

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

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

Abstrak—Sistem *Data-to-Text* menjadi salah satu pilihan untuk menerjemahkan data *non-linguistik* kedalam bentuk tekstual. Namun seiring dengan perkembangan teknologi, 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 *general*, sebagai solusi agar sistem *Data-to-Text* dapat menerima masukan 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 secara general. Sistem yang dikembangkan dapat menghasilkan informasi berupa ringkasan data, informasi data terkini dan informasi prediksi. Pengembangan sistem dilakukan dalam bahasa pemrograman R dengan memanfaatkan beberapa *packages* yang tersedia. Eksperimen dilakukan dengan mengukur tingkat *Readability* dari berita yang dibangkitkan, *Computation Time*, dan membandingkan hasil dengan penelitian terkait. Hasil eksperimen menunjukkan 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 menghasilkan 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 Corpora*; *Fuzzy Rule-based*; *Time-series Analysis*; *Linear Regression*; *Knuth-morris-pratt*

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, dengan 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 *general* dengan jangka waktu jam, harian, mingguan, bulanan atau bahkan tahunan namun dibatasi hanya untuk data yang berjenis eksak dan *time series*. Data *general* 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-general mungkin.

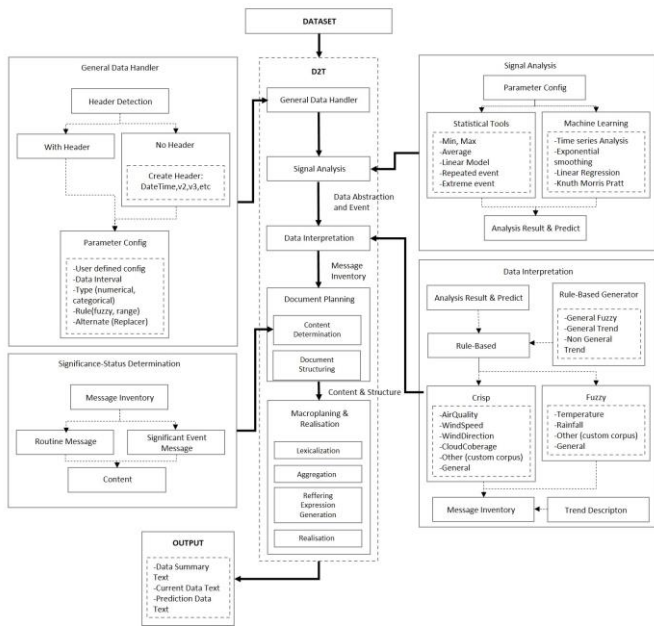
Setidaknya ada dua masalah utama yang harus diperhatikan dalam pembangunan sebuah *corpus*. Pertama, beragamnya jenis informasi yang disimpan dalam sebuah *corpus*, sehingga aspek sumber daya menjadi salah satu faktor yang penting dan harus diperhitungkan. Kedua, beragamnya jenis *user*, sehingga *corpus* 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 *general* lalu menghasilkan keluaran se-general 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 Regression*, *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 sistem *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 GENERAL

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 *general*, sehingga berfokus pada tahapan *Signal Analysis* dan *Data Interpretation*, lalu terdapat tambahan tambahan yaitu *General Data Handler* seperti pada gambar 1. Model ini menggambarkan ringkasan tahap pengembangan sistem.



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

A. General Data Handler.

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

Tabel I *Parameter Config*

	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], 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
V2	09/01/2015 00:00	14.657	177	05/01/2002 00:00	8.279	17	2174.527	10.3548904761905	+
V3	06/01/2018 00:00	13.03707	210	06/01/2002 00:00	6.913	18	2086.51976	9.93580838095238	+
V4	09/01/2015 00:00	22.2083	177	05/01/2001 00:00	12.87269	5	3517.31851	16.7491357619048	+
V5	09/01/2015 00:00	15.08156	177	05/01/2001 00:00	5.62349	5	2040.78856	9.7179452380952	+
V6	06/01/2018 00:00	10.52963	210	06/01/2002 00:00	4.716	18	1512.70414	7.20335304761905	+
V7	05/01/2009 00:00	3.948	101	05/01/2010 00:00	1.178	113	603.21457	2.87245033333333	+
V8	11/01/2016 00:00	3.03798	191	05/01/2002 00:00	1.062	17	280.78225	1.33710595238095	+
V9	06/01/2014 00:00	11.265	162	09/01/2001 00:00	4.897	9	1759.46607	8.37840985714286	+
V10	12/01/2013 00:00	11.443	156	12/01/2001 00:00	5.672	12	1858.65248	8.85072609523809	+

Gambar 2 Hasil model *signal analysis*

Selain itu, dalam proses ini terdapat proses penelusuran *Extreme Event*, pertama-tama dicari selisih antara data ke *i* dan *i+1*, 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.

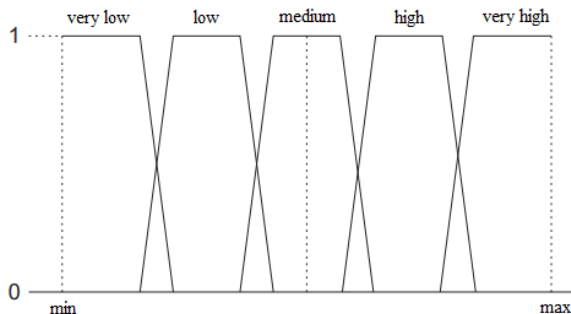
Sedangkan untuk suatu data dikategorikan *repeated event* jika jumlah baris data nilai yang sama secara berturut-turut melebihi *n* baris data, user dapat menentukan nilai *n* dengan mengubah value dalam folder *config*. Namun jika user tidak mendefinisikan, maka digunakan nilai default, yaitu: untuk 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 *general*, tahap ini terbagi menjadi dua proses utama yaitu, interpretasi *Rule-based* menggunakan himpunan *Crisp* dan interpretasi menggunakan himpunan *Fuzzy*. Namun untuk data *general* yang beum mempunyai *Rule* sebagai acuan interpretasi, maka ada proses khusus bernama *General Fuzzy Generator* dan *General Trend Generator*, dimana data yang bertipe *general* (yang tidak terdapat dan tidak didefinisikan dalam *Parameter Config*) akan menggunakan konsep *Rule-based* dengan menggunakan himpunan *fuzzy*. Data *general*

akan diinterpretasi berdasarkan *corpus* yang terdapat pada file “GeneralAdjective.csv”, dengan nilai keanggotaan yang merupakan modifikasi *Fuzzy membership function* pada *General 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 1 Fuzzy membership function for General Parameter

Selain untuk data *general*, terdapat interpretasi data yang sudah difenisikan pada *Parameter Config*, diantaranya *AirQuality* [16], *WindSpeed* [17], *WindDirection* [18], dan *CloudCoverage* [19] menggunakan *Crisp membership function* sedangkan *Temperature* [20], *Rainfall* [21], pengguna dapat mengkostumisasi proses ini dengan mengedit 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 keanggotaan 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], sedangkan *Document Structuring* dilakukan 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 menglompokkan 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 *Parameter Config*.

III. DESAIN EKSPERIMEN

Eksperimen dilakukan dengan cara membandingkan hasil dengan penelitian sebelumnya serta membangkitkan berita dengan 12 *test-case* [23] seperti pada tabel I yang kemudian setiap hasil eksperimen dilakukan pengukuran pada empat aspek yaitu, *Readability* dan *Computation Time* [23][24], serta perbandingan keluaran sistem *general* dibandingkan dengan sistem pada suatu bidang yang sudah spesifik, juga validasi *representative text*. Pengujian *Readability* dilakukan dengan menggunakan aplikasi *Readability Analyzer* pada situs *datayze* dan *grammarly*, 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
KB_WH	(Buy) Maret 2018	Situs web Bank Indonesia (https://www.bi.go.id/) kurs rupiah terhadap valuta asing dengan rentang bulanan, selama 1 Januari 2002 hingga 1 Juni 2018 dengan menggunakan header
KB_NH	(Buy) April 2018	Situs web Bank Indonesia (https://www.bi.go.id/) kurs rupiah terhadap valuta asing dengan rentang bulanan, selama 1 Januari 2002 hingga 1 Juni 2018 tanpa menggunakan header
KB_NHM	(Buy) Mei 2018	Situs web Bank Indonesia (https://www.bi.go.id/) kurs rupiah terhadap valuta asing dengan rentang bulanan, selama 1 Januari 2002 hingga 1 Juni 2018 dengan kustomisasi <i>Corpus</i> , tanpa menggunakan header
SD_WH	Juli 2016	Data interval harian dari situs web https://solargis.com . Selama satu tahun mulai dari bulan 01 Juli 2016 hingga 31 Desember 2016 dengan menggunakan header
SD_NH	Agustus 2016	Data interval harian dari situs web https://solargis.com . Selama satu tahun mulai dari bulan 01 Agustus 2016 hingga 31 Desember 2016 tanpa menggunakan header
SD_NHM	September 2016	Data interval harian dari situs web https://solargis.com . Selama satu tahun mulai dari bulan 01 September 2016 hingga 30 September 2016 dengan kustomisasi <i>Corpus</i> , tanpa menggunakan header
CL_WH	1 Januari 2016 - 31 Desember 2017	Data klimatologi pada situs web www.MeteoGalicia.gal , selama satu tahun pada periode 2016-2017 dengan menggunakan header
CL_NH	1 Januari 2016 - 31 Desember 2017	Data klimatologi pada situs web www.MeteoGalicia.gal , selama satu tahun pada periode 2016-2017 tanpa menggunakan header
CL_NHM	1 Januari 2016 - 31 Desember 2017	Data klimatologi pada situs web www.MeteoGalicia.gal , selama satu tahun pada periode 2016-2017 dengan kustomisasi <i>Corpus</i> , tanpa menggunakan header
AQ_WH	2016-2017	Data kualitas udara pada situs web www.MeteoGalicia.gal , selama satu tahun pada periode 2016-2017 dengan menggunakan header
AQ_NH	2016-2017	Data kualitas udara pada situs web www.MeteoGalicia.gal , selama satu tahun

Kode Dataset	Dataset	Detail dan Sumber
		pada periode 2016-2017 tanpa menggunakan header
AQ_NHM	2016-2017	Data kualitas udara pada situs web www.MeteoGalicia.gal , selama satu tahun pada periode 2016-2017 dengan kustomisasi <i>Corpus</i> , tanpa 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 [25], dan lainnya, peneliti menggunakan data pada penelitian DWP [7], yaitu data klimatologi dari stasiun MeteoGalicia. Perbandingan *output* dapat dilihat pada tabel II.

Tabel III Perbandingan Output Sistem

Penelitian	Ouput
CL_WH Output	<p>Regarding to the daily data, from 07/06/2016 00:00 to 07/06/2017 00:00, with parameters: Cloud Coverage, Temperature, Wind Speed, Wind Direction, and Rainfall. It is clear that, Temperature trend is decreased but the rest is increased. There were some repeating value more than 7 days: During 6-21 Jul 2016, 23 Jul - 3 Aug 2016, 5-18 Aug 2016, 26 Aug - 6 Sep 2016, 17-23 Sep 2016, 24 Oct - 2 Nov 2016, 25 Nov - 3 Dec 2016, 5-13 Dec 2016, 23 Dec 2016 - 2 Jan 2017, 16-24 Jan 2017, 13-20 Mar 2017, 2-24 Apr 2017, 14-21 Jun 2017 Rainfall stayed constant at 0 (no rain). Temperature increased greatly (10.3 points) from 25-29 Mar 2017, While Cloud Coverage decreased greatly (-87.6 points) from 7-10 Mar 2017, Temperature decreased greatly (-9.8 points) from 4-7 Jan 2017, Wind Speed decreased greatly (-27.03 points) from 3-8 Feb 2017, Wind Direction decreased greatly (-270 points) from 12-14 Jul 2016, and Rainfall decreased greatly (-64.1 points) from 3-5 Feb 2017.</p> <p>Today data (07/06/2017 00:00) illustrate that: Cloud Coverage normal change to mostly cloudy. Temperature keep stable at warm. Wind Speed stay stable at light Breeze. Wind Direction shifted to North West. Rainfall stay stable at no rain.</p> <p>According to the prediction result, it's predicted 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 moderately turn to partly cloudy. Temperature will constant at warm. Wind Speed will still stable at light Breeze. Wind Direction will normaly move to West. Rainfall will still stable at light rain. "</p>
CL_NH Output	<p>Regarding to the daily data, between 07/06/2016 00:00 to 07/06/2017 00:00, with parameters: V2, V3, V4, V5, and V6. It is clear that, V3 trend is decreased but the rest is increased. There were some repeating value more than 7 days: During 6-21 Jul 2016, 23 Jul - 3 Aug 2016, 5-18 Aug 2016, 26 Aug - 6 Sep 2016, 17-23 Sep 2016, 24 Oct - 2 Nov 2016, 25 Nov - 3 Dec 2016, 5-13 Dec 2016, 23 Dec 2016 - 2 Jan 2017, 16-24 Jan 2017, 13-20 Mar 2017, 2-24 Apr 2017, 14-21 Jun 2017 V6 stayed constant at 0 (low). V3 increased greatly (10.3 points) from 25-29 Mar 2017, While V2 decreased greatly (-87.6 points) from 7-10 Mar 2017, V3 decreased greatly (-9.8 points) from 4-7 Jan 2017, V4 decreased greatly (-27.03 points) from 3-8 Feb 2017, and V6 decreased greatly (-64.1 points) from 3-5 Feb 2017."</p> <p>"Today data (07/06/2017 00:00) describe that: V2 averagely change to medium. V3 begin to turn to high. V4, and V6 constant at low. V5 shifted to very high. "</p> <p>According to the prediction result, it's expected that V2 will averagely turn to medium. V3 will normaly move</p>

Penelitian	Ouput
	to high. V4 will steady at low. V5 will begin to turn to very high. V6 will stay stable at very low. "
CL_NHM Output	<p>Regarding to the daily data, between 07/06/2016 00:00 to 07/06/2017 00:00, with parameters: V2, V3, V4, V5, and V6. It is clear that, V3 trend is decreased but the rest is increased. V3 increased greatly (10.3 points) from 25-29 Mar 2017, While V2 decreased greatly (-87.6 points) from 7-10 Mar 2017, V3 decreased greatly (-9.8 points) from 4-7 Jan 2017, V4 decreased greatly (-27.03 points) from 3-8 Feb 2017, and V6 decreased greatly (-64.1 points) from 3-5 Feb 2017."</p> <p>"Today data (07/06/2017 00:00) describe that: V2 averagely change to medium. V3 begin to turn to high. V4, and V6 constant at low. V5 shifted to very high. "</p> <p>According to the prediction result, it's expected that V2 will averagely turn to medium. V3 will normaly move to high. V4 will steady at low. V5 will begin to turn to very high. V6 will stay stable at very low. "</p>
[6] Output	<p>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.</p> <p>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.</p>
[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	<p>Winds northwest 15 diminishingto light monday afternoon. Cloudy with occasional light snow. Fog patches. Visibilities 2 to 5 nm in snow.</p> <p>Belle isle.</p> <p>Northeast gulf northeast coast.</p> <p>Gale warning in belle isle and northeast gulf issued.</p> <p>Gale warning in northeast coast continued.</p> <p>Freezing spray warning continued.</p> <p>Winds southwest 15 to 20 knots increasing to west gales 35</p>
[22] Output	<p>-Light rian showers are likely</p> <p>-Sunny intervals with rain being possible – less likely than not.</p> <p>-Sunny with rain being unlikely</p>

Berdasarkan pada tabel II jumlah konten tentu semakin banyak, namun secara tekstual aplikasi ini tidak sebaik *output* DWP [7] pada penjelasannya, dikarenakan konsep aplikasi ini dibangun untuk data *general* 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 [21] sedangkan dalam penelitian Gkatzia hanya disampaikan terkait keadaan langit. Hal ini

menunjukkan bahwa D2T yang telah dibangun sebelumnya hanya untuk data yang spesifik, tidak *general*, sehingga ada kemungkinan sistem sebelumnya tidak berjalan jika diberikan data *time series* yang lain.

B. Hasil Eksperimen

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

Tabel IV Hasil Pengukuran Aspek *Readability*

Kode Dataset	<i>Flesch Reading Ease Score</i> (Readabilityformulas)	<i>Flesch Reading Ease Score</i> (Datayze)
CL_WH	79.3	67.65
CL_NH	90.3	82.27
CL_WHM	87.5	79.58
KB_WH	60.3	50.6
KB_NH	89.5	84.38
KB_NHM	91.5	86.77
SD_WH	82.1	78.1
SD_NH	83.7	74.3
SD_NHM	74.9	90.2
AQ_WH	87.2	81.5
AQ_NH	81.3	86.6
AQ_NHM	72.4	75.7
Rata-rata	81,7	78,1
Rata-rata Keseluruhan	80.0	

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 IV.

Tabel V Hasil Pengukuran *Computation Time*

Kode Dataset	<i>Running Time</i> (s)
CL_WH	2.01
CL_NH	1.98
CL_WHM	2.14
KB_WH	1.84
KB_NH	2.08
KB_NHM	2.08
SD_WH	1.98
SD_NH	2.03
SD_NHM	2.25
AQ_WH	2.44
AQ_NH	2.63
AQ_NHM	2.79
Rata-rata	2.235

V. KESIMPULAN

Pengembangan sistem *Data-to-Text* untuk data *general* dengan menggunakan *Machine 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 memperoleh nilai rata-rata keseluruhan 80.0 pada aspek *Readability* yang artinya keluaran dari sistem ini tergolong dalam kategori bacaan yang sangat mudah untuk dipahami bahkan oleh tingkat siswa sekolah dasar sekalipun. Sedangkan pada aspek *Computation Time* diperoleh rata-rata waktu komputasi 2.235 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 *Machine 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, General-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] K. Kukich, "Design of a knowledge-based report generator," *Proc. 21st Annu. Meet. Assoc. Comput. Linguist.*, p. 145, 1983.
- [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. R. Cliff and K. M. Billy, "Estimation of the Parameters of a Linear regression System Using the Simple Averaging Method," vol. 13, no. 11, pp. 7749-7758, 2017.
- [12] J. W. Crowder, J. G. Moore, L. DeRose, and W. J. Franek, "Air Pollution Field Enforcement," no. September 1999, 1999.
- [13] R. Rowlett, "Beaufort Scales (Wind Speed)," 2001. [Online]. Available: <https://www.unc.edu/~rowlett/units/scales/beaufort.html>. [Accessed: 20-May-2018].
- [14] J. Zandlo, G. Spoden, P. Bouley, and D. Ruschy, "Wind Direction and Degrees," *University of Minnesota*, 2001. [Online]. Available: <http://snowfence.umn.edu/Components/winddirectionanddegreeswithtable3.htm>. [Accessed: 20-May-2018].
- [15] J. Huby, "Cloud Coverage," 2010. [Online]. Available: <http://www.theweatherprediction.com/habyhints/189/>. [Accessed: 20-May-2018].
- [16] A. Ramos-soto, A. Bugarin, and S. Barro, "Fuzzy Sets Across the Natural Language Generation Pipeline," vol. c, pp. 1-16, 2016.
- [17] 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.

- [18] E. Reiter, "Building Natural-Language Generation Systems," pp. 91–93, 1996.
- [19] 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.
- [20] A. Belz, "Probabilistic Generation of Weather Forecast Texts," *Naacl-Hlt*, no. April, pp. 164–171, 2007.
- [21] 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.
- [22] D. Gkatzia, O. Lemon, and V. Rieser, "Natural Language Generation enhances human decision-making with uncertain information," 2016.
- [24] D. Bertram, "Likert Scales," *University of Calagary, Department of Computer Science*, 2013. [Online]. Available: <http://poincare.matf.bg.ac.rs/~kristina/topic-dane-likert.pdf>. [23]
- [25] D. Bertram, "Likert Scales," *University of Calagary, Department of Computer Science*, 2013. [Online]. Available: <http://poincare.matf.bg.ac.rs/~kristina/topic-dane-likert.pdf>.
- [25] J. Soehn et al, "Requirements of a User-Friendly, General-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.