World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:18, No:3, 2024

A Text Classification Approach Based on Natural Language Processing and Machine Learning Techniques

Rim Messaoudi, Nogaye-Gueye Gning, François Azelart

***Abstract*—**Automatic text classification applies mostly natural language processing (NLP) and other artificial intelligence (AI)-guided techniques to automatically classify text in a faster and more accurate manner. This paper discusses the subject of using predictive maintenance to manage incident tickets inside the sociality. It focuses on proposing a tool that treats and analyses comments and notes written by administrators after resolving an incident ticket. The goal here is to increase the quality of these comments. Additionally, this tool is based on NLP and machine learning techniques to realize the textual analytics of the extracted data. This approach was tested using real data taken from the French National Railways (SNCF) company and was given a high-quality result.

***Keywords*—**Machine learning, text classification, NLP techniques, semantic representation.

I. Introduction

PREDICTIVE maintenance is one of the applications of anomaly detection and it is in this context that the research project is articulated. The objective of the CALCHAS project is to continuously determine the state of health of several components through the supervision of the different parts that compose it. This supervision coupled with machine learning methods will allow us to perform several tasks: (a) Exploit past and future data to create a predictive model to extend the life of a component; (b) Classify the elements that cause failures in order of significance; and (c) Schedule maintenance operations that reduce the downtime of the component. For the predictive maintenance of components, we try to take advantage of knowledge from historical data to predict failures on continuous data flows. It is difficult to use parametric models in the context of predictive maintenance since these data do not necessarily obey a particular trend or a linear evolution.

The first part of the project is to work on textual data presenting incidents notes written by agents and administrators in the service now tool. The problem statement of this part is presented in Fig. 1.

The main objective is also to propose, first, an innovative solution for the analysis and qualification of these comments and resolution notes added by administrators at the end of the incidents' resolutions. It is important that the administrator, at the end of the ticket resolution, clearly describes the solution that he brought to solve the problem. The description of these sentences, must be clear, precise, and above all informative. Also, this description should generate business knowledge, on which employees will be able in the future, to solve similar tickets, and also to predict potential incidents. This solution will be tested in real data taken from the SNCF incident ticking system. This system is generated by the Service Now tool. The objective is then to use the proposed solution on other data using the same ticketing tool.

II. Related Work

This section is dedicated to describing related work describing the subject of textual data classification using machine learning algorithms.

A. Machine Learning Algorithms for Textual Analysis Review

Machine learning is part of AI. It has undergone a remarkable evolution in recent years and is still in continuous development. Also, machine learning is mainly interested in the design, development, optimization and implementation of several calculation methods that aim to analyze and process data structures and transform them into models. These models are then applied to solve different tasks (pattern recognition, classification of medical data, etc.) [1]. What characterizes machine learning compared to other traditional classification approaches is the fact of training computers to calculate or analyze problems and improve their ability to learn. Indeed, models based on machine learning are built using different models. They come from multidisciplinary currents. They refer to the analysis, processing and implementation of methods that allow a machine to evolve and perform tasks that are difficult to perform by more traditional algorithmic approaches. Moreover, machine learning incorporates statistical analysis to form data or values. It allows the construction of models from input data, which then helps to automate decision-making operations. Machine learning has been integrated into many fields of research and development, including natural sciences, facial recognition, video games, biology and robotics. We distinguish several types of machine learning, which are described below:

- Supervised learning: This involves guiding the algorithm with inputs that are previously labeled with predetermined outputs. In this case, the classes are known in advance. The objective of this method is to train the computer to find

Rim Messaoudi is with the Akkodis Research, Akkodis, France (corresponding author, phone: 0033650269699; e-mail: rim.messaoudi@akkodis.com).

Nogaye-Gueye Gning and François Azelart are with the Akkodis Research, France (e-mail: nogaye-gueye.gning@akkodis.com, francois.azelart@akkodis.com).

errors to adjust the classification model by modifying its parameters (for example the weights for an artificial neural network).

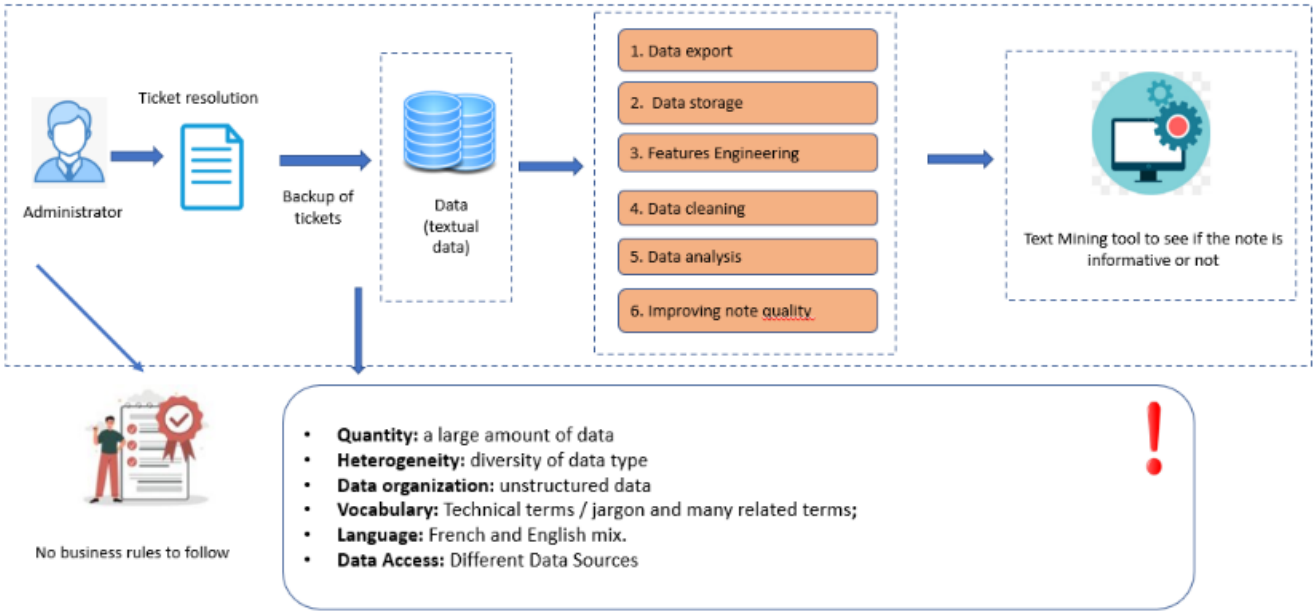

Fig. 1 Problem statement presentation

- Unsupervised learning: This one is based on unlabeled inputs and the outputs are not determined a priori. In this case, the learning algorithm will determine the classes in an autonomous way. The objective of this method is to allow the machine to reveal hidden patterns ensuring the automatic classification of raw data. This type of learning is useful for processing complex data and organizing it in a meaningful way.
- Semi-supervised learning: This type of learning uses a combination of labeled and unlabeled data and is between the two previous types of learning. This combination showed efficient classification accuracy but the problems of computation time and memory costs are still open.
- Reinforcement learning: This type of learning also sits between supervised and unsupervised learning. The objective of this method is to teach the algorithm to find the best solution through the construction of successive models. Reinforcement learning is concerned with the interaction with the environment, which makes it possible to reinforce the behaviors producing the adequate responses. It takes into account some key concepts based on the fact that the intelligent agent observes the effects of its actions, deduces from its observations the quality of its actions and improves its future actions. During the transmission phase to the network, the information can be reduced: it indicates whether the response generated from the network is correct or not.

For unsupervised algorithms, each must be applied according to its specificity. The SVM method corresponds to complex data sets. The isolation drill has a short execution time and is very easy to implement, however this method is not suitable for complex datasets. For methods based on deep learning such as ANN, they have a good anomaly detection performance in the context of predictive maintenance but are expensive in memory and difficult to implement. The preferred machine learning strategy in this project is to create a data model from the main unsupervised algorithms cited in the state of the art, One Class Support Vector Machine (OC-SVM, Local Outlier Factor (LOF) and IForest [2], [3]) and evaluate performance. Then it remains important to use optimization tools such as cross-validation or holdout. It would also be interesting to experiment with ensemble methods in view of their potential in terms of performance compared to the use of a single method. The XGboost method improves the performance of these different machine learning algorithms and presents encouraging results in the field of predictive maintenance [4]. In addition, the identification of the causes of breakdowns can be carried out first of all by a Gaussian mixture model using the EM (expectation maximization) algorithm in order to build clusters giving rise to a categorization of the breakdowns [5]. For the analysis of data flows, we take into account the categorization of failures resulting from the model built with historical data. For the detection of anomalies in the context of time series of different sensors, we can denote several approaches among which LSTM play an important role since this method is efficient even in a context of predictive maintenance [6].

Existing anomaly detection techniques are based on two important properties of anomalies: they behave very differently from others and they are rare. Among these techniques, we distinguish:

- Statistical techniques: can be parametric or non-parametric. Unlike the nonparametric approach, the parametric approach assumes a priori knowledge of the data distribution. Statistical methods build a model with a

confidence interval from existing data. New data that do not fit this pattern will be considered anomalous [7].

- Techniques based on proximity: include those based on nearest neighbors and those based on clustering. Nearest-neighbor techniques determine for an observation its k nearest neighbors by calculating the distance between all observations in the dataset. These methods require a preliminary calculation, and as a result, they are expensive in execution time. There are many approaches to nearest-neighbor-based methods: we found for example the distance-based approach [8].
- Clustering techniques: this type of techniques consider data tuples as objects. They partition the objects into different groups, or clusters, in this case objects within a cluster are "similar" to one another and "dissimilar" to objects in other clusters [17].
- Techniques based on deep learning: represent a class of supervised or unsupervised machine learning algorithms based on methods such as auto encoders (AE) and One-Class Neural Networks (OCNN) [9].
- Other techniques: exist such as those based on support vector machines neural networks [10], methods adapted to large dimensions by construction of subspaces or by dimension reduction.

B. NLP Algorithms for Textual Analysis

NLP, or TALN (automatic language processing), is defined as a discipline of AI that essentially allows machines to be able to understand human language as it is written or speak [11]. It is a discipline that deals with the manipulation and generation of this language via an interface linking computer science and linguistics. NLP is applied in various fields of application, namely: machine translation, spell checking, text classification and categorization, and voice recognition. Textual data, in the majority of cases, are unstructured and have semantic content where words, sentences and texts are not just a succession of letters, but they include a particular meaning. To process these textual data, it takes a phase of analysis and preparation to transform it into usable data. Subsequently, they will be applied with machine learning algorithms (e.g., Random Forest, Gradient Boosting, and deep learning).

With regard to the field of textual analysis, the NLP makes it possible to carry out various functionalities, of which we quote: the automation, simplification and acceleration of many tasks (translation, analysis of themes and feelings, categorization and detection of entities, etc.). It is very useful today to incorporate in many AI solutions this type of characteristics. In this context, [12] proposes an approach that processes repair and maintenance reports to help detect anomalies and differentiate between normal and damaged operating states of machines and their installations. This work uses NLP techniques to analyze these data and create a tool that optimizes predictive maintenance and supports the development of fault catalogs. Some details are indicated via the developed tool are we quote the definition of the time intervals of the work instructions as well as the troubleshooting guides. The study of [13] utilizes NLP techniques to analyze failure reports in complex power systems. The objective is to address the subject of predictive maintenance by reducing downtime and repair time and increasing the operational efficiency of equipment while reducing troubleshooting costs. The approach is also based on the analysis of textual data by integrating semantic techniques (ontologies).

In addition, [14] emphasizes the subject of textual data analysis via NLP techniques. These data can contain significant information to predict the duration of machine failures, potential causes of a problem, or the need to stop production to perform repair activities. This information can be viewed using machine learning (ML). However, these data are generally unstructured and require the exploitation of deep learning models, namely CamemBERT and FlauBERT. The proposed information extraction method was able to provide valuable information from a set of poorly structured maintenance reports. The best-known NLP algorithms processing textual data are summarized in the works of [15] and [16].

C. Discussion

The study of these approaches has shown the importance of algorithms linked to AI in improving several tasks related to predictive maintenance such as the diagnosis of systems within companies, and the automation of files and reports related to hardware failures. In addition, we noticed the richness of semantic models and their ability to specify different complex situations in the field of predictive maintenance. They thus suggest high-performance IT solutions.

Considerable attention has been paid to anomaly prediction in general. However, most of these studies, e.g. [14], do not discuss the process of managing incidents that can interfere with the normal operation of equipment in the company. In addition, some works only, e.g. [12], deal with the detection of failures based on machine learning techniques without giving additional information on the types of incidents, information on the resolution process carried out. Other works, e.g. [15] and [16], have dealt only with the use of ontologies and NLP to provide a decision-making tool that can be shared between users.

To recap, the advantage of this work is to enhance the application of NLP techniques for solving various problems related to predictive maintenance, especially the analysis of unstructured textual data. By studying this work, it has been found that we cannot use these algorithms to deal with tickets or incidents that may affect a company or a system. Hence the need for an algorithm that allows:

- Access the various SNCF databases and collect data;
- Clean and structure data through the application of semantic techniques (NLP and ontologies);
- Do the textual analysis and populate the ontology with the data;
- Realize classification models through the integration of machine learning;
- Test the proposed solutions and improve their performance.

The algorithm to be proposed takes data on incidents and tickets as input and outputs information related to the activities

of admins (notes, comment date, etc.) who worked on the ticket. The objective is to have an idea of the resolutions made. This work is carried out based on several quality requirements to be met. More details on the developed algorithm are mentioned in Section V.

III. Proposed Approach

In this section, we present our predictive maintenance solution dedicated to detecting anomalies and failures at SNCF (see Fig. 2). The approach we propose contains different interconnected parts, starting with the data collection and cleaning phase from SNCF, these data will be analyzed and represented semantically through the integration of NLP techniques and ontologies.

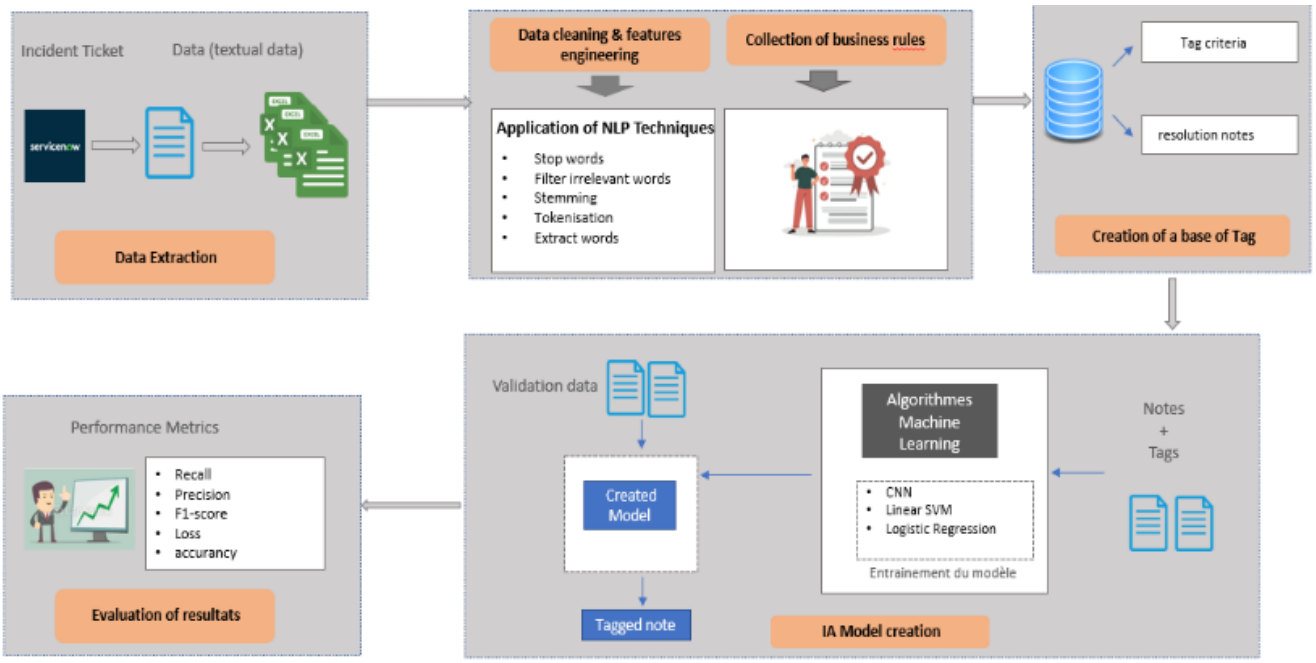

Fig. 2 Proposed approach for text treatment and classification

Subsequently, the next step is to build prediction models based on machine learning techniques (for example CNN models). Evaluation of results and optimization of performance are also part of our approach.

A. Automatic Tag of Admins' Notes

A good comment is defined as a comment that brings knowledge related to the resolution of the incident. The objectives of a well-documented trouble ticket are to:

- Serve as a knowledge base for the subsequent resolution of similar incidents;
- Allow the identification of problems, and help the investigations required to resolve the problems;
- Enable effective resolution of 'over-accidents' incidents caused by the resolution of a first incident.

However, an incident ticket is not a resolution procedure (i.e., KB (knowledge base instruction), even if it can occasionally serve as a basis for writing such an instruction. As a reminder, a comment is visible from the service now portal. On the other hand, a work note is only visible to those who have back-office type access to the incident ticket. The ticket documentation quality criteria are as follows:

- Concise: The resolution note should be kept concise. Notes/comments can be more detailed but still 'reasonable'. Moreover, to keep the resolution note concise, it is better to focus on the comments area for the details.
- Language: Only French in the resolution notes. In the comments it is more complicated because there may be traces of commands that will come out in English, for example.
- Clarity: Explanations should be understandable. Correct spelling and grammar also help to make the comment more understandable.
- Factual: As far as facts are possible. If possible, proof of resolution (e.g., disk volumes after a purge, return of service etc.). If approvals were required it is desirable to include these.

Table I summarizes the criteria selected for the annotation of resolution scores.

B. Applied Tools

In this section, we present the applied tools included to realize our tool. Table II gives an overview of these techniques and the name of the used API.

TABLE I
BUSINESS RULES AND CRITERIA' OF TAG

Criteria/Ratings	N/A	0	1	2
Description action included? (0-2)	Not applicable	No description	Brief description, or excessive description in details	Description giving the essential information without excess
Understandability. (0-2)	Not applicable	Reference to another source of information without further explanation	Reference to another source with insufficient explanation	No reference to another source or well-exposed reference
[NdR] Quality of the description of the result obtained	Not applicable	No description	Too brief or excessively detailed description	Complete but concise description
Timestamp included? (0-1)	Always applicable!	No timestamp	Presence of a date and time of the action	
Clarity? (0-2)	Always applicable!	For example, very obscure or amounting to jargon. Or even a simple log of the action	Understandable but with effort (comments diluted in a log, excessive length)	Clear and concise

TABLE II
APPLIED TOOLS

Name	Description
Keras	Open source library written in python
Tenserflow	Open source machine learning tool
Python	Programming language
FastAPI	High performance web framework for building APIs
Streamlit	Framework that allows you to create web applications that can easily integrate machine learning models and data visualization tools
Spacy	spaCy is a free and open source Python library released under the MIT License for NLP

C. Evaluation

To evaluate and our model, we reserved around 1000 notes to test the applied techniques. We used NLP techniques such as tokenization, removing accent and empty words to clean the data. The obtained results are presented in Table III. Regarding these results, we find that all models generated impressive preliminary results such as the "Understandability level" and "Timestamp" criteria.

TABLE III
PRESENTATION OF RESULTS

Tag Criteria (IA models)	Accuracy
Understandability (0-2)	83%
Description action included? (0-2)	75%
Quality of the description of the obtained result	74%
Clarity? (intelligible language) (0-2)	80%
Timestamp included? (0-1)	94%

IV. CONCLUSION AND PERSPECTIVES

In this paper, we presented an approach based on the CNN models that aims to qualify the resolution notes while providing additional information. Another work is devoted to developing a method that allows notes to be automatically annotated based on well-defined qualification criteria (e.g., clarity, timestamp, quality of description, etc.).

In the future, we aim to work on the optimization of our prediction method by exploiting other techniques such as SVM, linear regression and LLM models in order to improve it and gain in performance and profitability. We also intend to develop the incident tag architecture based on reinforcement learning. The objective is to compare the results with the supervised learning methods proposed this year.

We also plan to extract and work on the images attached to the comments.

REFERENCES

[1] Li, Y. (2022, January). Research and application of deep learning in image recognition. In 2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA) (pp. 994-999). IEEE.
[2] Roy, S. D., & Debbarma, S. (2022). A novel OC-SVM based ensemble learning framework for attack detection in AGC loop of power systems. Electric Power Systems Research, 202, 107625.
[3] Barbariol, T., & Susto, G. A. (2022). TiWS-iForest: Isolation forest in weakly supervised and tiny ML scenarios. Information Sciences, 610, 126-143.
[4] Bekar, E. T., Nyqvist, P., & Skoogh, A. (2020). An intelligent approach for data pre-processing and analysis in predictive maintenance with an industrial case study. Advances in Mechanical Engineering, 12(5), 1687814020919207.
[5] Obermair, C., Apollonio, A., Wuensch, W., Felsberger, L., Cartier-Michaud, T., Catalán Lasheras, N., ... & Millar, W. L. (2021). JACoW: Machine Learning Models for Breakdown Prediction in RF Cavities for Accelerators. JACoW IPAC, 2021, 1068-1071.
[6] Malakouti, S. M., Ghiasi, A. R., Ghavifekr, A. A., & Emami, P. (2022). Predicting wind power generation using machine learning and CNN-LSTM approaches. Wind Engineering, 46(6), 1853-1869.
[7] Kumar, D., Sarangi, P. K., & Verma, R. (2022). A systematic review of stock market prediction using machine learning and statistical techniques. Materials Today: Proceedings, 49, 3187-3191.
[8] Nuankaew, P., Chaising, S., & Temdee, P. (2021). Average weighted objective distance-based method for type 2 diabetes prediction. IEEE Access, 9, 137015-137028.
[9] Oza, P., & Patel, V. M. (2019, May). Active authentication using an autoencoder regularized cnn-based one-class classifier. In 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019) (pp. 1-8). IEEE.
[10] Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In Machine learning (pp. 101-121). Academic Press.
[11] Sharifani, K., Amini, M., Akbari, Y., & Aghajanzadeh Godarzi, J. (2022). Operating Machine Learning across Natural Language Processing Techniques for Improvement of Fabricated News Model. International Journal of Science and Information System Research, 12(9), 20-44.
[12] Öztürk, E., Solak, A., Bäcker, D., Weiss, L., & Wegener, K. (2022). Analysis and relevance of service reports to extend predictive maintenance of large-scale plants. Procedia CIRP, 107, 1551-1558.
[13] Carchiolo, V., Longheu, A., Di Martino, V., & Consoli, N. (2019, September). Power Plants Failure Reports Analysis for Predictive Maintenance. In WEBIST (pp. 404-410).
[14] Usuga-Cadavid, J. P., Lamouri, S., Grabot, B., & Fortin, A. (2021). Using deep learning to value free-form text data for predictive maintenance. International Journal of Production Research, 1-28.
[15] Gasparetto, A., Marcuzzo, M., Zangari, A., & Albarelli, A. (2022). A Survey on Text Classification Algorithms: From Text to Predictions. Information, 13(2), 83.
[16] Mehta, R., Jurečková, O., & Stamp, M. (2023). A Natural Language Processing Approach to Malware Classification. arXiv preprint arXiv:2307.11032.

Rim Messaoudi completed her Ph.D. from the University of Clermont Auvergne, France, and the University of Sfax, Tunisia (2021). She completed

her initial education from various reputed educational institutes in Tunisia (ISIMM, Enet'COM and FSEGS). She completed her master's degree in 2017 and her License (Computer Science) in 2014. She has experience in developing deep learning algorithms and writing research articles. She has been involved in the medical field and the detection of liver cancer lesions from MRI/CT scans images. Her areas of interest are machine learning (ML), artificial intelligence (AI), medical image processing, and data classification. She has published several research papers in journals of repute and in refereed international conferences published by Springer. She is also contributing as a reviewer in the editorial boards of a few reputed journals.