Voicemail Urgency Detection Using Context Dependent and Independent NLP Techniques

Asma Trabelsi[1], Séverine Soussilane[2] and Emmanuel Helbert[2]
[1]*Alcatel-Lucent Enterprise, ALE International, 32, avenue Kléber 92700 Colombes, Paris, France*
[2]*Master Data Science and Complex Systems, Université de Strasbourg, France*

Keywords: Voicemail Classification, Urgency Determination, BERT Embedding, Data Augmentation, Explainability.

Abstract: Business field has improved exponentially during the last two decades: working methods have changed, more and more users are connected to each other across the globe, same teams as well as different teams can be separated by countries in big companies. So, users need a way to select messages to treat in priority for a better business management and a better communication. In this paper, we implement an approach enabling to classify voicemail messages into urgent and non urgent. The problem of determining urgency being still vast and open, some criteria should be used to decide the importance of messages depending to one's necessity. Among these criteria, we can mention the sender position, the time of sending as well as the textual content. In this paper, we focus on classifying voicemail messages based on their contents. As there exist several Machine Learning approaches for text vectorization and classification, various combinations will be discussed and compared for the aim of finding the most performant one.

1 INTRODUCTION

Artificial Intelligence, in short AI, is largely used today to solve several real world problems. One of the benefits of AI is to let machines take over the recurrent tasks that human usually do with poor added value. In today's world, AI had and is still having a pivotal role in various domains like in the medical field (Vaishya et al., 2020), in the industry (Dopico et al., 2016), in education (Cheng et al., 2020), in communication (Cayamcela and Lim, 2018), etc. Machine Learning, which is a sub-field of AI, enables the automation of problem resolution through the learning process of known cases. Today, Machine Learning is used to handle simple to medium to complex problems. It is helpful in various domains like energy management (Veiga et al., 2021), mobile network analysis (Sevgican et al., 2020), or even in aerial maps creation through image processing (Mnih, 2013), text classification (Mujtaba et al., 2019), and even in voice recognition field. The later has been considered as an interesting search field but is viewed as a difficult task regarding the diversity of languages. It has to be noted that each language can have a different structure in the written form, moreover, the same terms' meaning from the same language can differ depending on the context (polysemy). This complexity has the origin of the exploration of human language and the introduction of Natural Language Processing (NLP) (Kang et al., 2020). NLP has also been considered as a tool for the identification and change of communication behaviour.

Communication is the way of exchanging information between a source and a destination. Humans' system of communication is flexible and precise (Krauss, 2002). They convey what they want to say in a meaningful way through writing, speaking or with signs. By this way of communication, humans are able to express their feelings, their needs, exchange information, etc. Communication is valuable between any group of individuals, particularly in the business field. Companies' staffs do need to exchange properly to lead a project in the best possible way. To date, all companies have the habit of using some communication platforms to connect their staffs, either SaaS (Software as a Service) or on-premises. These solutions provide instant messaging but also legacy voice interactions and in particular, voice messaging. Voice messaging services are very important to share critical information and request immediate action from the recipient. But as it requires manual consultation, there is no assurance about the delivery of the message and its acknowledgement, though the importance and the urgency of the mes-

Trabelsi, A., Soussilane, S. and Helbert, E.
Voicemail Urgency Detection Using Context Dependent and Independent NLP Techniques.
DOI: 10.5220/0011685800003393
In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART 2023) - Volume 3*, pages 450-456
ISBN: 978-989-758-623-1; ISSN: 2184-433X
Copyright © 2023 by SCITEPRESS – Science and Technology Publications, Lda. Under CC license (CC BY-NC-ND 4.0)

sage. On the other hand, some messages do not require immediate action and do not need to be listen to immediately. Thus, combining transcription and voice message classification can be very helpful to filter, prioritize and ease message reading. Basically, transcription will enable to transpose voice media into text for a quick silent consultation on a screen without the need of audio transducer, while message classification will enable to display only urgent message, lowering the tasks burden. AI and Machine Learning techniques may help to classify voicemails in compliance with their urgency and to display urgent voicemails transcribed to the user through text notification. A study was carried out in 2019 on urgency recognition in voicemail by analysing the speech directly (Kamiyama et al., 2019). Our approach is different as we propose to analyse the text semantics after the voicemail messages transcription.

Though, understanding the meaning of human natural language in a voice message or in a text for classification purposes is not an easy goal. Several studies already exist, not only in the context of voicemails classification, but also in prioritization of e-mails (Choudhari et al., 2020), help-desk tickets (Al-Hawari and Barham, 2021), etc. Some research scientists have already classified e-mails into relevant categories by extracting keywords that appear frequently, while others treat all the mail content (Gupta and Goyal, 2018 ; Bacchelli et al., 2012). It is important to highlight that previously mentioned studies use context independent vectorization to convert text to vector for a classification aim. The same kind of approach is used for news classification (Li et al., 2018). But, to our best knowledge, there are very few studies on voicemails treatments as well as there is no open source voicemail data that can be leveraged to test the performance of various solutions. Thus, the motive of this study is twofold: First, collect voicemail data. Second, build and test models enabling voicemail classification. In our approach, all processing will be performed on text from voicemail transcription. For the classification of voicemails into urgent and non urgent, we relied on both context dependant word embedding and context independent word embedding for the vectorization methods. We will then experiment several well-known Machine Learning classifiers like Logistic Regression, XGBoost, Support Vector Machine. The idea is to find the best combination between vectorization and classification algorithms for voicemail classification.

This paper will be organized as follows: Section 2 is dedicated to the state-of-the-art with some explanations on Machine Learning algorithms and vectorizers. In Section 3, we will present our approach. In Section 4, we will discuss the results. In Section 5, we draw our conclusions as well as our future works.

2 STATE OF THE ART

In this section, we highlights some well-known Machine Learning classifiers as well as existing vectorization techniques.

2.1 Machine Learning Algorithms

There exist several Machine Learning classifiers. In this section, we describe briefly well-known ones including KNN, SVM, XGBoost and Logistic Regression. In what follows, we describe in more details each of these algorithms.

2.1.1 Logistic Regression

For this algorithm, a linear threshold is used for classifying input data. It computes the relation between the output and the independent features (Tripepi et al., 2008). In the event of value too close from the threshold, the input can be misclassified. So, there is a risk for the predictions to be wrong (Pekhimenko, 2006). Interestingly, Logistic Regression has been used in several cases, among them, the detection of susceptible landslides causing human deaths in a part of Himalaya. It has also shown its efficiency in fashion trends forecasting for the textile domain as well as numerous others.

2.1.2 Support Vector Machines

It functions like the Logistic Regression algorithm but the threshold is now a hyperplane. The hyperplane which maximizes the distances between classes will be chosen to separate values by the algorithm itself. This aspect has been applied in many real world problems, such as, HIV peptides detection as well as for text classification because it works fine for high dimensional data.

2.1.3 *k*NN - *k* Nearest Neighbours

This method chooses one of the inputs as reference. It represents the reference point in a multi dimensional space. It will place other inputs in the space by computing a distance metric between each input and the reference input. According to the k number we choose, it will represent k other inputs as neighbour for the reference input and so form a cluster (a class). The remaining inputs will be in the other classes. This implies that we should choose the optimal k number

(Alshehri, 2020). This algorithm, like SVM can be easily generalized as it doesn't require any knowledge of the domain. It has been used in various fields, even in the speech recognition tasks in phonetics classification (Asaei et al., 2010).

2.1.4 XGBoost

Briefly written, XGBoost uses the creation of an ensemble of regression trees for minimizing the loss function. This algorithm has proved its efficiency compared to other methods such as Deep Neural Network or even *k*NN in gene expression prediction. Indeed, XGBoost is less expensive and more interpretable than Deep Neural Network. Interpretability is important in this domain to know the impact of each gene in diseases.

2.2 Vectorization

This method consists of converting textual data into numerical vectors to give them as an input to Machine Learning algorithms. This technique encloses two sorts of word embedding in the text domain: Context independent and context dependent methodes. In the following explanations, we will describe some details of these approaches.

2.2.1 Context Independent Approach

There are plenty of methods that are context independent. Among them, 'Bag of words' is a method which is commonly put into practice in context-independent methods. It represents words by their number of occurrences in the dataset. 'TF-IDF' is, however, a better method as it is based on each term's frequency in the input text as well as in the whole dataset. Context-independent embedding can be easier to implement and can also give acceptable performance results but the prediction on new data might be wrong in most cases. For example, the word "bank" does not have the same meaning in these two sentences: "The man went fishing by the bank of the river" and "The man was accused of robbing a bank". In a context-independent embedding tool, the word "bank" will have the same vector for the whole input corpus. Imagine we give the label not urgent to the first sentence and label urgent to the second one. The classification model will get confused for the usage of the word bank. As a consequence, if we want to predict the class of a new sentence containing "bank" in it, the output may be impacted. For example the prediction of "ALERT! The bank is under the control of thieves" can be not urgent. This is where we introduce context-dependent embedding techniques which differentiate the vectors attributed to a word according to the context.

2.2.2 Context Dependant Approach

To train an NLP model, we need millions of data because language is a complex tool of communication. In practice, when we work on a project and we do not have so much data available to learn every aspects of the concerned language, it is recommended to focus on pre-trained models using large corpora such as wikipedia pages or other books. Among them, there is BERT, a context-based model that reads sentences from right to left but also from left to right, which other encoders do not do. It is therefore bidirectional and helps to understand words in their context. It is well known for its outstanding performance in text classification. BERT uses a transformer to learn the linkage between words in the attention layer of the architecture. A transformer contains an encoder and decoder to predict. However BERT uses only the encoder part as it is a language model. Before the encoding part, the data is pre-processed by the algorithm, this step is called "tokenization". Regarding the later part, the input is a little bit transformed from what we have after a classic tokenization before fed to the model. BERT adds a token [CLS] at the beginning of each sentence and a token [SEP] to separate each sentence from each other. This is the first embedding layer. The second one is the segment layer which gives a marker token to know to which sentence each word belongs to. The third one is the positional layer indicating each word's position in the sentence. After the tokenization comes the encoding. Encoding is the technique used to learn relationships between words through the 'Masked Language'. It hides words randomly and tries to predict them in relation to other words surrounding them from their right and their left at the same time. The words therefore have a different vectorization depending on the context in which they are used in the sentence taken as input. To learn the relationship between sentences, the model is given pairs of sentences (x,y) and it tries to predict if the sentence y is the next sentence of x in the original input. Finally, words will have different vectors as per the context.

3 PROPOSED APPROACH

Alcatel-Lucent Enterprise already provides means in its collaboration platforms to indicate the level of urgency in message notification. Being able to extend this feature to voice messages would greatly improve

the quality of experience of the communication. Classifying voicemail as urgent or not urgent will indeed help the receiver focus on the most important messages and fasten their treatement. This could be an advantage in the customer care sector for example. For the caller, it gives an insurance that her message will be listen to with the right level of urgency. Our approach is presented in Figure 1. We can observe that in the company, voicemail data that are received from the sender have to be transcribed into written form, translated into English and passed to the classifier model. The idea behind the translation is to handle multi-language aspect by using only one model instead of one model per language. Knowing that translating tools such as DeepL are very efficient, we assume that there will net be much loss of information. After the classification, the message will be translated into the receiver's language and delivered. To date, there exist many vectorizers and classifiers as we have seen in the above sections. So, the choice is wide and difficult. In regards to voicemail messages, we do not have many studies that have been made on their treatments. This study is an additional research to the few that already exist on how to classify voicemail messages.

4 EXPERIMENTATION SETTINGS AND RESULTS

4.1 Experimentation Setting

In this section, we will present our experimentation settings including data collection, data labeling, data pre-processing as well as model explainability.

4.1.1 Data Collection and Labeling

Data collection is among one of the most pivotal steps in AI projects, only relevant data will make the program give accurate results. We have created our own data set composed of the same number of urgent messages and non-urgent messages. This choice has been made to deal with data imbalance during the classification task. Indeed, classification algorithms will learn one class better than others if it is represented by more data than the remaining classes. The dataset was filled by many different individuals in order to keep the data unbiased because not everyone talks the same way to express urgency. Our idea was to leverage our company resources by considering it as good representation of general business case. We simply asked employees working in all part of the organization to write five urgent voice messages and five non-urgent voice messages. Each contributor was asked to use her usual wording. The benefits of this method were numerous: The gathering of nearly real messages without data privacy issues, the de-facto labelling of all messages as urgent or not urgent, the taking into account of various business contexts. 80 persons answered to the request leading to the collection of 800 messages. 400 remaining messages were also written manually by the team to finally obtain a first data set of 1200 messages. We then used the library NLPaug (Deng and Shrestha, 2019) to augment the data. It will generate new data using the words that we have already in our dataset. This method is BERT friendly as it uses BERT model to generate full sentences. Our data is then composed of 1800 labeled voicemail messages.

4.1.2 Data Pre-Processing

Once the data is collected and labeled, we have to move to the next step which is data pre-processing. Firstly, we have transformed the text to lower-case. After that, we have denoted each word or combination of words of the content separately. The next step is to remove punctuation marks, names, greetings and gender related words from our voicemail data. However, applying the removal of these words will reduce the size of the dataset which can impact the accuracy of the model. In that case, it will be better to add stopwords later when based on some explainability model.

4.1.3 Model Explainability

Machine Learning algorithms are considered as black-boxes which are defined by their inexplicable decision making process due to non-linearity. We only know the output but we do not know how it has come to the final decision. In this study, we used the explainability method to know which are the words of the input text that have been used to make the decision. This will help us to adapt the data cleaning part by removing words which are driving to wrong predictions. There exist several approach allowing us to make explainable AI. LIME (Kadiyala and Woo, 2022) is one among well-known algorithms used to explain any classifier in an uniform way. For each prediction, it observes the neighbour inputs locally and tries to extract words that have helped the decision of those neighbour instances. Though LIME is a local explainability algorithm, it actually helped us increasing the model's accuracy.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

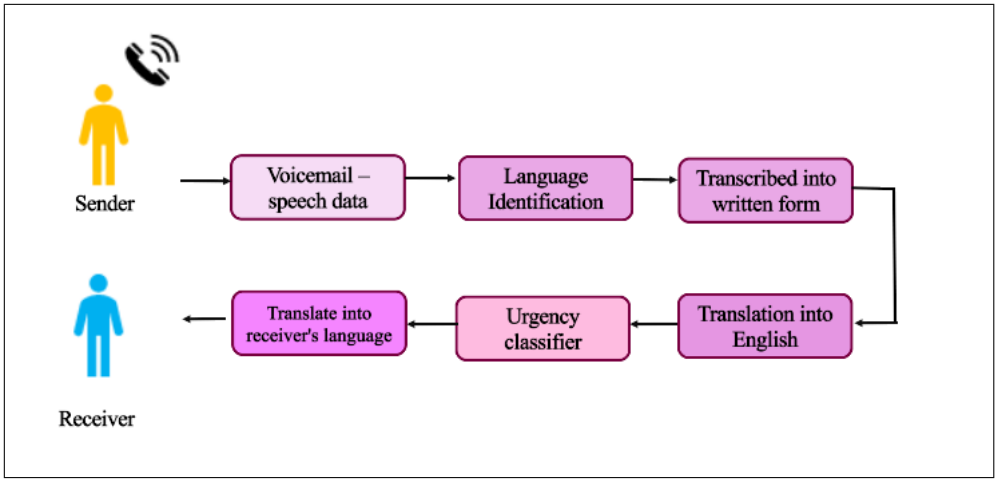

Figure 1: Approach of voicemail treatment at Alcatel-Lucent Enterprise.

4.2 Experimentation Results

This subsection is dedicated to the analysis of our experimentation results. For the comparison, we have made k-fold cross validation with k equals 10. The results are given in Table 1. We can observe that TF-IDF vectorization gets fair accuracy, a little bit better than BERT embedding. The reason for obtaining such results is explained by the urgency keywords that appear very frequently in our dataset. As TF-IDF is based on frequency of occurrences of the words, it gets a bit higher accuracy than BERT embedding. The problem with this vectorization technique is that it will perform less well when it comes to new data if the words of the new message do not contain the frequent words observed by the model. It can be less performant in the event of the urgency being conveyed indirectly. We tested the program with real cases to know if TF-IDF really performs poorly compared to BERT, and it is actually the case. Let us take for example the following sentence "Your attention is needed, your passport has to be renewed". This sentence has been classified as not urgent with TF-IDF vectorizer because the initial dataset did not contain words like "passport" or "renewal" while the message is actually urgent as it is classified using BERT embedding. So, despite the accuracy here being almost equal to TF-IDF, BERT embedding will be the best to use for any new voicemail messages. For our case, we conclude that we will keep the combination BERT-SVM as SVM has the highest average accuracy with BERT embedding. The reason why Logistic Regression is less efficient than other algorithms in all cases is that our features (words) are not totally independent whereas Logistic Regression computes the relationship between the output and independent features. SVM normally works well with any kind of data because the idea of finding an unique hyperplane separating the classes at maximum can be easily generalized to any case.

Table 1: Average Accuracy using various algorithms on dataset.

	KNN	XGBoost	SVM	LR
TF-IDF	89%	86%	**90%**	89%
W2V	73%	**75%**	33%	70%
BERT	84%	83%	**88%**	87%

5 CONCLUSION

In this paper, we have explored the problem of transcribed voicemail classification for business aims using Machine Learning tools. As for any text classification task, there exist different vectorization modes as well as different classification algorithms. The choice of the best combination is still an open question. The idea behind this study is to compare several vectorization and classification algorithms combinations for voicemail classification. Experimentally, we have shown that a combination of BERT and SVM as well as a combination of TF-IDF have given the best results. We retained BERT-SVM as it seems to be the best solution for classifying voicemails that differs completely from the train data. As a future work, we would like to collect more and more data in order to improve the model even more. We would like also to explore other combinations or other classification techniques like Neural Networks as well as models allowing to handle uncertain data through the evidence theory (Skowron, 1990) and use evidential machine learning classifiers such as Evidential KNN

(Jiao et al., 2015), Enhanced Evidential KNN (Trabelsi et al., 2017) and also evidential decision trees (Li et al., 2019). These kind of algorithms have been used for solving several real world problems when it is about uncertain data. We also have the idea of extending this model to other tasks like support tickets classification, and to vocal messages left in the inbox.

REFERENCES

Al-Hawari, F., & Barham, H. (2021). A Machine Learning based help desk system for IT service management. Journal of King Saud University-Computer and Information Sciences, 33(6), 702-718.

Chiusano, F. (2021). Two minutes NLP – 11 word embeddings models you should know. https://medium.com

TF-IDF for Document Ranking from scratch in python on real world dataset. https://towardsdatascience.com

Pekhimenko, G. (2006). Penalizied logistic regression for classification. Dept. Comput. Sci.., Univ. Toronto, Toronto, ON M5S3L1.

Joachims, T. (1998, April). Text categorization with support vector machines: Learning with many relevant features. In European conference on Machine Learning (pp. 137-142). Springer, Berlin, Heidelberg.

Text Classification Examples & How to Put Them to Work, https://monkeylearn.com/text-classification-examples/

BERT Explained: A Complete Guide with Theory and Tutorial, https://towardsml.wordpress.com/

Khedkar, Sujata, et al. "Explainable AI in healthcare." Healthcare (April 8, 2019). 2nd International Conference on Advances in Science & Technology (ICAST). 2019.

Krauss, R. M. (2002). The psychology of verbal communication. International Encyclopaedia of the Social and Behavioral Sciences. London: Elsevier, 16161-16165.

Gupta, D. K., & Goyal, S. (2018). Email classification into relevant category using neural networks. arXiv preprint arXiv:1802.03971.

Mozetič, I., Grčar, M., & Smailović, J. (2016). Multilingual Twitter sentiment classification: The role of human annotators. PloS one, 11(5), e0155036.

Bacchelli, A., Dal Sasso, T., D'Ambros, M., & Lanza, M. (2012, June). Content classification of development emails. In 2012 34th International Conference on Software Engineering (ICSE) (pp. 375-385). IEEE.

Alshehri, Y. A. (2020, March). Text mining for incoming tasks based on the urgency/importance factors and task classification using Machine Learning tools. In Proceedings of the 2020 the 4th International Conference on Compute and Data Analysis (pp. 183-189).

Bacchelli, A., Dal Sasso, T., D'Ambros, M., & Lanza, M. (2012, June). Content classification of development emails. In 2012 34th International Conference on Software Engineering (ICSE) (pp. 375-385). IEEE.

Oni, A. C., Ogude, U. C.,& Uwadia, C. O. Email Urgency Classifier Using Natural Language Processing and Naïve Bayes. Computer Networks, Infrastructure Management And Security (CoNIMS), 85.

Trabelsi, A., Elouedi, Z., & Lefevre, E. (2017, July). Ensemble enhanced evidential k-NN classifier through random subspaces. In European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty (pp. 212-221). Springer, Cham.

Vaishya, R., Javaid, M., Khan, I. H., & Haleem, A. (2020). AI (AI) applications for COVID-19 pandemic. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14(4), 337-339.

Choudhari, S., Choudhary, N., Kaware, S., & Shaikh, A. (2020). Email Prioritization Using Machine Learning. Available at SSRN 3568518.

Veiga, R. K., Veloso, A. C., Melo, A. P., & LamBERTs, R. (2021). Application of Machine Learning to estimate building energy use intensities. Energy and Buildings, 249, 111219.

Sevgican, S., Turan, M., Gökarslan, K., Yilmaz, H. B., & Tugcu, T. (2020). Intelligent network data analytics function in 5G cellular networks using Machine Learning. Journal of Communications and Networks, 22(3), 269-280.

Mnih, V. (2013). Machine Learning for aerial image labeling. University of Toronto (Canada).

Dayhoff, J. E. (1990). Neural network architectures: an introduction. Van Nostrand Reinhold Co..

Li, C., Zhan, G., & Li, Z. (2018, October). News text classification based on improved Bi-LSTM-CNN. In 2018 9th International conference on information technology in medicine and education (ITME) (pp. 890-893). IEEE.

Kamiyama, H., Ando, A., Masumura, R., Kobashikawa, S., & Aono, Y. (2019, November). Urgent Voicemail Detection Focused on Long-term Temporal Variation. In 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (pp. 917-921). IEEE.

Tripepi, G., Jager, K. J., Dekker, F. W., & Zoccali, C. (2008). Linear and logistic regression analysis. Kidney international, 73(7), 806-810.

Asaei, A., Bourlard, H., & Picart, B. (2010). Investigation of kNN classifier on posterior features towards application in automatic speech recognition (No. REP_WORK). Idiap.

Dopico, M., Gómez, A., De la Fuente, D., García, N., Rosillo, R., & Puche, J. (2016). A vision of industry 4.0 from an AI point of view. In Proceedings on the international conference on AI (ICAI) (p. 407). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

Cheng, C. T., Chen, C. C., Fu, C. Y., Chaou, C. H., Wu, Y. T., Hsu, C. P., ... & Liao, C. H. (2020). AI-based education assists medical students' interpretation of hip fracture. Insights into Imaging, 11(1), 1-8.

Cayamcela, M. E. M., & Lim, W. (2018, October). AI in 5G technology: A survey. In 2018 International Conference on Information and Communication Technology Convergence (ICTC) (pp. 860-865). IEEE.

Skowron A. The rough sets theory and evidence theory. Fundamenta Informaticae. 1990 Jan 1;13(3):245-62.

Mujtaba, G., Shuib, L., Idris, N., Hoo, W. L., Raj, R. G., Khowaja, K., ... & Nweke, H. F. (2019). Clinical text classification research trends: Systematic literature review and open issues. Expert systems with applications, 116, 494-520.

Al-Hawari, F., & Barham, H. (2021). A Machine Learning based help desk system for IT service management. Journal of King Saud University-Computer and Information Sciences, 33(6), 702-718.

Kadiyala, S. P., & Woo, W. L. (2022). Flood Prediction and Analysis on the Relevance of Features using Explainable AI. arXiv preprint arXiv:2201.05046.

Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised Machine Learning algorithms: classification and comparison. International Journal of Computer Trends and Technology (IJCTT), 48(3), 128-138.

Deng, A., & Shrestha, E. BERT-based Transfer Learning with Synonym Augmentation for Question Answering.

Kang, Y., Cai, Z., Tan, C. W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. Journal of Management Analytics, 7(2), 139-172.

Sgouropoulou, C., & Voyiatzis, I. (2021, July). XGBoost and Deep Neural Network Comparison: The Case of Teams' Performance. In Intelligent Tutoring Systems: 17th International Conference, ITS 2021, Virtual Event, June 7–11, 2021, Proceedings (Vol. 12677, p. 343). Springer Nature

Li, Mujin, Honghui Xu, and Yong Deng. "Evidential decision tree based on belief entropy." Entropy 21.9 (2019): 897.

Jiao, Lianmeng, Thierry Denœux, and Quan Pan. "Evidential editing k-nearest neighbor classifier." European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty. Springer, Cham, 2015.