Summary of "*Text-Aware Predictive Monitoring of Business Processes*." - Marco Pegoraro, Merih Seran Uysal, David Benedikt Georgi, and Wil M.P. van der Aalst

Section 1: Introduction

Digital transformation has advanced over the years in most industries.

Process mining involves the discovery of process models from historical data, the level of compliance between data and the model, and the process models with additional information taken from complete process cases.

Predictive monitoring refers to predicting features of partial process instances based on recorded information of complete process instances.

Currently, predictive monitoring techniques operate with process mining and machine learning, control-flow of perspective event data, and additional associated data, primarily in text format.

The aim is to assess the extent to which textual information can influence predictive monitoring.

Section 2: Related Work

In previous studies, researchers focused on "white-box", human-interpretable models drawn from statistics. They proposed various things such as decision trees, simulation through stochastic Petri nets (a powerful mathematical and graphical notation for model languages to analyze and design a wide range of discrete systems), process outcome prediction with forests and logistic regression, and process discovery.

In the 2010s, the focus shifted from "white-box" to "black-box", specifically recurrent neural networks, because they're more accurate thanks to the use of **Long Short-Term Memory (LSTM)**. It's capable of predicting the cycle time of process instances, and with additional attributes, it provides data-aware prediction.

Section 3: Preliminaries

4 different text models exist to structure text:

- **Bag of Words (BoW),** highlights words in a document by considering how often they appear in that document and how rare they are across all documents, without caring about the order of the words.
- **Bag of N-Grams (BoNG),** is like Bag of Words (BoW), but instead of looking at single words, it looks at sequences of words called **n-grams**. This helps it understand the importance of word sequences in a document, taking into account both their frequency and their order.
- **Paragraph Vector (Doc2Vec)** trains a neural network to understand the meaning of whole documents. It does this by learning to predict words based on their context and creating a special vector for each document that captures its meaning and relationships between words. This vector becomes a short summary of the document's content.
- **Latent Dirichlet Allocation (LDA)** is a model that identifies topics in a collection of documents. It assumes that documents are created based on a process where topics are

chosen and then words are chosen from those topics. LDA represents documents as mixtures of topics and doesn't take into account the order of words.

Section 4: Prediction Model Architecture

Predictive monitoring estimates a process instance's target feature using historical data. This involves analyzing partial traces, which are event-related at certain execution points. These partial traces, equivalent to all trace prefixes in the log, form the training base for the predictive model.

Categorical attributes are turned into binary features. **Numerical attributes** are adjusted to fit within a certain range based on historical data.

Section 5: Evaluation

- **Pure LSTM Approach**: This method, only uses information like the activity, timestamp, and other non-textual attributes of each event. It's considered very accurate in predictive monitoring.
- **Process Model-based Prediction**: This method builds an annotated transition system from an event log. Each state is annotated with historical data to predict new outcomes. Ongoing traces are matched to states for predictions. An improved version of this method predicts the next activity and outcome. Only the first 8 events of a trace are considered when creating the state space.
- **Horizon Length Experiments**: The experiments explored different lengths of prediction horizons (1, 2, 4, 16) but generally resulted in worse outcomes.

Section 6: Conclusion

New ways to predict BPM can happen by merging historical and textual data that are related with one another. This method yields encouraging outcomes and points towards potential areas of research that merge language comprehension with process analysis.

Pegoraro, M., Uysal, M. S., Georgi, D. B., & Van Der Aalst, W. M. P. (2021). Text-Aware predictive monitoring of business processes. *Business Information Systems*, 221–232. https://doi.org/10.52825/bis.v1i.62