Discovery of Corrosion Patterns using Symbolic Time Series Representation and N-gram Model

Shakirah Mohd Taib[1],
Zahiah Akhma Mohd Zabidi[2],
Izzatdin Abdul Aziz[3]
and Farahida Hanim Mousor[4]
Department of Computer and
Information Sciences
Universiti Teknologi Petronas
32610 Seri Iskandar, Perak, Malaysia

Azuraliza Abu Bakar[5]
Center for Artificial Intelligence
Technology
Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor, Malaysia

Ainul Akmar Mokhtar[6]
Department of Mechanical Engineering
Universiti Teknologi Petronas
32610 Seri Iskandar, Perak, Malaysia

Abstract—There are many factors that can contribute to corrosion in the pipeline. Therefore, it is important for decision makers to analyze and identify the main factor of corrosion in order to take appropriate actions. The factor of corrosion can be analyzed using data mining based on historical datasets collected from monitoring sensors. The purpose of this study is to analyze the trends of corroding agents for pipeline corrosion based on symbolic representation of time series corrosion dataset using Symbolic Aggregation Approximation (SAX). The paper presents the analysis and evaluation of the patterns using N-gram model. Text mining using N-gram model is proposed to mine trend changes from corrosion time series dataset that are transformed as symbolic representation. N-gram was applied for the analysis in order to find significant symbolic patterns that are represented as text. Pattern analysis is performed and the results are discussed according to each environmental factor of pipeline corrosion.

Keywords—*Pipelines corrosion analysis; Symbolic Aggregation Approximation (SAX) representation; corrosion patterns; corrosion factor*

I. Introduction

Time series information is substantial in numerous application areas such as financial market modeling, weather forecasting, sensor systems and motion tracking. The purpose of analyzing the time sequence is to discover hidden knowledge as to figure the future patterns. In the oil and gas industry, pipeline is the main transportation to deliver the products. Pipeline corrosion usually happens in oil and gas industry as a lot of equipment is made from steel. Besides, the natural existence of corroding agents can initiate the chemical reaction that accelerates the corrosion process. Corrosion is the degradation of the material that attacks every component at every stage in the oil and gas industry and could occur because of the chemical reaction with the environmental. Corrosion is a threat to the oil field structures in pipelines, casing and tubing [1]. This problem can be the cause of pipeline's leakage that brings a vast impact to the operation process and system infrastructure cost [2]. Therefore, effective management of tools and equipment maintenance is important to avoid high maintenance cost. Hence, the analysis and monitoring of the pipeline's system is required to discover the important corrosion patterns and predict the consequences of the system's failure. The failure can be measured from the data collection of the static equipment such as sensors. Different types of sensor devices in oil and gas industry produce large-scaled of data that can reach several terabytes per day [3]. This historical time series data requires space and implementation of big data strategy. Therefore, efficient and effective time series representation and similarity searching become one of important issues in analyzing data from sensor devices. Numerous dimensionality reduction methods have been proposed for sufficient time series data representation including Symbolic Aggregate Approximation (SAX). SAX is a symbolic representation method for time series data [4], [5] that offers simple and efficient dimensionality reduction. In this study, SAX is used to transform numerical time series data into a symbolic data representation. The time series data that were recorded by sensor devices in a pipeline contains sensor readings based on several agents that might contribute to the corrosion. In order to discover hidden patterns in the transformed symbolic time series, N-gram model was used as a tool to analyze pattern trends in the corroding agents behavior. The content of this paper was organized as follows. The background of study section contains a discussion on the corrosion issue and data pre-processing method used in this study. The pre-processing consists of data cleaning by fixing the missing values and sorting out the important data that required for analysis. This section includes a further discussion of SAX method, N-gram models as well as Markov chains assumptions. The next section describes a methodology for the study that starts from the data collection until evaluation. In Result and discussion section, the result of the analysis patterns from both SAX and N-gram model will be discussed based on symbolic patterns that are discovered using real pipeline sensors time series data. In the final section of the paper, some conclusive remarks and directions for future work are put forward.

## II. Background of Study

A. Corrosion in Pipelines

Many important electrochemical, chemical, hydrodynamic and metallurgical parameters have been identified as main corrosion factors in pipelines [6], [7]. The effect of main

factors such as pH, temperature and ion migration give major influence to the corrosion rates. Solution pH and chloride concentration have a significant relationship that affects the corrosion process [8]. Chloride often occurs in the pipeline as a negative ion when it dissolved in the water phase present in the pipeline [9]. This chloride ion will bond with other elements that allow the corrosion to occur especially in subsea pipelines. Thus, the acidic rain that contains chloride also might be a factor that contributes to the increasing of corrosion rate [10]. Besides that, high temperature in a pipeline can cause internal corrosion. Temperature is a platform to accelerate the chemical and electrochemical processes occurring in the pipeline [11]. A low temperature makes the corrosion rate slowly increases due to the continuous dissolution of ion in a pipeline.

Acquiring data from corroding agents using sensor devices is a practical way for corrosion monitoring and early warning of structural failure as well as prediction of pipelines life [12]. However, large stream data from sensor needs cleansing and analysis to extract meaningful information from it. The feasibility of real-time pipeline monitoring and inspection system using acoustic sensors has been investigated by [2]. They found that this system can provide early detection such as corrosion and leaks of the pipelines but needs improvement on the poor quality of signal measurement and noise that may lead to inaccurate data transmission.

B. Time Series Representation

Time series analysis has been widely used in various fields of research. Esling and Agon [13] defined time series as a collection of values obtained from sequential measurements over time and stored as large dataset which causes the major issue for the high dimensionality of data. Time series tasks can be categorized as prediction, clustering, classification and segmentation [14]. Time series data can be univariate or multivariate when several series simultaneously span multiple dimensions within the same time range. Therefore, a well-defined and approximated representation for the original data is very important in the analysis of time series [5]. There are many approaches that has been highlighted for time series data representation such as Discrete Fourier Transform (DFT) [15], Discrete Wavelet Transform (DWT) [16], Singular Value Decomposition (SVD) [17], Piecewise Aggregate Approximation (PAA) [18] and the symbolic representation approach which is Symbolic Aggregation Approximation (SAX)[5]. Symbolic representation allows the application of data structures and algorithms from the text processing and bio-informatics research. Lin at.al [4] have proposed SAX as a time series representation method by transforming numeric values into alphabet sequence. The data is transformed by PAA representation before it is being symbolize into a discrete string. Therefore, the algorithm extends the PAA-based approach acquiring the calculation and low computational many-sided quality while giving acceptable flexibility for data mining. SAX is the first symbolic representation that offers a dimensionality reduction and a lower bound of the Euclidean distance [4], [19]. Implementation of SAX method in this study is briefly discussed in the next section.
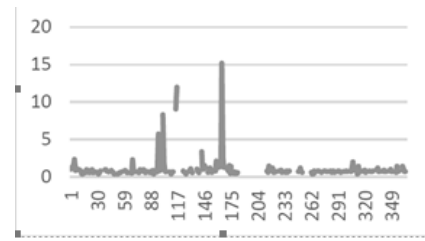
(a) Before Interpolation
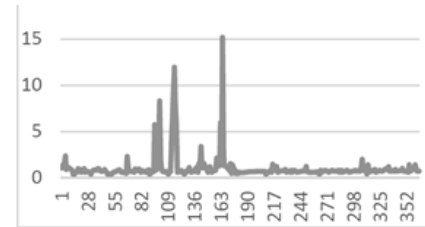
(b) After Interpolation

Fig. 1. Interpolation for missing values

C. Data Pre-proessing

In data mining, data pre-processing is an important stage to clean and transform raw data into appropriate format for further mining tasks. The basic pre-processing techniques used in this study are data cleaning and data normalization. Missing data is a common problem in time series dataset due to equipment failures during recording process. In order to transform the raw data into SAX, the missing data need to be filled in to maintain the consistency of the result. One of the missing data solutions for time series is interpolation technique. Linear interpolation can estimate the missing values based on the continuity in a single sequence [20]. Linear interpolation evaluates the estimation of a capacity between two known esteems. Linear interpolation requires assessing a new value by connecting two adjacent values with a straight linear as shown in (1).

$$y = y_1 + (x - x_1)\frac{y_2 - y_2}{x_2 - x_1} \qquad (1)$$

Fig 1. shows the example of time series data with missing values and a cleaned data after interpolation using [21] in Microsoft Excel.

D. N-gram Model

N-gram model is a simple text mining model that assigns probabilities to sentences and sequences of words. The concept of N-gram can be demonstrated from the chain rule of probability [22] as shown in (2).

$$P(W) = P(w_4 \mid w_1, w_2, w_3) \qquad (2)$$

Whereby, P(W) is a sequence of words and

$$P(w_4 \mid w_1, w_2, w_3)$$

is the conditional probability of word w_4 given the sequence w_1, w_2, w_3 .The sequence of N will be represented as $w_1...w_n$. Based on equation (2), it can be computed into (3).

$$P(w_1^n = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1^2)...P(w_n \mid w_1^{n-1}))$$
$$= \prod_{k=1}^{n} P(w_k \mid w_1^{k-1})$$

$$(3)$$

The chain rules outline the connection between the joint probability of a sequence and the conditional probability of a word given past words. From (2),the estimation of words can be calculated by multiplying a number of conditional probabilities[22]. However, using chain rule is not suitable for this study as the long sequence of symbolic data from SAX words cannot be computed the exact probability, $P(w_n \mid w_1^{n-1})$. Therefore, bigram model is used to predict the conditional probability of the next word. Instead of calculates the probability of the previous symbolic data $P(w_n \mid w_1^{n-1})$, it can be calculated using the conditional probability of the preceding word $P(w_n \mid w_{n-1})$. For example, instead of computing the probability into (4) it can be computed into (5).

$$P(c \mid cbccccacccc) \qquad (4)$$

$$P(c \mid c), P(c \mid b), P(c \mid a) \qquad (5)$$

Therefore, bigram model is used to predict the conditional probability of the next word [22] and the approximation as shown in (6).

$$P(w_n \mid w_1^{n-1}) \approx P(w_n \mid w_{n-1}) \qquad (6)$$

This assumption that depends on the previous frequency of words sequence is called Markov assumption. Markov models are the class of probabilistic model that can accept and anticipate the probability of some future unit without looking too far into the past [22]. Therefore, from bigram, it can derives to the trigram and to the N-gram which takes *N-1* words into the past. Hence, the common equation for N-gram estimation to the conditional probability of the next word in a sequence as shown in (7).

$$P(w_n \mid w_1^{n-1}) \approx P(w_n \mid w_{n-N+1}^{n-1}) \qquad (7)$$

III. METHODOLOGY

The methodology of this study consists of six stages as shown in Fig. 2.

This study was started by collecting time series data from the oil and gas company. The data is a record of sensor readings for five different corroding agents. Corroding agents in the dataset are described as Agent A, Agent B, Agent C, Agent D and Agent E. The agents are selected environmental factor that may contribute to corrosion rate of the pipeline in oil and gas industry. Each dataset was recorded from March 2010 to December 2016. The data was recorded for one-year analysis. However, as to standardize a count for one-year analysis, the data was selected between January 2011 to December 2016. Fig. 3 shows the actual data that was collected
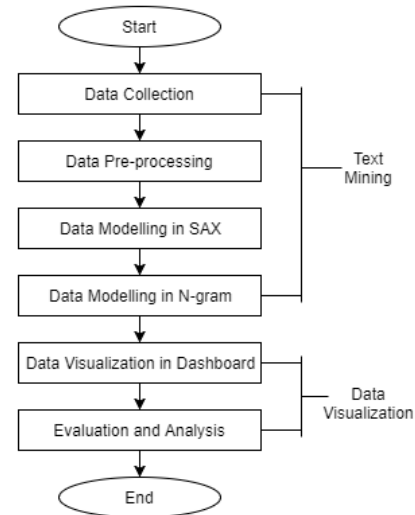

Fig. 2. Research Methodology

Fig. 3. Time series dataset of corroding agent

in a numeric form of time series. The data consists of daily sensor's reading for corroding agent A to E and the corrosion rate.

The second stage is data pre-processing where the raw data was analysed to detect errors and missing values. Linear interpolation method was applied to impute the missing values in the raw data. Data quality is important in order to achieve a better result in mining task. Other than imputation process, data was normalized and selected data was prepared for the next stage which is data representation using SAX. The result of the translated data is presented in Fig. 4 in the next section. The data are being standardized into one-year analysis for six years starting from 2011 to 2016. Next stage is the N-gram modeling where the transformed SAX data is clustered using N-gram model. N-gram model can assign probabilities to sentences and sequences of words. Therefore, N-gram was used to evaluate yearly trend of each corroding agent based on the co-occurrence of symbolic sequence patterns. In order to visualize the result of N-gram analysis, a dashboard was

TABLE I.   SAX REPRESENTATION OF CORRODING AGENT TIME SERIES
DATA.

| Corroding Agent	SAX Representation
A	b b b d b c b b b b b b b b b b b c c c b b b c b b b b b d d b b b b b b d c b c c b b b b b b c c c c b b b c b b b b b b b b b c c c b b c c c c c c c b c c c b b b
B	b b c c d c d b b a c c b b b b a b b d b c c d c c b b c b c c b b b c b b b b c c c c c c c b c b c c b c b c c c c c d b b c b c b b b b b b b b b c c d b c b c b b
C	b d c c c d d b b a a a b b c b c c c c c b a a b b b c c c d c b b b b b b b b c c d d b b b b c b b c c c b c b b b d c c c b b d b b b d b b
D	b b c b c c c b b b c c c c c c c b b c c b b a b b b b b b b b b b c d c c c b b c b b b b c b c b c c c c c a c c c c c c a c c c a c c c c c
E	c c c c c a c c c c c c c c c c c c c c c c a a b a a b b b d d b d c c

developed using R language. In the final stage, the overall
result for this study were evaluated and analysed.

### A. Result and Discussion

This study set out to investigate the hidden patterns from
symbolic time series of pipeline corroding agents using N-
gram model. The following discussion will focus on the sym-
bolic representation and analysis of corroding agents' behavior
based on N-gram results. A clean dataset was transformed
into symbolic representation after interpolation process was
completed. TABLE I shows symbolic representation result
after transforming numeric time series data using the SAX
algorithm. Each corroding agent has six strings that represents
the symbolic SAX patterns generated from 2011 to 2016.

From the SAX representation shown in TABLE I, it can be
analysed that each word represents a certain range value. This
is because SAX uses the concept of Piecewise Aggregate Ap-
proximation to get the range of breakpoints for each symbol.
For this study, the symbolic range is from *a* to *d* whereby *d*
is the highest range value and *a* is lower than 0. Fig. 4 shows
the SAX graph for Agent A and Agent C.

Both agents shown in Fig. 4 have different patterns in
2014. Agent A has minimal changes in the pattern throughout
2014 except for increase trend in the first quarter of the year.
Meanwhile Agent B has a decreasing trend in the last quarter of
2014. After transforming the data, the symbolic representation
data was used to classify the sequence of each symbolic pattern
using N-gram model. According to Bhakkad [23], to predict
for which word comes next for particular pattern of document
is based on the occurrence of different bigram frequency.
Therefore, bigram was used to cluster the pattern trend for each
corroding agent in order to discover co-occurrence behavior.
The pattern was analyzed by counting the frequency of each
bigram that yearly found in SAX graph. Fig 5 shows the
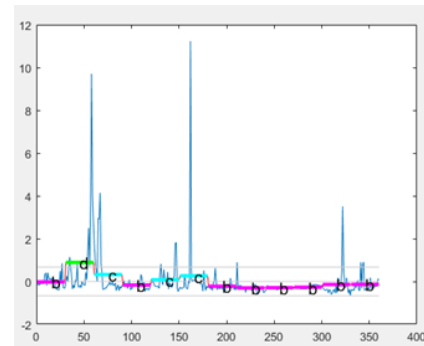
(a) Agent A
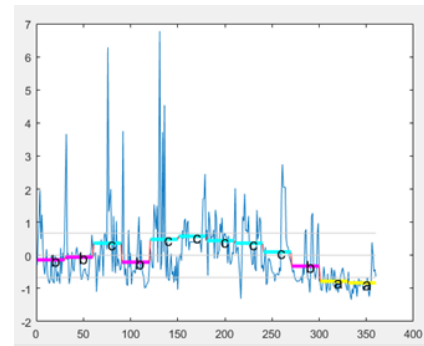
(b) Agent C

Fig. 4.   SAX patterns for Agent A and Agent C in 2014

bigram frequency for each corroding agent from Agent A to
Agent E within year of 2011 until 2016.

Term frequency-inverse document frequency (*tf-idf*) has
been conducted in order to evaluate the pattern for all the en-
vironmental factors of corrosion rate. Term frequency-inverse
document frequency (*tf-idf*) is a statistical measure that com-
monly used in information retrieval and text mining to evaluate
the important of a text document. Term frequency is used to
measure the frequently term that occurs in a document while
inverse document frequency measures how important a term
is. Term frequency took the more frequent words while inverse
document frequency also took along the rare words that occur.
In this study, term frequency-inverse document frequency is
used for pattern categorization. Term frequency in this study
will take account on the more occurrence pattern while inverse
document frequency will take the rare pattern less occur using
log to measure the *tf-idf*. The 0 value in *tf-idf* is consider as
the pattern is not very informative as the pattern often occurs
throughout the analysis. *Tf-idf* can be categorized into two
conditions that are high and low frequency. The high *tf-idf*
shows that the bigram frequency is more important compared
to othes. Fig. 6 illustrates the *tf-idf* patterns for each corroding
agent.

Based on the *tf-idf* result, it shows that Agent C and Agent
E have more significant bigrams compared to the other three
agents. Agent D has only one important bigram while Agent
A has two bigrams that have same *tf-idf* value.

B. Corrosion Rate Pattern

All corroding agents contribute to the corrosion in a
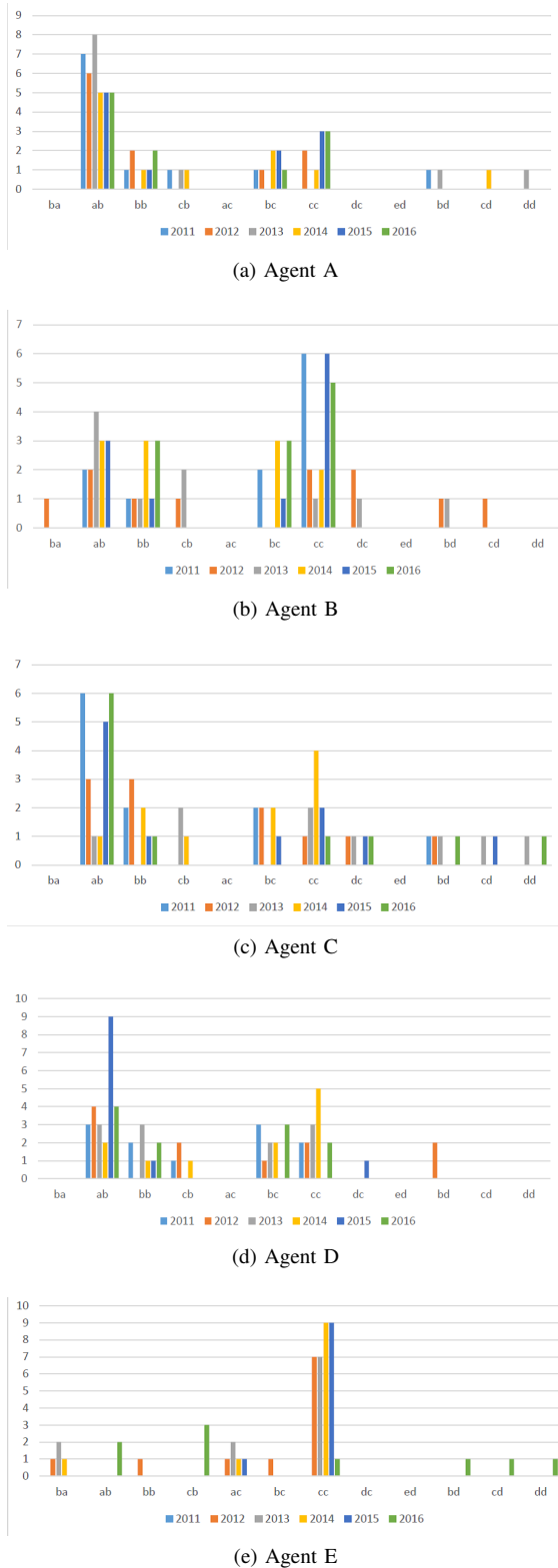pipeline throughout the year. Based on the SAX graphs, the

(a) Agent A

(b) Agent B

(c) Agent C

(d) Agent D

(e) Agent E

Fig. 5. Bigram Frequency of all corroding agents from 2011 - 2016

Fig. 6. Terms Frequency Inverse Document (tf-idf) Patterns

C. Relationship between the Bigram Pattern of Environmental Factor and Corrosion Rate

Based on the *tf-idf* result, it can be analyzed that Agent C and Agent E have more patterns than other environmental factors. From 2011 to 2016, there were changes in both agents that accelerated the corrosion rate in a pipeline. Agent D represents only *dc* pattern. From SAX graph, it shows the pattern has decreased from *d* to *c*. The similar pattern can be found in 2014, 2015 and 2016. This pattern shows that there was a decrease time series for Agent D in 3 years. Therefore, it can be concluded that Agent D has less contribution to the corrosion rate throughout the 3 years. Agent A pattern includes *cd* and *dd*. There is an increasing pattern from *c* to *d* and then the pattern remains unchanged in SAX graph. Therefore, the present of Agent A in a pipeline can cause an increasing in corrosion rate for year 2011, 2012, 2015 and 2016. As shown

corrosion rate has different trend for different year. Fig. 7 shows SAX graph of corrosion rate in 2011 and 2016.

(a) Corrosion rate in 2011
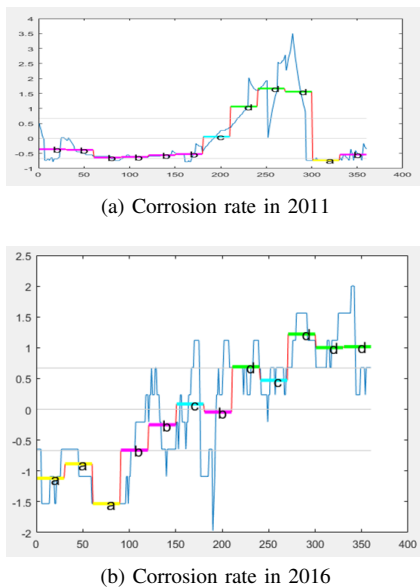

(b) Corrosion rate in 2016

Fig. 7. SAX patterns for Agent A and Agent C in 2014

in the previous section, SAX graph of corrosion rate for Agent B has decreased from *d* to *c* and significant bigrams based on *tdf-idf* is *dc* and *ba*.

Similar trend is represented by the other pattern, *ba*. Thus, it can be concluded that Agent B does not affect the corrosion rate. Based on the *tf-idf* result, the Agent E pattern is *ac ba cd dd*. The pattern shows an increasing in corrosion rate from *a* to *c* in year 2015. Therefore, it can be concluded that Agent E accelerated the corrosion process in 2015. The pattern shows a decreasing trend from *b* to *a* in 2013. However the it increased from *c* to *d* and become constant until end of year. This shows that environmental factor of Agent E contributes the most of corrosion rate throughout the 6-year analysis. The pattern for Agent C is constant from *c* to *b* in 2013, 2014 and 2016 but it shows a decreasing trend from *d* to *c* for year 2015 and 2016 while it started to increase again from *c* to *d* in year 2016. Throughout the six years analysis, symbolic pattern of Agent C shows that it accelerated the corrosion rate in some particular years while some other years the corrosion rate decreased.

Agent C and Agent E are the two-major environmental factors that contributed to the corrosion rate throughout the six years analysis. The pipeline might contain more substance from Agent C and Agent E that affect internal corrosion.SAX graph for Agent E in 2015 shows high values in Agent E and the increase of corrosion rate in the pipeline. According to [24], corrosion rate is low if the pipeline contains lower amount of some particular agents for several metals. For this case it might be because Agent E is a substance that accelerates the chemical and electrochemical processes occurring in the pipeline [11].

IV. Conclusion

This work describes a symbolic time series approach to analyse corrosion rate in the context of oil and gas industry. Five corroding agents were determined to investigate the measure of corrosion rate. Different types of corroding agents or factors give different rate of pipeline corrosion throughout 6-years observation. The original time series collected with

missing values must be processed and represented as symbolic time series representation using SAX. A text-based method which is N-gram is able to define important SAX sequences using term frequency-inverse document frequency (*tf-idf*). This technique allows the pattern to be more descriptive as *tf-idf* reflects the important trend that can occur in the data analysis. Further work on this topic is currently being carried out. Other types of imputation techniques suggested by other researchers may be investigated to improve data quality and to have more accurate result. The understanding of the complexity of the dataset particularly of corrosion process with multiple environmental factors is the most important direction for the future work. A method to predict the correlation between the multiple corrosion factors need to be identified based on the symbolic representation and text mining techniques .

References

[1] D. Brondel, R. Edwards, A. Hayman, D. Hill, S. Mehta, and T. Semerad, "Corrosion in the oil industry," *Oilfield review*, vol. 6, no. 2, pp. 4–18, 1994.

[2] M. Golshan, A. Ghavamian, and A. M. A. Abdulshaheed, "Pipeline monitoring system by using wireless sensor network," *IOSR J. Mech. Civ. Eng*, vol. 13, no. 3, pp. 43–53, 2016.

[3] R. M. Aliguliyev and Y. N. Imamverdiyev, "Big data strategy for the oil and gas industry: General directions," *Problems of information technology*, vol. 2017, no. 2, 2017.

[4] J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing sax: a novel symbolic representation of time series," *Data Mining and knowledge discovery*, vol. 15, no. 2, pp. 107–144, 2007.

[5] B. Lkhagva, Y. Suzuki, and K. Kawagoe, "Extended sax: Extension of symbolic aggregate approximation for financial time series data representation," *DEWS2006 4A-i8*, vol. 7, 2006.

[6] S. Nešić, "Key issues related to modelling of internal corrosion of oil and gas pipelines–a review," *Corrosion science*, vol. 49, no. 12, pp. 4308–4338, 2007.

[7] Z. Ahmad, *Principles of corrosion engineering and corrosion control*. Elsevier, 2006.

[8] Y. Wang, G. Cheng, W. Wu, Q. Qiao, Y. Li, and X. Li, "Effect of ph and chloride on the micro-mechanism of pitting corrosion for high strength pipeline steel in aerated nacl solutions," *Applied Surface Science*, vol. 349, pp. 746–756, 2015.

[9] A. Fu and Y. Cheng, "Effects of alternating current on corrosion of a coated pipeline steel in a chloride-containing carbonate/bicarbonate solution," *Corrosion science*, vol. 52, no. 2, pp. 612–619, 2010.

[10] M. Ilman *et al.*, "Analysis of internal corrosion in subsea oil pipeline," *Case Studies in Engineering Failure Analysis*, vol. 2, no. 1, pp. 1–8, 2014.

[11] D. Ghazali, "A study of corrosion to the carbon steel in the present of carbon dioxide," Ph.D. dissertation, University Malaysia Pahang, 2010.

[12] M. Tan, F. Varela, Y. Huo, F. Mahdavi, K. Wang *et al.*, "An overview of recent progresses in acquiring, visualizing and interpreting pipeline corrosion monitoring data," in *CORROSION 2018*. NACE International, 2018.

[13] P. Esling and C. Agon, "Time-series data mining," *ACM Computing Surveys (CSUR)*, vol. 45, no. 1, p. 12, 2012.

[14] C. Ratanamahatana, E. Keogh, A. J. Bagnall, and S. Lonardi, "A novel bit level time series representation with implication of similarity search and clustering," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2005, pp. 771–777.

[15] S. Wu, Y. Wu, Y. Wang, and Y. Ye, "An algorithm for time series data mining based on clustering," in *Communications, Circuits and Systems Proceedings, 2006 International Conference on*, vol. 3. IEEE, 2006, pp. 2155–2158.

[16] F. Mörchen, "Time series feature extraction for data mining using dwt and dft," 2003.

[17] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, *Fast subsequence matching in time-series databases*. ACM, 1994, vol. 23, no. 2.

[18] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra, "Dimensionality reduction for fast similarity search in large time series databases," *Knowledge and information Systems*, vol. 3, no. 3, pp. 263–286, 2001.

[19] P. Gao, S. Yan, L. Xie, and J. Wu, "Dynamic reliability analysis of mechanical components based on equivalent strength degradation paths," *Strojniški vestnik-Journal of Mechanical Engineering*, vol. 59, no. 6, pp. 387–399, 2013.

[20] "Linear interpolation with excel," http://www.datadigitization.com/dagra-in-action/linear-interpolation-with-excel/, accessed: 2018-09-30.

[21] "Method to calculate interpolation step value in excel," https://support.microsoft.com/en-us/help/214096/method-to-calculate-interpolation-step-value-in-excel., 2017, accessed: 2018-09-1.

[22] D. Jurafsky and J. H. Martin, "Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition," pp. 1–1024, 2009.

[23] A. Bhakkad, S. Dharamadhikari, and P. Kulkarni, "Efficient approach to find bigram frequency in text document using e-vsm," *International Journal of Computer Applications*, vol. 68, no. 19, 2013.

[24] B. S. Pijanowski and I. Mahmud, "A study of the effects of temperature and oxygen content on the corrosion of several metals," Catholic University Of America Washington Dc Inst. Of Ocean Science And Engineering, Tech. Rep., 1969.