# Taxonomical Modeling and Classification in Space Hardware Failure Reporting.

Daniel Palacios[1,2], Terry Hill[2]
[1]Department of Physics, University of Houston, Houston, TX.
[2]NASA, Engineering Processes Methods Branch, Johnson Space Center, Houston, TX

## Motivation

- Space hardware from previous space missions were moved from paper to digital versions.
- Due to budgetary constraints, there has been little enterprise-level effort put into how to use this information
- +54,000 documents where space hardware failures are reported.
- Successful analysis like trending, correlation and root cause identification of problematic engineering processes is necessary to support human space flight in future lunar missions.

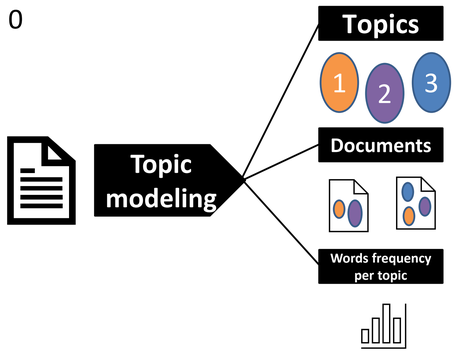
## Machine learning and Natural Language Processing

- Machine learning models are well optimized for numerical data.
- Text can be transformed into numbers by word-vector representations (Baevski et al., 2020) or by Term Frequency and Inverse Document Frequency matrix (Jalilifard et al., 2020).

## Topic Modeling: Latent Dirichlet Allocation

- Topic modeling refers to the unsupervised machine learning process of analyzing text to identify a common set of words (or topics) from a data set and using them to classify the data.
- Many commercial and open-source tools were tested, in data science, sometimes the challenge is to find the right tool. For this case LDA was preferred after several validation strategies (Jelodar et al., 2018).
- Amazon Web Services Comprehend, Corextopic (hierarchical topic modeling), BERTopic, and Customized TF-IDF Topic modeling.

0




**0.** Topic Modeling diagram. After performing topic modeling in a document topics can be analyzed to extract important keywords, phrases and knowledge from a text.

## Text Classification: Bidirectional Encoder Representations from Transformers

- In order to quantifying the quality of the groups generated by LDA and turning the unsupervised process to a supervised one, a text classification model had to be chosen. BERT was preferred for this use case (Devlin et al., 2018).
- Many tools were tested: Naïve Bayes Classifiers (Multinomial and Logistic Regressions), Convolutional Neural Networks, and Recurrent Neural Networks.

## References

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised370
learning of speech representations. CoRR abs/2006.11477371
Gawade, M., Mane, T., Ghone, D., and Khade, P. (2018). Text document classification by using wordnet379
ontology and neural network. International Journal of Computer Applications 182, 33–36. doi:10.5120/380
ijca2018918229381
Hoyle, A., Goel, P., and Resnik, P. (2020). Improving neural topic models using knowledge distillation.384
arXiv preprint arXiv:2010.02377385
Lewis, D. D. (1992). Feature selection and feature extraction for text categorization. In Speech and Natural394
Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992395
Lin, F. F., Muzumdar, K., Laptev, N. P., Curelea, M., Lee, S., and Sankar, S. (2019). Fast dimensional396
analysis for root cause investigation in large-scale service environment. CoRR abs/1911.01225397
Sharp, R., Pyarelal, A., Gyori, B., Alcock, K., Laparra, E., Valenzuela-Escˊarcega, M. A., et al.401
(2019). Eidos, INDRA, & delphi: From free text to executable causal models. In Proceedings of the402
2019 Conference of the North American Chapter of the Association for Computational Linguistics403
(Demonstrations) (Minneapolis, Minnesota: Association for Computational Linguistics), 42–47.404
doi:10.18653/v1/N19-4008405

## Acknowledgements

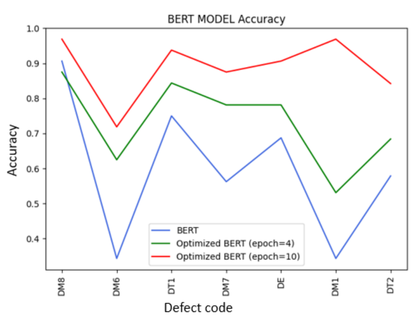
We thank NASA's Office of STEM Engagement The Minority University Research and Education Project (MUREP).

INSPIRE · ENGAGE · EDUCATE · EMPLOY
The Next Generation of Explorers

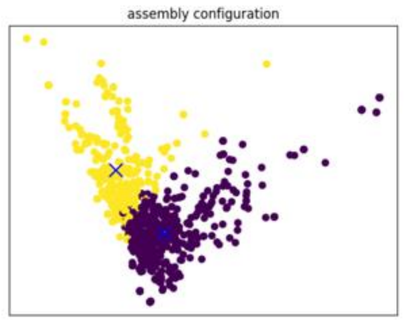
## Taxonomy

- After identifying sub-clustering from the groups generated by LDA, a taxonomy was created by running LDA again on the groups themselves.
- Defect Code (already existing classification), process labels (main grouping), sub-process labels (subgrouping)
- Several validation steps and improvement were taken with similarity matrices, feature extraction clustering (Lewis, 1992, a customized validation algorithm.
- Sub-process labels were mapped to old failure codes used in the past by NASA.

1A




1B

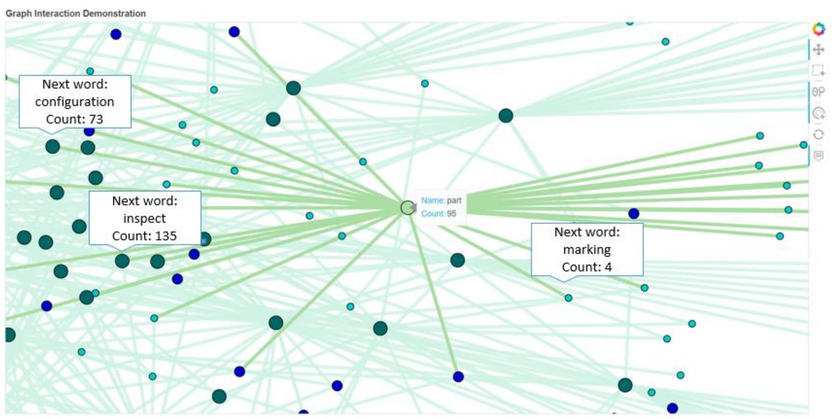



**1A.** BERT Accuracy plot across different sets of Space Hardware Discrepancy Reports. The blue line shows accuracy of BERT with default parameters. The green and red line correspond to BERT with optimal parameters at different epochs (computational cycles). **1B**. Feature Space Clustering Example. With this Unsupervised clustering technique, it was possible to identify number of clusters within topics in the text data set. Each color corresponds to a distinct subgroup.

## Cause-Effect Relationships and Markov Chains

- INDRA-Eidos: The Integrated Network and Dynamical Reasoning Assembler (INDRA) is an information assembler that extracts statements from text in molecular biological systems. Eidos is built on INDRA, and is main application is to extract statements from non-molecular biological systems (Sharp et al., 2019).
- INDRA-Eidos extracts events, influences (cause-effect), annotations, parameters, monomers, and rules.
- Other attempts include Factor Analysis and Apriori Algorithm.
- Markov chains were created to analyze relations of dominant and relevant keywords by proximity in the text.

2

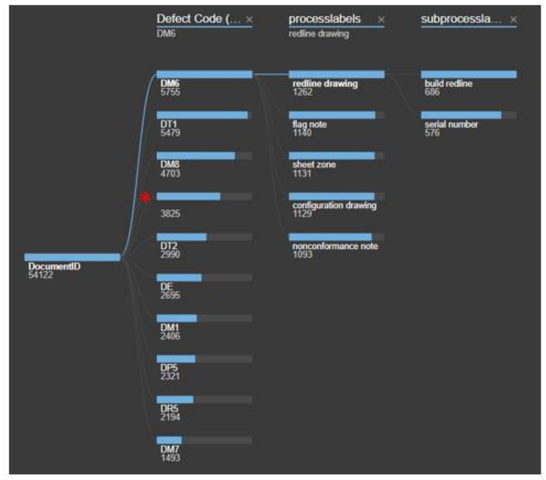



**2.** Markov Chain Tree Interactive Graphs examples from specific subgroups from LDA-BERT Taxonomy. Larger dark green nodes correspond to higher word count, medium blue nodes correspond to 10-25 word count, and smaller cyan nodes correspond to below 10 word count. Purple edges and orange nodes are highlighted from selection box tool.

## Results

- Improved classification capabilities, while identifying key engineering processes contributors.
- New classification will improve the report monitoring capabilities.

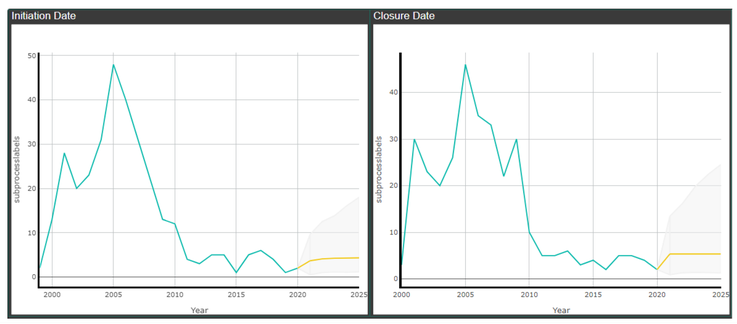
3




**3.** Taxonomical Tree for the classification of Space Hardware Discrepancy Reports. The first level is the Defect Code which is a label that already exist when filling the DRs information. Process labels correspond to the topics from LDA-BERT after running it with each respective Defect Code. Note it works also for DRs without Defect Code, which is very prominent in Open DRs. Subprocess labels is the last branch of the taxonomy, and it is obtained by running LDA-BERT again on each process label.

4

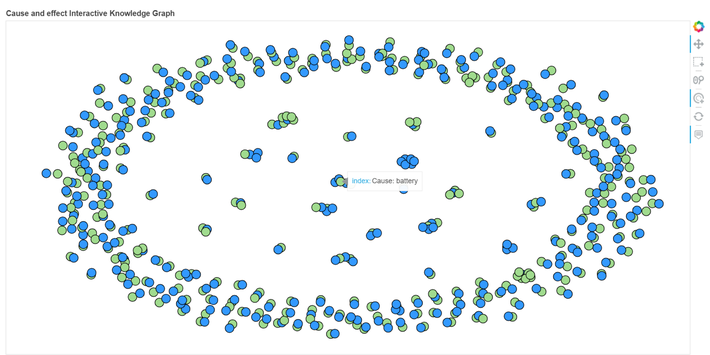
- Trending prediction with neural networks.




**4.** Forecast using Neural Network by MAQ Software on a subbranch group from LDA-BERT Taxonomy. Green section corresponds to the reports time series data, while the yellow section corresponds to the predicted trend for the future.

5

- Correlations analysis between processes.
- Common causes and effects from interactive knowledge graphs.




**5.** Interactive Knowledge Graph created with Bokeh-networkx from INDRA-Eidos results. Blue nodes corresponds to effects, green nodes correspond to causes. Icons in the right correspond to interactive tools like pan tool, selection box, zoom wheel, and reset. Some basic interactive tools were included like wheel zoom, highlighting tool, and hover display properties.