Med. Informatics Assoc.*, vol. 18, no. 5, pp. 621–624, 2011.

[11] K. Kukich, "Design of a knowledge-based report generator," *Proc. 21st Annu. Meet. Assoc. Comput. Linguist.* -, p. 145, 1983.

[12] A. Atilga, "Analysis of long-term temperature data using mann-kendall trend test and linear regression methods: the case of the southeastern anatolia region", Scientific Papers. Series A. Agronomy, Vol. LX, 2017.

[13] M. Regnier, "Knuth-Morris-Pratt algorithm: An analysis", Mathematical Foundations of Computer Science, pp. 431-444, 1989.

[14] E. Ostertagov, "A forecasting using simple exponential smoothing method", Acta Electrotechnica et Informatica, Vol. 12, No. 3, pp.62–66, 2012.

[15] R. Castillo-Ortega, N. Marín, C. Martínez-Cruz, and D. Sánchez, "A proposal for the hierarchical segmentation of time series. Application to trend-based linguistic description," *IEEE Int. Conf. Fuzzy Syst.*, pp. 489–496, 2014.

[16] J. W. Crowder, J. G. Moore, L. DeRose, and W. J. Franek, "Air Pollution Field Enforcement," no. September 1999, 1999.

[17] R. Rowlett, "Beaufort Scales (Wind Speed)," 2001. [Online]. Available: https://www.unc.edu/~rowlett/units/scales/beaufort.html. [Accessed: 20-May-2018].

[18] J. Zandlo, G. Spoden, P. Bouley, and D. Ruschy, "Wind Direction and Degrees," *University of Minnesota*, 2001. [Online]. Available: http://snowfence.umn.edu/Components/winddirectionanddegreeswithouttable3.htm. [Accessed: 20-May-2018].

[19] J. Huby, "Cloud Coverage," 2010. [Online]. Available: http://www.theweatherprediction.com/habyhints/189/. [Accessed: 20-May-2018]. [20] A. Belz, "Probabilistic Generation of Weather Forecast Texts," *Naacl-Hlt*, no. April, pp. 164–171, 2007.

[21] A. Ramos-soto, A. Bugarin, and S. Barro, "Fuzzy Sets Across the Natural Language Generation Pipeline," vol. c, pp. 1–16, 2016.

[22] E. Reiter, "Building Natural-Language Generation Systems," pp. 91–93, 1996.

[23] A. Ramos-Soto, A. Bugarín, and S. Barro, "On the role of linguistic descriptions of data in the building of natural language generation systems," *Fuzzy Sets Syst.*, vol. 285, pp. 31–51, 2016.

[24] A. Belz, "Probabilistic Generation of Weather Forecast Texts," *Naacl-Hlt*, no. April, pp. 164–171, 2007.

[25] R. I. Kittredge and N. Driedger, "Using Natural-Language Processing to Produce Weather Forecasts," *IEEE Expert. Syst. their Appl.*, vol. 9, no. 2, pp. 45–53, 1994.

[26] D. Gkatzia, O. Lemon, and V. Rieser, "Natural Language Generation enhances human decision-making with uncertain information," 2016.