DATA ANALYTICS 2023

The Twelfth International Conference on Data Analytics

ISBN: 978-1-68558-111-4

September 25 - 29, 2023

Porto, Portugal

DATA ANALYTICS 2023 Editors

Les Sztandera, Thomas Jefferson University, USA

Laura Garcia, Universidad Politécnica de Cartagena, Spain

DATA ANALYTICS 2023

Forward

The Twelfth International Conference on Data Analytics (DATA ANALYTICS 2023), held between September 25th and September 29th, 2023, continued a series of international events on fundamentals in supporting data analytics, special mechanisms, and features of applying principles of data analytics, application-oriented analytics, and target-area analytics.

Processing of terabytes to petabytes of data or incorporating non-structural data and multi-structured data sources and types require advanced analytics and data science mechanisms for both raw and partially processed information. Despite considerable advancements on high performance, large storage, and high computation power, there are challenges in identifying, clustering, classifying, and interpreting a large spectrum of information.

We take here the opportunity to warmly thank all the members of the DATA ANALYTICS 2023 technical program committee, as well as all the reviewers. The creation of such a high-quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and effort to contribute to DATA ANALYTICS 2023. We truly believe that, thanks to all these efforts, the final conference program consisted of top-quality contributions. We also thank the members of the DATA ANALYTICS 2023 organizing committee for their help in handling the logistics of this event.

We hope that DATA ANALYTICS 2023 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the field of data analytics.

DATA ANALYTICS 2023 Chairs

DATA ANALYTICS 2023 Steering Committee

Wolfram Wöß, Institute for Application Oriented Knowledge Processing | Johannes Kepler University, Linz, Austria
Kerstin Lemke-Rust, Hochschule Bonn-Rhein-Sieg, Germany
George Tambouratzis, Institute for Language and Speech Processing, Athena R.C., Greece
Les Sztandera, Thomas Jefferson University, USA
Ivana Semanjski, Ghent University, Belgium
Sandjai Bhulai, Vrije Universiteit Amsterdam, The Netherlands

DATA ANALYTICS 2023 Publicity Chairs

Laura Garcia, Universitat Politecnica de Valencia, Spain
Lorena Parra Boronat, Universitat Politecnica de Valencia, Spain

DATA ANALYTICS 2023
Committee

DATA ANALYTICS 2023 Steering Committee

Wolfram Wöß, Institute for Application Oriented Knowledge Processing | Johannes Kepler University, Linz, Austria
Kerstin Lemke-Rust, Hochschule Bonn-Rhein-Sieg, Germany
George Tambouratzis, Institute for Language and Speech Processing, Athena R.C., Greece
Les Sztandera, Thomas Jefferson University, USA
Ivana Semanjski, Ghent University, Belgium
Sandjai Bhulai, Vrije Universiteit Amsterdam, The Netherlands

DATA ANALYTICS 2023 Publicity Chairs

Laura Garcia, Universitat Politecnica de Valencia, Spain
Lorena Parra Boronat, Universitat Politecnica de Valencia, Spain

DATA ANALYTICS 2023 Technical Program Committee

Arianna Agosto, University of Pavia, Italy
Irfan Ahmed, Virginia Commonwealth University, USA
Raed Ibrahim Alharbi, University of Florida, USA
Madyan Alsenwi, Kyung Hee University, Global Campus, South Korea
Katie Antypas, Lawrence Berkeley National Laboratory, USA
Najet Arous, University of Tunis Manar, Tunisia
Abderazek Ben Abdallah, The University of Aizu, Japan
Sadok Ben Yahia, Tallinn University of Technology, Estonia
Soumia Benkrid, Ecole Nationale Supérieure d'Informatique, Algeria
Flavio Bertini, University of Parma, Italy
Nik Bessis, Edge Hill University, UK
Sandjai Bhulai, Vrije Universiteit Amsterdam, Netherlands
Jean-Yves Blaise, UMR CNRS/MC 3495 MAP, Marseille, France
Jan Bohacik, University of Zilina, Slovakia
Ozgu Can, Ege University, Turkey
Wandercleiton Cardoso, University of Genoa, Italy
Julio Cesar Duarte, Instituto Militar de Engenharia, Rio de Janeiro, Brazil
Junghoon Chae, Oak Ridge National Laboratory, USA
Richard Chbeir, Université de Pau et des Pays de l'Adour (UPPA), France
Daniel B.-W. Chen, Monash University, Australia
Yujing Chen, VMware, USA
Tushar Chugh, Google, USA
Giovanni Costa, ICAR-CNR, Italy
Bi-Ru Dai, National Taiwan University of Science and Technology, Taiwan
Mirela Danubianu, University "Stefan cel Mare" Suceava, Romania
Monica De Martino, National Research Council - Institute for Applied Mathematics and Information Technologies (CNR-IMATI), Italy

Kerstin Lemke-Rust, Hochschule Bonn-Rhein-Sieg, Germany
Clement Leung, Chinese University of Hong Kong, Shenzhen, China
Yuening Li, Texas A&M University, USA
Ninghao Liu, Texas A&M University, USA
Weimo Liu, Google, USA
Fenglong Ma, Pennsylvania State University, USA
Ruizhe Ma, University of Massachusetts Lowell, USA
Massimo Marchiori, University of Padua, Italy / European Institute for Science, Media and Democracy, Belgium
Mamoun Mardini, College of Medicine | University of Florida, USA
Miguel A. Martínez-Prieto, University of Valladolid, Spain
Alfonso Mateos Caballero, Universidad Politécnica de Madrid, Spain
Archil Maysuradze, Lomonosov Moscow State University, Russia
Abbas Mazloumi, University of California, Riverside, USA
Gideon Mbiydzenyuy, Borås University, Sweden
Ryan McGinnis, Thomas Jefferson University, USA
Letizia Milli, University of Pisa, Italy
Yasser Mohammad, NEC | AIST | RIKEN, Japan / Assiut University, Egypt
Thomas Morgenstern, University of Applied Sciences in Karlsruhe (H-KA), Germany
Lorenzo Musarella,University Mediterranea of Reggio Calabria, Italy
Azad Naik, Microsoft, USA
Roberto Nardone, University Mediterranea of Reggio Calabria, Italy
Alberto Nogales, Universidad Francisco de Victoria | CEIEC research center, Spain
Panagiotis Oikonomou, University of Thessaly, Greece
Ana Oliveira Alves, Polytechnic Institute of Coimbra & Centre of Informatics and Systems of the University of Coimbra, Portugal
Riccardo Ortale, Institute for High Performance Computing and Networking (ICAR) - National Research Council of Italy (CNR), Italy
Moein Owhadi-Kareshk, University of Alberta, Canada
Yu Pan, University of Nebraska-Lincoln, USA
Massimiliano Petri, University of Pisa, Italy
Hai Phan, New Jersey Institute of Technology, USA
Antonio Pratelli, University of Pisa, Italy
Yiming Qiu, Rice University, USA
Víctor Rampérez, Universidad Politécnica de Madrid (UPM), Spain
Andrew Rau-Chaplin, Dalhousie University, Canada
Ivan Rodero, Rutgers University, USA
Sebastian Rojas Gonzalez, Hasselt University / Ghent University, Belgium
Antonia Russo,University Mediterranea of Reggio Calabria, Italy
Gunter Saake, Otto-von-Guericke University, Germany
Bilal Abu Salih, Curtin University, Australia
Burcu Sayin, University of Trento, Italy
Andreas Schmidt, Karlsruher Institut für Technologie (KIT), Germany
Ivana Semanjski, Ghent University, Belgium
Sina Sheikholeslami, EECS School | KTH Royal Institute of Technology, Sweden
Patrick Siarry, Université Paris-Est Créteil, France
Angelo Sifaleras, University of Macedonia, Greece
Joaquim Silva, 2Ai - School of Technology | IPCA, Portugal

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

XAI for Semantic Dependency

How to Understand the Impact of Higher-Level Concepts on AI Results

Holger Ziekow
Faculty of Business Information Systems
Furtwangen University
Germany
e-mail: holger.ziekow@hs-furtwangen.de

Peter Schanbacher
Faculty of Business Information Systems
Furtwangen University
Germany
e-mail: peter.schanbacher@hs-furtwangen.de

***Abstract*—Explainable Artificial Intelligence (XAI) methods, such as partial dependency plots, or individual conditional expectation plots, help to understand the impact of feature values on the output of an Artificial Intelligence (AI) model. However, these techniques can only analyze the concepts manifested in a single feature. This makes it hard to investigate the impact of higher-level concepts, spanning across multiple features (for example, a model prediction may depend on the morbidity of a patient, while morbidity is only indirectly reflected through features about symptoms). In this paper, we present and test a concept for getting insight into model dependency on aspects on a higher semantic level. This enables an understanding of how a model output changes based on meaningful higher-level concepts and aids data scientists in analyzing machine learning models.**

Keywords-Interpretability, Understandability; Explainability; explainable AI; XAI; human-centered AI; black-box models.

I. INTRODUCTION

Due to increasing computational power, improving algorithms and access to big-data, Artificial Intelligence (AI) models gained popularity in recent years. Applications range from healthcare (Lee et al. [15]; Chen et al. [6]), credit risk (Szepannek and Lübke [23]), autonomous driving (Grigorescu et al. [14]; Feng et al. [9]), image classifications (Sahba et al. [20]), audio processing (Panwar et al. [19]), among others.

The large number of parameters and complex interactions make most AI models (in particular deep neural networks) hard to understand and difficult to interpret the results. For many applications, it is required not only to have a model with high accuracy but also to explain the outcomes. Regulators (European Commission [8]) require the understandability of these models, in particular to increase their trust (Lui and Lamb [17]) and assess potential biases (Challen et al. [5]).

What "explainability" means is not well defined and might be misleading (Rudin [27]). It further depends on the context of the application. For an MRI scan, the explanation might be a heat map of relevant areas for the model. For sentiment analysis of user feedback, the explanation might be relevant words of the text. Surrogate models such as decision trees may give an insight into more complex models.

In general, explainability methods can be distinguished into either global explainability on the model level such as variable importance (Breiman, [4]), Partial Dependency Plots (PDP) (Friedman [10]), or Accumulated Local Effects (ALE) (Apley and Zhu [1]), or local explainability on the level of individual predictions such as Shapley values (SHAP, Shapley [21] or Strumbelj and Kononenko [22]), or Local Interpretable Model Explanations (LIME) (Ribeiro et al. [24]).

We lean on the notion of Partial Dependency Plots (PDP). However, unlike PDPs, we capture the dependency on a higher-level concept, and not a single feature (e.g., a concept that manifests in many features or the combination of many feature values). The analysis shows the model output if a certain concept is more or less present. E.g., one may analyze if a medical model leans more or less towards a certain recommendation, dependent on the morbidity of a patient. Yet, morbidity may not be an explicit input of the model but indirectly reflected in a set of features about certain symptoms. Another example is an image classifier. Existing methods analyze the impact of pixels or regions in specific figures (see, e.g., Bulat and Tzimiropoulos [3]). However, reasoning about the semantics of these regions is up to the analyst and must be done instance by instance. With our method, one gains an understanding of how presence of a certain concept impacts the model output. To the best of our knowledge, this constitutes a new approach. In this context, we refer to the approach as semantic dependency analysis (not to be confused with semantic dependency in NLP). As an illustrative example, we analyze how the presence of vegetation impacts the classification of an image as showing a city or rural area.

Our main contributions are the following

- We present a new general concept which we call Semantic Dependency Analysis (SDA).
- We provide formalisms to define two fundamental ways of implementing SDA.
- We describe a specific implementation along a sample case.
- We present experimental results that demonstrate the working and utility of the approach.

The remainder of the paper is structured as follows. The introduction is followed by Section 2 presenting the current state of literature and how our approach fits into the related work. Section 3 defines the concept of Sematic Dependency Analysis (SDA) and presents a possible implementation for generators as well as prediction models. Section 4 shows how

SDA can be used for an illustrative image classification example. Section 5 summarizes and provides conclusions.

II. Related Work

White box models, also known as transparent or interpretable models, offer humans a clear understanding of the underlying decision-making process. White box models are algorithms such as linear regression, decision trees, or logistic regression. On the other hand, there are black box models such as deep neural networks. They have a vast number of parameters and can therefore account for complex interactions. While these models often exhibit remarkable performance, their decision-making processes are difficult to understand. This lack of interpretability raises concerns regarding trust, fairness, accountability, and potential biases within the model (see Riberio et al. [24]). Explainable Artificial Intelligence (XAI) is needed to establish trust of the user and the AI model (Arrieta et al. [2]). Users want to have information why the model proposed a certain decision (Wang et al. [28] or Gandi and Mishra [12]). Further, XAI is needed to detect and mitigate biases to promote fairness (Ridley [25]). Recent regulation requires the "right to explanation" for individuals affected by AI-driven decisions (Gallese [11]). Another prominent application area of XAI is the medical domain, due to the often-sensitive nature of AI decisions (see [29] for a survey). PDPs (see Friedman [10]) have long been used to understand the impact of a certain feature. PDPs have computational advantages and are easier to understand for a layman compared to most alternative XAI methods (Dwivedi et al. [7]). However, PDPs do not properly take feature interactions into account (Linardatos et al. [16]). To account for the interaction effects, individual conditional expectation plots (ICE, see Goldstein et al. [13]) were developed. An alternative approach is Accumulated Local Effect plots (ALE) (Apley and Zhu [1]). While PDPs are based on marginal distribution, ALE plots are based on conditional distribution. All those PDPs related methods show the impact of a certain feature given in the dataset. Higher-level concepts which are often of interest but not included in the data, therefore, cannot be analyzed. Consider, for example, patient data, such as age, sick days, therapy, income. The higher-level concept of interest "morbidity" is however not the data. To analyze the impact of such a higher-level concept, we introduce the semantic dependency analysis.

III. Semantic Dependency Analysis

In this section, we introduce the concept of Semantic Dependency Analysis (SDA). We lean on the notion of partial dependency plots that are defined as follows (see Molnar [18]):

$$\hat{f}_S(x_S) = E_{X_C}\left[\hat{f}(x_S, X_C)\right] = \int \hat{f}\,(x_S, X_C) dP(X_C).$$

Here, x_S is the feature value of the analyzed feature S, X_C are the other features in the model, and $\hat{f}(x_S, X_C)$ the AI model applied on the complete feature vector (containing x_S and X_C). Intuitively, the partial dependence function represents the average prediction if all data points have the given feature value x_S.

In SDA, we do not analyze a single feature S but a higher-level concept H where $x_H \in H$ are values reflecting the presence (or degree of presence) of that concept (i.e. for elements in H we expect an order relation with respect to the presence of the semantic concept H). We define the analysis for a given higher-level concept H (SD_H) as

$$SD_H(x_H) = E_X\left[\hat{f}\left(g(x_H, X)\right)\right].$$

Here, $g(x_H, X)$ is a random variable that returns feature vectors for the model $\hat{f}$ in accordance to x_H, and in compliance with X. That is, the resulting values stem from the distribution of X and also have concept x_H. We subsequently discuss two concepts of the implementation of g.

A. Implementation with generators

One way to implement g is to use synthetic data generators. Values of x_H in H and the distribution of X drive the generation of data points in accordance with x_H. For instance, x_H may be mapped to a prompt in a text to image model. The distribution of X may be reflected by further elements of the prompt (i.e. a prompt describing X). Note that there are further options to account for the distribution of X. This includes the use of image-to-image models, taking samples from X as input, or training the generator on X.

Alternative implementations may use rule-based data generators (in particular for tabular data) or 3D engines for image generation. The feasibility of different data generation approaches depends heavily on the use case. In our sample case, we use a diffusion model for illustration (see Section 4).

B. Implementation with prediction models

Another way of implementing g is to use a prediction model $d(x, x_H)$ that can detect the presence of x_H in a data point x in X. The advantage is that data points can be sampled from real data with the distribution of X. Assuming that d returns a score for the presence of x_H in x, we can implement $g(x_H, X)$ by drawing from $\{x \in X \mid d(x, x_H) \geq t\}$, where t is a threshold denoting the minimum probability that x_H is present. Note, that using d also allows for an alternative definition of SD_H as continuous function, dependent on the ε-environment around the presence score s, denoting the presence of X_H in a data point x:

$$SD_H(x_H, s) = E_X\left[\hat{f}(\{x \mid d(x, x_H) \in [s-\varepsilon, s+\varepsilon]\})\right].$$

IV. Experiments

In this section, we describe experiments that demonstrate along an example the viability of the approach and illustrate a possible instantiation of the concept.

A. Experimental Setup

We use a sample application as proof of concept and test for the SDA approach. As sample task, we use a binary classification task for images. That is, we aim to classify images in either class A (showing a city) or class B (showing a rural landscape).

For our test application we used synthetic data, generated with stable diffusion version 2 (see Rombach et al. [26]). We used default parameters, except for the sampling steps of 100. The prompts for generating the different classes are:

Class A:
Positive prompt: *Photograph a city, high quality photography, Canon EOS R3*
Negative prompt: *digital art, drawing*

Class B:
Positive prompt: *Photograph of a rural landscape, high quality photography, Canon EOS R3*
Negative prompt: *digital art, drawing*

For the training data, we generated images of 512×512 pixels. Figures 8 and 9 show examples from the training set of both classes. For the experiments, we use an arbitrary classification network generated by ChatGPT 4.0. The network architecture is shown in Figure 1. In each test run, we trained the model using 5 epochs and batch size 32. We then analyzed the resulting networks with SDA.

For the SDA, we implement $g(x_H, X)$ by adding to the prompt of class A. That is, we limit the analysis to the recognition of class A. By keeping the prompt for class A, we realize the compliance to X in the data generation. Note that this is only an approximate solution, as the control over the diffusion model's output is limited. By adding to the prompt of A, we implement the presence of x_H. By using in total 4 different prompt additions, we implement an order over the total of 4 elements x_H in H. Specifically, we used the following prompts.

Data set "cityNoTrees":
Same prompt as for class A but with "trees" in negative prompt. See
Figure **10.** Images in this data set contain some trees but less than the data sets "City", "cityTrees", "TreesCity".
Positive prompt: *Photograph a city, high quality photography, Canon EOS R3*
Negative prompt: *digital art, drawing, trees*

Data set "City":
Same prompts as for as class A. Images in this data set contain more trees than the data sets "cityNoTrees", but less than in "cityTrees" and "TreesCity".

Data set "cityTrees":
Same as class A but with "trees" added after "city" to the positive prompt (giving more importance to "city" than to "trees"). See Figure 11. Images in this data set contain more trees than the data sets "cityNoTrees", and "City", but less than in "TreesCity".
Positive prompt: *Photograph a city, trees, high quality photography, Canon EOS R3*
Negative prompt: *digital art, drawing*

Data set "TreesCity":
Same as class A but with "trees" added before "city" to the positive prompt (giving more importance to "trees" than to "city"). See Figure 12. Images in this data set contain more trees than all the other data sets but are still generated to show locations within cities.

Positive prompt: *Photograph trees, city, high quality photography, Canon EOS R3*
Negative prompt: *digital art, drawing*

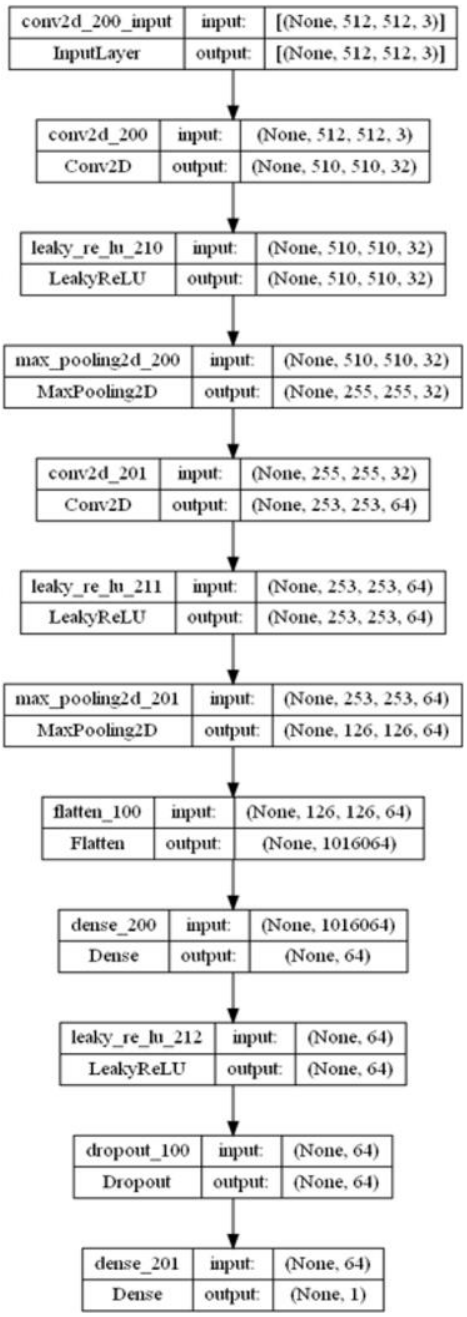

Figure 1. Network architecture for experiments.

The prompts for the different data sets are given in the order of presence of the concept "trees". Although the exercised control over the diffusion networks is limited, manual inspection of the resulting images verifies the intended outcome. (The test set for "city" is not shown, as the prompt

was the same as for the training set and the results are of similar appearance).

We trained 6 different networks with the given architecture on the training set containing classes A and B. (The choice of 6 networks is arbitrary and driven by the length of the paper). We used 800 images for each class. For the SDA we used 800 images for each of the 4 prompts “cityNoTrees”, “City”, “cityTrees” and “TreesCity”. With the SDA, we aim to investigate if the presence of the concept “trees” makes the model less confident about class A, that is, if increased presence of trees leads to a lower probability for detecting class A (city).

B. Results

Figures 2-7 show the results of the tests for six different networks. The results show box plots of the probabilities assigned by different models to the images for the data sets “cityNoTrees”, “City”, “cityTrees” and “TreesCity”. The probability shown is the determined probability of *not* showing a city. The mean values in the box plots are the values intended for the SDA, according to our definition in Section 3. However, the additional information from the box plots gives additional insights about the distribution.

If the presence of trees causes models to deem the label “city” less likely, we expect to see increasing means from left to right in the plots. That is because the “cityNoTrees”, “City”, “cityTrees” and “TreesCity” are ordered according to that presence of the higher-level concept “Tree”. We observe that this is the case for all analyzed networks. (Note that, although “City” used the same prompt as the training set, the models have an even higher confidence for “cityNoTrees“ than for “City”).

Analysts learn from the SDA that the trained models are impacted by the higher-level concept of trees, as well as the nature of that impact. The behavior of the networks and the SDA results are plausible. Hence, the experiments verify the viability and utility of our approach.

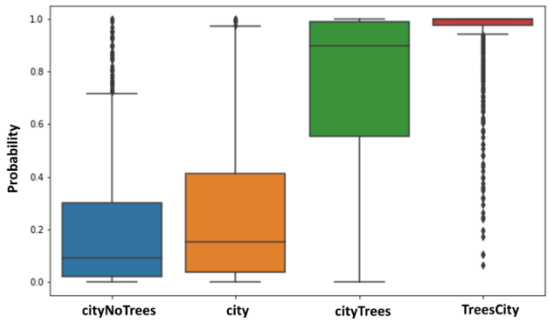

Figure 2. SDA for network 1.

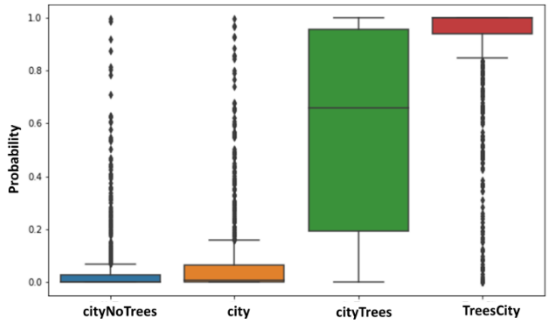

Figure 3. SDA for network 2.

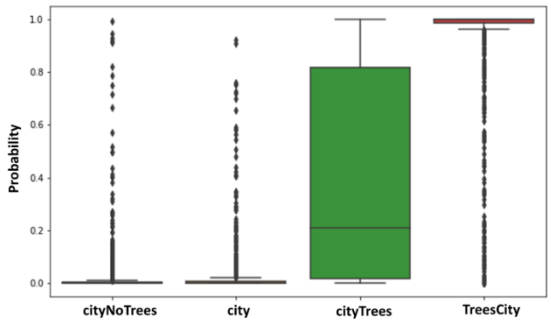

Figure 4. SDA for network 3.

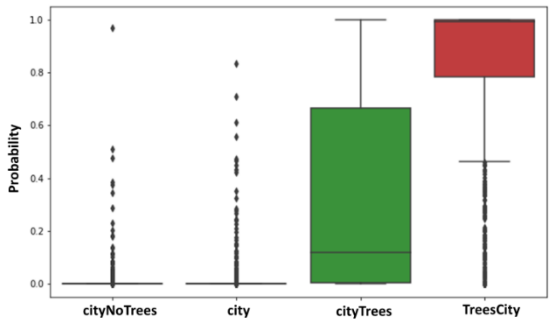

Figure 5. SDA for network 4.

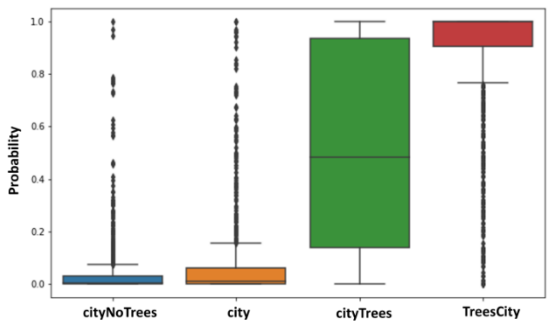

Figure 6. SDA for network 5.

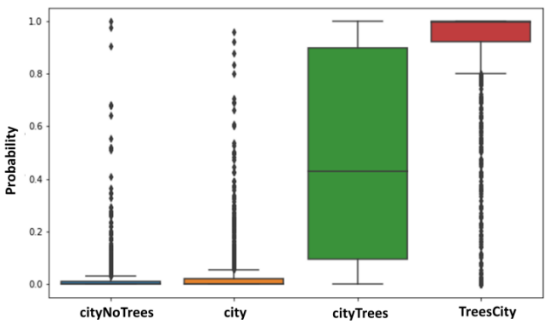

Figure 7. SDA for network 6.

V. CONCLUSIONS

Our article introduces the concept of Semantic Dependency Analysis (SDA), which goes beyond traditional Partial Dependency Plots (PDPs) by capturing dependencies on higher-level concepts rather than individual features. The analysis showcases how the output of a model changes based on the presence or absence of a specific concept.

As an illustrative example, we analyzed how the presence of vegetation affects the classification of an image as a city or rural area. The results of the analysis demonstrate that the trained models are influenced by the higher-level concept of trees and provide insights into the nature of this impact. The observed behavior of the networks aligns with expectations and supports the viability and utility of the approach employed in the experiments. Analysis can use such insight to reason about the working of their models.

Future work will address further options for implementing $g(x_H, X)$ and the challenge that implementations may only approximate the intended behavior. In particular, with the generative approach, reflecting x_H to the desired degree is a challenge in implementing g. However, our illustrative example shows its viability.

By moving beyond individual features and focusing on broader concepts, SDA provides valuable insights into how a higher-level concept influences predictions or classifications. The formalisms and implementation described in the text provide a foundation for conducting SDA and analyzing various domains, ranging from medical models to image classifiers. Overall, SDA has the potential to enhance interpretability and decision-making in AI systems, contributing to advancements in explainable AI and fostering trust in AI-driven solutions.

ACKNOWLEDGEMENT

We would like to thank Valentin Göttisheim for feedback and support in the creation of this paper.

Figure 8. Samples from training set for label "City".

Figure 9. Sample from training data set for label “Landscape”.

Figure 10. Sample from test set with “City” in positive and “Trees” in negative prompt (cityNoTrees).

Figure 11. Sample from test set with first “City” and then “Trees” in positive prompt (cityTrees).

Figure 12. Sample from test set with first “Trees” and then “City” in positive prompt (TreesCity).

REFERENCES

[1] D. W. Apley and J. Zhu, "Visualizing the effects of predictor variables in black box supervised learning models". J. R. Stat. Soc. Ser. B, 2020.

[2] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". Information fusion, vol. 58, pp. 82-115, 2020.

[3] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression", In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14, pp. 717-732, Springer International Publishing, 2016.

[4] L. Breiman, "Random Forests. Machine Learning", vol. 45, pp. 5-32, 2001.

[5] R. Challen, J. Denny, M. Pitt, L. Gompels, T. Edwards, and K. Tsaneva-Atanasova, "Artificial intelligence, bias and clinical safety". BMJ Quality & Safety, vol. 28, pp. 231-237, 2019.

[6] R. Chen, L. Yang, S. Goodison, and Y. Sun, "Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data", Bioinformatics, vol. 36, pp. 1476–1483, 2020.

[7] R. Dwivedi et al., "Explainable AI (XAI): Core ideas, techniques, and solutions", ACM Computing Surveys, vol. 55, pp. 1-33, 2023.

[8] European Commission, "proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)". https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206, 2021 (access on 12.6.2023).

[9] D. Feng et al., "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges", IEEE Transactions on Intelligent Transportation Systems, vol. 22, pp. 1341-1360, 2020.

[10] J. Friedman, "Greedy function approximation: A gradient boosting machine", Annals of Statistics, vol. 29, pp. 1189-1232, 2001.

[11] C. Gallese, "The AI Act proposal: a new right to technical interpretability?", arXiv preprint arXiv:2303.17558, 2023.

[12] N. Gandhi and S. Mishra, "Explainable AI for healthcare: A study for interpreting diabetes prediction", In Machine Learning and Big Data Analytics (Proceedings of International Conference on Machine Learning and Big Data Analytics, ICMLBDA 2021, pp. 95-105, Springer International Publishing, 2022.

[13] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation", J. Comput. Graph. Stat. 2015, vol. 24, pp. 44–65, 2015.

[14] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving". Journal of Field Robotics, vol. 37, 362-386, 2020.

[15] S. M. Lee et al., "Deep Learning Applications in Chest Radiography and Computed Tomography: Current state of the Art", Journal of Thoracic Imaging, vol. 34, 2019.

[16] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI : A review of machine learning interpretability methods", Entropy, vol. 23, pp. 18, 2020.

[17] A. Lui and G. W. Lamb, "Artificial intelligence and augmented intelligence collaboration : regaining trust and confidence in the financial sector", Information & Communications Technology Law, vol. 27, pp. 267-283, 2018.

[18] M. Christoph, "Interpretable machine learning", Lulu.com, 2020.

[19] S. Panwar, A. Das, M. Roopaei, and P. Rad, "A deep learning approach for mapping music genres", In 2017 12th System of Systems Engineering Conference (SoSE), pp. 1-5, IEEE, 2017.

[20] A. Sahba, A. Das, P. Rad, and M. Jamshidi, "Image graph production by dense captioning", In 2018 World Automation Congress (WAC), pp. 1-5, IEEE, 2018.

[21] L. S. Shapley, "Notes on the n-Person Game", Santa Monica, RAND Corporation, 1951.

[22] E. Strumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions", Knowledge and Information Systems, vol. 41, pp. 647-665, 2014.

[23] G. Szepannek and K. Lübke, "How much do we see? On the explainability of partial dependence plots for credit risk scoring", Argumenta Oeconomica, vol. 1, pp. 137-150, 2023.

[24] M. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135-1144, 2016.

[25] M. Ridley, "Explainable AI (XAI): Confronting Bias, Discrimination, and Fairness in Machine Learning", In Access Conference Proceedings, 2019.

[26] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models", In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684-10695, 2022.

[27] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead", Nature Machine Intelligence, vol. 1, pp. 206–215, 2019.

[28] Y. C. Wang, T. C. T. Chen, and M. C. Chiu, "An improved explainable artificial intelligence tool in healthcare for hospital recommendation", Healthcare Analytics, vol. 3, pp. 100147, 2023.

[29] R. K. Sheu and M. S. Pardeshi, "A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach", Human Interaction and Scoring System. Sensors, vol. 22, pp. 8068, 2022.

Comparison between Surrogate Safety Assessment Models (SSAM) and Accident Models on Unconventional Roundabouts

Antonio Pratelli
Department of Civil and Industrial Engineering
University of Pisa
Pisa, Italy
e-mail: antonio.pratelli@unipi.it

Lorenzo Brocchini
Department of Civil and Industrial Engineering
University of Pisa
Pisa, Italy
e-mail: lorenzo.brocchini@phd.unipi.it

Pietro Leandri
Department of Civil and Industrial Engineering
University of Pisa
Pisa, Italy
e-mail: pietro.leandri@unipi.it

Rosaria Aiello
Department of Civil and Industrial Engineering
University of Pisa
Pisa, Italy
e-mail: aiello.sara91@gmail.com

***Abstract*— This paper describes the comparison between the Surrogate Safety Assessment Model (SSAM) of the Federal Highway Administration (FHWA) and the predicted number of accidents calculated through analytical models, regarding Unconventional Roundabouts. The novelty of this comparison lies precisely in the fact that the 3 roundabouts analyzed fall into the category of so-called Unconventional Roundabouts, i.e., arrangements with "roundabout circulation", which do not fall within the types listed in the Italian Legislation (Ministerial Decree 19-04-2006). In roundabout intersections, among the various types of accidents that may occur, those of the rear-end collision type occur more frequently, for which it was decided to use the formulas of the accident models relating to this type of conflict. In particular, the conflicts type "Approach" for the Maycock & Hall model and the conflict type "Rear end" for the Arndt & Troutbeck model were taken into consideration. As mentioned, in addition to the application of analytical models, possible points of conflict (of the same category, i.e., "Rear end") were evaluated using dynamic simulation models. In particular, the dynamic simulation software Aimsun™ was used as a means to obtain the necessary inputs for the evaluation of the surrogate safety carried out through SSAM, a software application that reads the trajectory files generated by the simulation programs. In the final part of this paper, the conclusions on the comparison and some possible future ideas for further research developments have been included.**

Keywords- Unconventional Roundabouts; Microsimulations; Aimsun; SSAM; Accidents Models.

I. INTRODUCTION

This paper starts from the idea of the authors to develop the work carried out by Vasconcelos et al. in the article "Validation of the Surrogate Safety Assessment Model for Assessment of Intersection Safety" [1]. In particular, the authors have decided to resume the research work carried out and extend it with their contribution, starting from their conclusion that the Surrogate Safety Assessment Model is a quite promising approach to assessing the safety of new facilities, innovative layouts and traffic regulation schemes. Then, the present work started from the fact that it is difficult to calculate the possible number of accidents in roundabouts with innovative layouts, because, unlike the conventional ones which are "geometrically identifiable", they have highly variable geometric parameters and therefore it is difficult to able to describe their road safety with a single model. So, this research tried to describe the comparison between the Surrogate Safety Assessment Model (SSAM) of the Federal Highway Administration (FHWA) and the predicted number of accidents calculated through analytical models, regarding Unconventional Roundabouts. The extension of the work of Vasconcelos et al. and therefore the novelties lie precisely in the fact that the 3 roundabouts analyzed fall precisely into the category of so-called Unconventional Roundabouts, i.e., arrangements with "roundabout circulation", which do not fall within the types listed in the Italian legislation (Ministerial Decree 19-04-2006: "Functional and geometric rules for the construction of road intersections" [2]). These roundabouts have shapes and dimensions that are out of the ordinary concept of roundabout intersection. As regards the accident models, it was decided to consider the formulas of the conflict type "Approach" for the Maycock & Hall [3] model and those of the conflict type "Rear end" for the Arndt & Troutbeck [4] model. This choice is based on the fact that among the various types of accidents that can occur in roundabout intersections, rear-end collisions occur more frequently (literature the values vary from 20% to 25%). As far as the surrogate safety evaluation is concerned, it was carried out using SSAM (a software application that reads the trajectory files generated by the simulation programs) [5]. It was decided to use Aimsun™ as a dynamic microsimulation software, with which it was possible to obtain the ".trj files", i.e., the trajectory files, essential for calculating the possible points of conflict, which, by definition, are the points where two vehicles can potentially collide with each other at road intersections.

Also, in this case, the points of conflict of the "Rear end" category have been taken into consideration. Finally, to improve the visualization style of the points of conflict extrapolated from SSAM, it was decided to use the software Quantum Geographic Information System (QGIS); in this application, the files extrapolated from SSAM were inserted and geolocated. The following sections will follow: a first more theoretical section which will deal with the Italian Unconventional Roundabouts with some examples that are taken into consideration; two sections concerning the SSAM approach from FHWA and the existing roundabouts accident models; the final section, followed by the conclusions and the future research work, which will explain the comparison of the two approaches.

II. ITALIAN UNCONVENTIONAL ROUNDABOUTS

The subsections that follow will primarily deal with the theory of the so-called Unconventional Roundabouts, with reference to the Italian Legislation; and then move on to some practical instances.

A. Unconventional Roundabouts Theory and Italian Legislation

First of all, it is appropriate to specify what is meant by Unconventional Roundabouts [6] and why the authors decided to develop their research on them. In the Italian legislation (Ministerial Decree 19-04-2006 [2]), there can be three basic types of roundabouts based on the diameter of the outer circumference: Conventional Roundabouts with an outer diameter between 40 and 50 m; Compact Roundabouts with outside diameters between 25 and 40 m; Mini Roundabouts with external diameter between 14 and 25 m. For arrangements with "roundabout circulation", which do not fall within the above typologies, we, therefore, speak of Unconventional Roundabouts and for them, the geometric dimensioning and verification must be adapted. When we talk about Unconventional Roundabouts must be considered both the so-called "new generation roundabouts" (Raindrop Roundabouts; Turbo Roundabouts [7] [8]; Two-Geometry Roundabouts [9] [10]), which are currently being built for the purpose of fulfilling safety and performance objectives in cases where classic roundabouts are unable to work well; both the so-called "old roundabouts" which had dimensions and geometries suitable for when precedence was on the branches instead of on the ring (first generation roundabouts) [11]. In Italy, there are many Unconventional Roundabouts of both "typologies", both because in terms of space there is the need to adopt solutions that are not conventional, and because for the moment there are always obsolete roundabouts on the national territory which have not been adapted and which in fact are often poor in terms of security. Precisely for this last consideration, in this discussion the authors have decided to take into consideration 3 Unconventional Roundabouts of the latter type and have decided to analyse them in terms of safety, also because from this point of view there are no in-depth studies for them. A final introductory consideration concerns the type of accidents that the authors decided to analyse, i.e., rear-end collisions.

They are the conflicts/accidents that occur on the entrance branches more frequently at "roundabout" intersections and for this reason they were chosen as a study parameter.

B. Territorial framework and O/D Matrices of the 3 identified Roundabouts

This short paragraph lists the 3 Unconventional Roundabouts analyzed by the authors. All 3 roundabouts are situated in Italy, in the Tuscany region and are located in urban areas, therefore the speed referred to during the calculations is equal to 50 km/h [12]. In particular, in Fig. 1, Fig. 2 and Fig. 3, the 3 aerial images extracted from Google Earth are reported, where the progressive numbers of the branches of the roundabouts are also reported. Reference is made to them for the reconstruction of the Origin/Destination (O/D) matrices, reported in turn in Table I, Table II and Table III.

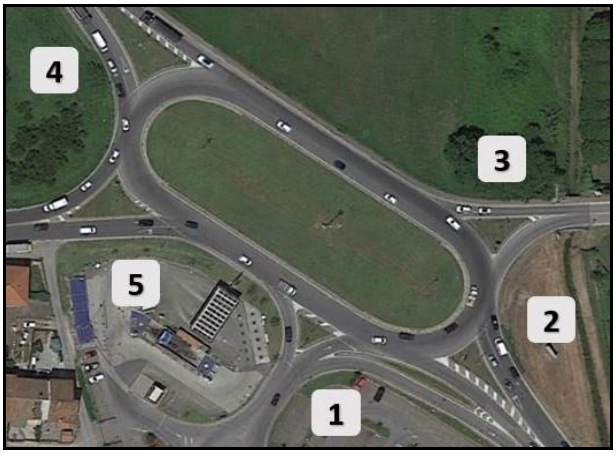

Figure 1. Territorial framework of the 1st Unconventional Roundabout located on SP61-Lucchese-Romana in Lucca, Tuscany, Italy (*source*: Google Earth Pro)

Figure 2. Territorial framework of the 2nd Unconventional Roundabout located on Viale Nazario Sauro in Livorno, Tuscany, Italy (*source*: Google Earth Pro)

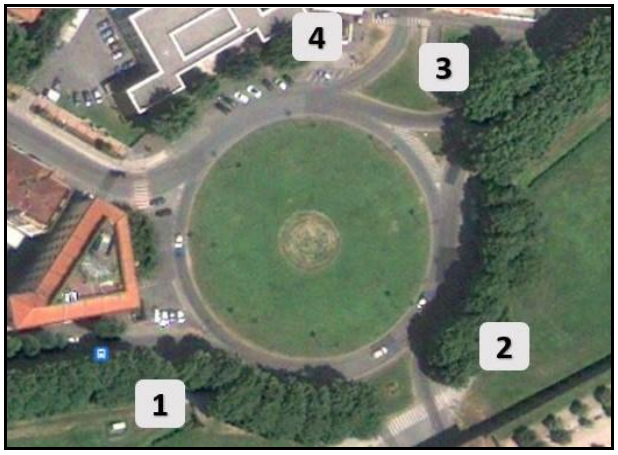

Figure 3. Territorial framework of the 3rd Unconventional Roundabout located on Porta Santa Maria in Lucca, Tuscany, Italy (*source*: Google Earth Pro)

TABLE I. O/D MATRIX OF THE 1ST UNCONVENTIONAL ROUNDABOUT

Roundabout 1 - SP61 Lucchese-Romana (Lucca, Tuscany, Italy)						
Matrice O/D	1	2	3	4	5	TOT
1	0	142	60	36	72	310
2	36	0	140	346	812	1334
3	44	204	0	114	76	438
4	58	320	56	0	280	714
5	58	794	184	372	0	1408
TOT	196	1460	440	868	1240	4204

TABLE II. O/D MATRIX OF THE 2ND UNCONVENTIONAL ROUNDABOUT

Roundabout 2 - Viale Nazario Sauro (Livorno, Tuscany, Italy)				
Matrice O/D	1	2	3	TOT
1	0	390	517	907
2	443	0	691	1134
3	476	541	0	1017
TOT	919	931	1208	3058

TABLE III. O/D MATRIX OF THE 3RD UNCONVENTIONAL ROUNDABOUT

Roundabout 3 - Porta Santa Maria (Lucca, Tuscany, Italy)					
Matrice O/D	1	2	3	4	TOT
1	181	299	1749	0	2229
2	253	0	195	0	448
3	951	52	12	0	1015
4	263	51	12	0	326
TOT	1648	402	1968	0	4018

These matrices were elaborated starting from the data surveys carried out on the 3 roundabouts through the use of Sony DCR-SX34 digital cameras, positioned at specific points of the intersections, during the peak periods of the week [13].

III. SSAM APPROACH FROM FHWA

This concise section has been included to define what is meant by surrogate security assessment and how it is possible to carry out such an assessment. Safety analysis is a decisive aspect in the evaluation of design choices both for the new road system and for the adaptation of the existing road network. The Federal Highway Administration (FHWA) has developed and made available the Surrogate Safety Assessment Model (SSAM) program, through which it aims to offer designers, researchers and companies specializing in road design and construction a tool for assessing the safety of an intersection by estimating the frequency of conflicts [14] [15].

The concept of surrogate safety derives from the desire to develop alternative tools to the existing ones to evaluate the accident frequency of road infrastructure: in particular, while the ordinary methods derive from statistical evaluations based on accidents that have occurred, the surrogate safety methods are instead based on factors that do not require years of accident statistics. The SSAM program elaborates the trajectory files (.trj files) obtained in output from a dynamic simulation program (in the case of the present research it is decided to use the Aimsun™ program, but in general VISSIM™, TEXAS™, etc.). In detail, SSAM evaluates every single vehicle-vehicle interaction according to criteria with which it can establish whether there is a point of conflict and to which category it belongs. At the end of the elaborations, SSAM presents the results in tables, allowing the user to filter them according to parameters of his choice. As regards the classification of conflicts, the program contemplates four types: Rear end, Lane changing, Crossing and Unclassified. To classify them, the program evaluates the crossing angle of the trajectories, if this angle is less than 20° the conflict is of the Rear end type. In the present research, the latter have been taken into consideration, since, as already explained, they are the ones that occur most frequently in roundabout intersections. Their unit of measurement is expressed in conflicts/day.

IV. EXISTING ROUNDABOUTS ACCIDENT MODELS

Roundabouts, in general, are considered to be the safest road junctions as they have several advantages including reduction of points of conflict and lower movement and departure speeds. However, accidents can also occur on them and in particular, several studies state that the most common accident that can occur is a rear-end collision. To study the safety characteristics of the elements of the road system, there are several models for predicting accidents [16]. The authors have decided to use in this research two of the most used models, namely those of the Maycock & Hall model and the Arndt & Troutbeck model. They were chosen because they allow the number of accidents to be calculated taking into consideration both the traffic demand, the geometric characteristics of the intersection, and the dynamic ones (such as speed, for example). With these models, it is possible to calculate various types of accidents, but clearly, as explained above, it was decided to use the formulas of the Conflicts Type "Approach" for the Maycock & Hall [3] model (1) and those of the Conflict Types "Rear end" for the Arndt & Troutbeck [4] model (2), which indicate precisely rear-end collisions. Both models make it possible to estimate the number of accidents over a period of time and therefore their unit of measurement is expressed in accidents/years [17]. The two formulas (1) and (2) used are therefore reported below, specifying that the coefficients of these formulas are the standard ones calibrated for conventional roundabouts. In fact, another of the interesting aspects of this research was precisely that of verifying whether these coefficients could also work for Unconventional Roundabouts. To answer this question, see the next section.

$$A_2 = 0.0057 \times Q_e^{1.7} \times \exp(20C_e - 0.1e) \quad (1)$$

where:

- Q_e = entering flow, respectively (1000s of vehicles/day);
- C_e = entry curvature [Ce = 1/Re and Re = entry path radius for the shortest vehicle path (m)];
- e = entry width [m].

$$A_r = C_1 \times Q_a^x \times Q_c^y \times S_a^z + C_2 \quad (2)$$

where:

- Q_a = average annual daily traffic (AADT) on the approach;
- Q_c = various AADT flows on the circulating carriageway adjacent to the approach;
- S_a = 85th percentile speed on the approach curve (the potential relative speed between approaching vehicles) [km/h];
- $C_1 = 9.62 \times 10^{-11}$; $C_2 = 0$; x = 1; y = 0.5; z = 2. [4]

V. COMPARISON OF THE TWO APPROACHES

The following section presents the results of the research. First of all, a summary table (Table IV) of the calculations carried out is shown which served to reconstruct the graphs on which most of the considerations will be made.

TABLE IV. SUMMARY TABLE OF THE CALCULATIONS MADE

Roundabout	Approach	Qe [veh/d]	Arndt & Troutbeck Rear-end [acc/y]	Maycock & Hall Approach [acc/y]	SSAM (TTC = 1.5 s) [conflicts/d]
1	1	3100	0,10	0,07	24
	2	13340	0,28	0,33	383
	3	4380	0,14	0,13	63
	4	7140	0,19	0,23	165
	5	14080	0,29	0,34	207
2	1	9070	0,16	0,15	120
	2	11340	0,20	0,37	203
	3	10170	0,16	0,27	119
3	1	22290	0,18	0,55	160
	2	4480	0,15	0,13	82
	3	10150	0,16	0,32	104
	4	3260	0,09	0,07	36

Furthermore, the authors considered it necessary to also report an explanatory image of the surrogate safety assessment. In detail, the following image (Fig. 4) shows an extract of the QGIS software of one of the roundabouts chosen as an example (Roundabout 2), where the points of conflict have been inserted, georeferenced (with TTC = 1.5 s) extracted from the SSAM software after processing the ".trj file", which in turn was obtained from the Aimsun™ simulation software. The Time to Collision (TTC) is one of the SSAM software parameters and expresses the minimum collision time [18]. It can range from an infinite maximum value, when two vehicles never meet, to a minimum value of 0 seconds when an accident occurs. Various studies have been conducted to identify a threshold value of the TTC, such as to separate major accidents from minor and negligible or without consequences accidents [19]. This value, depending on the study, was identified as a fixed value or as the result of a function dependent on the speed or deceleration of the vehicles. The authors have decided to keep the default value of the SSAM program which assumes the value TTC = 1.5 s.

Figure 4. Example of Number of Conflicts obtained by SSAM software and reported on QGIS of 2nd Roundabout

Below are the graphs (Fig. 5, Fig. 6 and Fig. 7) which summarize most of the research results. In particular, each graph refers to one of the 3 roundabouts and is structured as follows: the Q_e (entrance vehicular flow) expressed in vehicles per day is shown on the abscissa axis; while there are two different y axes. The left y-axis is incident models (Arndt & Troutbeck / Maycock & Hall) and is expressed in accidents per year, while the right is the SSAM results and is expressed in conflicts per day.

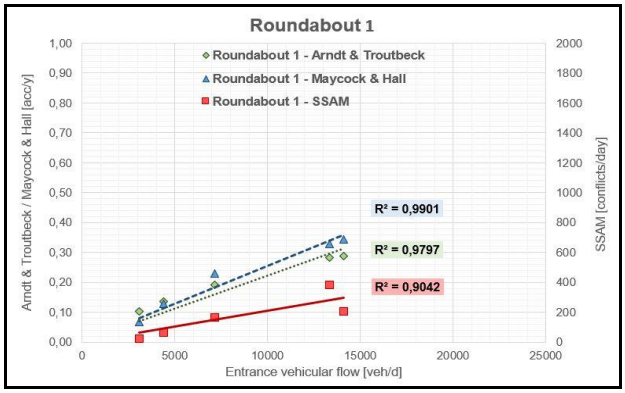

Figure 5. Graph of Results for the 1st Unconventional Roundabout

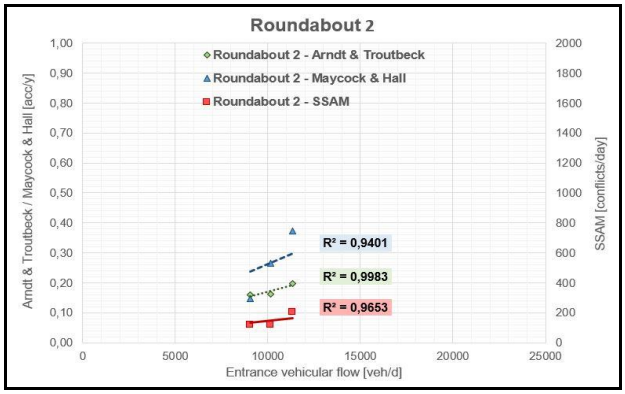

Figure 6. Graph of Results for the 2nd Unconventional Roundabout

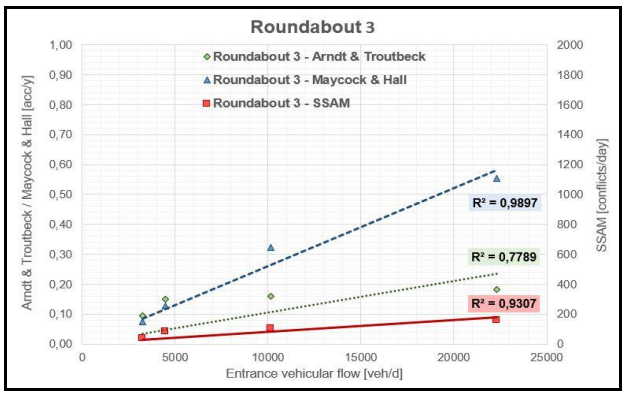

Figure 7. Graph of Results for the 3rd Unconventional Roundabout

On the graphs, as many points have been reported as there are entrance arms of the roundabout in question and a linear trend line passing through the origin (0; 0) has then been created for them. After that, the authors decided to calculate the coefficient of determination R^2 for each trend line. It is a statistical value that allows us to understand whether a linear regression model can be used to make predictions. Its value is always between 0 and 1, or between 0% and 100% if you want to express it in percentage terms. $R^2 = 0$ indicates a model whose predictor variables do not explain the variability of y around its mean at all. $R^2 = 1$ indicates a model whose independent variables fully explain the variability of y around its mean; that is, knowing the values of the independent variables one can predict exactly what the value of y will be. Clearly, the values 0 and 1 are limit values, what emerges is that the greater the value of R^2, the more the model has high predictive power, i.e., the better the ability of the explanatory variables to predict the values of the dependent variable. Usually, we talk about high R^2 values, when they are higher than 0.7. At this point, after having explained the type of graphs used and the reference values, it is possible to go into detail on the considerations relating to the actual results. For all the graphs, i.e., for all the roundabouts, the R^2 values are generally excellent (they are always higher than 0.9, except for one case), both as regards the accident models and as regards the values of the conflicts obtained with SSAM. This is an excellent result as the 3 roundabouts to which the models have been applied are Unconventional Roundabouts, i.e., "different" intersections from the ones on which the models have been calibrated. Therefore, as a first result, it is certainly possible to state that the accident models used (Arndt & Troutbeck / Maycock & Hall), which are already valid and validated for conventional roundabouts, can also be used for Unconventional Roundabouts, using the same formulations and the same coefficients. Also, with regard to the SSAM results, the R^2 values are always higher than 0.9 and despite the different scales it is possible to state that the trend of the trend lines of the points deriving from SSAM is very similar to that relating to the accident models. This is therefore another excellent result that the authors have arrived at, namely that even for Unconventional Roundabouts there is a correspondence between the accident models and the calculation of the conflicts carried out with SSAM.

Finally, the authors also noted a further fact regarding Fig. 7, i.e., the graph referring to roundabout number 3. The trend line of the Arndt & Troutbeck model has an R^2 that is always acceptable, but clearly lower than all the others (0.7789). The explanation that the authors came up with is the following: roundabout number 3, in addition to being of an unconventional type, is also atypical from the point of view of the approaches, since, as can be seen from the territorial framework (Fig. 3) and the corresponding O/D matrix (Table III), the approach 4 is formed only by the input branch and not the output branch. This, together with the particular geometry of the roundabout, has led to a high difference between the incoming flow rate Q_e and the circulating flow rate Q_c of the adjacent approach 1 (this difference is underlined in Table V). So, another result that the authors have reached is the consideration that the model of Arndt & Troutbeck does not adapt perfectly to Unconventional Roundabouts in which there is, for some branches, a high difference between the incoming flows and circulating flows.

TABLE V. EXTRACT FROM THE CALCULATION TABLE, WHERE THE DIFFERENCE BETWEEN QE AND QC CAN BE SEEN

Roundabout	Approccio	Qe [veh/d]	Qc [veh/d]	Delta (Qe-Qc/Qc)
3	1	22290	1150	18,38
	2	4480	19540	0,77
	3	10150	4340	1,34
	4	3260	14490	0,78

A final comparison was also made for the 3 Unconventional Roundabouts as a whole. In fact, a last graph (Fig. 8), of the same typology as the previous ones, was constructed however by taking into consideration the roundabouts as a whole and no longer approach by approach. In this way, it was possible to compare the 3 roundabouts on a single graph and this led to the following consideration.

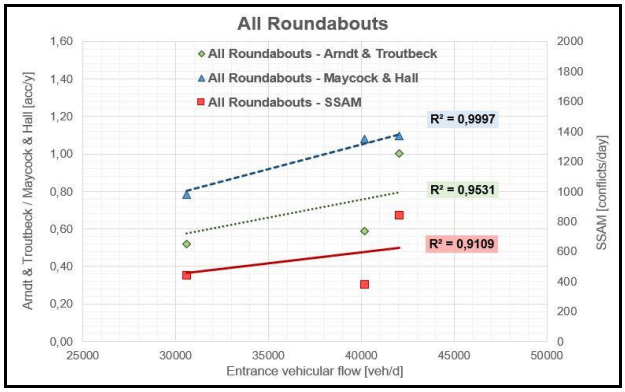

Figure 8. Graph of Results for the three Unconventional Roundabouts together

The values of R^2 are excellent and also the roundabout 3 which had a deficit on the Arndt & Troutbeck model due to the difference between the incoming flows and the circulating flows at one of the approaches, if it is considered as a whole, it is possible to homogenize with the other results.

VI. CONCLUSIONS AND FUTURE RESEARCH WORK

This article describes the comparison between the Federal Highway Administration (FHWA) Surrogate Safety Assessment Model (SSAM) and the predicted number of accidents calculated using the Arndt & Troutbeck and Maycock & Hall analytical models, as concern the Unconventional Roundabouts [20] [21]. 3 Unconventional Roundabouts located on the Italian territory that have different shapes and sizes from the regulatory standards were analysed. Other works and articles have been published regarding the comparison between the models mentioned, however, the novelty of this research proposed by the authors lies precisely in the different base data, i.e., the Unconventional Roundabouts. The type of accident and conflict chosen for the comparison made is that of rear-end collisions, as it is the most common present on roundabout intersections. In the sections of the article, various initial considerations follow one another which deepen the concepts of Unconventional Roundabouts, surrogate safety analysis models (SSAM) and accident models; up to section V where the results of the entire research were clearly explained. Summarizing these results, the authors found that: 1) the accident models used (Arndt & Troutbeck / Maycock & Hall) already valid and validated for conventional roundabouts, can also be used for Unconventional Roundabouts, using the same formulations and the same coefficients also because a certain correspondence was also found between them in terms of the number of accidents per year; 2) also for Unconventional Roundabouts there is a correspondence between the accident models and the calculation of the conflicts carried out with SSAM; 3) Arndt & Troutbeck model is not perfectly suited to Unconventional Roundabouts in which there is, for one or more branches, too high a difference between incoming flows and circulating flows. Before concluding the work, the authors decided to also propose some ideas of the possible future development of this research. First of all, this work can certainly be expanded by analysing further case studies and thus obtaining more points to use on the graphs obtained. Furthermore, the accident models utilised were used in the first analysis without the recalibration for the Unconventional Roundabouts; therefore another next steps could be proper to go and search for the actual accident data and thus verify whether the parameters used can be further improved and better recalibrated for Unconventional Roundabouts (it is emphasized that however, as explained in section V, the accident models used, can already be used also for Unconventional Roundabouts, given the statistical results obtained by the authors).

REFERENCES

[1] L. Vasconcelos, L. Neto, A. M. Seco, and A. B. Silva, "Validation of the Surrogate Safety Assessment Model for Assessment of Intersection Safety", Transportation Research Board, No. 2432, pp. 1-9, Washington, D.C., 2014.

[2] Italian Ministry of Infrastructures & Transport, "Norme funzionali e geometriche per la costruzione delle intersezioni stradali", DM n. 1699 of 19/04/2006, Rome, 2006.

[3] G. Maycock, and R. D. Hall., "Accidents at four-arm roundabouts", PTRC Summer Annual Conference, Brighton, United Kingdom, 1984.

[4] O. K. Arndt, and R. J. Troutbeck., "Relationship Between Roundabout Geometry and Accident Rates", Transportation research circular, 1998.

[5] M. S. Ghanim, and K. Shaaban, "A Case Study for Surrogate Safety Assessment Model in Predicting Real-Life Conflicts", Arabian Journal for Science and Engineering, Vol. 44 pp. 4225-4231, 2019.

[6] T. Tollazzi, "Alternative Types of Roundabouts. An informational guide". Springer, Berlin, Germany, 2015.

[7] A. B. Silva, L. Vasconcelos, and S. Santos, "Moving from Conventional Roundabouts to Turbo-Roundabouts", EWGT2013 – 16th Meeting of the EURO Working Group on Transportation. Procedia - Social and Behavioral Sciences, Vol. 111, pp. 137-146, 2014.

[8] G. Tesoriere, T. Campisi, A. Canale, and T. Zgrablić, "The Surrogate Safety Appraisal of the Unconventional Elliptical and Turbo Roundabouts", Journal of Advanced Transportation, pp. 1-9, 2018.

[9] A. Pratelli, R. R. Souleyrette, and L. Brocchini, "Two-Geometry Roundabouts: Design Principles", Transportation Research Procedia, Vol. 64, pp. 299-307, 2022.

[10] A. Pratelli, and L. Brocchini, "Two-Geometry Roundabouts: Estimation of Capacity", Transportation Research Procedia, Vol. 64, pp. 232-239, 2022.

[11] A. R. Alozi, and M. Hussein, "Multi-criteria comparative assessment of unconventional roundabout designs", International Journal of Transportation Science and Technology, Vol. 11, pp. 158-173, 2022.

[12] D. Ciampa, M. Diomedi, F. Giglio, S. Olita, U. Petruccelli, and C. Restaino, "Effectiveness of unconventional roundabouts in the design of suburban intersections", European Trasnport, Vol. 80, 2020.

[13] A. Pratelli, P. Sechi, and R. R. Souleyrette, "Upgrading Traffic Circles to Modern Roundabouts to improve Safety and Efficiency – Case Studies from Italy", Promet – Traffic&Transportation, Vol. 30, No. 2, pp. 217-229, 2018.

[14] D. Gettman, "Surrogate Safety Measures from Traffic Simulation Models", Transportation Research Record Journal of the Transportation Research Board, No. 1840, 2003.

[15] O. Giuffrè, A. Grana, M. L. Tumminello, T. Giuffrè, and S. Trubia, "Surrogate Measures of Safety at Roundabouts in AIMSUN and VISSIM Environment", 15th Scientific and Technical Conference Transport, Katowice, Poland, 2018.

[16] A. P. Tarko, "Use of crash surrogates and exceedance statistics to estimate road safety", Accident; Analysis and P revention, Vol. 45, pp. 230-240, 2012.

[17] S. Daniels, T. Brijs, E. Nuyts, and G. Wets, "Extended prediction models for crashes at roundabouts", Safety Science - SAF SCI, Vol. 49, pp. 198-207, 2011.

[18] A. Abdelhalim, and M. Abbas, "An Assessment of Safety-Based Driver Behavior Modeling in Microscopic Simulation Utilizing Real-Time Vehicle Trajectories", 2022.

[19] A. Pratelli, L. Brocchini, and N. Francesconi, "Estimating and Updating Gap Acceptance Parameters for Hcm6th Roundabout Capacity Model Applications", WIT Transactions on Ecology and the Environment, Vol. 253, pp. 477-486, 2021.

[20] TRB, "Roundabouts: An Informational Guide - Second Edition". Washington, D.C.: The National, 2010.

[21] TRB, "Highway Capacity Manual 6th Edition: A Guide for Multimodal Mobility Analysis", Washington, D.C.: The National, 2016.

Employing HDF5 File Format for Marine Engine Systems Data Storage

Giuseppe Giannino, Michelangelo Tricarico, Andrea Orlando
Innovation and Development Centre
Isotta Fraschini Motori (IFM)
Bari, Italy
E-mail: {giuseppe.giannino, michelangelo.tricarico, andrea.orlando}@isottafraschini.it

***Abstract*—This early-stage paper describes a proposal for a standard approach to store data acquired from communication protocols mostly used in marine engines control systems for propulsion and power generation. The reasons to use Hierarchical Data Format version 5 (HDF5) are explained, and some concrete applications are presented based on real assets.**

Keywords - HDF5; Marine engines; Big data; Communication protocols; File format.

I. INTRODUCTION

A marine engine can be considered the heart of a vessel, independently from its specific application, i.e., propulsion or power generation for civil or institutional usage such as cruise or warship. Due to the high complexity of these engines and the very high reliability requirements as well as the large number of sensors, a lot of data is generated and exchanged via communication protocols.

These data can be considered the base for the development of innovative and always increasing diagnostic and predictive strategies, generally based on Knowledge-based models, Artificial Intelligence, and Statistical models. All of them are aimed to fault identification by replacing the simplest and largely used maintenance approaches [1].

These latter are mostly based on scheduled maintenance operations and threshold diagnostic and don't consider all historian data, so the result is not an optimized strategy.

In this context, Isotta Fraschini Motori (IFM) [2], a Fincantieri company, is investigating and developing custom diagnostic models to be integrated within the next generation of marine engine automation and control systems.

The developments aim to be suitable not just for classical and most used diesel Internal Combustion Engines (ICE) but also for new engines based on green fuels like methanol, hydrogen, and so on.

The main purpose of this research work is to explore the applicability and usage of the HDF5 file format as base for data storage.

Many current on-board storage systems provide ASCII/.txt/.csv files as output of the acquisition and storage process. These files guarantee easy access to data, but they were never designed for the massive scale of big data and tend to eat up resources unnecessarily (e.g., CPU-intensive) The reading and writing processes are not efficient and practical, especially when high performances are required during data processing phase [3] or real-time analysis.

The main idea behind this proposal is to define a standard approach for storing data acquired across multiple communication protocols installed on board of an engine for marine purposes that can help to make the information agnostic to the specific communication protocol and then speed up and facilitate the development of diagnostic and predictive algorithms and data analysis tools.

The structure of the paper is as follows: Section II introduces general material useful to understand the next sections, therefore an overview of HDF5 file format and the mainly used and investigated communication protocols is given. Section III explains why and how HDF5 can be a suitable solution for storing data acquired from marine engines over different communication protocols. Section IV briefly presents a benchmarking analysis, in terms of amount of used storage memory, between the proposed HDF5 and .csv file format, typically used in these contexts. Section V presents a real example, developed by IFM, of a software tool for visualizing and analyzing data stored into HDF5 files and acquired from marine engines. Conclusions and considerations for future steps are reported in Section VI.

II. GENERAL CONCEPTS

In this section, the main concepts orbiting around this proposal and useful to better understand the next results are reported.

A. HDF5 file format

Hierarchical Data Format (HDF) with the version HDF5 represents the most upgraded version of this data model. The word "hierarchical" in HDF refers to its tree-like structure.

When working with huge amounts of data, the availability of a conceptual model can help to organize and manage data, by visualizing mentally the structure itself especially in case of correlations among them [4].

More specifically, HDF is a data model, file format and I/O library designed for storing, exchanging, managing, and archiving complex data including scientific, engineering, and remote sensing data. An HDF file format has two main data objects, Groups and Datasets.

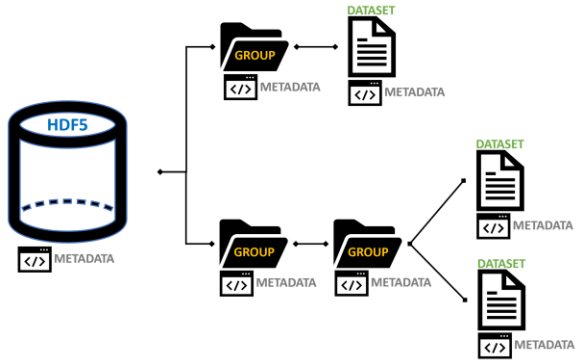

Figure 1. HDF5 - Basic components overview

Groups are the overarching structures aimed at creating and maintaining collections of related objects. Every file starts with a root group (*Root Group/)* [5].

Datasets, usually stored within groups, include a multidimensional array of elements together with additional information describing the dataset. This allows the HDF file format to be self-describing, which means that all groups and datasets can have additional information that describes their content [6]. This information is called Attributes and contain user-defined metadata, usually used to describe the nature (e.g., unit of measurement), and intended usage of dataset or group [5]. Attributes are defined based on the paradigm *key-value*, with a unique name for the specific object and a value associated. A very simple overview of the main HDF5 component is depicted in Figure 1.

An HDF file can store data consisting of different data types. Ten families or classes of HDF5 datatypes are currently supported: integer, floating-point, string, bitfield, opaque, compound (a kind of tuple type), reference, enum, variable-length sequence and array [7].

HDF5 (released in 1998 and identified as .hdf5 or .h5) has allowed to overcome the size limit of files and the number of objects in a file. HDF5 can have a flexible layout strategy defined on user needs. This can be done by using some important features of HDF5 file format as (i) specialized data storage options based on chunking, data compression, extendable arrays, and split files, (ii) Virtual File Layer (VFL) for different types of storage, such as single, multiple files, local memory, network protocol and files on parallel file system, (iii) Parallel IO through Message Passing Interface (MPI-IO). A good reference for all HDF5 details can be found on [8].

It is noteworthy to acknowledge that in the realm of big data other file formats are available, such as Parquet [9], ORC [10] and Avro [11]. These file formats provide column-wise or row-wise serialization of data, instead HDF5 stores multi-dimensional arrays, then by changing the arrangement of the arrays, the users can effectively store data either column-wise or row-wise. Matching the right storage file format is crucial since it can have impact on various fronts. In general, a good approach could be starting by analyzing the nature of data, then the data semantics and the purposes of the storage in terms of future operations on them.

In our specific application, here presented, the analysis starts from the HDF5 applicability for marine engine data storage but the idea behind this choice retains the possibility to extend the approach to the whole vessel system.

The complex concept of a vessel can be easily broken down in a combination of smaller sub-systems mounted on board of a self-consistent asset capable of operating far from mainland. Each sub-system has its own role and generates data.

In the context of building the smart vessel of the future, a common aggregation of all these data is needed to implement high level applications (e.g., fuel consumption optimization, unmanned asset, system maintenance optimization and remote assistance, safety improvements,

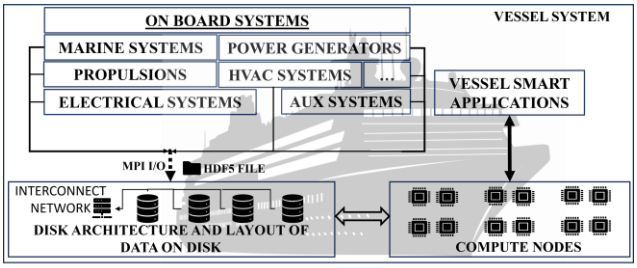

Figure 2. Centralized storage architecture for data of the future vessel sub-systems

strategic functionalities for war scenarios), as shown in Figure 2.

The hierarchical data organization, attributes storage, the multi-dimension capabilities and the intuitive data access based on the path-to-resource approach make the HDF5 a great candidate for supporting all previous concepts, especially for who will work at vessel application development level. Here, the main requirement is developing the smart functionalities with an easy and intuitive interface to data.

Further, the remarkable capabilities of parallel IO exposed by HDF5 can play a crucial role in implementing this kind of centralized system for on-board vessel data storage and processing.

Certainly, with the intention of expanding the study presented here it is not excluded that additional considerations will be done by studying and testing different data file formats, gradually the topic is explored.

B. CAN bus

Controller Area Network (CAN) is a serial communication protocol. It acts as a very common real-time communication protocol for control systems.

CAN communication has some unique characteristics that are distinct: robust real-time capability, reliable data transmission, and strong resistance to interference.

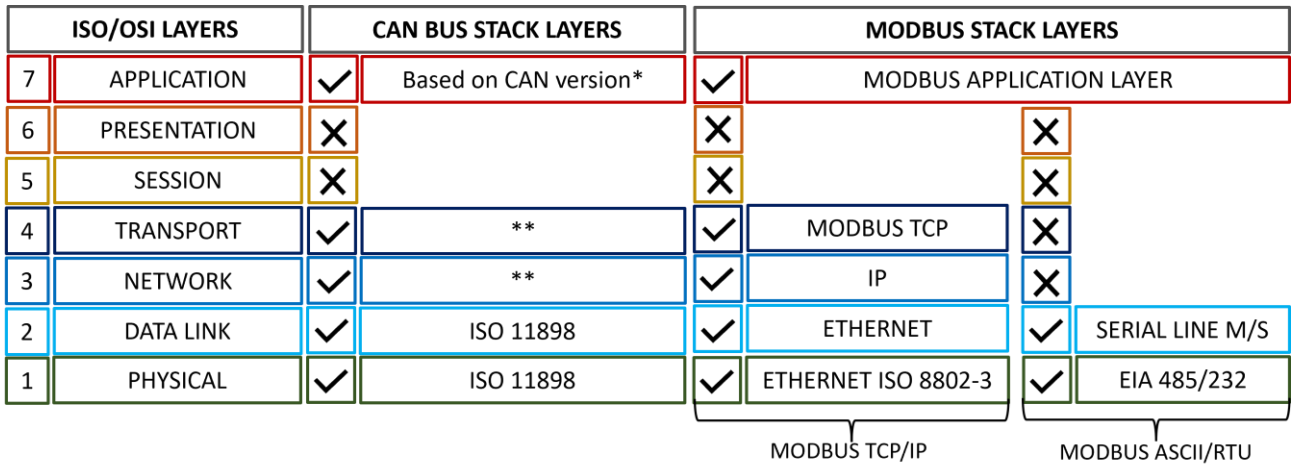

Figure 3. CAN bus and MODBUS ISO/OSI layers

It was initially introduced and used in automotive applications, then industrial automation, marine vessels, medical equipment, and industrial machinery have all utilized CAN due to its high performance and reliability.

Figure 3 is a custom IFM representation where the International Organization for Standardization (ISO) layers distribution for CAN protocol is shown.

Communication over CAN protocol is based on the principle that devices can transmit and receive messages using a shared bus, and a reliable transmission is ensured by a conflict detection mechanism. A good reference for the CAN bus protocol is the standard indicated in [12].

The primary limitation of CAN is the trade-off between transmission rate and distance. The maximum transmission rate allowed by CAN is 1 Mbps, but this rate is applicable only for shorter communication distances (less than 40 meters). If the communication distance exceeds this range, the transmission rate needs to be lowered to ensure reliable data transfer. Consequently, in long-distance communication, the transmission rate of CAN may correspondingly decrease. Hence, while CAN possess advantages such as reliability and real-time capability, it requires a balance between transmission rate and distance in long-distance communication. This trade-off is the main limit of CAN protocol.

The Society of Automotive Engineers (SAE) has developed a family of standards that pertain to the design and use of devices that transmit electronic signal and control information among vehicle components. In the field of engines, SAE has released the standard J1939, developed to exploit CAN protocol physical layer and much of the standard CAN data-link layer [13]. The maximum data rate of CAN J1939 is 250 Kbps and messages include a 29-bit identifier which defines (i) the message priority, (ii) who is the sender and (iii) what kind of data is contained within it.

C. MODBUS

MODBUS is a serial communication protocol originally developed in 1979 by Modicon for its Programmable Logic Controllers (PLC). Over the years, MODBUS has emerged as one of the most prevalent communication protocols in industrial control systems.

It operates as Master-Slave protocol, where there is a designated device acting as the master and other devices as slaves. The master device can both read and write data to the slave devices. The slave devices only transmit data upon receiving request from the master. The master assumes the role of communicator, sending inquiries and making requests to the servant. These requests may involve reading or writing data or performing specific operations. The slave provides appropriate responses accordingly. Details on MODBUS can be found in [14].

MODBUS primarily falls within layer 7 of the OSI Reference Model (the so-called “Application Layer”) and therefore is compatible with any lower-level communication protocols including EIA/TIA-232, EIA/TIA-485, Ethernet (via TCP/IP), as shown in Figure 3. Modbus data formats typically fall into two categories: ASCII and RTU format. In the first data format, ASCII characters are transmitted, where each character is represented by two bytes. With the latter, data format is transmitted in binary form.

D. Marine engine's automation & control systems topology

In Figure 4, a typical IFM automation and control system architecture for an ICE (aimed to genset and propulsion applications) is shown. The engine is a real asset produced by Isotta Fraschini Motori, the *16V170 G*.

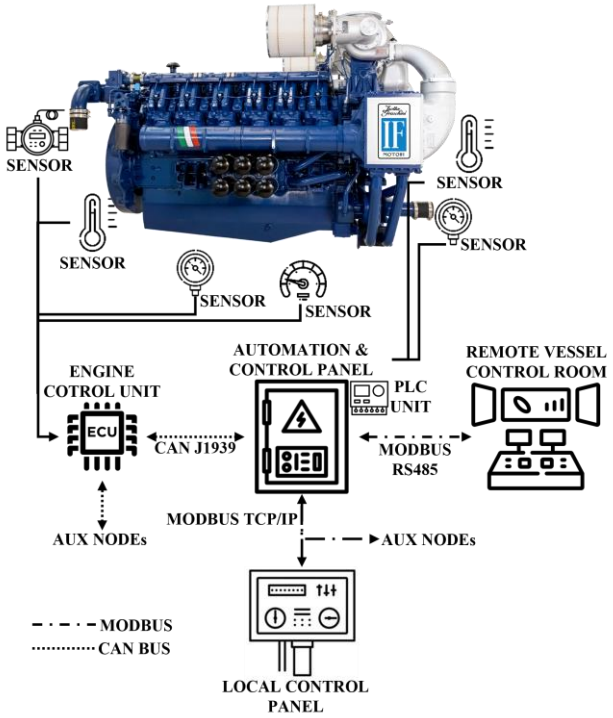

Figure 4. Marine Engine's high level topology

Recognizable are (i) automation & control panel where PLC unit and data-gathering system are located, (ii) the Engine Control Unit (ECU), (iii) local control panel where a Human Machine Interface (HMI) is positioned to facilitate the local asset control and settings, (iv) sensors and (v) communication protocols.

All these components contribute to the realization of a complex system aimed to guarantee a high level of efficiency of the asset in terms of performance, reliability, safety, and integration with vessels.

The ECU item, based on signal acquired from lots of sensors mounted on board the engine, is the heart of the full system and it's in charge of handling the injection cycles in terms of duration, fuel quantity and so on. Here, all the engine control logics and calibration maps are located and continuously executed. Then, the PLC based unit oversees all additional functionalities of the system, such as: communication with ECU, interfacing the engine with vessel central control room, running security logics based on additional sensors, managing the generated output power by interfacing the engine with the electrical machine (in case of genset application), and so on.

Generally, data can be of different nature and ranges: temperatures, pressures, speeds, flows and electrical measurements. Their treatment is not different just in terms of control logics but, as evidenced in the referenced figure, also for the communication protocols where they travel.

A so called data-gathering system is usually located in the control panel or, depending on customer requirements, it could be externally and remotely located. Its main goal is to automatically record, scan and retrieve the data with high speed and greater efficiency during asset running. A data-gathering system is designed, from the hardware and software perspective, i) to be highly and easily configurable from users based on specific application, ii) to be able to communicate with different and, at the same time, several communication protocols, iii) to be able to recover after a malfunction (e.g., node unavailable over a communication network or an anomalous condition that frozen the acquisition system), iv) to store high amount of data and cyclically delete oldest or not useful data based on user configuration.

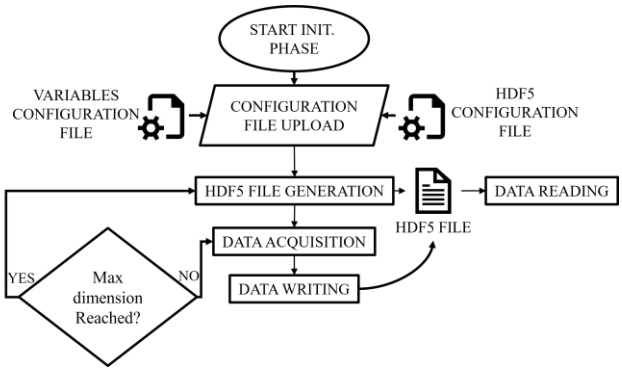

Figure 5. Working principle for data storage within HDF5 on data-gathering system

Depending on customer requirements, data could be stored locally or remotely on a cloud infrastructure. Both solutions must allow us to easily analyze data in case of troubleshooting or during scheduled maintenance operations. Further, independently from the adopted local or remote solution, the software component running on the data-gathering system, after a first check of the right connection with remote nodes (e.g., PLC, ECU, etc.) over different communication protocols, starts acquiring data, pre-processing them and storing data. If for some reason the connection with remote node is lost, then data source is down, the data-gathering system must try to restore communication and handle the storage processes.

III. HDF5 FOR COMMUNICATION PROTOCOLS DATA STORAGE

Based on previous concepts, this proposal aims to present the idea to use the HDF5 file format for storing the big amount of data acquired from the multiple communication protocols used on board marine engines.

The design of the HDF5 files starts by looking at the data organization provided for each communication protocol. Due to their differences at each level of the ISO/OSI stack the HDF5 file associated will be shaped differently.

Figure 5 shows the working principle of the data storage within an HDF5 file format, proposed, and adopted by IFM. The software entity, based on the specific user configuration where the enabled communication protocols and its parameters (Hardware interface, baud-rate, etc.) are

defined, initiates the HDF5 file uploading a so-called *variable configuration file*. The latter is a sort of matrix where variable names, addresses, units, descriptions, and other features are defined. The HDF5 file is configured by considering all parameters handled by the software library wrapping the HDF5 definitions, such as compression, chunk, dimensions and so on.

Data starts to be acquired over communication protocols, processed, and moved into the specific position of the referenced HDF5 file. The addressing of data is acted by means of a simple POSIX-style strategy with "/" separators indicating the hierarchy level.

The design of the software entity responsible for handling the specific HDF5 file structure and transferring the data is based on the design of an appropriate HDF5 architecture, in which the groups and datasets are defined according to the specific communication protocols and system requirements.

Based on this approach, two HDF5 architectures have been designed and here proposed. The adopted strategy, to define the HDF5 architecture, starts from considering how the variables sent over the protocols are statically mapped to be later decoded.

Generally, for the kind of asset under study, marine engines, we can consider the following two variables configuration files and requirements:

- CAN J1939 - Variables are filled into a so called .dbc file [15] that allows to decode the information needed to understand a vehicle's CAN bus traffic. Here, the main information reported are frame name, frame ID, format, length, and description. Each time a new frame is acquired over CAN bus, it is firstly processed by means of .dbc file and then the extracted information is written into HDF5 file within the specific assigned position. Frames (identified in Figure 6 as groups in the HDF5 file under the group ROOT_FOLDER/) can travel with different raster time in the network (*DATASET_Time* in each group stores this information). Internally CAN bus frames are filled with information called *signals* (e.g., group CANbus_PKT_1 contains *DATASET_SGN_1* to *DATASET_SGN_n*). Signals can have different lengths and for each the .dbc file contains information about datatype, length, multiplication factor, offset, minimum and maximum values, unit, and description stored within the HDF5 as attributes.
  The CAN J1939 HDF5 architecture can be also employed for different CAN version, such as standard CAN 2.0 A or B with very low impact.
- MODBUS - Variables are filled into a configuration matrix (usually a .csv file) where name, address, datatype, multiplication factor and unit are defined. Each sampling instant (a single GROUP *Timestamp/* stores this information under the main group ROOT_FOLDER/ as shown in Figure 7) the data-gathering system reads the whole MODBUS memory addresses from the remote node. In the HDF5 file dedicated to MODBUS communication protocol, a GROUP named *MODBUS_VARIABLES/* is initialized with a certain number of DATASET equal to the number of MODBUS variables to be acquired (1 ÷ *g* in Figure 6). Acquired values are firstly

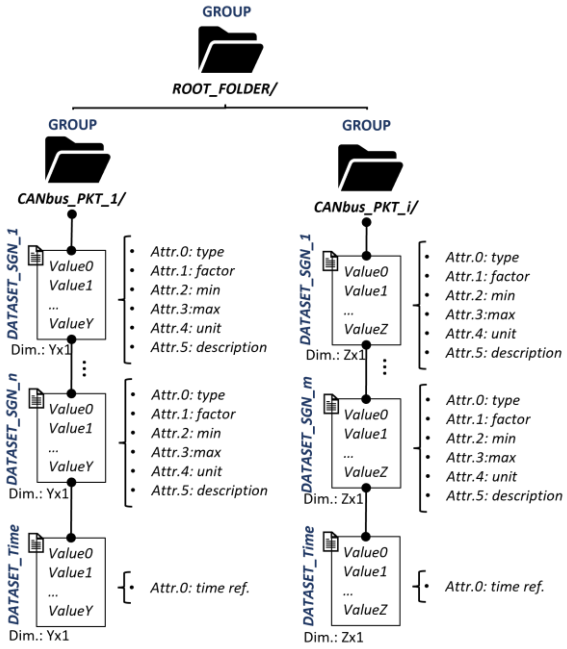

Figure 6. HDF5 proposed architecture for CAN-J1939

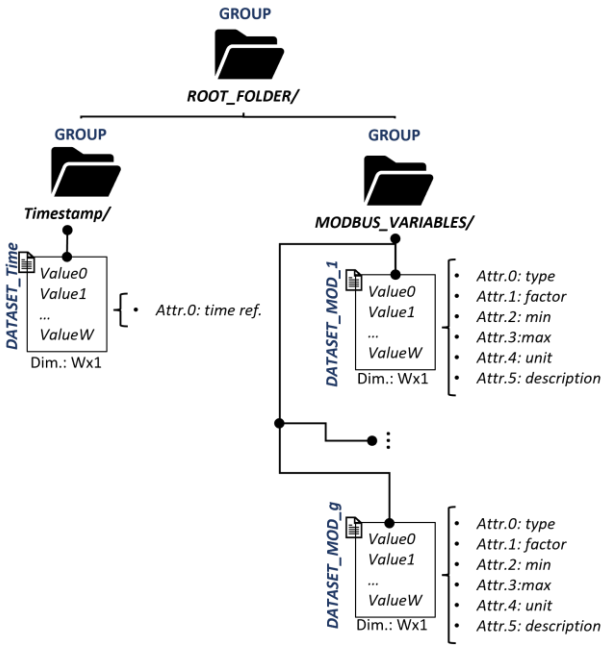

Figure 7. HDF5 proposed architecture for MODBUS

decoded by means of .csv configuration matrix and then stored into the HDF5 file within the specific dataset.

Starting from these considerations, the development has brought to obtain the HDF5 architectures reported in Figure 6 and Figure 7.

It is important to highlight that thanks to the HDF5 capabilities to store metadata under attributes associated to groups or datasets, all the information surrounding the data itself, such as units and descriptions, are not lose but filled into the HDF5 file also. This allows to keep each HDF5 file self-consistent in each moment of its life.

IV. FILE FORMATS BENCHMARKING ANALYSIS

Some preliminary considerations on the data storage and file opening time optimization are reported here. In Figure 8, a comparison with a typical file format, .csv, for storing diagnostic data on board marine engines is reported.

Within same conditions in terms of number of variables (457), sample rate (about 3.5 sec.), quantity of metadata stored (6 per variable), number of samples stored (17043 per each variable) and acquisition duration (about 16 hours) is easy to appreciate the improvement that HDF5 files allow to get. The opening time has been calculated by using the open-source HDFView software tool [8] for HDF5 file and Excel for the .csv file.

The optimization of about 20% in file dimension and more than 2 seconds less in opening time is not negligible considering that a huge amount of HDF5 files could be cyclically generated and stored on the same data-gathering system for the next analysis. Further improvements can be investigated in terms of data compression by exploring some change in the HDF5 architecture without losing the advantages related to having a well-defined data organization which allows us to easily retrieve data.

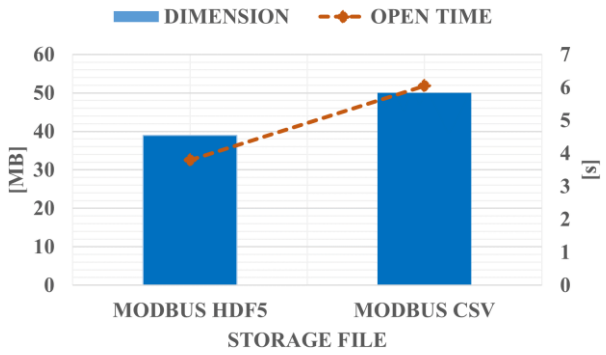

Figure 8. HDF5 vs CSV

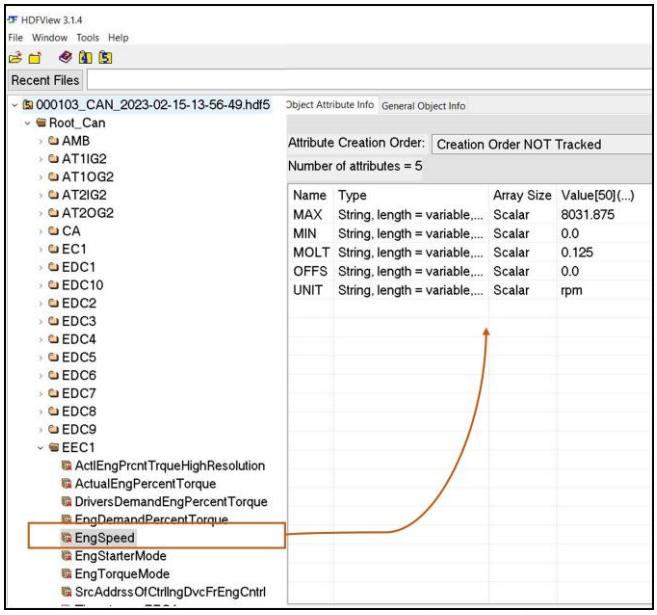

Figure 9. HDFView representation of an HDF5 file

V. EXAMPLE OF HDF5 FILE VISUALIZATION AND ANALYSIS TOOL

After data are stored into an HDF5 file to be able to visualize and analyze them a tool is required. Some open-source SW tools are available on the web. In Figure 9 a screenshot of the popular tool HDFView shows how an HDF5 file, compliant to the architecture proposed in this publication and described in section III, appears. In particular, the figure is referred to the specific CAN bus protocol based on the tree architecture illustrated in Figure 6.

Looking at the left pane, identifiable are (i) group ROOT_FOLDER/ named as *Root_Can*, (ii) group CANbus_PKT_i/ named as *EEC1* in the example and (iii) dataset DATASET_SGN_n/ named as *EngSpeed* (squared) with its specific attributes shown in the right pane of the tool. The dataset EngSpeed is filled with the data collected over CAN bus protocol (visible in HDF5View by clicking on the dataset itself).

In our case, we have developed internally a custom visualization and analysis tool capable of running on Windows and Linux-kernel based machines. The idea is to have the SW tool installed on the assistant operator laptop and/or on the local HMI of the automation and control systems to always give the possibility to study historical data stored into the HDF5 files. A time synchronization process has been implemented so that the data-gathering system is time referenced with PLC unit in turn synchronized with the vessel control systems.

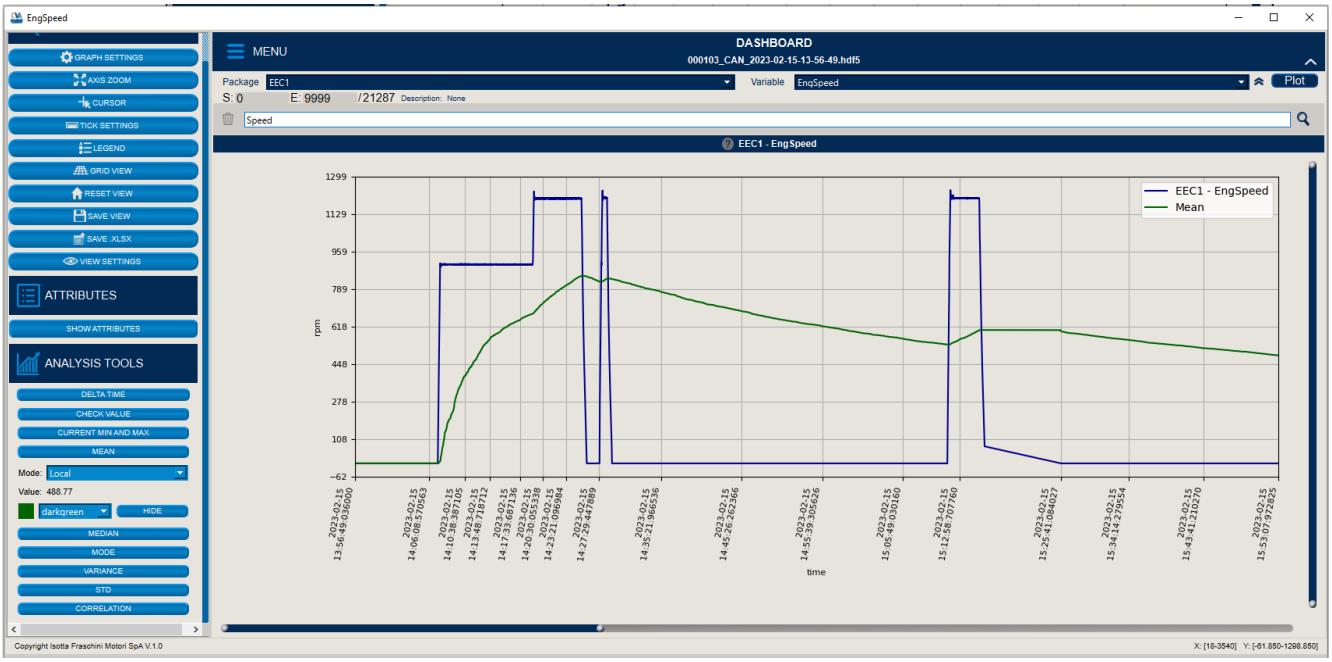

Figure 10. HDF5 - Visualization and Analysis Tool (IFM custom)

Figure 10 shows a sample screenshot of the SW tool developed in IFM. It allows us to view the stored data but also to study some statistical parameters and compare variables by time references for correlation studies. Every time a new HDF5 file is uploaded the software tool first identifies the communication protocols and recognizes the architecture. Then, when the user selects specific variables, the SW tool accesses the specific path and uploads data and metadata.

The specific data plotted in Figure 9 is the EngSpeed dataset extracted from the same HDF5 file analyzed in Figure 9.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we propose an approach to standardize the storage of data acquired over simple or complex systems where single or multiple communication protocols are involved to share functional and diagnostic information. This could speed up the process of democratizing the data usability among different vessel subsystem providers and facilitate the data processing and analysis for different end-users (e.g., maintenance operators, engineers, etc.). The advantages of using HDF5 files are mainly (i) the capability to organize data in user-friendly architecture where data can be written and read using the easy approach of path-to-resource and (ii) optimizing the data storage in terms of required memory.

The usage presented in this paper is primarily focused on marine engines, where on-board data storage and high level applications need to be implemented and connection to remote resources is not always and easily feasible for security or strategic reasons. With a wider-ranging look, it is possible to bring the same approach in different contexts such as automotive and micro-mobility.

Subject to future work further developments have been planned in IFM to optimize and deepen the usage of HDF5 file format within the specific scenario. Extended on-field tests will be executed to evaluate the impact of HDF5 file format in terms of storage optimization and ease and speed of data usage by high level applications.

ACKNOWLEDGMENT

This project is supported (was supported) by founds of *Fondo Europeo di Sviluppo Regionale Puglia* (Italy) – *POR Puglia 2014-2020*, grant number PBJMCM8 (MIR: A0101.168).

REFERENCES

[1] J.A. Pagan, “Marine Diesel Engine Diagnosis System Based on Thermodynamic Model and Artificial Intelligence", *PhD thesis, Polytechnic University of Cartagena*, 2017.

[2] Isotta Fraschini Motori Website [Online]. Available from www.isottafraschini.it (accessed August 24, 2023).

[3] A. Pfeiffer, I. Bausch-Gall, and M. Otter “Proposal for a Standard Time Series File Format in HDF5”, *Proceedings of the 9th International Modelica Conference*, pp. 495-506, September 2012, DOI 10.3384/ecp12076495.

[4] K. Läufer and K. Hinsen “Five Good Reasons to Use the Hierarchical Data Format”, Copublished by the IEEE CS and the AIP, September/October 2010, DOI 10.1109/MCSE.2010.107.

[5] M. Yang, R. E. McGrath, and M. Folk, “HDF5 – A High Performance Data Format for Earth Science”, *21st International Conference on Interactive Information*

Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology, January 2005.

[6] S. Ambatipudi and S. Byna, “A Comparison of HDF5, Zarr, and netCDF4 in Performing Common I/O Operations”, Publisched on www.arxiv.org, February 2023 (accessed August 24, 2023).

[7] M. Folk, Q. Koziol, and E. Poirmal, “An Overview of the HDF5 technology suite and its application”, Proc. EDBT/ICDT 2011 Workshop on Array Databases, pp. 36-47, March 2011, DOI 10.1145/1966895.1966900.

[8] HDFgroup Website. [Online]. Available from www.hdfgroup.org (accessed August 24, 2023).

[9] Apache Parquet Website. [Online]. Available from https://parquet.apache.org/ (accessed August 24, 2023).

[10] Apache ORC Website. [Online]. Available from https://orc.apache.org/docs/ (accessed August 24, 2023).

[11] Apache AVRO Website. [Online]. Available from https://avro.apache.org/ (accessed August 24, 2023).

[12] ISO Website. [Online]. Available from www.iso.org (accessed August 24, 2023).

[13] SAE Website. [Online]. Available from www.sae.org (accessed August 24, 2023).

[14] MODBUS Website. [Online]. Available from www.modbus.org (accessed August 24, 2023).

[15] Open Vehicles Website. [Online]. Available from https://docs.openvehicles.com/en/latest/components/vehicle_dbc/docs/dbc-primer.html (accessed August 24, 2023).

What Do You Call Your Analytical Endeavours?

An Analysis of Term Usage in German Job Openings from 2017 to 2022

Finja Below
Department of Computer Science
NORDAKADEMIE
Elmshorn, Germany
finja.below.i16a@nordakdemie.org

Uwe Neuhaus
Department of Media Education
Europa-Universität Flensburg
Flensburg, Germany
uwe.neuhaus@uni-flensburg.de

Michael Schulz
Department of Computer Science
NORDAKADEMIE
Elmshorn, Germany
michael.schulz@nordakademie.de

***Abstract*—This paper examines which terms are used for analytical information systems in business practice and how their use has developed over time. For this purpose, a total of 3,000 German job openings from 2017 to 2022 are subjected to a frequency analysis. Eight relevant generic terms are identified and divided into first- and second-order generic terms based on their definition of relevance to practice. A detailed examination of the common occurrence of the generic terms and the change in their use over time shows that the traditional term Business Intelligence continues to dominate in practice and that the terms Artificial Intelligence and Data Science are becoming increasingly established in a more technical context. More specific terms, such as Advanced Analytics, Machine Learning and Data Mining, on the other hand, are becoming less important for describing analytical information systems.**

Keywords: Analytical Information Systems; Business Intelligence; Artificial Intelligence; job openings; frequency analysis.

I. INTRODUCTION

Procedures used for data analysis to support decision-making in a business context have existed for decades [4] [6]. Over time, different terms for such Analytical Information Systems (AIS) have emerged. These emphasize certain aspects, focus on a specific application area, or pick up on currently "fashionable" terms [11]. When new terms emerge, it usually takes a while for standardized definitions and a common image to take hold [8]. In addition, the use of terms often differs in the scientific literature and in business practice [11].

This article analyzes which terms are used for AIS in business practice. German job openings from 2017 to 2022 are used for the analysis. It is assumed that job openings are a suitable source of data, as (a) they contain current data and (b) companies need to formulate tasks and requirements for applicants realistically in order to find suitable employees [14].

The aim of this work is to identify the Generic Terms (GT) used currently in business practice. These are determined with the help of a frequency analysis. In addition, the relationship between the GT and the changes in term usage over time are examined. Not part of this article is the explanation of the identified practice definition and relevance shifts over time. These will be analyzed in later research.

This paper is divided into five sections. Following this introduction, Section 2 defines the terms relevant to this article. Next, Section 3 explains the data basis used for the study. Subsequently, Section 4 presents the results of the data analysis. The article concludes with an outlook on future research in Section 5.

II. DEFINITIONS OF TERMS

The job openings studied are from the AIS sector. With such systems, analyses can be carried out to support management in decision-making. They are usually based on large volumes of data that have to be processed depending on the analysis objective [7]. The examination of job openings described in the following sections revealed that a variety of GT is used in business practice. In particular, the terms used emphasize the scope of the analyses, their degree of complexity, or their methodology. In the following, the most frequently occurring GT are presented.

The term Business Intelligence (BI) describes concepts and methods for supporting managers in decision-making with the help of IT systems [11]. The main objective of BI is to make data available and prepare it for analysis [9]. As a result, there is a close connection to the term Business Analytics (BA). It is often regarded as an extension of BI to include more advanced forms of analysis [9]. According to other sources, on the other hand, the term represents an addition to BI of these advanced analysis methods (cf. e.g. [1]).

To emphasize the increasing complexity of analytics, some authors use the term Advanced Analytics (AA) [3]. With its mathematically and algorithmically sophisticated methods, AA goes beyond the traditional data analysis of BI and BA [1]. The term Data Analytics (DA), on the other hand, emphasizes the importance of increasingly large, multifaceted, and frequently changing data sets for decision making. DA encompasses the entire process from data collection and analysis to the presentation of results [13].

While the terms BI and BA emphasize the economic context, the terms Data Science (DS), Artificial Intelligence (AI), Data Mining (DM) and Machine Learning (ML) focus more on the methods used. DS is an interdisciplinary field in which insights are extracted semiautomatically from sometimes complex data [12]. It thus has great intersections with AI. This term covers methods that enable a system to

interpret data and use it to learn patterns [10]. Subfields of AI are DM, which provides methods and algorithms to discover previously unknown relationships in large data sets [2], and ML, which allows computers to improve automatically based on experience [5].

III. DATA BASIS

This study continues the work of [14] and adds a temporal component to it.

Job openings collected once a year from 2017 to 2022 on the largest job portals in Germany serve as the data basis. Since the goal of this study is to identify GT used in business practice, a search that was too narrow - e.g., for terms such as BI or AI - could have resulted in relevant information being overlooked. For this reason, "analy*" was used as the search term, with the asterisk replaced by any string that combined to form a German- or English-language word. Only job openings in which the entered term was mentioned in the job title and which were written in German were considered.

Furthermore, job openings that clearly do not fall into the area of analysis with the goal of decision support were eliminated for the following investigations. This eliminated primarily job openings from the natural sciences and those that focused on classic software development tasks. This elimination step was carried out independently by all three authors in order to minimize the degree of subjectivity in the compilation of the data basis. Divergent decisions were discussed to arrive at a mutually accepted result. As a final data set, 500 job openings were used for each year under consideration.

Based on the resulting 3,000 job openings, a frequency analysis was performed. In this work, terms were chosen as the unit of analysis, formed either by single words or by n-grams.

Only terms that appear in more than 5% of all job openings were considered relevant GT. Candidates that are mentioned in 3-5% of the job openings and thus just miss the threshold are the terms Predictive Analytics, Digital Analytics, Web Analytics, Big Data Analytics, and Data Engineering.

IV. DESCRIPTIVE DATA ANALYSIS

Figure 1 shows the development of the GT under investigation from 2017/2018 (starting point of the arrows) to 2021/2022 (end point of the arrows). For the analysis, the data of two consecutive years were combined. Data of the years 2019/2020 are not shown for reasons of clarity. The value on the x-axis indicates what proportion of sole use of GT is accounted for by the respective term. Sole use means that none of the other GT considered was mentioned in the job openings. In 2021/2022, sole mention of GT was identified in 298 job openings (2017/2018: 293, 2019/2020:353).

The y-axis shows the proportion of the respective terms in job openings in which more than one GT was mentioned.

The area below the diagonal (Area I) thus contains those GT whose share of sole use is greater than the share of their use in combination with other GT. They can be defined as GT of the first order. They are GT that are sufficiently comprehensive and precise to fully describe an analytics area.

The area above the diagonal (Area II), on the other hand, contains those GT whose share of common use with other GT is higher than the share of their stand-alone use. They are defined as second-order GT; depending on the context, they often cannot stand alone but require other GT for specification.

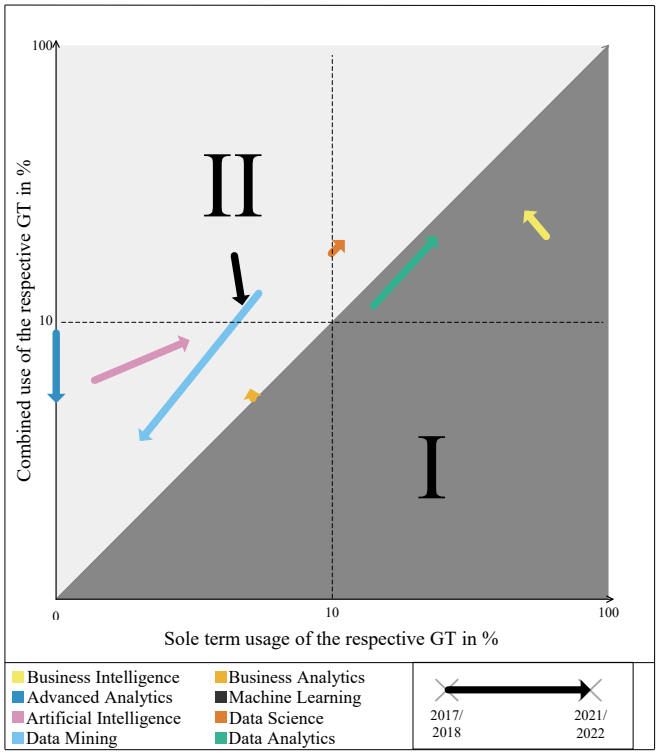

Figure 1. GT development from 2017/2018 to 2021/2022.

The diagram shows the practice definition and practical relevance shift of GT in the period under consideration:

- The *practice definition* of GT is said to be stable if the ratio between sole and joint use does not change, i.e., there is no shift over time or a shift parallel to the diagonal.
- An increase (decrease) of the *practical relevance* can be determined by the fact that the distance of an GT to the origin of the coordinate system increases (decreases) with respect to Figure 1.

Area I of Figure 1 contains the two terms that form GT of the first order: DA and BI. Even though BI alone has lost shares, it remains the most frequently used term overall. In 2021/2022, it was included in 32.5% of job openings (2017/2018: 32.8%, 2019/2020: 39.3%). DA, on the other hand, did not change in its practice definition during the period under consideration, but its relevance did: While the term was included in 12.6% of job openings in 2017/2018, it was already 21.4% in 2021/2022.

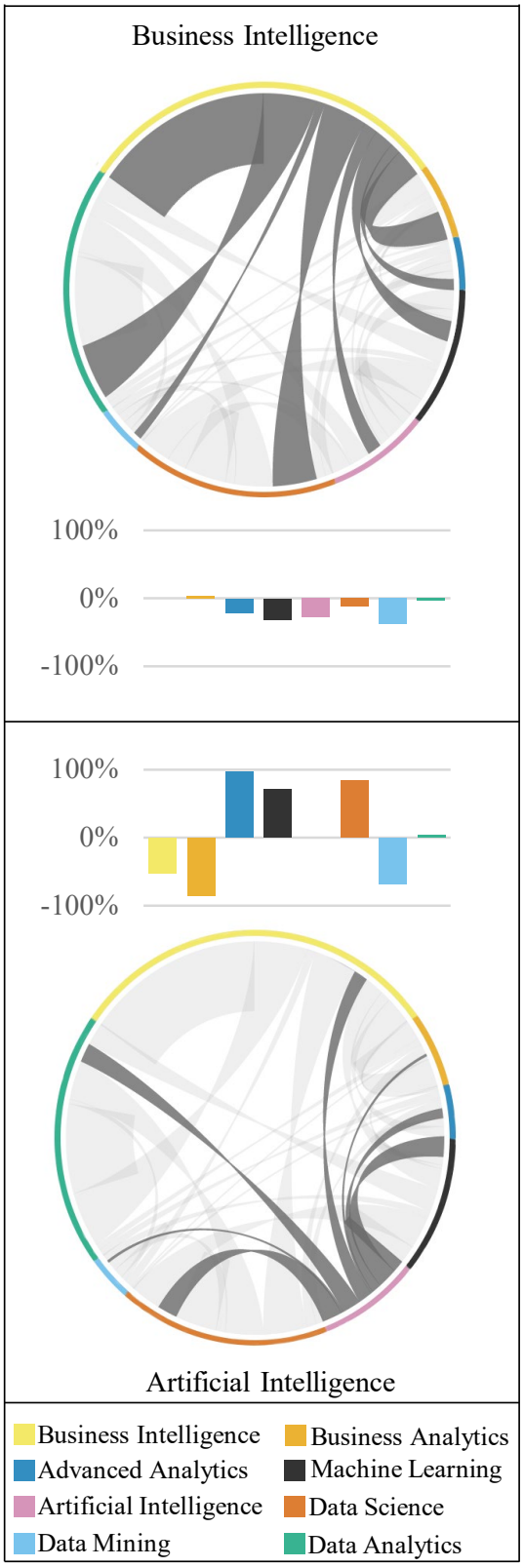

Figure 2. Interrelationships of the GT.

The greater number of GT examined are positioned in Area II of Figure 1 and are thus second-order GT. Only the practice definition and relevance of the term BA remained almost stable over the six years under consideration. The terms AA, which was insignificant as a stand-alone term during the entire period under consideration, and DM have lost relevance while the practice definition has remained unchanged.

The term ML has also lost importance overall, but has gained slightly as a stand-alone term; its practice definition has shifted accordingly. The terms AI and DS, on the other hand, gained overall relevance. The practice definition of DS has remained stable, while AI has become more established as a stand-alone term.

Figure 2 uses Chord diagrams for the years 2021/2022 to show which GT appear in the job openings in combination with which other GT. The bar charts show the deviation of the identified individual relationships from the relationships when all GT are considered (a priori probability). The first part of the figure shows the BI relationships, since this is the most frequently used first-order GT, and the second part the AI relationships, since the relevance of this second-order GT increased the most in the period under review.

The frequency with which GT is mentioned is made clear by the size of the respective occupied circle section in the Chord diagram. In the case of BI, it can be seen that the term is most frequently mentioned on its own (34%). If it occurs together with other terms, these are in particular DA (20%), DS (16%) and BA (10%), where the correlation is slightly weaker in each case than would have been expected when considering the a priori probability. Mention of BI together with more specialized and more technical generic terms such as ML (7%), AI (5%) and DM (3%), on the other hand, are significantly less frequent.

The generic term AI rarely stands alone (6%); it is very often accompanied by other GT that specify the focus of the job openings. The most frequent combinations are DS (22%), ML (22%) and DA (21%). The first two also occur significantly more frequently than might have been expected (DS +84%, ML +71%). At -52% significantly less frequently than would have been expected when considering the a priori probability, AI is mentioned in the job openings together with BI (15%). The term AA occurs 97% more frequently in combination with AI than would have been expected.

V. Conclusion and Future Research

In this article, the terms most intensively used by companies at present for AIS were identified. For this purpose, 3,000 job openings serves as analytical data basis. Furthermore, it was examined to what extent the practical definition and practical relevance of GT changed from 2017 to 2022. Using the terms Business Intelligence and Artificial Intelligence as examples, it was shown which GT are frequently or rarely used in combination with other GT.

In the future, more in-depth text analytics will be applied based on the described data set from 2017 to 2022. The job openings contain further information that allows a better understanding of the focus and developments of the use of AIS in operational practice. This includes the qualifications required of applicants in the various phases of data analysis,

the analysis software skills needed, and the planned areas of application and types of tasks. Frequently, the ads also name the driving developments underlying the analytics desire (e.g., digitalization, Industry 4.0, or Internet of Things). It is planned to extract and cluster this information so that its relevance can be determined and significant changes over time can be shown.

This more detailed investigation will also make it possible to work out more concretely how the GT under consideration are used in operational practice. On this basis, an attempt can then be made to explain the shifts in practice definition and relevance presented in this article and to determine whether there has been a fundamental change in data analysis for decision support or whether it is simply a shift in the trend of terms used.

REFERENCES

[1] P. Chamoni, "Data Science - Entwicklungslinien und Trends" [in English: "Data Science - Lines of Development and Trends"] in D. Frick, A. Gadatsch, J. Kaufmann, B. Lankes, C. Quix, A. Schmidt and U. Schmitz, (Eds.), Data Science: Konzepte, Erfahrungen, Fallstudien und Praxis [in English: Data Science: Concepts, Experiences, Case Studies and Practice], Springer Fachmedien, pp. XII–XVI, 2021.

[2] J. Cleve and U. Lämmel, Data Mining (3. ed.). De Gruyter, 2020.

[3] C. Dittmar, "Die nächste Evolutionsstufe von AIS: Big Data" [in English: The next evolutionary stage of AIS: Big Data] in P. Gluchowski and P. Chamoni (Eds.), Analytische Informationssysteme: Business Intelligence-Technologien und -Anwendungen [in English: Analytical Information Systems: Business Intelligence Technologies and Applications], (5. ed., vol. 51). Springer Gabler, pp. 55-65, 2016.

[4] G. A. Gorry and M. S. S. Morton, "A framework for management information systems", Sloan Management Review, vol. 13, no. 1, pp. 56-79, 1971.

[5] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects", *Science*, Vol 349, Issue 6245, pp. 255-260, 2015.

[6] P. G. Keen, "Decision support systems: a research perspective" in Decision support systems: Issues and challenges: Proceedings of an international task force meeting, pp. 23-44, 1980.

[7] H. Kempter, Betriebliche Informationssysteme: Datenmanagement und Datenanalyse [in English: Business Information Systems: Data Management and Data Analysis]. Kohlhammer Verlag, 2017.

[8] J. H. Kroeze, M. C. Matthee and T. J. D. Bothma, "Differentiating data- and text-mining terminology" in J. Eloff (Ed.), ACM Conferences. Proceedings of the 2003 annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology, South African Institute for Computer Scientists and Information Technologists, pp. 93-101, 2003.

[9] E.-P. Lim, H. Chen and G. Chen, "Business Intelligence and Analytics", ACM Transactions on Management Information Systems, vol. 3, no. 4, pp. 1–10, 2013.

[10] P. Mikalef and M. Gupta, "Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance", Information & Management, vol. 58, no. 3: 103434, 2021.

[11] S. Negash, "Business Intelligence", Communications of the Association for Information Systems, vol. 13, no. 15, pp. 177-195, 2004.

[12] M. Schulz, et al., "Introducing DASC-PM: A Data Science Process Model", ACIS 2020 Proceedings, vol. 45, 2020.

[13] C.-W. Tsai, C.-F. Lai, H.-C. Chao amd A. V. Vasilakos, "Big data analytics: a survey", Journal of Big Data, vol. 2, no. 1, pp. 1-32, 2015.

[14] U. Neuhaus and M. Schulz, "Wie nennt ihr, was ihr da tut? Der Begriff "Analytics" in der betrieblichen Praxis" [in English: "What do you call what you're doing? The term 'Analytics' in operational practice", NORDBLICK, no. 6, pp. 4-29, 2018.

Text Classification Using a Word-Reduced Graph

Hiromu Nakajima
Major in Computer and Information Sciences
Graduate School of Science and Engineering,
Ibaraki University
Hitachi, Ibaraki, Japan
e-mail: 22nm738g@vc.ibaraki.ac.jp

Minoru Sasaki
Dept. of Computer and Information Sciences
Faculty of Engineering, Ibaraki University
Hitachi, Ibaraki, Japan
e-mail: minoru.sasaki.01@vc.ibaraki.ac.jp

Abstract— **Text classification, which determines the label of a document based on cues such as the co-occurrence of words and their frequency of occurrence, has been studied in various approaches to date. Conventional text classification methods using graph structure data express the relationship between words, the relationship between words and documents, and the relationship between documents in terms of the weights of edges between each node. They are then trained by inputting into a graph neural network. However, text classification methods using those graph-structured data require a very large amount of memory, and therefore, in some environments, they do not work properly or cannot handle large data. In this study, we propose a graph structure that is more compact than conventional methods by removing words that appear in only one document and are considered unnecessary for text classification. In addition to save memory, the proposed method can use a larger trained model by utilizing the saved memory. The results showed that the method succeeded in saving memory while maintaining the accuracy of the conventional method. By utilizing the saved memory, the proposed method succeeded in using larger trained models, and the classification accuracy of the proposed method was dramatically improved compared to the conventional method.**

Keywords- text classification; graph convolutional neural network; semi-supervised learning.

I. INTRODUCTION

Text classification is the task of estimating the appropriate label for a given document from a predefined set of labels. This text classification technique has been applied in the real world to automate the task of classifying documents by humans. Many researchers are interested in developing applications that take advantage of text classification techniques, such as spam classification, topic labeling, and sentiment analysis.

Recently, Graph Convolutional Neural networks (GCNs) [1], which can take advantage of data in graph structures, have been used to solve text classification tasks. TextGCN [2], VGCN-BERT [3], and BertGCN [4] are examples of text classification methods that utilize data from graph structures. In TextGCN [2], word and document nodes are represented on the same graph (heterogeneous graph), which is input into GCNs for learning. VGCN-BERT [3] constructs a graph based on the word embedding and word co-occurrence information in Bidirectional Encoder Representations from Transformers (BERT), and learns by inputting the graph into Vocabulary Graph Convolutional Network (VGCN). BertGCN [4] is a text classification method that combines the advantages of transductive learning of GCNs with the knowledge obtained from large-scale prior learning of BERT. The graphs produced by these graph-based text classification methods use relations between words and between words and documents, but do not use relations between documents, and are prone to topic drift. Therefore, in [5], we proposed a graph structure that uses relations between documents to solve this problem. The method of [5] boasts the best performance among existing methods for text classification on three datasets (20NG, R8, and Ohsumed). However, a new problem arises from the addition of relationships between documents to the graph, which increases the size of the graph and requires a lot of memory space. Therefore, we considered that compacting the size of the graph would reduce the memory requirement and allow the use of larger data and larger trained models.

The purpose of this study is twofold. The first is to successfully save memory by constructing a graph structure that removes words considered unnecessary in text classification to solve the problem of insufficient memory. The second is to improve classification accuracy over conventional methods by utilizing the reduced memory and using larger trained models. Specifically, words that appear in only one document are removed from the graph, reducing both the weights of edges between word nodes and the weights of edges between word nodes and document nodes, thereby saving memory. We believe that this will result in a graph that is more compact than the graphs created by conventional methods, saving memory and improving the accuracy of text classification by using a larger trained model.

This paper is organized as follows. In Section 2, we first describe graph neural networks used for text classification and existing research on text classification using graphs. After that, the structure of graphs created in conventional methods is described. In Section 3, we describe the graph structure of the proposed method and the processing after graph construction. In Section 4, we describe the experiments we conducted to evaluate the proposed method and show the experimental results. In Section 5, we discuss the experimental results presented in Section 4 and conclude in Section 6.

II. RELATED WORKS

A. Text Classification Using Graph Neural Networks

Graph Neural Network (GNN) [6] is a neural network that learns relationships between graph nodes via the edges that connect them. There are several types of GNNs depending on their form. Graph Convolutional Neural networks (GCNs) [1]

is a neural network that takes a graph as input and learns the relationship between nodes of interest and their neighbors through convolutional computation using weights assigned to the edges between the nodes. Graph Autoencoder (GAE) [7] is an extension of autoencoder, which extracts important features by dimensionality reduction of input data, to handle graph data as well. Graph Attention Network (GAT) [8] is a neural network that updates and learns node features by multiplying the weights of edges between nodes by Attention, a coefficient representing the importance of neighboring nodes. GNNs are used in a wide range of tasks in the field of machine learning, such as relation extraction, text generation, machine translation, and question answering, and have demonstrated high performance. The success of GNNs in these wide-ranging tasks has motivated researchers to study text classification methods based on GNNs, and in particular, many text classification methods based on GCNs have been proposed. In TextGCN [2], document and word nodes are represented on the same graph (heterogeneous graph), which is input into GCNs for training. In recent years, text classification methods that combine large-scale pre-trained models such as BERT with GCNs have also been studied extensively. VGCN-BERT [3] constructs a graph based on word co-occurrence information and BERT's word embedding, and inputs the graph into GCNs for learning. In BertGCN [4], a heterogeneous graph of words and documents is constructed based on word co-occurrence information and BERT's document embedding, and the graph is input into GCNs for learning. In [5] we propose a graph structure that exploits relationships between documents.

The detailed description of BertGCN and the graph structure proposed in [5] is given in the Section II-B.

B. Graph Structure of Conventional Methods

BertGCN is a text classification method that combines the knowledge of BERT obtained by large-scale pre-training utilizing large unlabeled data with the transductive learning of GCNs, and was proposed by Lin et al. in July 2021. In BertGCN, each document is input into BERT, the document vector is extracted from its output, and it is input into GCNs as an initial representation of document nodes together with a heterogeneous graph of documents and words for training.

Lin et al. distinguish the names of the training models according to the pre-trained model of BERT and the type of GNN used. The correspondence between pre-trained models and model names is shown in Table I. In this study, RoBERTaGCN, a model employing roberta-base and GCNs, is the target of improvement.

TABLE I. NAMES OF THE MODELS.

Pre-Trained Model	GNN	Name of Model
bert-base	GCN	BertGCN
roberta-base	GCN	RoBERTaGCN
bert-base	GAT	BertGAT
roberta-base	GAT	RoBERTaGAT

RoBERTaGCN defines weights between nodes on heterogeneous graphs of words and documents as in (1).

$$A_{i,j} = \begin{cases} PPMI(i,j), & i,j \; are \; words \; and \; i \neq j \\ TF-IDF(i,j), & i \; is \; document, j \; is \; word \\ 1, & i = j \\ 0, & otherwise \end{cases} \quad (1)$$

Positive point-wise mutual information (PPMI) is used to weight edges between word nodes. PPMI is a measure of the degree of association between two events and can be viewed as word co-occurrence in natural language processing. Term frequency-inverse document frequency (TF-IDF) values are used for the weights of edges between word nodes and document nodes; TF-IDF values are larger for words that occur more frequently in a document but less frequently in other documents, i.e., words that characterize that document.

In [5], weights between nodes on heterogeneous graphs of words and documents are defined as in (2).

$$A_{i,j} = \begin{cases} COS\_SIM(i,j), & i,j \; are \; documents \; and \; i \neq j \\ PPMI(i,j), & i,j \; are \; words \; and \; i \neq j \\ TF-IDF(i,j), & i \; is \; document, j \; is \; word \\ 1, & i = j \\ 0, & otherwise \end{cases} \quad (2)$$

As shown in (1), RoBERTaGCN considered relations between words and relations between words and documents in the form of weights of edges between nodes, but did not consider relations between documents. Therefore, we have improved RoBERTaGCN to consider the relationship between documents by expressing the relationship between documents in the form of weights of edges between document nodes. $COS\_SIM(i,j)$ in (2) is the weight of edges between document nodes and represents the cosine similarity. Each document is tokenized and input into BERT to obtain a embedding for each document. The Cosine similarity is calculated between the obtained vectors, and if the Cosine similarity exceeds a predefined threshold value, the Cosine similarity is added as weights of edges between corresponding document nodes.

III. METHOD

This section describes the details of the proposed method. Figure 1 shows a schematic diagram of the proposed method. First, a heterogeneous graph of words and documents is constructed from documents. Next, the graph information (weight matrix and initial node feature matrix) is input into the GCN, and the document vector is input into the feed-forward neural network. Finally, a linear interpolation of the two predictions is computed and the result is used as the final prediction.

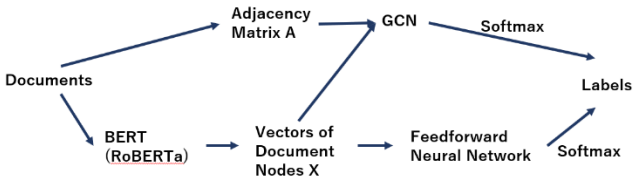

Figure 1. Schematic Diagram of the Proposed Method.

A. Build Heterogeneous Graph

First, a heterogeneous graph containing word and document nodes is constructed. The weights of edges between nodes i and j are defined as in (3). df indicates the number of documents in which the word appears. The difference between (2) and (3) is that the words that appear in only one document are removed. This reduces both the weights of edges between word nodes and the weights of edges between word nodes and document nodes, thus saving memory. PPMI is used for the weights of edges between word nodes, and TF-IDF values are used for the weights of edges between word and document nodes.

$$A_{i,j} = \begin{cases} COS\_SIM(i,j), & i,j \text{ are documents and } i \neq j \\ PPMI(i,j), & i,j \text{ are words and } i \neq j \\ & df(i) > 1, df(j) > 1 \\ TF-IDF(i,j), & i \text{ is document}, j \text{ is word} \\ & df(j) > 1 \\ 1, & i = j \\ 0, & otherwise \end{cases} \quad (3)$$

The process from Section B to E below is based on RoBERTaGCN [4].

B. Creating the Initial Node Feature Matrix

Next, we create the initial node feature matrix to be input into the GCNs. We use BERT to obtain the document embeddings and treat them as the input representations of the document nodes. The embedded representation X_{doc} of a document node is represented by $X_{doc} \in \mathbb{R}^{n_{doc} \times d}$, where n_{doc} is the number of documents and d is the number of embedding dimensions. Overall, the initial node feature matrix is given by (4).

$$X = \begin{pmatrix} X_{doc} \\ 0 \end{pmatrix}_{(n_{doc}+n_{word}) \times d} \quad (4)$$

C. Input into GCN and Learning by GCN

The weights of the edges between nodes and the initial node feature matrix are input into GCNs for training. The output feature matrix $L^{(i)}$ of layer i is computed by (5).

$$L^{(i)} = \rho\left(\tilde{A} L^{(i-1)} W^{(i)}\right) \quad (5)$$

ρ is the activation function and $\tilde{A}$ is the normalized adjacency matrix. $W^i \in \mathbb{R}^{d_{i-1} \times d_i}$ is the weight matrix at layer i. $L^{(0)}$ is X, the input feature matrix of the model. The dimension of the final layer of W is (number of embedded dimensions) × (number of output classes). The output of the GCNs is treated as the final representation of the document node, and its output is input into the softmax function for classification. The prediction by the output of the GCNs is given by (6). g represents the GCNs model. The cross-entropy loss in labeled document nodes is used to cooperatively optimize the parameters of BERT and GCNs.

$$Z_{GCN} = softmax\left(g(X, A)\right) \quad (6)$$

D. Input into Feedforward Neural Network and Learning by Feedforward Neural Network

Optimizing GCNs with an auxiliary classifier that directly handles BERT-embedded representations leads to faster convergence and improved performance. Specifically, a document embedded representation X is input into a Feedforward Neural Network. The output is then fed directly into a softmax function with a weight matrix W to create an auxiliary classifier with BERT. The prediction by the auxiliary classifier is given by (7).

$$Z_{BERT} = softmax(WX) \quad (7)$$

E. Interpolation of Predictions with BERT and GCN

A linear interpolation is computed with Z_{GCN}, the prediction from RoBERTaGCN, and Z_{BERT}, the prediction from BERT, and the result of the linear interpolation is adopted as the final prediction. The result of the linear interpolation is given by (8).

$$Z = \lambda Z_{GCN} + (1 - \lambda) Z_{BERT} \quad (8)$$

λ controls the trade-off between the two predictions. $\lambda = 1$ means using the full RoBERTaGCN model, while $\lambda = 0$ means using only the BERT module. When $\lambda \in (0, 1)$, the predictions from both models can be balanced, making the RoBERTaGCN model more optimal. Experiments by Lin et al. using the graph structure in (1) show that $\lambda = 0.7$ is the optimal value of λ.

IV. EXPERIMENTS

In this study, two experiments were conducted.

Experiment 1: Experiment to confirm the effectiveness of the graphs of the proposed method.

In Experiment 1, the classification performance of the proposed method using compact graphs was compared with other methods. The parameter λ, which controls the balance between predictions from BERT and predictions from GCNs, was fixed at 0.7. Preliminary experiments were conducted on the validation data, and the optimal values of the threshold of the cosine similarity for each dataset are shown in Table II. We used the values in Table II as our threshold values. The trained model used was roberta-base. Accuracy was used to evaluate the experiment. Positive is the label of the correct answer, negative is the label of the incorrect answer, and negative is all the remaining labels except the correct label.

TABLE II. OPTIMAL VALUE FOR COSINE SIMILARITY THRESHOLD.

Dataset	Optimal Threshold Value
20NG	0.99
R8	0.975
R52	0.96
Ohsumed	0.965
MR	0.97

Experiment 2: Experiment to check classification accuracy when changing to a larger trained model.

In Experiment 2, we take advantage of the memory savings and check the accuracy of the proposed method by applying a larger trained model. Specifically, the learned model is changed from roberta-base to roberta-large. λ and cosine similarity values are set to the same values as in Experiment 1.

A. Data Set

We evaluated the performance of the proposed method by conducting experiments using the five data sets shown in Table III. We used the same data used in RoBERTaGCN. Each dataset was already divided into training and test data, which we used as is. The ratio of training data to test data is about 6:4 for 20NG, about 7:3 for R8 and R52, about 4.5:5.5 for Ohsumed, and about 6.5:3.5 for MR.

TABLE III. INFORMATION OF EACH DATA SET.

Dataset	Number of Documents	Average of Words
20NG	18846	206.4
R8	7674	65.7
R52	9100	69.8
Ohsumed	7400	129.1
MR	10662	20.3

・20-Newsgroups (20NG)

20NG is a dataset in which each document is categorized into 20 news categories, and the total number of documents is 18846. In our experiments, we used 11314 documents as training data and 7532 documents as test data.

・R8, R52

Both R8 and R52 are subsets of the dataset provided by Reuters (total number is 21578). R8 has 8 categories and R52 has 52 categories. The total number of documents in R8 is 7674, and we used 5485 documents as training data and 2189 documents as test data. The total number of documents in R52 is 9100, and we used 6532 documents as training data and 2568 documents as test data.

・Ohsumed

This is a dataset of medical literature provided by the U.S. National Library of Medicine, and total number of documents is 13929. Every document has one or more than two related disease categories from among the 23 disease categories. In the experiment, we used documents that had only one relevant disease category, and the number of documents is 7400. We used 3357 documents as training data and 4043 documents as test data.

・Movie Review (MR)

This is a dataset of movie reviews and is used for sentiment classification (negative-positive classification). The total number of documents was 10662. We used 7108 documents as training data and 3554 documents as test data.

B. Experimental Environment

The experiments were conducted using Google Colaboratory Pro+, an execution environment for Python and other programming languages provided by Google. The details of the specifications of Google Colaboratory Pro+ are shown in Table IV.

TABLE IV. DETAILS OF THE SPECIFICATIONS OF GOOGLE COLABORATORY PRO+.

GPU	Tesla V100（SXM2） / A100（SXM2）
Memory	12.69GB（standard） / 51.01GB（CPU / GPU (high memory)） / 35.25GB（TPU (high memory)）
Disk	225.89GB（CPU / TPU） / 166.83GB（GPU）

C. Result of Experiment

TABLE V. CLASSIFICATION PERFORMANCE OF THE PROPOSED METHOD.

	20NG	R8	R52	Ohsumed	MR
Text GCN	86.34	97.07	93.56	68.36	76.74
Simplified GCN	88.50	-	-	68.50	-
LEAM	81.91	93.31	91.84	58.58	76.95
SWEM	85.16	95.32	92.94	63.12	76.65
TF-IDF +LR	83.19	93.74	86.95	54.66	74.59
LSTM	65.71	93.68	85.54	41.13	75.06
fastText	79.38	96.13	92.81	57.70	75.14
BERT	85.30	97.80	96.40	70.50	85.70
RoBERTa	83.80	97.80	96.20	70.70	89.40
RoBERTa GCN	89.15	98.58	94.08	72.94	88.66
[5]	89.82	**98.81**	94.16	74.13	89.00
Proposed method (base)	**90.02**	98.58	**96.88**	73.53	89.65
Proposed method (large)	89.95	98.58	96.81	**76.08**	**91.50**

Table V compares the classification performance of the proposed method with the conventional methods. [5] shows the classification performance when using the graph structure in (2). proposed method (base) is the result of Experiment 1, and proposed method (large) is the result of Experiment 2.

Comparing the results of the Proposed method (base) with the other methods, the accuracy of 20NG, R52, and MR improved. The accuracy of the other datasets also maintains a high level. Even with a compact graph in which words that appear only in one document are removed, the classification performance remains high. Therefore, it can be said that the proposed method succeeds in saving memory.

Comparing the results of the Proposed method (large) with the other methods, the accuracy is significantly improved for Ohsumed and MR. The classification performance of Ohsumed was 76.08%, 1.95% higher than that of [5], and that of MR was 91.50%, 1.85% higher than that of the Proposed method (base).

V. DISCUSSION

Table VI shows the number of word types that appear in each dataset and the number of words that are removed in the graph structure of (3). Table VII shows the number of PPMI edges added in the original graph structure and the number of PPMI edges removed in the graph structure of (3). Table VIII shows the number of TF-IDF edges added in the original graph structure and the number of TF-IDF edges removed in the graph structure of (3). Since TF-IDF edges are added between word and document nodes, the number of edges removed is the same as the number of words removed. From these three tables, it can be seen that the graph of the proposed method reduces the number of edges by 1 to 20%. Experimental results show that the classification performance of the proposed method maintains performance of the method using the original graph structure. Therefore, it can be said that the proposed method succeeds in saving memory because it reduces the number of edges on the graph while maintaining the accuracy.

We believe that the reason why the accuracy was maintained even with a compact graph is because the words to be removed were limited to words that appear only in a single document. Words that appear in only one document do not propagate document topic information through the word node, and thus text classification performance is maintained even if those words are removed.

This study also confirmed the document classification performance when the trained model was changed to a larger one, taking advantage of the memory savings. When the learned model was changed from roberta-base to roberta-large, the accuracy improved significantly. It is thought that the change to roberta-large improved the accuracy because it was able to acquire embedded representations that better reflect the characteristics of the documents.

TABLE VI. NUMBER OF WORDS REMOVED.

Dataset	Number of Words	Number of Words Removed
20NG	42757	755
R8	7688	225
R52	8892	245
Ohsumed	14157	851
MR	18764	8687

TABLE VII. NUMBER OF PPMI EDGES REMOVED.

Dataset	Number of PPMI Edges	Number of Edges Removed
20NG	22413246	127662
R8	2841760	32954
R52	3574162	36138
Ohsumed	6867490	129938
MR	1504598	314950

TABLE VIII. NUMBER OF TF-IDF EDGES REMOVED.

Dataset	Number of TF-IDF Edges	Number of Edges Removed
20NG	2276720	755
R8	323670	225
R52	407084	245
Ohsumed	588958	851
MR	196826	8687

VI. CONCLUSION AND FUTURE WORK

To solve the memory-consuming problem of conventional text classification methods based on graph structures, this paper proposes the text classification method using compact graphs in which words that appear only in one document are removed. Experiments confirmed that the proposed method can maintain the accuracy of the conventional method while saving a lot of memory. Experiments also showed that the acuracy of text classification improves when the learned model is changed to a larger one, taking advantage of the saved memory.

Future work includes comparing the accuracy with the proposed method when other features are used instead of cosine similarity and optimizing the parameter λ for each data.

REFERENCES

[1] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks.” In ICLR, 2017.

[2] L. Yao, C. Mao and Y. Luo, “Graph convolutional networks for text classification.” In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 7370-7377, 2019.

[3] Z. Lu, P. Du and J. Y. Nie, “Vgcn-bert: augmenting bert with graph embedding for text classification.” In European Conference on Information Retrieval, pp. 369-382, 2020.

[4] Y. Lin, Y. Meng, X. Sun, Q. Han, K. Kuang, J. Li and F. Wu, “BertGCN: Transductive Text Classification by Combining GCN and BERT” In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 1456–1462, 2021.

[5] H. Nakajima and M. Sasaki, “Text Classification Using a Graph Based on Relationships Between Documents.” In Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation, pp. 119–125, Manila, Philippines. De La Salle University. 2022.

[6] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner and G. Monfardini, "The graph neural network model," IEEE Transactions on Neural Networks, vol.20, no.1, pp.61-80, 2008.

[7] T. N Kipf and M. Welling, “Variational graph auto-encoders.” arXiv preprint arXiv:1611.07308. 2016b.

[8] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio and Y. Bengio, “Graph attention networks.” arXiv preprint arXiv:1710.10903. 2017.

[9] J. Bastings, I. Titov, W. Aziz, D. Marcheggiani and K. S. An, “Graph convolutional encoders for syntax-aware neural machine translation.” In Proceedings of the 2017 Conference on Empirical Methods in

Natural Language Processing, pp. 1957-1967, Copenhagen, Denmark. Association for Computational Linguistics. 2017.

[10] L. Huang, D. Ma, S. Li, X. Zhang and H. Wang, "Text level graph neural network for text classification." In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3444–3450, 2019.

[11] R. K. Bakshi, N. Kaur, R. Kaur and G. Kaur, "Opinion mining and sentiment analysis." In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 452-455, IEEE. 2016.

[12] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu and J. Gao, "Deep Learning Based Text Classification: A Comprehensive Revie." ACM Computing Surveys, vol.54, Issue 3, no.62, pp.1-40, 2021.

[13] X. Liu, X. You, X. Zhang, J. Wu and P. Lv, "Tensor graph convolutional networks for text classification" arXiv:2001.05313v1. pp.8409-8416, 2020.

[14] G. Wang, C. Li, W. Wang, Y. Zhang, D. Shen, X. Zhang, R. Henao and L. Carin, "Joint embedding of words and labels for text classification." In ACL, pp.2321–2331, 2018.

[15] D. Shen, G. Wang, W. Wang, M. R. Min, Q. Su, Y. Zhang, C. Li, R. Henao and L. Carin, "Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms." In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (vol. 1: Long Papers), pp. 440–450, 2018.

[16] P. Liu, X. Qiu and X. Huang, "Recurrent neural network for text classification with multi-task learning." In IJCAI, pp.2873–2879, AAAI Press. 2016.

[17] A. Joulin, E. Grave, P. Bojanowski and T. Mikolov, "Bag of tricks for efficient text classification." In EACL, pp.427–431, Association for Computational Linguistics. 2017.

[18] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." Proceedings of NAACL-HLT 2019, pp. 4171–4186, 2019.

[19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach." arxiv:1907.11692v1. 2020.

Analysis of Antithetical Elements in English Literary Passages Using Stochastic Models

Chenjie Zeng
School of Humanities and Social Science
The Chinese University of Hong Kong, Shenzhen
Shenzhen, China
E-mail: chenjiezeng@link.cuhk.edu.cn

Clement H. C. Leung
School of Science and Engineering &
Guangdong Provincial Key Laboratory of Future Networks of Intelligence
The Chinese University of Hong Kong, Shenzhen
Shenzhen, China
E-mail: clementleung@cuhk.edu.cn

***Abstract*—Thesis and antithesis are rhetorical figures often employed by well-known authors in English literary passages. Antithesis refers to the juxtaposition of contrasting words or ideas, which often, although not always, appears in the form of parallel structures. Such rhetorical figure is used to contrast opposing ideas, creating a sense of tension and urgency, as well as heightening the emotional impact of the speech. Contrasting the antithesis is the thesis, which refers to a statement, assertion, or tenet. A thesis also is a proposition laid down or stated, especially as a theme to be discussed and proved, or to be maintained against attack. It is observed that the points of occurrence of thesis and antithesis are random, yet they occur with remarkable regularity alternating each other. In this study, we represent the lengths of thesis and antithesis by using an alternating renewal stochastic process. We find that the underlying parameters of the alternating renewal process are able to usefully characterize the writing style of individual authors and help us to quantify and understand their linguistic technique and intention. Using the present method in conjunction with other techniques, such as sentiment analysis, and word embedding models, it is possible to gain a deeper understanding of the literature and its underlying themes and structures.**

Keywords-stochastic alternating renewal process; antithesis; literary passages; thesis; renewal function.

I. INTRODUCTION

Mathematical models provide a useful analytic mechanism for the quantitative study of the characteristics of literary work. For example, Markov models have been used extensively in studying English literature and language. One notable application is in the area of computational stylometry, which is a field that uses statistical methods to identify the authorship of anonymous or disputed literary texts. The use of Markov models in literature and language analysis has been a fruitful area of research, providing new insights into the structure and patterns of language use in literature, as well as enabling new methods for identifying core features of individual writers and texts.

For example, in Shakespeare's play *Henry V*, Henry's speech to his men before the Battle of Agincourt (Act 4, Scene 3) is a masterful example of rhetorical and dramatic effectiveness. The speech has its historical and literary value and importance. On the one hand, the Battle of Agincourt is a triumph for the English in the Hundred Years' War. The battle occurred on Saint Crispin's Day, October 25th, 1415. Despite being outnumbered by the French, the English emerged victorious, and this unexpected win had a significant impact on English morale and reputation. It also dealt a severe blow to France and began a new phase of English superiority that would last for 14 years. On the other hand, Henry's speech is filled with rhetorical figures, such as antithesis, rhetorical questioning, allusion, parallelism, and so on. These stylistic devices create a stirring call to arms that inspires the English soldiers to fight with courage and conviction.

Therefore, the stylistic and rhetorical figures serve the theme of Henry's speech to his men before the Battle of Agincourt in Shakespeare's play Henry V. The theme is related to the power of leadership and inspiration in the face of adversity, and the responsibility of leaders to motivate and uplift their followers. The speech also shows personal responsibility, and the ability to rise to significant challenges. In this speech, Henry seeks to inspire his soldiers to fight and win against overwhelming odds, despite their fears and doubts. He does this by appealing to their patriotism, sense of honor and duty, and faith in God. He also shows them he is with them, leading the charge into battle and sharing their risks and struggles.

Through this speech, Shakespeare portrays the idea that leadership is not just about commanding others but inspiring and motivating them to be their best selves. It is about connecting with people on a deeper emotional level and giving them a sense of purpose and direction, even in the face of great danger and uncertainty. Henry's speech also highlights the importance of personal responsibility and accountability for leaders and those they command. By taking personal responsibility for his soldiers' well-being and showing that he is willing to share in their burdens and take on the risks of battle, Henry earns their respect and loyalty, and inspires them to fight with all their might.

In Section II, previous and relevant studies in the vast field of English literature and computer science will be introduced, so to portray a landscape of the related work. In Section III, we define the notation and function of "thesis" and "antithesis" as rhetorical devices in literary passages, including Shakespeare's *Sonnet* and *Henry V*, where King

 ISBN: 978-1-68558-111-4

Henry delivered a speech to his men before the Battle of Agincourt. More importantly, in this section, we explain how literary rhetoric is associated with mathematic and stochastic models, which bridges the gap between the disciplines of English Literature and Computer Science. Later in Section IV, as the nature of this interdisciplinary paper, we model the rhetoric thesis and antithesis by applying alternating renewal processes. Finally, in Section V, we test the model by conducting experiments. We analyze two literary passages in vastly different genres, including Leo Tolstoy's novel "War and Peace" and Barack Obama's 2008 presidential speech "Yes We Can".

II. RELATED WORKS

Antithesis, known as confrontation or contradiction, refers to the juxtaposition of opposing thoughts, concepts, meanings, and images that are logically comparable [14]. It is a literary device characterized by the presentation of contrasting words or meanings. Many literary scholars have studied the function of antithesis. For example, in their work, Raximovna and Mirusmanovna [14] elucidate the practical characteristics of antithesis within literary texts, employing examples from both English and Uzbek fiction for analysis. It is commonly employed in literary texts, particularly in works of fiction, and holds significant value for in-depth examination [14].

Moreover, mathematical models have been used extensively in studying English literature and language. To be more specific, researchers have used Markov models to analyze the writings of Shakespeare, examining the patterns of word choice and syntax in his plays and sonnets to identify his distinctive writing style. By using Markov models to identify the probability of certain words or phrases occurring in a particular order, researchers have been able to detect subtle features of Shakespeare's writings [1]–[3]. Markov models have also been applied to study narrative structure in English literature. Researchers have examined the patterns of plot elements and character development in novels and other works, using Markov models to identify the most common transitions between different states in the narrative. This has allowed them to identify the key events and turning points in a story, as well as to analyze how different characters influence the course of the plot. For instance, Markov models are commonly used for statistical learning applications to capture the sequential patterns of data over time. While Hidden Markov models are widely researched, Devesh [7] explores an approach inspired by symbolic dynamics. This approach involves two main steps for successfully representing time-series data in a discrete space: first, continuous attributes are discretized, and second, the size of the temporal memory of this discretized sequence is estimated. Both steps are crucial for an accurate and concise representation of time-series data. The first step, discretization, determines the information content of the resulting sequence. The second step, memory estimation, is essential for extracting predictive patterns in the discretized data. The effectiveness of using a discrete Markov process for signal representation is determined by these two steps.

Another area of research has focused on using Markov models in natural language processing, which is the field of computer science that involves using computers to process and analyze human language. Markov models have been used to analyze the structure and syntax of sentences, as well as to identify patterns of co-occurrence between different words or phrases in a text. For example, Eder et al. [6] discuss R in the context of Stylometry with R, used for analyzing writing styles in stylometry. Stylometry is a field that studies writing styles quantitatively, such as authorship verification, which can be useful in forensic contexts and historical research. The paper presents the potential applications of stylometry for computational text analysis, using several example case studies from English and French literature. The package is particularly effective in exploratory statistical analysis of texts, especially with regard to authorial writing style. The package has an appealing graphical user interface for novices without programming skills, such as those in the Digital Humanities. Experienced users can benefit from the package's standard pipelines for text processing and different similarity metrics.

III. THESIS AND ANTITHESIS IN LITERARY PASSAGES

One of the key rhetorical figures used in the speech is the antithesis. Antithesis refers to the juxtaposition of contrasting words or ideas, which often, although not always, in parallel structure. For example, Shakespeare's Sonnet says, 'Before, a joy proposed; behind, a dream' [16]. This figure is used to contrast opposing ideas and create a sense of tension in "Henry V". The contrast of antithesis is thesis. According to Oxford English Dictionary [11], a "thesis" refers to a statement, assertion, or tenet. Thesis also is a proposition laid down or stated, especially as a theme to be discussed and proved, or to be maintained against attack. Thesis, in *Logic*, is sometimes as distinct from hypothesis, while, in *Rhetoric*, it is contrasted with antithesis. The formal connection between these figures demonstrates that despite their discrete, distinct, or potentially conflicting natures, they have the capacity and necessity to contribute to the larger entity of creation or society, existing alongside each other in a meaningful manner [15]. For example, when Henry says (in stanza 1, Act 4 Scene 3, *Henry V*),

X 2 3 4
If we are marked to **die**, we are enough

5 6 7 8 9 10 11 12 -X
To do our country loss; and if to **live**,
X 2 3 -X
The **fewer** men, the **greater** share of honor

Henry highlights the high stakes of the battle and contrasts the potential for death with the potential for glory. The highlights and contrasts create a sense of urgency and heighten the emotional impact of the speech. Shakespeare explores the concept of patriotism as a thematic element juxtaposed with profoundly unsettling notions, striving for a delicate equilibrium [12]. We use "X" to represent the first occurrence of rhetoric "die/fewer," and use "-X" to represent the follow-up antithesis "live/greater". Given the condition that an "X" appears in a poetic line, it is expected to have an

"-X" sooner or later. In the above lines that contain antithesis in Henry's speech, when we come across "X," such as "die" in the third line, we expect to have an antithesis "-X" like "live" later. Likewise, when we have the word "fewer", which is an "X". We will have its antonym "greater", which can be symbolized as "-X".

In order to analyze this phenomenon, let us count "X" and "-X" as 1, and then we can determine the exact length between them. The numbering starts from "X" and ends just before "-X" occurs. For instance, the beginning of the first "X" is "die", which is number 1. The length in terms of the number of words between "X" (die) and its antithesis "-X" (live) is thus 12 words. The same length-counting goes on when we find the next "X" (fewer) and "-X" (greater). And the length between "fewer" and "greater" is 3 words. Given a simple calculation, we will have the average length (12+3)/2=7.5 words in the first stanza of Henry's speech to his men before the Battle of Agincourt.

Another example would be in Stanza 7 of Henry's speech to his men before the Battle of Agincourt in *Henry V*:

X
Old men forget: yet all shall be **forgot**,
2 3 4 5 6
But he'll **remember** with advantages
7 8 9 10 11 12
What feats he did that day.
13 14 15 16
Then shall our names,
17 18 19 20 21 22 23
Familiar in their mouths as household words,
24 25 26
Harry the king,
27 28 29
Bedford and Exeter,
30 31 32
Warwick and Talbot,
33 34 35
Salisbury and Gloucester,
36 37 38 39 40 41 -X
Be in their flowing cups freshly **remember'd**.

In stanza 7, the length between the first occurrence of the thesis X "forgot" in the first line and "-X" (remember'd) in the tenth line is 41 words. Provided that there is no other prominent antithesis in Henry's speech to his men before the Battle of Agincourt, we can have an average length of the two stanzas: (7.5+41)/2=24.25 words.

Furthermore, the same pattern of "X" and a follow-up "-X" can also be seen in rhetorical questions. The rhetorical question is usually defined as any question asked for a purpose other than to obtain the information the question asks. For example (at the beginning of stanza 1),

X
What's he that wishes so?
-X X -X
My cousin, Westmoreland? No, my fair cousin;

Literarily speaking, "What's he that wishes so?" and "Westmoreland?" are likely to be statements regarding one's opinion of the person addressed rather than a genuine request to know. Similarly, when someone responds to the question by saying, "My cousin" is more likely to be an expression of feeling than a realistic request for information. The technical term for rhetorical questions, in general, such as "No, my fair cousin", is erotema, which is a rhetorical question to affirm or deny a point strongly by asking it as a question.

If we call the whole sentence "What's he that wishes so?" as "X," then we are expecting an answer like "My cousin", which is an "-X". The same pattern goes in the second pair of "questioning" and "answering" followed up. The second "questioning" is "Westmoreland?", which is a "X", and the second "answering" is "No, my fair cousin", which is another "-X".

Furthermore, allusion is also an effective rhetorical device used in Henry's speech. An allusion is a literary technique that makes a short reference to a well-known thing that the audience will likely know. This technique enables writers and speakers to convey a lot of meaning and importance in a concise manner. When Henry references the feast of Crispian, he draws on historical and religious symbolism to motivate his troops. The allusion serves to create a sense of continuity between the present moment and the past, and to elevate the importance of the battle in the minds of the soldiers. Nevertheless, the effectiveness of allusions depends on readers' comprehension and recognition of them, and their correct interpretation of the associated significance. If an allusion is perplexing or misconstrued, it can reduce its effectiveness by perplexing the reader.

While comprehending the comedic aspect may not be entirely achievable, it is evident that a significant source of amusement for the audience lies in encountering familiar and unpretentious traditions presented in a manner typically associated with religious or moral contexts, thereby alleviating the need for restraint [13]. For example, Crispin and Crispinian, are twin brothers born to a noble Roman family, having been beheaded during Diocletian's reign. They were executed for their religious beliefs on October 25th, probably in the year of 285 or 286. While fleeing religious persecution, the brothers worked diligently at their cobbler business in secret during night-time hours. By utilizing their trade, the brothers were able to support themselves and aid those in need. Their success led to Rictus Varus, governor of Belgic Gaul, becoming hostile towards them, resulting in the brothers being tortured and heaved into the river with millstones tied to their necks. Despite surviving this ordeal, they were killed by the emperor in 286. Later, October 25th becomes the feast day of Saints Crispin and Crispinian. Given the above historical background, it is expected that the twin brother "Crispin and Crispinian" should always appear in pairs. If we make "Crispin" as "X" and "Crispinian" as "-X", we can see a similar and repetitive pattern of "X" and "-X" in the speech. For instance,

This story shall the good man teach his son;
-X X
And **Crispin Crispian** shall ne'er go by,

Similarly, the repetitive pattern is shown as follows as well.

X -X
Bedford and **Exeter**,
X -X
Warwick and **Talbot**,
X -X
Salisbury and **Gloucester**,

Bedford and Exeter, Warwick and Talbot, and Salisbury and Gloucester are all pairs of names of places in England or surnames of people. Specifically, in Shakespeare's "Henry V", these names refer to specific people. Namely, the Duke of Gloucester and Duke of Bedford are brothers to the King; the Duke of Exeter is uncle to the King. And the following characters should be officers in King Henry's army, including Earl of Salisbury, Earl of Westmoreland, and Earl of Warwick. They are important figures and contrasts in the events leading up to and during the Battle of Agincourt. The next section develops a general stochastic model for these situations.

IV. Modelling Thesis and Antithesis Using Alternating Renewal Processes

The alternating recurrence of Thesis and Antithesis can be modelled as an alternating renewal stochastic process [8]. Upon the occurrence of the thesis, a certain length of words (X_i) elapsed before the occurrence of the antithesis; for such words, we shall refer to them as being under the dominance of the thesis. Likewise, upon the occurrence of the antithesis, a certain length of words (Y_i) also elapsed before the occurrence of the thesis; for such words, we shall similarly refer to them as being under the dominance of the antithesis. By focusing on such rhetorical devices, we have the following model of a passage.

$$S = X_1 + Y_1 + X_2 + Y_2 + \cdots + \cdots X_N + Y_N = \sum_{k=1}^{N} X_k + \sum_{k=1}^{N} Y_k \ ,$$

where $\{X_i\}$ are independent, identically distributed positive random variables representing the length of words under the dominance of the thesis pending resolution by the antithesis, and $\{Y_i\}$ are also independent, identically distributed positive random variables representing the length of words under the dominance of the antithesis awaiting the next occurrence of a new thesis or until the end of the passage is reached. In general, $\{X_i\}$ and $\{Y_i\}$ have different distributional properties. We shall refer to X_i as the thesis length and refer to Y_i as the antithesis length. The lengths of thesis and antithesis will form an important basic characterization of a piece of literary work.

Certain authors prefer to deploy a relatively lengthy X_i so that greater tension can be built up until its eventual resolution; other authors may choose to adopt a medium or short length for X_i to bring about a sharper impact. Thus, the $\{X_i\}$ and $\{Y_i\}$ often provide a useful mechanism to characterize the stylistics of different writers. For an arbitrary word under the dominance of a thesis, it can therefore be classified as either positive, which corresponds to a continuation of the thesis dominance pending resolution or negative, which corresponds to a resolution that switches from a thesis dominance to an antithesis. Under the dominance of a thesis, the propensity of a word being positive is represented by the probability p, while the probability that it is negative is represented by the probability q, where $p + q = 1$. On the other hand, for an arbitrary word under the dominance of an antithesis, it can be classified as either negative, which corresponds to a continuation of the antithesis dominance pending the arrival of a fresh thesis or positive, which corresponds to the arrival of a new thesis. For a given word under the dominance of an antithesis, the probability of it being negative is given by the probability q, while the probability that it is being positive is given by the probability p.

Thus, given a thesis occurs, then the thesis length has a probability distribution

$$\Pr[\, X_i = n] = qp^{n-1} \qquad n = 1, 2, \ldots$$

This can be seen as follows. Given the occurrence of a thesis, then it must consist of at least one positive word (the first word of the thesis), for otherwise, it would not be a thesis. For long passages, the number of words is large and may be mathematically approximated by infinity. Following the same reasoning, we also have

$$\Pr[\, Y_i = n] = pq^{n-1} \qquad n = 1, 2, \ldots$$

The probability generating function $F(z)$ of X_i is therefore given by

$$F(z) = \sum_{k=0}^{\infty} \Pr[X_i = k]\, z^k = \frac{q}{p} \sum_{k=0}^{\infty} p^k z^k = \frac{q}{p(1-pz)}.$$

The probability generating function $G(z)$ of Y_i is given by

$$G(z) = \sum_{k=0}^{\infty} \Pr[Y_i = k]\, z^k = \frac{p}{q} \sum_{k=0}^{\infty} q^k z^k = \frac{p}{q(1-qz)}.$$

The mean thesis length is obtained by differentiation

$$\mathrm{E}(X_i) = F'(1) = \frac{1}{q},$$

and similarly, the mean antithesis length is

$$\mathrm{E}(Y_i) = G'(1) = \frac{1}{p}.$$

The variance of the thesis length is

$$\mathrm{Var}\,(X_i) = F''(1) + F'(1) - F'(1)^2 = \frac{p}{q^2},$$

and the variance of the antithesis length is

$$\text{Var}\,(Y_i) = G''(1) + G'(1) - G'(1)^2 = \frac{q}{p^2}.$$

Another common way of characterizing a passage is determining the total length *L* of words in relation to a given length *r* of the thesis. Now, *L*=*k* ($\geqslant r$), iff the *k*th word encountered coincides with the *r*th positive word, and this happens with probability

$$\Pr[L = k] = \binom{k-1}{r-1} p^r q^{k-r}$$

for *k* = *r*, *r*+1, *r*+2, …. This can be seen by noting that in order for *L*= *k* to be true, we must have (*r*—1) positive words among the first (*k*—1) words encountered, and this has the binomial distribution

$$\binom{k-1}{r-1} p^{r-1} q^{k-r}$$

On multiplying this by the probability that the *k*th word is positive, we obtain Pr[*L* = *k*]. The mean value of *L* is given by

$$\mathrm{E}[L] = r\left(1 + \frac{q}{p}\right).$$

On identifying 1/*q* as the mean thesis length and that, similarly, 1/*p* as the mean antithesis length, we obtain the following relation

$$\mathrm{E}[L] = r\left(1 + \frac{\text{mean length of antithesis}}{\text{mean length of thesis}}\right).$$

Or in normalizing by *r*, we have

$$\frac{\mathrm{E}[L]}{r} = 1 + \frac{\text{mean length of antithesis}}{\text{mean length of thesis}},$$

which can be utilized to characterize the style of a passage.

Another useful characterization is represented by the renewal function [8], which in the present context is the number of distinct thesis episodes *N* within r positive words. The *r* positive words have at most (*r*—1) gaps among them, since between any two successive positive words, there may or may not be intervening negative words. The probability of having no such intervening words is of course *p* and that of having one or more such words is *q*. Hence, the number of distinct episodes of the thesis equals *k* iff there are (*k*—1) interventions among the (*r*—1) gaps, i.e.,

$$\Pr[N = k] = \binom{r-1}{k-1} q^{k-1} p^{r-k}$$

where 1 $\leqslant k \leqslant r$. This has mean

$$\mathrm{E}[N] = 1 + (r-1)q.$$

If *r* is large and *q* small, then we can use the Poisson approximation

$$\Pr[N = k] \cong \frac{[(r-1)q]^{k-1} e^{-(r-1)q}}{(k-1)!}.$$

V. EXPERIMENTS AND ILLUSTRATIONS

Leo Tolstoy's novel "War and Peace" and Barack Obama's speech "Yes We Can" are two works that belong to vastly different genres. "War and Peace" is a classic novel that was first published in 1869 and is considered one of the greatest literary works in history. It is a work of historical fiction set against the backdrop of the Napoleonic Wars and explores themes of power, war, love, and society in 19th-century Russia. On the other hand, Barack Obama's speech "Yes We Can" is a political speech delivered by the former US President during his 2008 presidential campaign. It is a work of oratory and aims to persuade and motivate the American people to vote for Obama and support his vision for the country's future. These two works may be vastly different in their genres, but both demonstrate the power of language and storytelling to explore important themes and ideas that resonate with readers and audiences.

A. Russian Novel

The first chapter of Leo Tolstoy's "War and Peace" sets the stage for the novel's exploration of power, war, and social norms in 19th-century Russia. Chapter One begins with a discussion of the state of Russian society in the early 1800s, with a focus on the importance of social hierarchy and wealth. The reader is introduced to several aristocratic families and their relationships with each other, including the Bolkonskys, the Rostovs, and the Bezukhovs. When Anna Pavlovna talked about family members, Prince Vasili said,

> "What would you have me do?" he said at last. "You know I did all a father could for their education, and they have both turned out fools. Hippolyte is at least a **quiet** fool, but Anatole is an **active** one. That is the only difference between them." He said this smiling in a way more natural and **animated** than usual so that the wrinkles round his mouth very clearly revealed something unexpectedly coarse and **unpleasant**.

Therefore, we have

```
                                X    2
Hippolyte is at least a quiet fool,
 3    4        5 6    -X
but Anatole is an active one.

He said this smiling in a way
                                X       2      3
more natural and animated than usual,
 4   5   6       7        8      9   10
so that the wrinkles round his mouth
11    12       13         14
very clearly revealed something
15                  16   17      -X
unexpectedly coarse and unpleasant.
```

The first "X" is "quiet" and its follow-up "-X" is "active". The length between "quiet" and "active" is 6 words. Then we have the second pair of "X" (animated) and "-X" (unpleasant), and the length is 17 words. The average length of these two sentences is (6+17)/2=11.5 words. Compared with Shakespeare's "Henry V," Leo Tolstoy's "War and Peace" has a shorter length of antithesis. To summarize, we have:

$E(X_i) = 11.5$
$E(Y_i) = 12$
For $r = 10$, $E(L) = 21$

B. Presidential Speech

Barack Obama delivered his New Hampshire Primary Concession Speech entitled "Yes We Can" on January 8th, 2008, in Nashua, New Hampshire. The speech was delivered in the aftermath of his defeat in the New Hampshire Democratic primary by Hillary Clinton. Obama had previously won the Iowa caucus, but his defeat in New Hampshire was seen as a major setback for his presidential campaign. The primary was a crucial moment in the race for the Democratic nomination, and Obama's defeat was unexpected, given the momentum he had gained after Iowa. Obama's concession speech was a pivotal moment in his campaign, one that would ultimately lead him to win the presidency. In the speech, Obama addressed his supporters, acknowledging the difficult road ahead while rallying them to continue the fight of his campaign. The speech was widely praised for its emotional appeal and demonstration of Obama's resilience and determination. It is often seen as a turning point in his campaign, ultimately leading to his victory in the Democratic primaries and his ultimate election as the 44th President of the United States. In the speech, Obama goes,

"And whether we are **rich** or **poor**, **black** or **white**, Latino or Asian, whether we hail from Iowa or New Hampshire, Nevada or South Carolina, we are ready to take this country in a fundamentally new direction.
……
We can bring **doctors** and **patients**, **workers** and **businesses**, **Democrats** and **Republicans** together, and we can tell the drug and insurance industry that, while they get a seat at the table, they don't get to buy every chair, not this time, not now."

Therefore, we have the following antithesis:

X 2 -X X 2 -X
And whether we are **rich** or **poor**, **black** or **white**
X 2 -X
We can bring **doctors** and **patients**,
X 2 -X
workers and **businesses**,
X 2 -X
Democrats and **Republicans** together

$E(X_i) = 2$
$E(Y_i) = 1+4+1+1=1.75$
For $r = 10$, $E(L) = 17$

C. Shakespearean Classic

X
What's he that wishes so?
-X X -X
My cousin, Westmoreland? No, my fair cousin;

$E(X_i) = (5+1)/2 = 3$
$E(Y_i) = 2$
For $r = 6$, $E(L) = 8$

We see that $E(X_i)$, $E(Y_i) = 2$, and $E(L)$ for these three categories of literary styles are distinctly different.

The antithesis is a literary device in which opposites are put close to one another in a sentence or phrase for contrasting effects. In the opening of Shakespeare's "Henry V" (before Act 1 Scene I), there are several examples of antithesis when Chorus enters and delivers the following speech.

Stanza 1:

O for a Muse of fire, that would ascend
X 2 3
The brightest **heaven** of invention,
4 - X X
A **kingdom** for a stage, princes to **act**
2 3 4 - X
And monarchs to **behold** the swelling scene!
Then should the warlike Harry, like himself,
Assume the port of Mars; and at his heels,
Leash'd in like hounds, should famine, sword and fire
Crouch for employment.
But pardon, and gentles all,
X
The flat unraised spirits that have **dared**
2 3 4 5 6 7 8
On this unworthy scaffold to bring forth
9 10 11 12 13 14 15 16
So great an object: can this cockpit hold
17 18 19 20 21 22 23 24 25
The vasty fields of France? or may we cram
26 27 28 29 30 31 32
Within this wooden O the very casques
33 34 -X
That did **affright** the air at Agincourt?

There are at least four pairs of thesis and antithesis in the above literary passages, such as heaven and kingdom, act and behold, scaffold and cockpit, dared and affright. These examples of antithesis in stanza 1 serve to emphasize the

tension in Shakespeare's play, particularly in regard to the relationship between the grand ambitions of artistic creation and the limitations of the physical world.

First, the length between "heaven" and "kingdom" is 4. Regarding heaven and kingdom, the thesis in this contrast is the idea that the kingdom is like a stage, a place of grandeur and excitement worthy of the presence of princes and monarchs. The stage is an unworthy scaffold, a place that is limited in scope and unable to contain the vastness of the world outside like heaven. However, the antithesis of the kingdom is heaven, which is limitless because it usually refers to the expanse in which the celestial bodies, such as the sun, moon, and stars are observed, which was historically considered to resemble a large vault stretching over the earth, also known as the sky or the firmament.

Second, in terms of "act" and "behold", the length between them is 4 as well. The thesis is that the stage is a place for actors to take on roles, an opportunity to "act" and perform for audiences. The antithesis is that the monarchs and audience members who watch the play are passive spectators who "behold" the action, emphasizing their potential power and leashed engagement.

Third, the length between "dared" and "affright" is 34 words, which is largely distinct from the previous two pairs of thesis and antithesis. According to [11], "affright" means "to frighten, terrify" and "to be or become afraid", while "dare" refers to "having boldness or courage". The thesis here is that the actors have dared to bring forth an object of great significance, symbolized by the "great object" they are attempting to create through their performance. The antithesis is that the objects, such as the "casques" (or helmets) used in the battle scene which "affright" the air at Agincourt, are so terrifying that they cannot be contained within the "wooden O" symbolizing the limited confines.

Stanza 2:

O, pardon! since a crooked figure may
X 2 3 4
Attest in **little** place a million;
5 6 7 8 9 10 - X
And let us, ciphers to this **great** accompt,
On your imaginary forces work.
Suppose within the girdle of these walls
Are now confined two mighty monarchies,
Whose high upreared and abutting fronts
X 2 3
The perilous narrow **ocean** parts asunder:
4 5 6 7 8 9 10
Piece out our imperfections with your thoughts;
11 12 13 14 15 16 17
Into a thousand parts divide on man,
18 19 20 21
And make imaginary puissance;
22 23 24 25 26 27 28 29 30 31
Think when we talk of horses, that you see them
32 33 34 35 36 37 38 -X
Printing their proud hoofs i' the receiving **earth**;
For 'tis your thoughts that now must deck our kings,
X 2 -X
Carry them **here** and **there**; jumping o'er times,
Turning the accomplishment of many years
Into an hour-glass: for the which supply,
Admit me Chorus to this history;
Who prologue-like your humble patience pray,
Gently to hear, kindly to judge, our play.

The thesis and antithesis in the passage are ocean and earth, little and great, here and there. These antithesis and thesis emphasize the grand ambition of the playwright, who attempts to transport the audience's imagination from the small confines of the stage to vast distances and times. The use of antithesis serves as a reminder of the limitations that can constrain even the most ambitious attempts to capture the epic grandeur of history on the stage.

Between "little" and "great", the length is 10 words. Here, the thesis is that a small place can be of great importance, while the antithesis is that something great can also be imperfect and have flaws. The line "a crooked figure may / Attest in little place a million" highlights the significance of small things in the grand scheme of things.

Regarding "ocean" and "earth", the length is 38 words. The thesis is the idea that two mighty monarchies are separated by the narrow and perilous "ocean" that lies between them. It suggests a vastness and distance that capture the imagination, allowing the audience to visualize and join the drama. The antithesis of "earth" is that these monarchies are confined within the walls of the stage, reducing the scale of the conflict and the drama.

"Here" and "there" share a relatively shortest length of 2 words. The thesis aims to transport the audience's imagination across vast distances and times, from "here" to "there", referring to the distant monarchies and their battles. The line "Carry them here and there; jumping o'er times" emphasizes the idea of travelling through time and space. The antithesis is more subtle but can be seen in the line "Piece out our imperfections with your thoughts," which suggests that imagination is needed to fill in the gaps and make the performance more complete.

These antitheses in the above two stanzas highlight the grandeur and magnitude of the events depicted in the play and contrast the lofty aspirations of the characters with the limitations of the stage and the mortal world. The use of antithesis also adds rhetorical complexity and depth to the language, highlighting Shakespeare's skill as a writer and his ability to convey powerful emotions through his use of language.

VI. SUMMARY AND CONCLUSION

Antithetical elements are often used by authors in English literary writings and speeches. Antithesis refers to the juxtaposition of contrasting words or ideas, which often, although not always, in parallel structure. Such rhetorical

figure is used to contrast opposing ideas, creating a sense of tension and urgency, as well as heightening the emotional impact of the speech. Contrasting the antithesis is the thesis, which refers to a statement, assertion, or tenet. Thesis also is a proposition laid down or stated, especially as a theme to be discussed and proved, or to be maintained against attack. It is observed that the points of occurrence of thesis and antithesis are random, yet they occur with remarkable regularity alternating each other. In this study, we represent the lengths of the thesis and antithesis by using an alternating renewal stochastic process. We find that the underlying attributes, such as the length of the thesis, the length of the antithesis, and the renewal function, of the alternating renewal process are able to usefully characterize the writing style of individual authors and help us to quantify and understand their linguistic technique and intention. Experiments are carried out to illustrate the present approach. Using the present method in conjunction with other techniques, such as sentiment analysis, and word embedding models, it is possible to gain a deeper understanding of the literature and its underlying themes and structures.

REFERENCES

[1] C. H. C. Leung and C. Zeng, "The Use of Multi-Step Markov Chains in the Characterization of English Literary Works," in Proceedings of the 11th International Conference on Data Analytics, pp. 43-48, 2022.

[2] C. Zeng and C. Leung. "The Use of Stochastic Models in the Analysis of Vast English Literary Data Corpora," in 2020 6th International Conference on Big Data and Information Analytics (BigDIA), pp. 282-288, 2020.

[3] C. H. C. Leung and C. Zeng. "Pattern Discovery and Stylometric Analysis in English Literature and Literary Translation Through State Integration in Markovian Representations," in *Journal on Advances in Software*, vol. 16, no. 1&2, June 2023, in press.

[4] E. Smith. *The Cambridge Introduction to Shakespeare*. Cambridge University Press, 2007.

[5] W. Shakespeare. *Henry V*. Oxford University Press, 2008.

[6] M. Eder, M. Kestemont, and J. Rybicki, "Stylometry with R: A package for computational text analysis," in *R Journal*, 8(1), pp. 107-121, 2016.

[7] D. K. Jha, "Markov Modeling of Time-Series Data using Symbolic Analysis," in arXivLabs, 2021. Available from: https://doi.org/10.48550/arXiv.2103.11238

[8] R. Bass. *Stochastic Processes*. Cambridge University Press, 2012.

[9] L. Tolstoy. *War and Peace*. Duke Classics, 2013.

[10] B. Obama. Yes We Can, 2008 Presidential Campaign Speech.

[11] Oxford English Dictionary.

[12] P. Honan. *Shakespeare: A Life*. Oxford University Press, 1999.

[13] C. L. Barber. *Shakespeare's Festive Comedy*. Princeton University Press, 2011.

[14] R. N. Raximovna and M. M. Mirusmanovna, "Comparison of Pragmatic Value of Antithesis in Fiction Texts." *Turkish Journal of Computer and Mathematics Education* 12.6, pp. 1195-198, 2021.

[15] V. S. Bailey, 'Pope and Antithesis: "Law and War with Words,"' in *Studies in English Literature 1500-1900*, 27(3), pp. 437–454, 1987.

[16] M. Dobson, W. Wells, W. Sharpe, and E. Sullivan. *The Oxford Companion to Shakespeare.* Oxford University Press, 2016.