SUPPLEMENT | EXTENDED ABSTRACT
https://doi.org/10.1071/AJ21118

The APPEA Journal

# Extracting events from Daily Drilling Reports using Fuzzy String Matching

Mariana S. Oliveira[A,B,*], Adriano Mourthe[A,B] and Maria Clara Duque[A,C]

For full list of author affiliations and declarations see end of paper

*Correspondence to:
Mariana S. Oliveira
Intelie, Rio de Janeiro, RJ, Brazil
Email: mariana.oliveira@intelie.com.br

## ABSTRACT

Continuous monitoring of oil drilling operations reduces process interruptions and equipment failure. It also contributes to the development of Key Performance Indicators, which leads to more efficient resource management. Daily Drilling Reports (DDRs) have long been the primary way of recording noticeable events, such as stuck pipe. DDRs came to constitute a valuable information base for most oil drilling companies. However, the task of extracting knowledge from DDRs can be costly and time-consuming. This work proposes an approach to recognise drilling events in DDRs using a rule-based language processing method called Fuzzy String Matching (FSM). We applied the FSM algorithm to search for a set of predefined keywords and key phrases to extract possible Invisible Lost Time (ILT) events from DDRs that may indicate risks or low operational efficiency. The fuzzy part of the algorithm allows the identification of terms or expressions that match the pre-established ones approximately rather than exactly, accounting for typos and different suffixes or prefixes. The proposed solution was applied on a data set of 392 real-world DDR records from a drilling company using a set of six ILT event's key phrases annotated by Subject Matter Specialists. This process can be readily replicated to other events. The results show that in 116 reports tagged as normal, 92 records were identified as possible ILT events, which represents, in hours, 56% of the total drill normal time. Such promising results can lead to very significant improvements in identifying and extracting drilling events within DDRs.

**Keywords:** Daily Drilling Reports, drilling, Fuzzy String Matching, Invisible Lost Time, keyword extraction, natural language processing, oil and gas, textual data.

## Introduction

The oil and gas industry has seen an explosion of available digital text in the last decade. Massive amount of textual data is generated through Daily Drilling Reports (DDRs). A DDR is a 24-h report of all key events that took place on a drilling rig. It is written by operators, constituting a rich source of information. Moreover, the content present in these reports can be very useful in identifying Invisible Lost Time (ILT) events, which can contribute to higher operational efficiency and cost reduction.

However, manually analysing such large amount of data is a costly and time-consuming process that is prone to human error. As a result, automatic processing of text data has emerged, where one common approach is to use Natural Language Processing (NLP) techniques. These techniques can be used to automate information retrieval and assist the decision-making and risk prevention process.

In this context, we applied a method based on NLP to automatically identify six types of ILT events in the DDRs called Fuzzy String Matching (FSM). The solution proposed was tested in a data set of 392 real-world DDR records from a drilling company and revealed 92 records as possible ILT out of 116 normal reports. Considering the duration of the reports, these ILT descriptions returned represent 56% of the total drill normal time. These promising results can impact operator's workflow by automating the search for ILT events within the DDRs, resulting in a faster and more reliable analysis of hidden lost operating time.

**Accepted:** 4 March 2022
**Published:** 13 May 2022

**Cite this:**
Oliveira MS *et al.* (2022)
*The APPEA Journal*
**62**(S1), S158–S161. doi:10.1071/AJ21118

## Methodology

This paper aims to automatically extract possible ILT events from DDRs that may indicate risks or low operational efficiency. To achieve this goal, we apply a rule-based language processing methodology with two main phases: preprocessing and keyword extraction with a FSM algorithm.

Preprocessing is initially performed to reduce noise from free format texts and improve the FSM model ability. As text noises, we considered all irrelevant characters, words, terms that do not add important information and can affect the model performance. The noise removal procedure consists of: lowercase conversion; symbols removal; tokens with less than two characters removal; and blank spaces removal.

The resulting preprocessed DDR records are the input to the FSM algorithm. The FSM algorithm is a rule-based approach that searches for keywords and expressions predefined by a user. The technique can find the expressions in text when they occur, allowing a limited number of errors in the matches (Navarro 2001). Thus, it is possible to find matches in texts even with misspelled words and different suffixes or prefixes.

We can use a distance metric in FSM to quantify the dissimilarity between two strings. In this work, we apply the Normalised Damerau-Levenshtein (NDL) distance (Damerau 1964; Levenshtein 1966). For a match, we considered a NDL distance of a maximum of 30%. We apply the FSM methodology to retrieve DDRs with risk conditions or ILT. Fig. 1 summarises the entire procedure.

Initially, the Subject Matter Experts (SMEs) read samples of DDR texts and defined a set of keywords or expressions related to problems to be searched in the documents. The list is composed of six main problems usually found at the drilling operations: vibration issues, circulation loss, equipment failure, geological risk, hole conditioning issues and directional drilling issues.

For each main problem, SMEs separate the keywords or expressions found during the search. For instance, the annotated expression ‘erratic torque’ found in the documents is strongly associated with vibration issues events. After annotation, raw documents are preprocessed to reduce noise, and then the FSM algorithm is applied using the previously annotated expressions. When the algorithm finds an approximated expression, the record is classified with the proper label. In the previous example, if ‘eraatic torque’ is found at a certain document, the expression is approximated by ‘erratic torque’ and the document is classified as vibration issues.

The final data set is validated by the SMEs. New expressions can be added to the keywords annotated database and the search can be further improved. Fig. 2 shows an example of the applied methodology.

## Results and discussion

### Data set

To validate the proposed method, a data set of real-world DDRs from a prominant drilling company was used. Each observation represents an entry of a DDR drill operation, and the features are composed of a textual description and the entry’s metadata, such as report date. In total, the data set contains 392 observations and 14 features. Although only the textual data was used as input for the proposed method, the metadata columns were important for the keyword definitions by the SMEs.

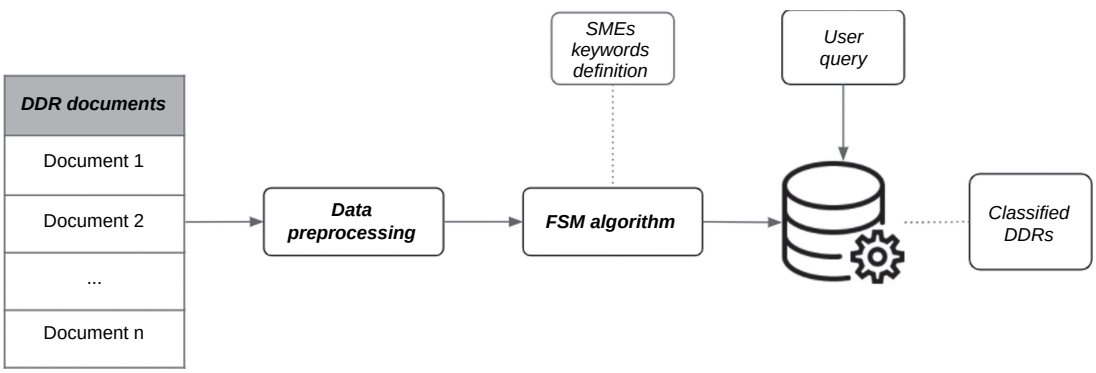



**Fig. 1.** Methodology diagram.

| Original record | Preprocessed record | Record with expressions |
|---|---|---|
| *Control ROP 75–115 fph due to erraatic torque* | *Control rop fph due to erraatic torque* | *Control rop fph due to erraatic torque*<br>Match: Erratic torque<br>Label: Vibration issue |

**Fig. 2.** Example application of the methodology at a document record.

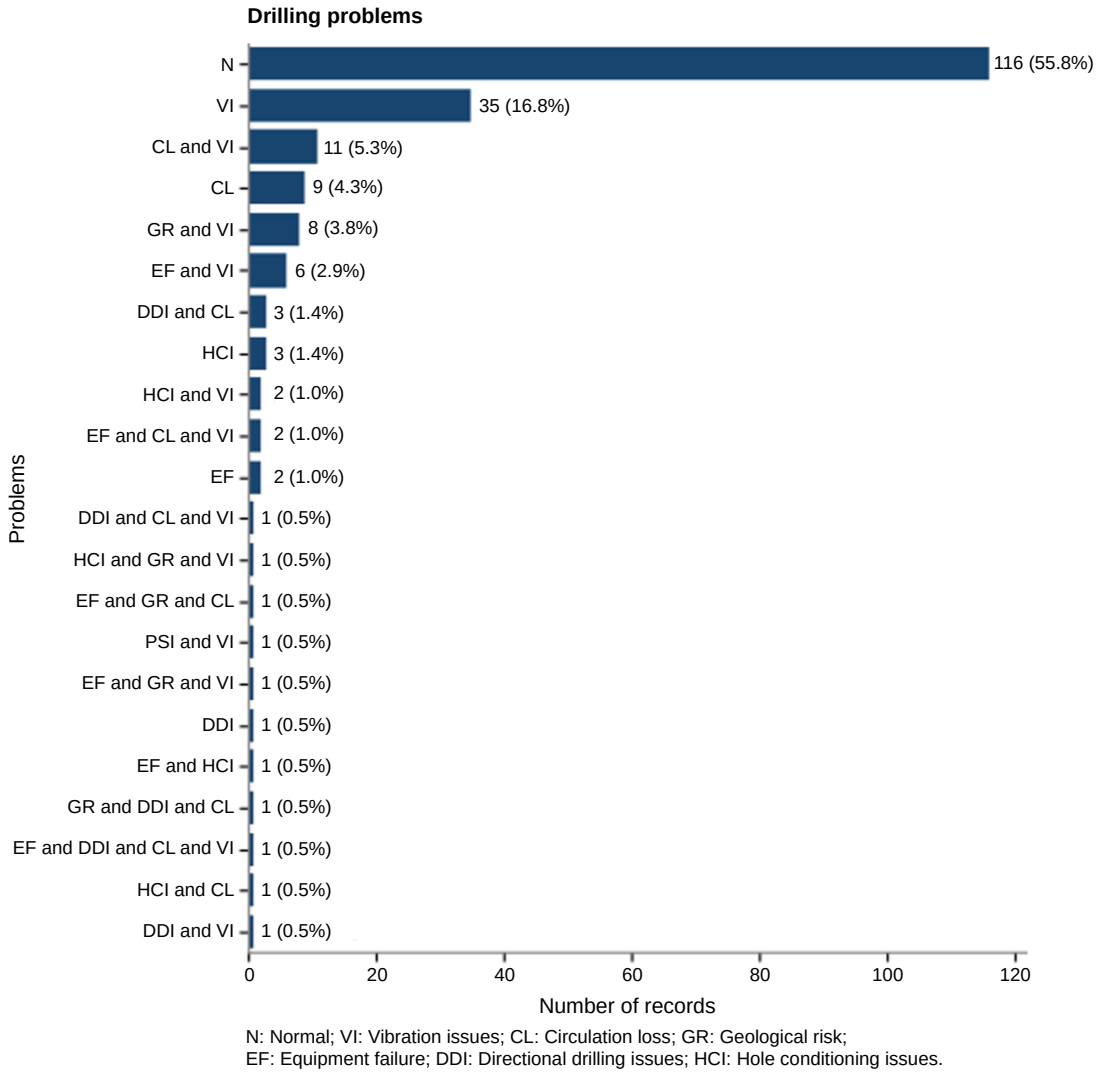



**Fig. 3.** Distribution of ILT categories identified by the FSM model, and the total count of remaining normal operations.

The data also contains a label that indicates whether a DDR entry concerns a normal or lost time operation. The labels are distributed in the following way: 208 (53%) observations are normal operations, total of 2568.5 h, and 184 (47%) are lost time operations, total of 1415.5 h.

## Experimental results

Given the data set and the set of keywords for each ILT category, we applied the methodology to the normal observations to reveal possible ILT events. Out of 208 normal observations, the model could extract 92 (44%) possible ILT descriptions, which sum up to 923 h of potential non-productive time.

Due to the nature of the method, a single DDR entry can have multiple labels, since it depends on the keywords identified. Therefore, the model could, for example, assign vibration issues and equipment failure labels to one entry if such entry contains at least one keyword from each category. Fig. 3 shows the distribution of the combinations of ILT categories identified by the proposed method.

## Conclusion

DDRs have vast importance to the drilling industry and can be very useful in identifying ILT events. However, the process of manually analysing DDRs to search for these events is a time- and resource-consuming process. To address this problem, the FSM method is proposed. It combines the specialists' knowledge with NLP techniques to find expressions taking into account typos, and different suffixes or prefixes. We applied the method to identify possible ILT

events in normal real-world DDR records and it was able to identify a sum of 923 h of potential non-productive time as well as determine the ILT category. These results show that the methodology was able to identify problems even with a lot of misspelling in the text, facilitating the search for problematic DDRs, reducing time and automating user workflow.

## References

Damerau FJ (1964) A technique for computer detection and correction of spelling errors. *Communications of the ACM* **7**, 171–176. doi:10.1145/363958.363994

Levenshtein V (1966) Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady* **10**, 707.

Navarro G (2001) A guided tour to approximate string matching. *ACM Computing Surveys* **33**, 31–88. doi:10.1145/375360.375365

**Data availability.** The data cannot be shared for privacy reasons.

**Conflicts of interest.** All authors confirm there are no conflicts of interest.

**Declaration of funding.** No funding from external organisations was received for this research.

**Author affiliations**

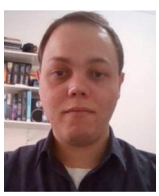
[A]Intelie, Rio de Janeiro, RJ, Brazil.

[B]Federal University of the State of Rio de Janeiro (UNIRIO), Rio de Janeiro, RJ, Brazil.

[C]Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, RJ, Brazil.

## The authors

**Mariana Oliveira** is a Data Scientist at Intelie, a Viasat Company, for 3 years. She works with development and implementation of Machine Learning and Natural Language Processing models in the Oil and Gas area. She graduated with a bachelor's degree in Computer Information Systems from the Federal University of the State of Rio de Janeiro (UNIRIO). Currently, she is a master's student in Computing at UNIRIO.



**Adriano Mourthe** is a Data Scientist at Intelie, a Viasat Company, for 5 years. He works with development and implementation of Machine Learning models for the Oil and Gas area. He graduated with a bachelor's degree in Computer Information Systems from the Federal University of the State of Rio de Janeiro (UNIRIO). Currently, he is a doctoral student in Informatics at UNIRIO.

**Maria Clara Duque** is a Data Scientist at Intelie, a Viasat Company, since August 2021. She works with development and implementation of Machine Learning and Natural Language Processing models in the Oil and Gas area. She graduated with a bachelor's degree in Petroleum Engineering from the Federal University of Rio de Janeiro (UFRJ). She also graduated with a master's degree in Industrial Engineering. Currently, she is a PhD candidate in Industrial Engineering at UFRJ.