Journal of Computers for Society

Journal homepage: https://ejournal.upi.edu/index.php/JCS

# Development of a Data-To-Text (D2T) System to Generate News on Streaming Data

*Ahmad Zainal Abidin*, Enjang Ali Nurdin, Lala Septem Riza*

Department of Computer Science Education, Universitas Pendidikan Indonesia, Indonesia
*Correspondence: E-mail: ahmad.zainal565@student.upi.edu

**ABSTRACT**

This research aims to develop a Data-to-Text system with input in the form of streaming data in batch form, to generate news in general. The development of a Data-to-Text system model is carried out by applying Machine Learning to overcome Streaming data, with the Piecewise Linear Approximation technique using the Least Square method. The developed system produces data summary information, current data information, and prediction information. System development is carried out in the R programming language by utilizing several available packages. The experiment was conducted by measuring the Readability level of the news raised, Computation Time, and comparing the results with related research. The experimental results show that the information produced is proven to represent the data provided and can be understood by the student level or above, and the computational time is quite good. The system can generate information based on meteorological data, climatological data, and financial data.



**ARTICLE INFO**

***Article History:***
*Submitted/Received 03 Mar 2021*
*First Revised 22 Apr 2021*
*Accepted 06 Jun 2024*
*First Available Online 07 Jun 2024*
*Publication Date 15 Sep 2024*

***Keyword:***
*Data-to-text,*
*Least square method,*
*Machine learning,*
*Natural language generation,*
*Picewise linear approximation,*
*Streaming,*
*Time-series.*

## 1. INTRODUCTION

Advancements in modern technology, systems have been developed that can generate text-based (linguistic) information from non-linguistic data (raw data measured by sensors or resulting from a series of events). This process, known as Natural Language Generation (NLG), transforms raw data into human-readable text (van der Lee *et al*., 2021; Zubair & Jafir, 2021; Vayre *et al*., 2017; Yagamurthy *et al*., 2023). Such data can come from various sources, including survey results, transaction records, sensor readings, financial statistics, weather data, and sales transactions, among others.

The architecture of Natural Language Generation (NLG) consists of four primary components: macroplanning, microplanning, linguistic realization, and presentation (Vicente *et al*., 2018; Barros *et al*., 2021; Kondadadi *et al*., 2013; Ramos *et al*., 2020; Barros *et al*., 2019; Garcia-Mendez *et al*., 2019). Each of these components includes its own subsections. For instance, macroplanning encompasses content planning, text planning, and Rhetorical Structure Theory (RST), while microplanning involves lexicalization.

Data to-text (D2T) is a Characteristic Language Age (NLG) framework intended to make message from non-phonetic information inputs, for example, sensor readings and successions of occasions (Reiter, 2011). D2T works inside the NLG structure, making an interpretation of information into text under the suspicion that the info information is generally precise and solid (Gkatzia *et al*., 2017).

The engineering of data to-text (D2T) is very like that of Regular Language Age (NLG), containing four principal parts: signal investigation, information translation, archive arranging, and microplanning and acknowledgment (Reiter, 2011). D2T fills in as a powerful answer for changing over non-phonetic information into message that the overall population can undoubtedly comprehend without losing the first significance of the information.

This research focuses on utilizing D2T for generating news articles derived from streaming data, covering various time intervals such as daily, weekly, monthly, or even yearly periods, specifically targeting precise data and time series. A data stream represents a series of events that can only be accessed once or a few times due to resource constraints in terms of computing power and storage capacity. This type of data source is distinguished by its rapid flow of information and is generated within a dynamic environment where the distribution is non-stationary.

Systems dealing with streaming data inputs must construct models that encapsulate each processed data point since past data cannot be retrieved, and it's crucial to develop models that encompass the entirety of the data. This forms the basis for advancing Data-to-Text models tailored for streaming data, leveraging Machine Learning techniques to address these challenges.

Streaming data can come from sensors (weather, air quality, water quality), periodic surveys (population growth, visitors to tourist attractions), or routine data (finance, buying and selling transactions), automatic data can be processed as input data, because computers enter data into other computers or directly from sensors without having to wait for humans to enter data (Muthukrishnan, 2005).

In the evolution of this D2T system, it's not solely focused on producing news summaries based on streaming data; rather, it incorporates real-time data updates and forecasts of future streaming data. These predictions are shaped by models derived from each streaming batch. To accomplish this, the D2T system development in this research employs a Machine Learning strategy for forecasting, specifically utilizing the Piecewise Linear Approximation

DOI: https://doi.org/10.17509/jcs.v5i2.70799
e- ISSN 2723-4088

(PLA) method with Least Squares. To expedite development, the author utilizes various R packages.

A few Information to-Text frameworks have been created in earlier examination. These incorporate the data to-text Climate Expectation (DWP) framework, equipped for creating climatological and climate news outlines for a one-month time span while likewise giving prescient bits of knowledge to the next day. Another model is the BabyTalk framework, which produces text outlines of neonatal information at regular intervals, filling in as choice help material for progressing modalities. Also, the BT-Attendant framework sums up occasions happening during nursing shifts, drawing from electronic clinical records of patients. Ultimately, the Information Based Report Generator is capable in creating stock reports in view of non-semantic item stock information from a market.

## 2. METHODS

The Information to-Text model involves four essential parts: signal examination, information translation, report arranging, and micro planning and acknowledgment (Reiter, 2011). In this examination, an Information to-Text model custom-made for streaming information was contrived, consolidating an extra part known as the Realtime Peruser Dataset, as portrayed in **Figure 1**. This model outlines the critical phases of framework improvement.

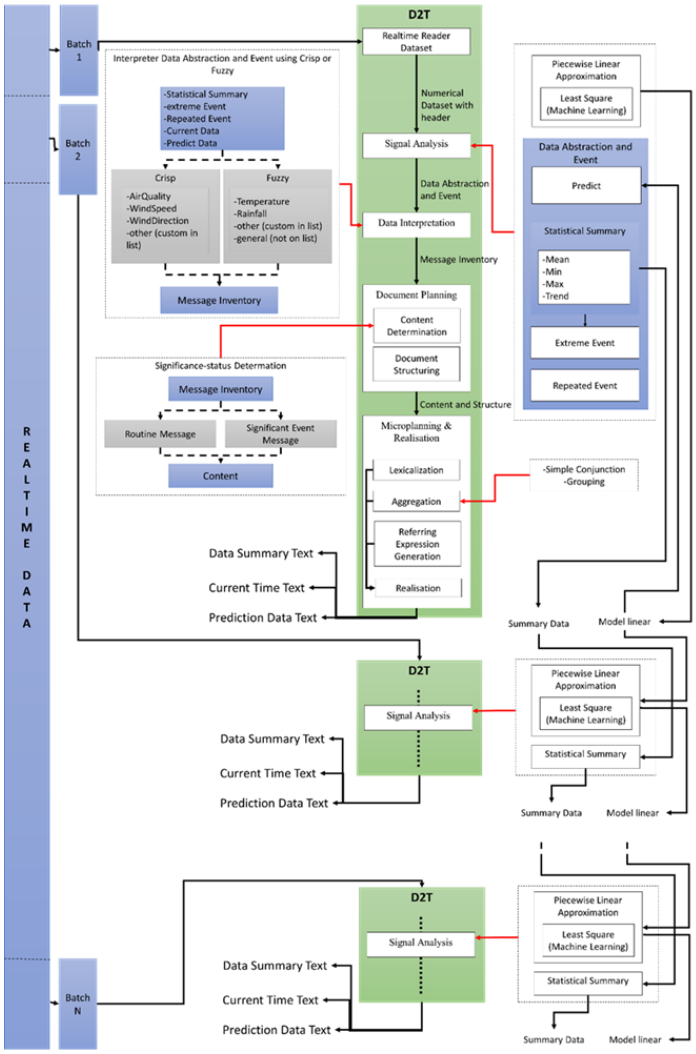


**Figure 1**. Data-to-text model on streaming data.

DOI: https://doi.org/10.17509/jcs.v5i2.70799
e- ISSN 2723-4088

### 2.1. Realtime Reader Dataset

In this process, checking the dataset is carried out in the specified folder, if there is a new dataset, then the Data-to-Text program in R language will be run. This process is done with the AJAX programming language to check the data, then send signals into the PHP program to run the D2T program that has been developed. With this process, the problem of very fast data flow speed can be solved.

### 2.2. Signal Analysis

In this process, the application of Machine Learning is carried out to overcome data storage problems with PLA techniques using the Least Square method. This deployment is done to get a model from each batch of data, so that if interpretation of the data is needed, it can generate it using the model that has been stored. In each batch, a combination of previously created models is carried out, so that one model is obtained that represents the entire batch. This model is also used to predict future data, with the process model as shown in **Figure 2**.

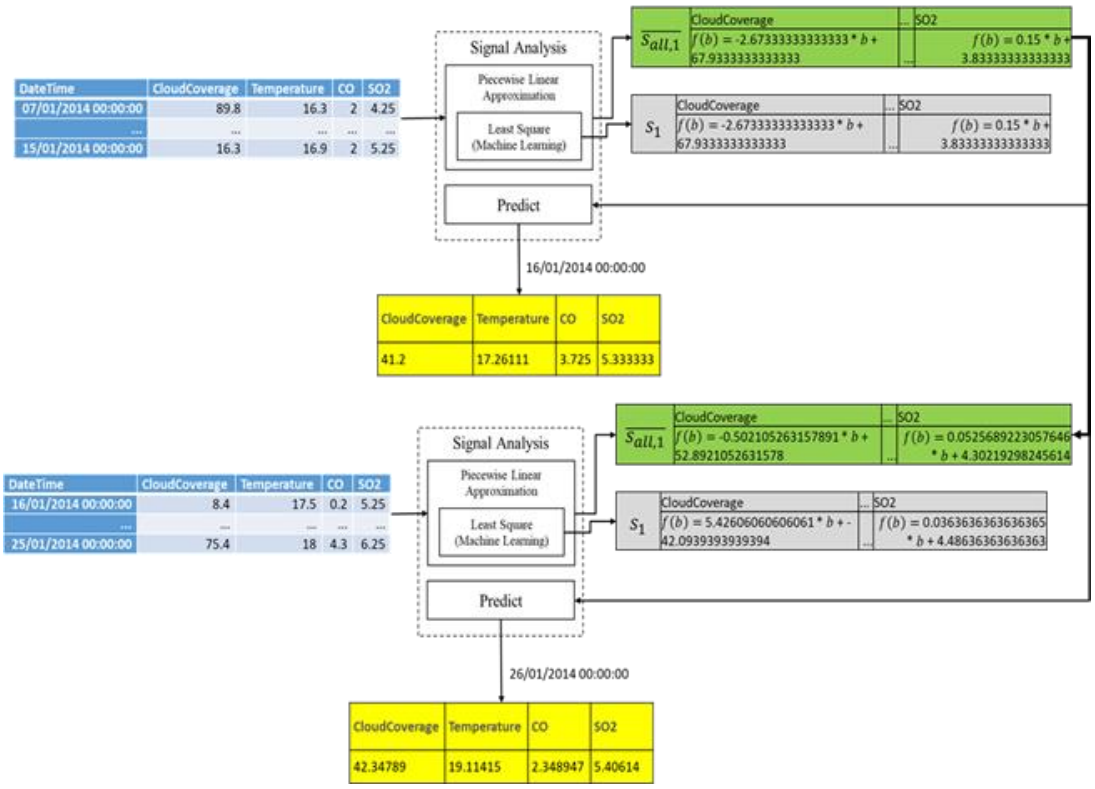


**Figure 2**. PLA process models in D2T and data prediction.

In Signal Analysis, another procedure involves data summarization, where each summary of data influences the subsequent one, as depicted in **Figure 3**. This is undertaken due to the potential occurrence of parameter anomalies within a batch. For instance, if the minimum temperature for a particular day is recorded as 20°C and the maximum temperature as 30°C, the calculated range might only be 10°C. However, the temperature range could surpass 50°C.

In the rundown cycle there is likewise a measurable occasion following interaction, where any adjustment of the scope of greatest or least qualities, trailed by an adjustment of the typical worth in a boundary will be put away and shown in the news. In view of the outline

DOI: https://doi.org/10.17509/jcs.v5i2.70799
e- ISSN 2723-4088

results, outrageous sign following is performed. An occasion is supposed to be outrageous when the worth of increment or lessening between the information to I and I + 1 surpasses 82% of the worth of the information range.

Then again, information is delegated a rehashed occasion on the off chance that the sequential event of information lines with a similar worth surpasses 10% of the complete number of information columns. For example, in a dataset traversing one year with everyday spans (365 columns), information falling into the rehashed occasion classification comprises exclusively of values that happen continuously for more than 36.5 information lines.

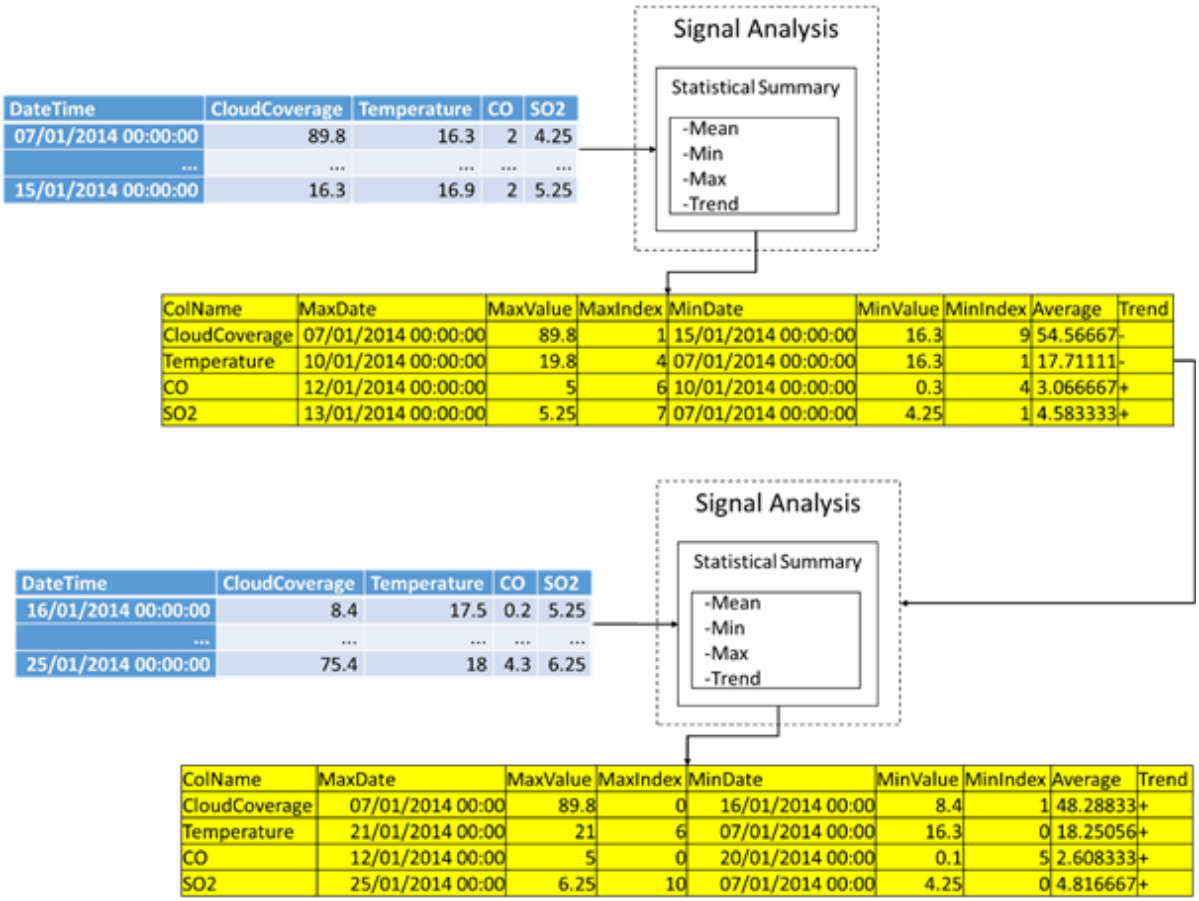


**Figure 3**. D2T data compact process model.

### 2.3. Data Interpretation

Because the news generated is general news where any data can be input (if the data is time series and follows the input data format), then at the Data Interpretation stage, any user can customize this process, but with limitations can only be interpreted data with Fuzzy membership function and Crisp membership function. Where users only need to enter parameters into the ParameterList.csv file in the corpus folder and enter the membership function file in the folder with the format name [parameter]Adjective.csv.

In this examination, the translation of accessible information, for example, AirQuality, WindSpeed, WindDirection, and CloudCoverage, utilizes the Fresh enrollment capability. In the meantime, Temperature (Ramos-soto *et al*., 2016), Precipitation (Ramos-soto *et al*., 2016), and vague boundaries (general) are deciphered utilizing the Fluffy participation

DOI: https://doi.org/10.17509/jcs.v5i2.70799
e- ISSN 2723-4088

capability. The understanding of general information depends on the GeneralAdjective.csv corpus, where every enrollment not entirely settled by the reach between the base and most extreme qualities got from the information synopsis. This alteration to the Fluffy participation capability for patterns, represented in **Figure 4**, involves setting the base and greatest upsides of the enrollment capability as the base and most extreme qualities in the outline. In this manner, the participation is partitioned by the quantity of classifications in the overall corpus.

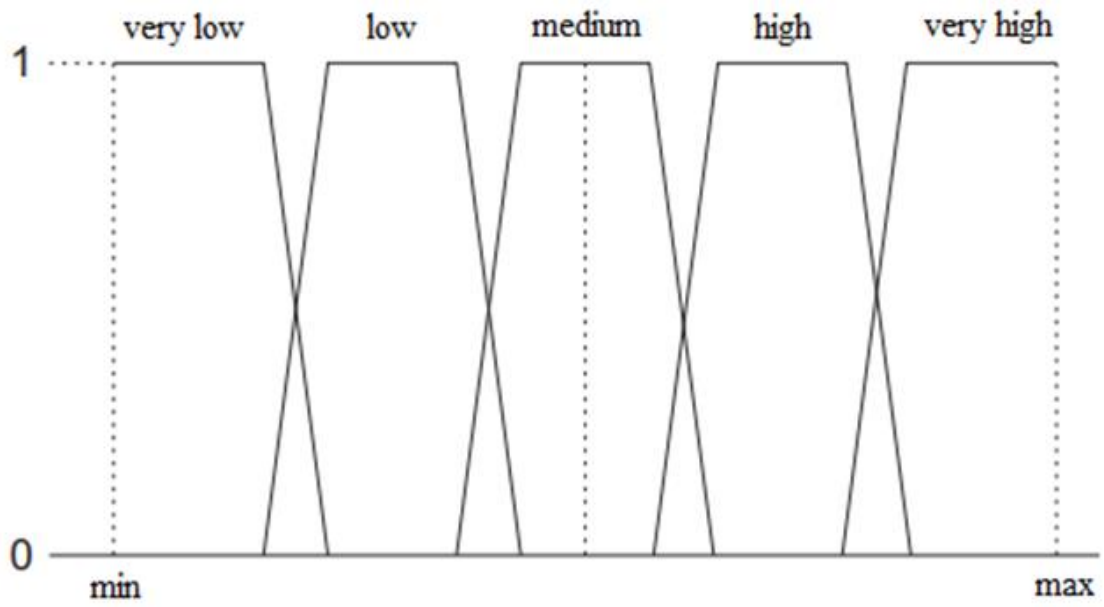


**Figure 4**. Fluffy participation capability for General Boundary.

### 2.4. Document Planning

During this phase, content selection (Content Determination) and text structure formation (Document Structuring) take place. Content selection involves categorizing content into two distinct groups: Routine Message and Significant Event Message. On the other hand, Document Structuring entails devising a framework based on the Target Text generated. In the summary text and description of the current data, content is categorized into these two groups, whereas prediction information exclusively utilizes the Routine Message category.

### 2.5. Document Planning Second

During this stage, there are somewhere around four fundamental undertakings to be finished: Lexicalization, Conglomeration, Alluding Articulation Age, and Design Acknowledgment. In the Lexicalization stage, the cycle includes addressing changes in information, for example, "expanded very from" or "diminished to". Collection happens when various messages are consolidated into a solitary unit utilizing Basic Combination Alluding to Difference Worth. Alluding Articulation Age involves haphazardly producing articulations considering the laid-out corpus. At long last, Construction Acknowledgment includes putting together all satisfied into a predefined structure (Rieter, 2011), trailed by saving the message in JSON design for show on the site.

DOI: https://doi.org/10.17509/jcs.v5i2.70799
e- ISSN 2723-4088

### 2.6. Experiment Design

The trial included contrasting the outcomes and past exploration and producing news from 15 experiments (Ramos-soto *et al*., 2016), as displayed in **Table 1**. Each exploratory outcome was assessed considering four perspectives: Coherence, Calculation Time (Ramos-soto *et al*., 2016), correlation of factual outline results with generally measurable information, and approval of delegate text. Meaningfulness was evaluated utilizing the Lucidness Analyzer application on the Datayze and Grammarly sites. Calculation Time was estimated utilizing the system.time() capability in R. Agent text approval was performed by contrasting the produced data and information perceptions.

**Table 1**. Test-case eksperimen.

| Kode Dataset | Dataset | Source |
|---|---|---|
| KB1 | (Buy) March 2018 | Website Bank Indonesia (https://www.bi.go.id/) kurs buy period March 2018 |
| KB2 | (Buy) April 2018 | Website Bank Indonesia (https://www.bi.go.id/) Course white periods April 2018 |
| KB3 | (Buy) May 2018 | Site Bank Indonesia (https://www.bi.go.id/) purchasing rate period May 2018 |
| SD1 | July 2016 | Day to day stretch information from the site https://solargis.com. For one year from 01 July 2016 to 31 July 2016 |
| SD2 | August 2016 | Day to day span information from the site https://solargis.com. For one year from 01 August 2016 to 31 August 2016 |
| SD3 | September 2016 | Day to day span information from the site https://solargis.com. For one year from 01 September 2016 to 30 September 2016 |
| SD4 | October 2016 | Day to day span information from the site https://solargis.com. For one year from 01 October 2016 to 31 October 2016 |
| SD5 | November 2016 | Day to day span information from the site https://solargis.com. For one year from 01 November 2016 to 30 November 2016 |
| SD6 | December 2016 | Day to day span information from the site https://solargis.com. For one year from 01 December 2016 to 31 December 2016 |
| KK1 | 2014-2015 | Site www.MeteoGalicia.gal, for one year in the period 2014-2015 |
| KK2 | 2015-2016 | Site www.MeteoGalicia.gal, for one year in the period 2015-2016 |
| KK3 | 2016-2017 | Site www.MeteoGalicia.gal, for one year in the period 2016-2017 |

## 3. RESULTS AND DISCUSSION

### 3.1. Comparison of System Output with related research

To facilitate comparison of output with related studies such as DWP (Muthukrishnan, 2005), Ramos study, and others, researchers used data on DWP study Muthukrishnan, 2005),

DOI: https://doi.org/10.17509/jcs.v5i2.70799
e- ISSN 2723-4088

namely climatological data from MeteoGalician stations. A comparison of outputs can be seen in **Table 2**.

Based on **Table 1**, obviously how much satisfied is expanding. In any case, literarily, this application doesn't proceed as well as DWP's result with regards to clarification. The explanation is that this application is intended for general information, permitting it to produce news from any informational index if the information follows the info design. Conversely, DWP's exploration requires explicit info information (boundaries). DWP's review utilizes two kinds of information: climatology and air quality, bringing about satisfied that has two areas. Ramos' exploration, then again, centres exclusively around air quality and wind speed. Goldberg's review presents messages about wind and snow independently (Kittredge & Driedger, 1994), while Gkatzia's exploration relates just to sky conditions. This demonstrates that recently evolved D2T frameworks were intended for explicit information types instead of general information, which proposes that these frameworks probably won't work as expected whenever furnished with various time series information.

**Table 2.** System output comparison.

| Research | Ouput |
|---|---|
| *Output* | Twists northwest 15 diminishingto light monday evening. Overcast with incidental light snow. Mist patches. Visibilities 2 to 5 nm in snow.<br>Beauty isle.<br>Upper east inlet<br>Upper east coast.<br>Storm cautioning in beauty isle and upper east bay gave.<br>Storm cautioning in upper east coast proceeded.<br>Freezing shower cautioning proceeded.<br>Twists southwest 15 to 20 bunches expanding to west storms 35 |
| *Output* | As for the air quality state, it will be variable despite the fact that is supposed to improve to great, inclined toward by the breeze during the next few days |
| *Outptut* | - Light rian showers are likely<br>- Radiant spans with downpour being conceivable - more uncertain than not.<br>-Sunny with rain being unlikely |
| *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.<br>As indicated by the month to month rundown result, this month was cooler and wetter than normal. With normal number of downpour days, appropriately the absolute downpour so far is well beneath the normal. There was downpour on regularly for 7 days from 02nd to 08th and serious downpour was dropped in 06th. The breeze for the month was light breeze in normal. Normal air quality was permissible. Normal temperature was expanded however 05 th was the coldest day of the month with 13.3 celcius degree temperature. |

DOI: https://doi.org/10.17509/jcs.v5i2.70799
e- ISSN 2723-4088

**Table 2 (continue).** System output comparison.

| Research | Ouput |
|---|---|
| GNG *Output* #1 | As indicated by the dataset (06 Jul 2016 (00:00:00) - 06 Jul 2017 (00:00:00)), with boundaries: Cloud Inclusion, Temperature, Wind Speed, Wind Bearing and Precipitation. Obviously Temperature pattern is diminished and Precipitation pattern is steady yet the rest is expanded Temperature, and Wind Speed boundaries are more Vary than the year before.<br>The Condition of today can be described 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 moderate condition.<br>Based on prediction result Wind Speed, and Rainfall will keep stable at Light Breeze, and moderate, Cloud Coverage, Temperature, and Wind Direction will decrease to Partly cloudy, cold, and South West. A conclusion of the predicted result is half variable will decrease slowly. |
| GNG *Output* #2 | According to the dataset, between 07 Jun 2015 (00:00:00) to 25 May 2016 (00:00:00), with parameters: Cloud Inclusion, Temperature, Wind Speed, Wind Heading and Precipitation. It demonstrates that Precipitation pattern is expanded yet Temperature pattern is diminished and the rest is consistent.<br>In today described that Cloud Coverage in Mostly cloudy condition. Temperature in warm condition. Wind Speed in Light Breeze condition. Wind Direction in South condition. Rainfall in moderate condition.<br>Based on the result of prediction Temperature will decrease significantly to cold, Wind Direction will increase to South West, Cloud Coverage, Wind Speed, and Rainfall will keep stable at Mostly cloudy, Light Breeze, and moderate. A conclusion of the predicted result is half variable will kept stable. |

### 3.2. Experimental Results

As far as intelligibility, an assessment was directed using the Flesch Perusing Straightforwardness Score acquired through the Meaningfulness Analyzer device accessible on the site www.datayze.com, alongside the Grammarly application. Thusly, the discoveries were classified in **Table 3**.

**Table 3**. Clarity perspective estimation results.

| Kode Dataset | *Flesch Readang Straightforwardness Score (Grammarly)* | *Flesch Readang Simplicity Score (Datayze)* |
|---|---|---|
| **KB1** | 44 | 42.13 |
| **KB2** | 42 | 37.07 |
| **KB3** | 33 | 29.11 |
| **SD1** | 28 | 19.59 |
| **SD2** | 29 | 15.23 |
| **SD3** | 30 | 17.3 |
| **SD4** | 27 | 15.72 |
| **SD5** | 23 | 13.09 |
| **SD6** | 28 | 17.29 |
| **KK1** | 60 | 57.77 |
| **KK2** | 59 | 55.17 |
| **KK3** | 63 | 60.39 |
| **Average** | 38.83 | 31.66 |
| **All Average** | 35.245 | |

DOI: https://doi.org/10.17509/jcs.v5i2.70799
e- ISSN 2723-4088

The Computation Time result is derived by executing the system.time() function within the R language, for instance, using syntax like system.time (source ("D2T_Main.R")). This process yields outcomes depicted in **Table 4**.

**Table 4**. Computation time measurement results.

| Kode Dataset | *Running Time (s)* |
|---|---|
| KB1 | 2.01 |
| KB2 | 1.98 |
| KB3 | 2.14 |
| SD1 | 1.84 |
| SD2 | 2.08 |
| SD3 | 2.08 |
| SD4 | 1.98 |
| SD5 | 2.03 |
| SD6 | 2.25 |
| KK1 | 2.44 |
| KK2 | 2.63 |
| KK3 | 2.79 |
| Average | 2.235 |

## 4. CONCLUSION

The production of an Information to-Text framework for streaming information utilizing AI is exceptionally profitable. This system operates autonomously, relying directly on sensor inputs or data from other computers, eliminating the need for user intervention. Additionally, prolonged usage of the system does not significantly increase hard drive usage, as processed data is replaced by models generated through Piecewise Linear Approximation, thus conserving storage space.

This research answers the shortcomings of previous research, namely DWP [7] where the development of the User Interface does not use shiny R packages, but a combination of the Codeigniter framework with PHP, JavaScript, AJAX, and html programming languages with JSON output intermediaries.

The conclusion drawn from the comprehensive analysis of the experiments indicates that the system output effectively reflects the provided data. The study achieved an average Readability score of 35.425, suggesting that the system output is comprehensible to a student level or higher. As to Time, the typical handling time was estimated at 2,235 seconds. Also, the exactness of forecasts using direct models, explicitly PLA with the Most un-Square, still up in the air to be 44%.

For future research, corpus development can be carried out for general cases, or adding corpuses for special cases, such as in Data Interpretation and adding features to detect parameter relationships such as the Association Rule. It used other algorithms to predict streaming data to compare with this study.

DOI: https://doi.org/10.17509/jcs.v5i2.70799
e- ISSN 2723-4088

## 5. AUTHORS' NOTE

The authors declare that there is no conflict of interest regarding the publication of this article. The authors confirmed that the paper was free of plagiarism.

## 6. REFERENCES

Barros, C., Lloret, E., Saquete, E., and Navarro-Colorado, B. (2019). NATSUM: Narrative abstractive summarization through cross-document timeline generation. *Information Processing and Management*, *56*(5), 1775-1793.

Barros, C., Vicente, M., and Lloret, E. (2021). To what extent does content selection affect surface realization in the context of headline generation?. *Computer Speech and Language*, *67*, 101179.

García-Méndez, S., Fernández-Gavilanes, M., Costa-Montenegro, E., Juncal-Martínez, J., González-Castaño, F. J., and Reiter, E. (2019). A system for automatic English text expansion. *IEEE Access*, *7*, 123320-123333.

Gkatzia, D., Lemon, O., and Rieser, V. Information to-text age further develops dynamic under vulnerability. *IEEE Computational Intelligence Magazine, 12*(3), 10-17.

Kittredge, R. I. and Driedger, N. (1994). Utilizing normal language handling to create weather conditions estimates. *IEEE Master. Syst. their Appl., 9*(2), 45-53.

Kondadadi, R., Howald, B., and Schilder, F. (2013, August). A statistical NLG framework for aggregated planning and realization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics 1(Long Papers)*, 1406-1415.

Muthukrishnan, S. Information streams: Calculations and applications. *Computer Science, 1*(2), 117-236.

Ramos-Soto, A., Bugarín, A., and Barro, S. On the job of phonetic depictions of information in the structure of normal language age frameworks. *Fluffy Sets Syst., 285*, 31-51.

Reiter, E. (2011). A Design for information to-text frameworks. *Computational Intelligence, 27*(1), 23-40.

S. Ramos, R. M., Monteiro, D. S., and Paraboni, I. (2020). Personality-dependent content selection in natural language generation systems. *Journal of the Brazilian Computer Society*, *26*(1), 2.

van der Lee, C., Gatt, A., van Miltenburg, E., and Krahmer, E. (2021). Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech and Language*, *67*, 101151.

Vayre, J. S., Delpech, E., Dufresne, A., and Lemercier, C. (2017). Communication mediated through natural language generation in big data environments: the case of Nomao. *Journal of Computer and Communications*, *5*(6), 125-148.

Vicente, M., Barros, C., and Lloret, E. (2018). Statistical language modelling for automatic story generation. *Journal of Intelligent and Fuzzy Systems*, *34*(5), 3069-3079.

DOI: https://doi.org/10.17509/jcs.v5i2.70799
e- ISSN 2723-4088

Yagamurthy, D. N., Azmeera, R., and Khanna, R. (2023). Natural Language Generation (NLG) for Automated Report Generation. *Journal of Technology and Systems*, *5*(1), 48-59.

Zubair, F., and Jafir, A. (2021). Natural Language Generation: Recent Advances and Applications. *International Journal of Advanced Engineering Technologies and Innovations*, *1*(1), 50-74.

DOI: https://doi.org/10.17509/jcs.v5i2.70799
e- ISSN 2723-4088