Auto-generating Textual Data Stories Using Data Science Pipelines

Ruilin Wang
University of Aberdeen
r.wang2.20@abdn.ac.uk

Somayajulu Gowri Sripada
University of Aberdeen
yaji.sripada@abdn.ac.uk

Nigel A Beacham
University of Aberdeen
n.beacham@abdn.ac.uk

ABSTRACT

Understanding a dataset directly is challenging but transforming the results of data analysis into data stories could help people build mental models and understand the dataset easily. In this paper, we present a new framework for data-to-text NLG to generate data stories for specific personas. In order to understand the feasibility of this method and if the human generated story is consistent with the story generated by the data science pipelines, we present two experiments: a data story study with 3 financial experts, 4 Ph.D. students, and 20 Amazon Mechanical Turk workers, which offers several data stories generated by humans; and a validation study involving 39 Amazon Mechanical Turk workers who conducted usability and understandability assessments for 9 high-quality data stories, written by humans and machine. We conduct a qualitative analysis of human-written data stories to determine what people consider when writing data stories and if the human generated story is consistent with the one generated by the data science pipeline. The experimental results show that readers comprehend machine-written data stories as well as they comprehend human-written data stories.

CCS CONCEPTS

• **Computing methodologies** → Artificial intelligence; Natural language processing; Natural language generation.

KEYWORDS

Data science, NLP, Data sensemaking, Data Storytelling

ACM Reference Format:
Ruilin Wang, Somayajulu Gowri Sripada, and Nigel A Beacham. 2021. Auto-generating Textual Data Stories Using Data Science Pipelines. In *2021 4th International Conference on Algorithms, Computing and Artificial Intelligence (ACAI'21), December 22–24, 2021, Sanya, China.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3508546.3508642

1 INTRODUCTION

As digital technology progressed, the availability of data has greatly increased, and data is becoming more and more important in many fields [1]. For example, open research data are heralded as having the potential to increase effectiveness, productivity, and reproducibility in science [2].

However, understanding a dataset is difficult, because some people lack the required data analysis skills, some others lack the background knowledge required to comprehend the dataset and finally most people simply lack the time to work directly with the dataset. There are studies which show that data presented as stories can help people understand data faster because stories convey information in a psychologically efficient format [3]. Currently, data storytelling involves combining data science with information visualization. While information visualization is an excellent presentation medium, complementing visualizations with textual narratives is likely to be more effective. For example, Calliope is a system that can automatically generate visual data stories from spreadsheets, but there are some performance bottlenecks in its design and implementation [4]. When data understanding is realized, people's mental models are modified. A mental model is considered to be a cognitive structure that forms the basis of reasoning, decision-making, and behavior [5]. The work reported in this paper hypothesizes that data stories expressed in plain English enable people to build effective mental models.

There has been a body of prior work known as data-to-text, in the field of Natural Language Generation (NLG), to automatically generate textual data stories [6] [7] [8] [9] [10] [11] [12]. The two main contributions of these prior works are (1) a rule-based data-to-text pipeline architecture [13] and (2) a methodology to acquire the required rule-based knowledge. The data-to-text methodology employs primarily a corpus-based knowledge engineering process; human-written data stories are aligned with their underlying datasets and rules that map data to text are manually extracted from such parallel data-to-text corpus. Applying this data-to-text methodology in practice is challenging because it involves a steep learning curve for developers, and it involves high manual costs associated with acquiring and maintaining all the required knowledge. More importantly, this methodology fails to take advantage of the recent developments in data science. Because a significant part of the data-to-text pipeline is devoted to extracting insights from data, data science should inform both the data-to-text pipeline as well as the data-to-text methodology.

The data-to-text pipeline is made up of two parts, data-to-information and information-to-text. This paper focuses more on the data-to-information part, a standard template-based solution is adopted for information-to-text. This study will bring data science and NLG together to auto-generate data stories. To make the data stories relevant to the audience, persona types and associate questions these personas ask of the underlying data were defined. The data stories will answer these questions for the specific personas.

ACAI'21, December 22–24, 2021, Sanya, China

ACM ISBN 978-1-4503-8505-3/21/12…$15.00
https://doi.org/10.1145/3508546.3508642

Table 1: Sample of the Dataset About Advertising Costs on Different Media and the Sales

TV advertising in $	Radio advertising in $	Newspaper Advertising in $	Sales in hundreds of $
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
...	...	...	...

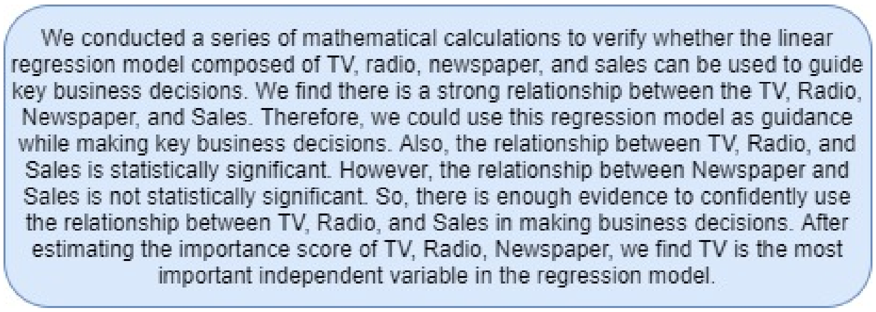

Figure 1: An Example of the Outcome from the Generation Processing

Thus, our work ties datasets to specific users (personas) through reusable data science pipelines that build specific models to answer specific questions. Our goal is to automatically create practical, user-centered data stories for data, reflecting the needs and expectations of different user personas. An example input dataset to our data story generator is shown in Table 1 and Figure 1 shows its corresponding data story.

2 A CONCEPTUAL FRAMEWORK FOR INTEGRATING DATA SCIENCE WITH DATA-TO-TEXT NLG

The consensus data-to-text pipeline architecture is shown in Figure 2. The first two modules (signal analysis and data interpretation) in the architecture are responsible for computing information from data and the last two modules are responsible for linguistically expressing this information. Since data science pipelines are designed to compute insightful information from data, the first part of the data-to-text pipeline should be replaced with data science pipelines to take advantage of the abundant data science tools and libraries of algorithms that are being developed in the data science community. Moreover, a consensus methodology emerged for carrying out data science and data-to-text methodology needs to be adapted to take advantage of the data science methodology. This section describes how the current work integrates data science with data-to-text NLG both at the pipeline level as well as at the level of methodology.

2.1 Adapting the Data-to-Text Pipeline

The consensus data-to-text pipeline architecture is shown in Figure 2-a. Figure 2-b shows a typical data science pipeline adapted from the pipeline suggested by Jones [5].

More recently with the advent of interpretable machine learning (IML) methods, it is possible to query models. Using these IML methods, we propose the new data-to-text pipeline shown in Figure 2-c. The new data-to-text pipeline could exploit the wide range of data science tools and libraries developed in the data science community in the first three modules of the new data-to-text pipeline. More importantly data science pipelines developed for one project could be reused across several data-to-text projects. This reusability is the main advantage of the proposed new pipeline in comparison to the traditional pipeline which often involves bespoke modules with poor reusability.

2.2 Adapting the Data-to-Text Methodology

Methodologically, data-to-text pipelines are designed using guidance from human-written data stories. A human written data story is made up of several sentences and paragraphs (discourse structure). Each of the sentences expresses a message, a unit of information. An analysis of all the sentences in a data story reveals the informational content of a data story. The first part (signal analysis and data interpretation) of the traditional data-to-text pipeline is then designed to extract the required information from the input data. In practice, algorithms used for data analysis tend to be aligned with individual messages (units of information) in the data story making the development very complicated. On the other hand, data science methodology to insight discovery from data is driven by business questions which express informational needs of business users. Taking this question-driven data science methodology into account we propose a new data-to-text methodology where data stories are modelled in terms of answers to business questions rather than linguistic units of sentences and paragraphs. This means, the new proposed data-to-text methodology makes business questions as the central concept driving the design of pipelines which allows the design of the data science part and the linguistic part of the pipeline to be decoupled. Since each of these parts is designed to provide answers to the same set of questions, the new methodology ensures consistency among all the modules of the pipeline. In addition, the new methodology has the advantage of deploying two

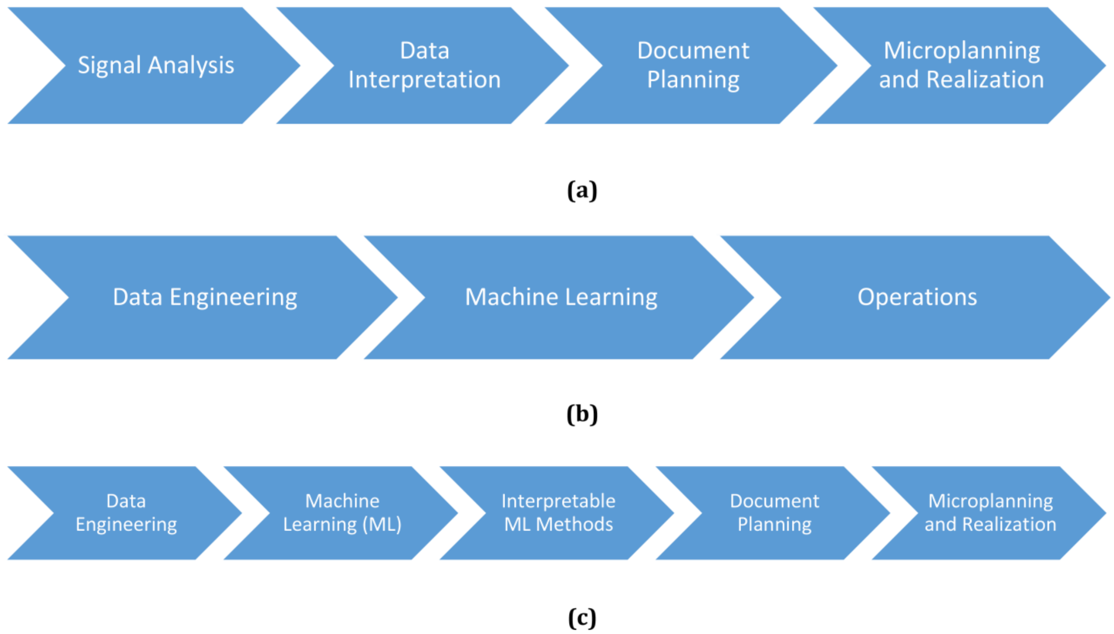

Figure 2: (a)The Data-to-Text Pipeline [13]-(b) The Data Science Pipeline [5] (c). The Proposed Data-to-Text Pipeline

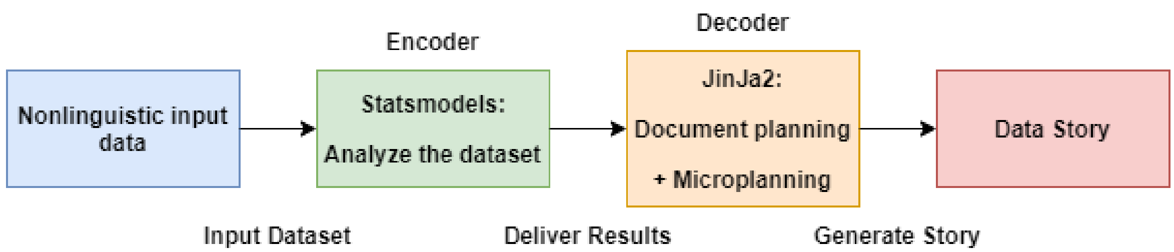

Figure 3: Prototype Implementation

different groups of developers, data scientists and natura language engineers.

3 A PROTOTYPE IMPLEMENTATION OF THE NEW DATA-TO-TEXT FRAMEWORK

We first propose a prototype based on natural language generation, as Figure 3 shows, the data will be read by the automatic story generation system. We use Statsmodels as the encoder because it is a very comprehensive package for statistical and economic research when using python [14] [15], it will try to fit the data with a suitable model and calculate the corresponding important information, such as R-square and P-value. The analysis results will be received by JinJa2 templates, as the decoder of the system, in the document planning stage, it will plan the content and choose the outline based on the fitting model from Statsmodels, in microplanning stage, it will refer expression, choose suitable words based on the personas and other settings, then aggregate them together to convert specifications to a real textual data story.

4 EVALUATION

For solving the above problem, it is necessary to understand how the mental model exactly works. Figure 4 is a general cognitive process, when we comprehend the dataset analysis results, we will first extract important information, then build a mental model to create easy to understand stories based on our own knowledge and experience, and finally judge whether to accept or dismiss the results [16]. However, as mentioned above, this is a long and energy-consuming process, and whether readers can understand the analysis results depends largely on readers' knowledge/experience and the time available.

Naturally, some studies have found people can understand the analysis results faster if they have a summary or story of the data because a good data summary can represent the core idea of data and effectively convey the meaning of it [17], and stories are formed by specific facts supported by data through appropriate logical connections, which can clearly highlight and emphasize key information and avoid ambiguity [18]. However, most of the data stories are written by human experts, and they still need to go through

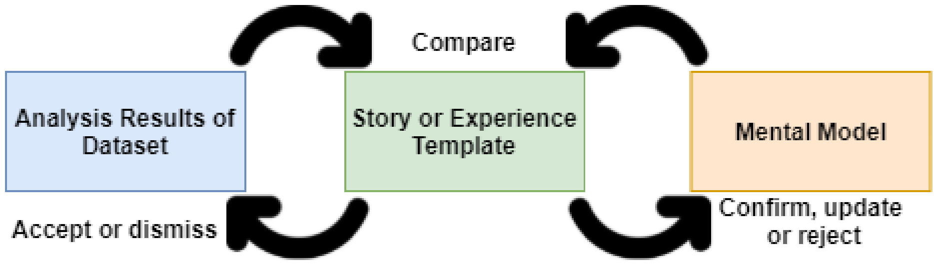

Figure 4: Sensemaking Processing

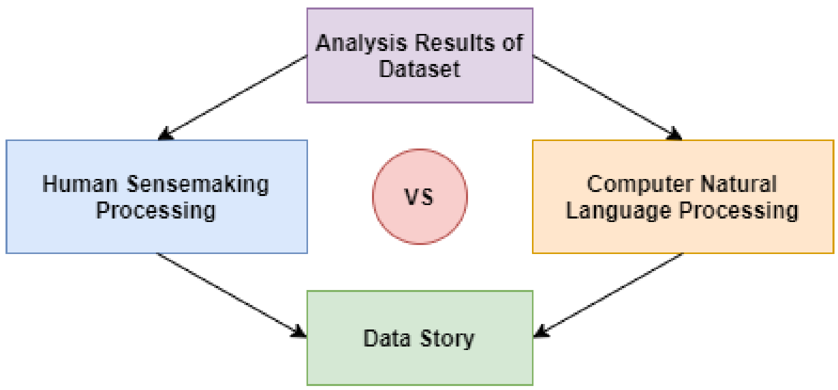

Figure 5: Two Methods of Data Story Generation Processing

the manual process to write the data story. Therefore, as Figure 5 shows, to accelerate this process and reduce the cost of learning or reviewing the knowledge needed to understand the data, we propose a new framework for data-to-text NLG that aims to automatically generate data stories by computer to accelerate the construction of mental models, and design experiments to verify our main hypothesis - could data stories generated by computers be no different from data stories written by humans in helping people build mental models and understand data?

In summary, we want to use computers to automatically generate data stories, it aims to help different personas quickly understand datasets since data stories could help people build mental models quickly. For the implementation of this method, we use a new framework that combines IML methods with traditional data-to-text pipelines. The experiments in next section will verify the feasible of the method.

5 DESIGN OF THE EXPERIMENT

Figure 6 is an overview of the two experimental processes. Experiment 1 not only provides us with materials for human-generated data stories but also allows us to better analyze if the attributes of human-generated data stories are consistent with computer pipeline templates. Experiment 2 verifies whether computer-generated data stories could be the same as human-generated data stories in terms of ease of understanding.

5.1 Experiment 1

This experiment invites 3 financial experts, 4 Ph.D. students, and 20 Amazon Mechanical Turk workers to participate, the main method of Experiment 1 is questionnaire investigation. The purpose for the participants is to read an analysis result of a dataset and then answer three questions which will guide participants to write a data story that aims to help the persona quickly understand the key information of the dataset.

To help participants write a good data story, we use words that are easier to be understood by most people instead of some professional words, such as using 'short report' instead of 'data story', 'main audience' instead of 'persona'. We also produce a guideline that consists of three questions that affect people's decision-making. Generally, the questions are about whether the relationship between independent variables and dependent variables is strong, whether independent variables have statistical significance to dependent variables, and which independent variable is the most important.

5.2 Experiment 2

This experiment invites 39 Amazon Mechanical Turk workers to participate, we build three computer pipeline templates to generate three data stories based on the same dataset and the same persona as Experiment 1 used, also select six high-quality data stories from Experiment 1, one from a Ph.D. student, two from Amazon Mechanical Turk workers, three from Experts. Then we let each participant randomly read one from the nine data stories and ask them to answer several questions. The first question for providing

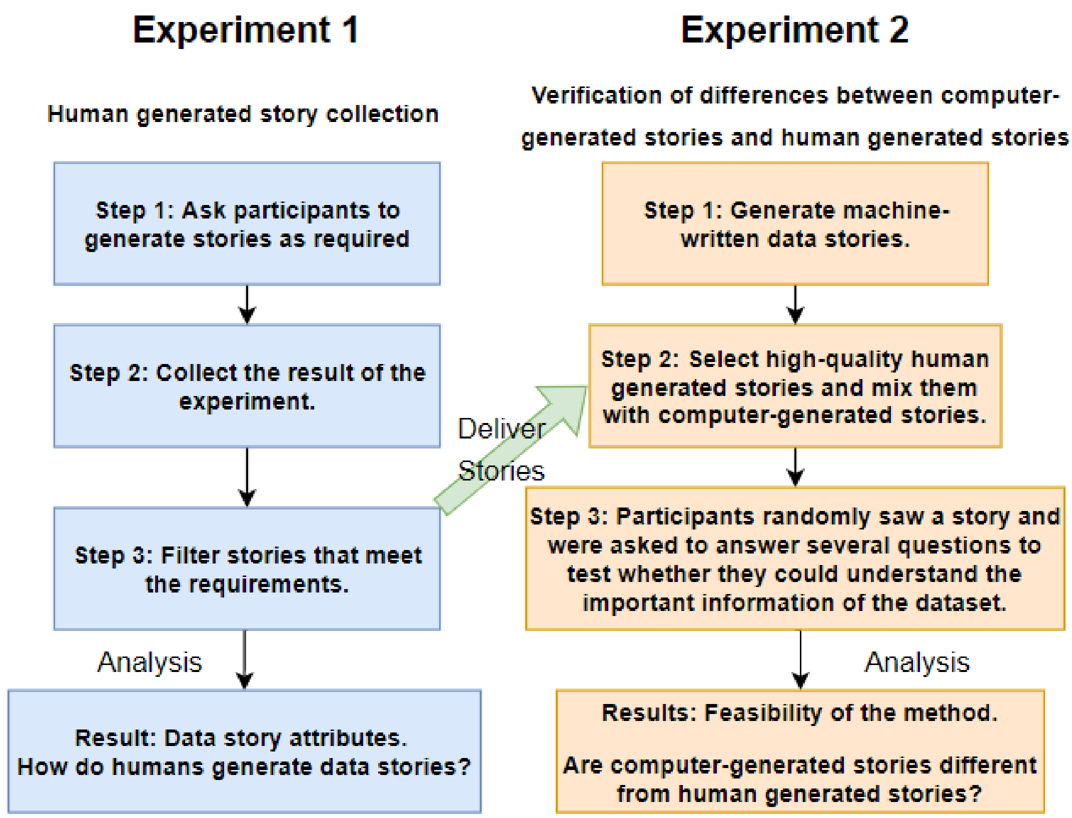

Figure 6: Experimental Process

the data story number, then answer four questions to test if participants can understand the key information of the dataset analysis result by just reading the data story without knowing the analysis results. The four questions are whether the relationship between independent variables and dependent variables is strong, which independent variables have statistical significance to dependent variables, which independent variable is the most important, and who you think is the main audience of this data story. This will verily if computer pipelines generated data stories could be as good as human-generated data stories.

6 RESULTS FROM THE EXPERIMENTS

6.1 Results from Experiment 1

After collecting the results from Experiment 1, we first screened the experimental results, removed the unqualified stories, then, we made statistics of each story length, and statistics on the number of words described each part. Also, we counted each word in all the stories, we exclude punctuation and common words such as "it", "and", "I", then listed the high-frequency words, to determine whether human-generated stories are consistent with data science pipelines, we also listed the high-frequency words of computer pipeline templates. Last but not the least, we found most people will write an introduction for the data story to briefly describe the purpose of the story, but few of them will write a conclusion to summarize the main idea of the story. After comparison, we find data story attributes written by humans are highly consistent with machine-written data stories.

6.2 Results from Experiment 2

Answers from the participants of Experiment 2 have been analyzed to create the data, we classify the statistical results according to the author type as Table 2 shows. Obviously, the error rate of computer-generated stories is lower than the other two. However, whether there are significant differences between the data stories generated by different authors needs further verification in the following section.

6.3 Statistical Significance and Effect Size

To verify could the data story generated by computers is no different from the data story written by humans, we ran the one-way Anova test and Tukey HSD to confirm whether the reader's understanding errors are related to the author of the data story. The data we used is the experiment result we collected from Experiment 2, in short, the data is the detailed content of Table 2, recording which data story each participant read and whether their answer to each question is correct or not.

Table 2: Error rate of each type of data story

Report Author	Read times	Error rate on Question 2 (%)	Error rate on Question 3 (%)	Error rate on Question 4 (%)	Error rate on Question 5 (%)
MTworkers and PhD student	16	25%	25%	0%	68.75%
Computer pipelines	16	6.25%	25%	0%	18.75%
Experts	7	28.57%	57.14%	0%	42.86%

Table 3: One-way Anova test Results on R-squared question

group1	group2	meandiff	p-adj	lower	upper	reject
MTworkers	Computers	0.1875	0.4413	-0.1829	0.5579	False
MTworkers	Experts	0.0268	0.9	-0.448	0.5016	False
computers	Experts	-0.1607	0.6757	-0.6355	0.3141	False

Table 4: One-way Anova test Results on P-value question

group1	group2	meandiff	p-adj	lower	upper	reject
MTworkers	Computers	0.0	0.9	-0.4303	0.4303	False
MTworkers	Experts	-0.3393	0.3018	-0.8908	0.2123	False
computers	Experts	-0.3393	0.3018	-0.8908	0.2123	False

Table 5: One-way Anova test Results on Important Score question

group1	group2	meandiff	p-adj	lower	upper	reject
MTworkers	Computers	0.0	0.5566	0.0	0.0	False
MTworkers	Experts	0.0	0.5566	0.0	0.0	False
computers	Experts	0.0	0.5566	0.0	0.0	False

The null hypothesis used in the one-way Anova test is "A data story generated by Experts, Amazon Mechanical Turk Workers, and Computer pipelines is no significant difference for readers to understand", and the alternate hypothesis is "A story generated by one or some of the authors has significant differences from a story generated by others."

The results as the Table 3, Table 4, Table 5, and Table 6 shows, column 'group1' and 'group2' means the author group, column 'p-adj' shows the adjusted P-value, column 'lower' and 'upper' are the lower and upper bounds of the 95% confidence interval, on the 'reject' column, 'False' means the test result agrees with the null hypothesis, 'True' means the test result rejects the null hypothesis. As we can see, for people misunderstanding the information of R-squared (Question 2), P-value (Question 3), and Important Score (Question 4), the one-way Anova test agree with the null hypothesis which is a data story generated by Experts, Amazon Mechanical Turk Workers, and Computer pipelines is no significant difference for readers to understand, however, for people misunderstanding the information of who is the main audience (Question 5) of that data story, the one-way Anova test shows that computer pipelines are no different from experts, but computer pipelines have significant differences from Amazon Mechanical Turk worker.

Although, we have verified the null hypothesis by one-way Anova test. But further, we also need to verify the effect size between the three groups. Because we only know there is a significant difference or not, but we do not know if the difference is small, medium, or large [19] [20]. More specifically, the effect size allows people to understand the size of the difference between groups [21]. Effect size also can increase the comparison ability of results, and make the results have practical significance [22].

Therefore, through the Cohen D test, we get the effect sizes between the three groups as Table 7. Each column shows which two groups are comparing, and each row shows which problem they compared for. Then we can compare Table 7 to Table 8. 'Relative size' represents the size of the real difference, 'Cohen D effect size' represents the effect size we got from the Cohen D test (Table 7).

For example, the Cohen D effect size on Table 7 first column and first row is 0.056 which is less than 0.2 in Table 8. So, we could make a conclusion: Let one group read experts generated data stories, and another group read Amazon Mechanical Turk workers generated

Table 6: One-way Anova test Results on Main Aduience question

group1	group2	meandiff	p-adj	lower	upper	reject
MTworkers	Computers	0.5	0.0107	0.1033	0.8967	True
MTworkers	Experts	0.2589	0.4371	-0.2496	0.7674	False
computers	Experts	-0.2411	0.4866	-0.7496	0.2674	False

Table 7: Cohen D test Results

	Experts VS MT Workers	Computers VS MT Workers	Computers VS Experts
Number of correct answers on Question R-Squared	0.056	0.451	0.413
Number of correct answers on Question P-Value	-0.683	0.000	0.683
Number of correct answers on Question Main Audience	0.523	1.130	0.542

Table 8: Relative Size and Effect Size

Relative Size	Cohen D Effect Size
Very Small	0.0
Small	0.2
Medium	0.5
Large	0.8
Very Large	1.4

data stories. After let two groups answer the question about R-squared. The difference in the number of correct answers between the two groups is very small.

7 DISCUSSION OF RESULTS

The main conclusion from our experiments is that machine-written data stories work as well as human-written (experts as well as Amazon Mechanical Turk workers) ones in effectively communicating the required information (answers to key questions) from an input dataset to the readers. Although several earlier studies too demonstrated this using rule-based data-to-text technology, the proposed work achieves this by following a new data-to-text framework that integrates recent advancements in data science and interpretable machine learning into data-to-text NLG both at the pipeline level as well as at the methodology level. Because the proposed pipeline design is driven by business questions major portions of the pipeline could be reused in contexts where similar questions are answered. Moreover, the link between questions and pipelines is extended to include user personas which enable developers to modularize pipeline design to link users to questions, questions to pipelines that compute the required answers from input data. Our experiments show that machine-written stories help users identify the audience for whom the stories have been written.

The experiments also provide evidence that Amazon Mechanical Turk workers could write good quality data stories. Although the current work uses these human-written data stories as part of experiments, they could also be used to train neural NLG models.

8 CONCLUSION AND FUTURE WORK

This paper presents a new framework for data-to-text NLG that takes advantage of the recent developments in data science both at the pipeline level as well as at the methodology level. Our evaluation experiments involving a prototype developed following the new framework show that users comprehend machine-written data stories as well as they comprehend human-written data stories. Because the pipelines, following the new framework, are designed under the guidance of user personas and their questions, the new framework improves the reusability of pipelines. Our future work will focus on strengthening the framework further by experimenting with a broad range of business questions seeking answers from larger datasets using a rich set of data science pipelines. We aim to run more experiments to confirm the equivalence of machine-written and human-written data stories. In addition, we intend to collect experimental evidence for improved reusability of data science modules across disparate data-to-text applications.

REFERENCES

[1] Manyika, M. G. I. J., Chui, M., Groves, P., Farrell, D., Kuiken, S. V., Doshi, E. A., 2013. Open data: Unlocking innovation and performance with liquid information. https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information
[2] Gregory, K., Cousijn, H., Groth, P.T., Scharnhorst, A., Wyatt, S., 2018. Understanding data retrieval practices: a social informatics perspective. https://pure.knaw.nl/ws/portalfiles/portal/6148431/1801.04971.pdf
[3] Gershon, N. and Page, W. 2001. What storytelling can do for information visualization. Commun. ACM, 44(8):31–37, 2001.
[4] Danqing Shi, Xinyue Xu, Fuling Sun, Y ang Shi and Nan Cao. 2020. Calliope: Automatic Visual Data Story Generation from a Spreadsheet. IEEE TRANSACTIONS ON VISUALIZA TION AND COMPUTER GRAPHICS, VOL. 27, NO. 2 https://arxiv.org/abs/2010.09975
[5] Jones, M. Tim. 2018. "Data, structure, and the data science pipeline." An introduction to data science, Part 1, IBM developerWorks, February 1. Accessed

2021-10-22.

[6] Dimitra Gkatzia, Oliver Lemon, and Verena Rieser. 2016. Natural language generation enhances human decisionmaking with uncertain information. In ACL

[7] Portet, F., Reiter, E., Hunter, J., and Sripada, S. 2007. Automatic Generation of Textual Summaries from Neonatal Intensive Care Data. In: Bellazzi, Riccardo, Ameen Abu-Hanna and Jim Hunter (Ed.), 11th Conference on Artifical Intelligence in Medicine (AIME 07), pp. 227-236.

[8] Reiter, E. Sripada, S. Hunter, J. Yu, J. and Davy, I. 2005. Choosing words in computer-generated weather forecasts. Artificial Intelligence, 167:137– 169.

[9] Riedl, M. O., and Young, R. M. 2010. Narrative planning: Balancing plot and character. InJAIR.

[10] Sripada, S. and Gao, F. 2007. Summarizing dive computer data: A case study in integrating textual and graphical presentations of numerical data. In Proceedings of Workshop on Multimodal Output Generation, volume CTIT Proceedings of the Workshop on Multimodal Output Generation, pages 149–157.

[11] Sripada, S. Reiter, E. Davy, I. 2003. SumTime-Mousam: Configurable marine weather forecast generator Expert Update, pp. 4-10

[12] Turner, R. Sripada, S. and Reiter, E. 2009. Generating approximate geographic descriptions. In European Workshop on Natural Language Generation, pages 42–49.

[13] Reiter, E. 2007. An Architecture for data-to-text systems.Proceedings of the European Workshop of Natural Language Generation 2007, 97–104.

[14] Seabold, Skipper and Perktold, Josef. 2010. Statsmodels: Econometric and Statistical Modeling with Python. Proceedings of the 9th Python in Science Conference.

[15] Wickham *et al.*, 2019. Welcome to the Tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

[16] Grolemund, G., Wickham, H. 2014. A cognitive interpretation of data analysis. Int. Stat. Rev. 82(2): 184– 204.

[17] Koesten, L., Simperl, E., Kacprzak, E., Blount, T., Tennison, J. 2018. Everything you always wanted to know about a dataset: studies in data summarisation. CoRR arXiv:1810.12423

[18] Lee, B. Henry Riche, N. Isenberg, P. and Carpendale, S. 2015. "More than telling a story: Transforming data into visually shared stories", IEEE Comput. Graph. Appl., vol. 35, no. 5, pp. 84-90, Sep./Oct. 2015, [online] Available: http://dx.doi.org/10.1109/MCG.2015.99

[19] Cohen, J. 1988. Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

[20] Cohen, J. 1994. The earth is round (p< .05). American Psychologist, 49, 997-1003.

[21] McLeod, S. A. 2019. What does effect size tell you? Simply psychology: https://www.simplypsychology.org/effect-size.html

[22] Kotrlik, J. W., & Williams, H. A. 2003. The incorporation of effect size in information technology, learning, and performance research. Information Technology, Learning, and Performance Journal, 21(1), 1–7.