Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 5 Iss. 4 October 2025, pp: 1287-1293
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

From Comments to Insight: Predictive Classification of Organizational Cultural Entropy Using SBERT, K-Means, and Logistic Regression

Sentri Indah Mayasari[1*], Imam Yuadi[2]

[1]Master of Human Resource Development, Graduate School, Airlangga University, Indonesia
[2]Department of Information and Library Science, Faculty of Social and Political Sciences, Airlangga University, Indonesia

E-Mail: [1]sentri.indah.mayasari-2024@pasca.unair.ac.id, [2]imam.yuadi@fisip.unair.ac.id

Received Jul 06th 2025; Revised Sep 18th 2025; Accepted Sep 22th 2025; Available Online Oct 30th 2025
Corresponding Author: Sentri Indah Mayasari

Abstract

This study aims to develop a machine learning-based predictive model based on clustered data to identify cultural entropy in organizations through the analysis of open-ended comments on employee perception surveys of superiors. energy used for unproductive activities in a work environment. Entropy shows the level of conflict, friction and frustration in the environment. With a text mining approach, answers to open-ended questions in the cultural entropy survey were processed with Sentence-BERT and clustered using the K-Means algorithm into two categories, namely cultural entropy and non-cultural entropy. The dataset that already has labels from the clustering results is used to develop a classification model. The algorithms used are Random Forest, Logistic Regression, and Support Vector Machine (SVM), which are evaluated through accuracy, precision, recall, and F1-score metrics and a confusion matrix. The results show that Logistic Regression provides the best performance with an accuracy of 0.985, a precision of 1.00, and an F1-score of 0.978 without any classification errors. These findings indicate that the clustering approach followed by machine learning-based predictive is effective in identifying organizational cultural entropy. This can be used to design appropriate interventions and as an early detection system for cultural entropy in human resource management.

Keyword: Cultural Entropy, Linear Regression, Machine Learning, SBERT, Text Mining.

1. INTRODUCTION

One of the important indicators that influences the effectiveness and sustainability of an organization is the aspect of organizational cultural health, where strong organizational cultural management plays an important role in creating innovation and driving the long-term sustainability of the company [1]. The parameter used to measure corporate culture is cultural entropy, which is the energy used for unproductive activities in a work environment. Entropy shows the level of conflict, friction and frustration in the environment [2][3]. In addition, the excessive bureaucracy, unclear communication, and low employee engagement are the main causes of cultural entropy [4]. Cultural entropy has an impact on decreasing employee motivation and job dissatisfaction, especially if there is a negative perception of superiors who are not in line with organizational values [5]. Another important factor that influences the increase or decrease in cultural entropy is the behavior of the leader [3]. Therefore, employee's perceptions of leader, especially on open-ended questions, need to be analyzed so that they can be utilized systematically. Research has found that excessive bureaucracy, unclear communication, and low employee engagement are the main causes of cultural entropy.

In this study, several machine learning algorithms were used to classify text data into relevant categories. Random Forest is an ensemble learning method that builds multiple decision trees during the training process and produces a final class based on the mode of all trees for classification tasks [6]. This algorithm is resistant to overfitting and is able to handle high-dimensional data effectively. Logistic Regression is a statistical model used for binary and multi-class classification, by estimating the probability of an outcome based on one or more independent variables through a logistic function [7]. Support Vector Machine (SVM) is a supervised learning algorithm that attempts to find the optimal hyperplane to separate data into classes with maximum margin and has shown good performance in high-dimensional spaces [8]. These algorithms are widely used in text classification tasks due to their effectiveness in handling both structured and unstructured data.

In some previous studies, the entropy of organisational culture has been measured from perspective of open and dissipative systems through agent and mathematical simulation approaches [9], [10]. However, these studies only focus on discrete values, not on an open narrative perspective. The relationship between organizational culture, innovation, and sustainability, concluding that organizational culture is an active resource that plays an important role in managing a company's innovative and sustainable development. But this study did not discuss qualitative data from open-ended comments [11]. Through text mining and sentiment analysis approaches, perceptions can be explored in the narrative of open comments submitted by employees [12]. In addition, there was a previous study conducted by T. Şimşek and A.B. Şimşek in 2025 which examined open-ended comments, but it focused more on measuring employees' emotions, satisfaction, and engagement in general, and has not specifically investigated cultural entropy using the Sentence-BERT (SBERT) approach combined with clustering and supervised learning models as carried out in this study [13]

In contrast to some previous studies that used a score-based approach, this study developed an analytical narrative approach to explore hidden sentiments that could have an effect on cultural entropy. This research aims to analyse employees' perceptions of their superiors through open-ended questions in a cultural entropy survey using text mining and sentiment analysis approaches. The analysis model developed will cluster comments into positive sentiment or entropy and negative sentiment or not entropy. Then the clustering results are developed with a supervised learning model to predict employee perceptions in the cultural entropy or non-entropy category.

2. MATERIALS AND METHOD

The stages in this research are (1) Business Understanding (2) Data Understanding; (3) Data Preparation (4) Modeling (5) Model Evaluation as shown in the Figure 1.

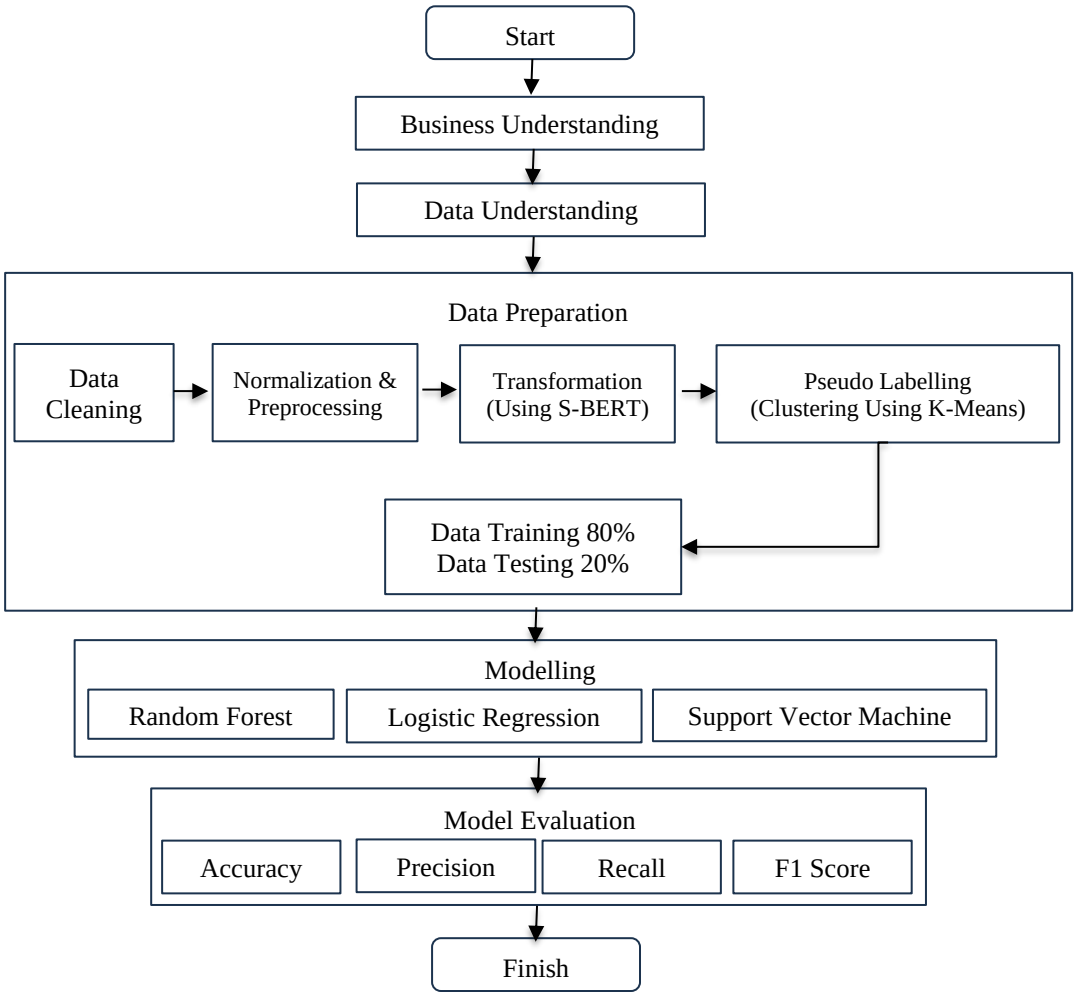

Figure 1. Methodology Flow Chart

2.1. Business Understanding

This stage aims to gain a deep understanding of the business problems to be solved through the data mining process [14]. In this study, the problem that occurs in the company is that there is still cultural entropy that can have an impact on employee productivity, one of which is related to employee perceptions of superiors.

So it is necessary to do a thorough analysis, including the results of the open questions of the cultural entropy survey. Open-ended survey questions can enable companies to gain insights beyond more commonly used closed-ended question formats by allowing respondents an opportunity to provide information with few constraints and in their own words [15]. Thus, it is necessary to identify clusters of positive sentiment or non-entropy and negative sentiment or cultural entropy using a text mining approach. The results of the clustering are continued in the prediction analysis so that it is more efficient and effective in analysing new comments. Thus, the model can be integrated in HR surveys or dashboards as an early detection tool for dysfunctional work culture.

2.2. Data Understanding

The Data Understanding stage aims to identify the quality and patterns in the data, so as to guide the selection of appropriate modelling techniques [16].The data in this study were obtained from answers to one of the open questions in the 2024 cultural entropy survey specifically related to employee perceptions of their superiors' leadership in a company in the energy sector in Indonesia. There are 338 comments in the form of free text in the form of sentences or short paragraphs in Indonesian and do not have an initial label.

2.3. Data Preparation

The data preparation stage aims to convert raw data into a clean format that can be used by analytical models, because good data quality directly affects the performance of machine learning models [16]. The steps taken at this stage are:

1. Initial data cleaning which is removing empty rows and ensuring comments are in text format.
2. Normalization and preprocessing by converting all text to lower case, removing special characters such as numbers, punctuation marks, and excess white space.
3. Transformation of text into vector representation by using the sentence BERT (SBERT) paraphrase-multilingual-MiniLM-L12-v2 model to convert each comment into a 384-dimensional vector. This embedding can capture the semantic meaning of each comment. SBERT is a modification of the Bidirectional Encoder Representations from Transformers (BERT) architecture designed to produce a fixed-dimensional and semantically meaningful sentence representation, thus enabling efficient calculation of similarity between sentences through methods such as cosine similarity [17] .
4. Pseudo labeling clustering is done by applying the K-Means algorithm (n=2) to the embedding results to divide the comments into two clusters, namely cultural entropy and non-cultural entropy. Pseudo labeling is a technique in semi-supervised learning that automatically labels unlabeled data using the results of unsupervised models or methods such as clustering (for example K-Means). These pseudo labels are then used as ground truth to train supervised models. The combination of K-Means and pseudo labeling is widely used in cases where labeled data is very limited, and this approach has been shown to improve model performance in classification PS I [15], [18]. This technique helps to automatically expand the training dataset and improve model performance without the need for ineffective and inefficient manual annotation [18], [19]
5. Supervision label creation is based on the clustered dataset and divides it into 80% training data and 20% testing data.

2.4. Modelling

This stage aims to select, train, and optimize machine learning algorithms according to the characteristics and objectives of the analysis by utilizing data that has been prepared in the previous stage to build an accurate predictive model [20]. Based on the clustering data, a prediction analysis is then carried out using supervised learning algorithms:

1. Random Forest
   Random Forest is an ensemble learning algorithm that constructs multiple decision trees during training and combines their results through majority voting for classification or averaging for regression. This approach improves predictive accuracy while reducing the risk of overfitting [21]. The algorithm is chosen because it can effectively handle high-dimensional data, is robust to noise, and provides stable performance across different types of datasets.

2. Logistic Regression
   Logistic Regression is a statistical method commonly used for binary and multi-class classification by estimating the probability of a categorical outcome based on one or more independent variables. It applies the logistic (sigmoid) function to map input values into probabilities between 0 and 1, making it particularly effective for predicting dichotomous outcomes [7]. This algorithm is chosen because of its simplicity, interpretability, and strong performance as a baseline model in classification tasks.

3. Support Vector Machine
   SVM is a supervised learning algorithm that aims to find the optimal hyperplane which maximally separates data into different classes. By focusing on support vectors, the algorithm ensures that the decision boundary is robust and generalizes well to unseen data, especially in high-dimensional spaces [22]. This algorithm is chosen because of its effectiveness in handling complex classification problems, strong theoretical foundation, and high accuracy in text classification tasks.

In this stage, the three algorithms are tested and tuned using GridSearchCV to find the best combination of hyperparameters.

2.5. Model evaluation

This stage aims to objectively measure the performance of the predictive model using appropriate metrics, so that the model with the best generalization to new data can be selected [20]. Evaluation is done to assess the extent to which the model can correctly classify comments. The data processing and evaluation of the classification models were carried out using the Python programming language with the scikit-learn library. The evaluation matrix used consists of accuracy, precision, recall, and F1 score, with the results presented in Table 1.

Table 1. Evaluation Matrix

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0,98529	1,00000	0,95833	0,97872
SVM	0,97059	0,92308	1,00000	0,96000
Random Forest	0,95588	0,95652	0,91667	0,93617

Based on Table 1, Logistic Regression provides the highest value in accuracy of 0.985, precision 1.00, and F1 Score 0.978. The next best performing algorithm is SVM with an F1 Score of 0.960, followed by Random Forest with an F1 Score of 0.936. Based on these results, Logistic Regression can be concluded as the best model in classifying comments into entropy and non-entropy in this study.

In addition, a comparison of the model's positive and negative predictions with the actual labels using a confusion matrix is carried out to show in detail false positives and negatives as well as true positives and negatives, as shown in Figure 2.

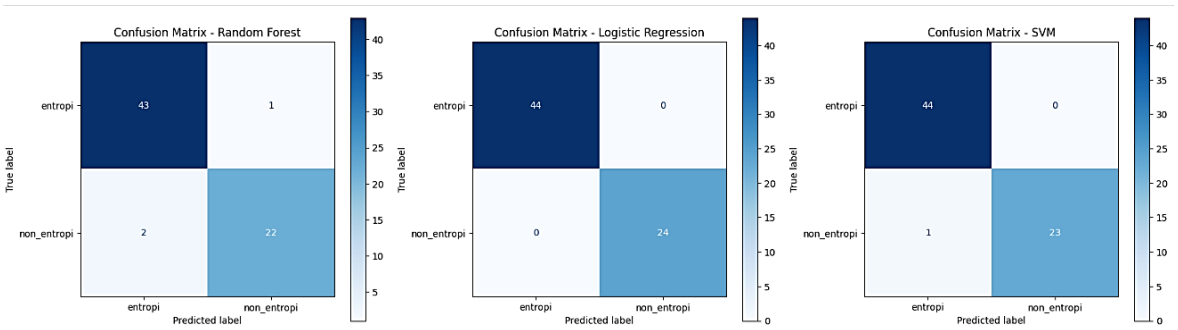

Figure 2. Confusion Matrix

Based on Figure 2, it can be seen that the Logistic Regression algorithm shows the most optimal performance by producing no classification errors (false positive and false negative values = 0), while SVM produces one error in non-entropy class prediction, and Random Forest produces one error in entropy and 2 errors in non-entropy.

Based on these two evaluation matrices, Logistic Regression is the most balanced and accurate model in mapping employee comments into the two categories. The model with the Logistic Regression algorithm is equipped with an automatic prediction function for new data.

3. RESULTS AND DISCUSSION

In line with the research methodology, the analysis was conducted through the stages of processing cultural entropy survey data, specifically from open-ended questions regarding employee perceptions of superior leadership. Text data was processed through cleaning, tokenization, and lemmatization, then converted into numerical representations using SBERT. The resulting representations were analyzed using the K-Means clustering algorithm to identify patterns and grouped into entropy and non-entropy categories as a basis for further classification. The K-Means clustering algorithm was applied after the comments were transformed into 768-dimensional vectors using SBERT, resulting in a matrix with the shape of [n×768]. The algorithm

grouped the embeddings into two clusters, representing entropy and non-entropy perceptions. For visualization, the dimensions were reduced using PCA into an [n×2] matrix, which clearly showed the separation between the clusters. This clustering stage also generated labels that were later used for supervised classification.

The data from the cultural entropy survey, especially the open-ended questions related to employee perceptions of supervisor leadership, were analyzed by clustering to obtain labels. The clustering process was carried out using the Python programming language with the scikit-learn library, and the results of the clustering are shown in Figure 3.

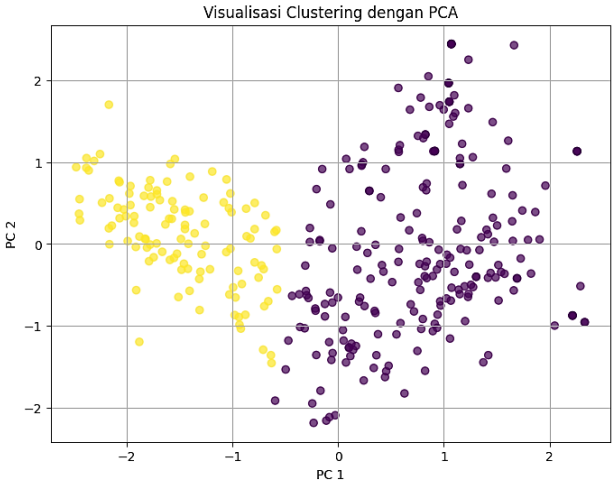

Figure 3. Clustering Result

Figure 3 shows the visualization of the clustering results of employee comments using the K-Means method, visualized with the PCA (Principal Component Analysis) dimension reduction technique. Each point represents one comment that has been converted into a vector through SBERT embedding, then projected into two main dimensions, namely entropy and non-entropy. Two different colors mark the two main clusters formed, with distributions that are quite separate from each other. This visualization reinforces the validity of the unsupervised approach used before proceeding to the supervised classification stage.

Furthermore, the clustering results are used as labels to build the supervised classification model. The three algorithms applied, namely Logistic Regression, SVM, and Random Forest, showed excellent classification performance. Based on the metric evaluation, the Logistic Regression model obtained the highest performance with F1 Score of 0.978 with no misclassification, followed by SVM (F1 Score: 0.960) and Random Forest (F1 Score: 0.936). The results of the confusion matrix also strengthen that the three models are able to classify comments that fall into cultural entropy and non-cultural entropy with high accuracy.

The classification results show that the algorithm can detect comments that show indicators of organizational cultural entropy, such as internal conflict, perceptions of leader inconsistency or excessive dominance. This strengthens the understanding that machine learning can be used as a tool to automatically detect and predict organizational cultural entropy by analyzing open-ended comment data. The clustering approach followed by classification is an efficient method in structuring qualitative text data into information that can be processed for strategic decision making, such as increasing leadership effectiveness, openness of communication, and harmonization of organizational values. In addition, the application of predictive analysis as an automatic detection of cultural entropy based on machine learning, as developed in this study, allows organizations to monitor cultural dynamics in real time and develop preventive intervention strategies [23]. By conducting narrative analysis in employee cultural entropy surveys, companies can explore issues that are not captured in quantitative data. Thus, this analysis is able to design more responsive and contextual work culture policies [24]. These findings not only provide methodological contributions in the application of text mining for HR analytics but also expand the use of predictive technologies in diagnosing cultural entropy risks earlier and systematically.

4. CONCLUSION

This study successfully developed a text mining-based machine learning pipeline to detect comments from cultural entropy survey data, especially perceptions of leader, and developed a prediction model based on the clustered dataset. The clustering process into 2 was carried out using SBERT embedding and the K-means algorithm with a fairly clear semantic separation. The resulting clusters, namely cultural entropy and non-cultural entropy, were then used as labels for training the classification model using three algorithms: Logistic Regression, SVM, and Random Forest. Based on the evaluation matrix, it was found that the Logistic

Regression model provided the best performance, with an F1 Score of 0.978 without misclassification on the test data. This approach has proven effective in converting unstructured sentence data into information that can be processed, interpreted, and used to make predictions systematically. Thus, organizations can monitor and evaluate work culture conditions more objectively and make appropriate interventions based on data.

REFERENCES

[1] I. Gorzen-Mitka, "A Green Approach on Risk Management: Exploring Constructs in a Concept Mapping Framework," *Scientific Papers of Silesian University of Technology. Organization and Management Series*, vol. 2024, no. 213, pp. 101–123, 2024, doi: 10.29119/1641-3466.2024.213.8.

[2] R. Barrett, *The Values-Driven Organization: Cultural Health and Employee Well-Being as a Pathway to Sustainable Performance*, Second. New York: Routledge, 2017.

[3] A. Rahman, F. Naufal, and S. G. Partiwi, "Measuring the entropy of organizational culture using agent-based simulation," in *Managing Learning Organization in Industry 4.0*, Routledge, 2020, pp. 109–115. doi: 10.1201/9781003010814-19.

[4] A. M. Rani and S. H. Senen, "Navigating Cultural Entropy: Leadership Strategies and Organizational Dynamics in Higher Education Institutions," *West Science Journal Economic and Entrepreneurship*, vol. 1, no. 12, pp. 511–520, Dec. 2023, doi: 10.58812/wsjee.v1i12.473.

[5] U. Udin, R. D. Dharma, R. Dananjoyo, and M. Shaikh, "The Role of Transformational Leadership on Employee Performance Through Organizational Learning Culture and Intrinsic Work Motivation," *International Journal of Sustainable Development and Planning*, vol. 18, no. 1, pp. 237–246, Jan. 2023, doi: 10.18280/ijsdp.180125.

[6] L. Breiman, "Random Forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[7] D. W. . Hosmer, Stanley. Lemeshow, and R. X. . Sturdivant, *Applied logistic regression*, 3rd ed. Wiley, 2013.

[8] C. Cortes, V. Vapnik, and L. Saitta, "Support-Vector Networks Editor," Kluwer Academic Publishers, 1995.

[9] B. Alla, B. Natalia, B. Sergey, and O. Svitlana, "Modelling of Creation Organisational Energy-Entropy," in *International Scientific and Technical Conference on Computer Sciences and Information Technologies*, Institute of Electrical and Electronics Engineers Inc., Sep. 2020, pp. 141–145. doi: 10.1109/CSIT49958.2020.9321997.

[10] A. Rahman, S. G. Partiwi, and R. S. Dewi, "Continuous Vector-Based Entropy Measurement on the Organizational Culture Evaluation," *SHS Web of Conferences*, vol. 189, p. 01011, 2024, doi: 10.1051/shsconf/202418901011.

[11] Z. Mingaleva, E. Shironina, E. Lobova, V. Olenev, L. Plyusnina, and A. Oborina, "Organizational Culture Management as an Element of Innovative and Sustainable Development of Enterprises," *Sustainability (Switzerland)*, vol. 14, no. 10, May 2022, doi: 10.3390/su14106289.

[12] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New Avenues in Opinion Mining and Sentiment Analysis," *IEEE Intell Syst*, vol. 28, no. 2, pp. 15–21, Mar. 2013, doi: 10.1109/MIS.2013.30.

[13] T. Şimşek and A. B. Şimşek, "Sentiment Analysis in Employee Experience Using Natural Language Processing and Machine Learning," 2025, pp. 309–346. doi: 10.4018/979-8-3693-7848-9.ch012.

[14] D. S. Putler and R. E. Krider, "A Process Model for Data Mining—CRISP-DM," in *Customer and Business Analytics: Applied Data Mining for Business Decision Making Using R*, 1st ed., Boca Raton, FL, USA: Chapman and Hall/CRC, 2012. doi: 10.1201/b12040-8.

[15] K. Cibelli Hibben, Z. Smith, B. Rogers, V. Ryan, P. Scanlon, and T. Hoppe, "Semi-Automated Nonresponse Detection for Open-Text Survey Data," *Soc Sci Comput Rev*, vol. 43, no. 1, pp. 166–190, Feb. 2025, doi: 10.1177/08944393241249720.

[16] P. Chapman *et al.*, *CRISP-DM 1.0: Step-by-Step Data Mining Guide*. Chicago: The CRISP-DM Consortium / SPSS Inc., 2000.

[17] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 3980–3990. doi: 10.18653/v1/D19-1410.

[18] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning," in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, Jul. 2020, pp. 1–8. doi: 10.1109/IJCNN48605.2020.9207304.

[19] W. Yang, R. Zhang, J. Chen, L. Wang, and J. Kim, "Prototype-Guided Pseudo Labeling for Semi-Supervised Text Classification," Long Papers. doi: 10.18653/v1/2023.acl-long.904.

[20] Jiawei. Han, Micheline. Kamber, and Jian. Pei, *Data mining : concepts and techniques*, 3rd ed. Elsevier/Morgan Kaufmann, 2012.

[21] L. Breiman, "Random Forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
[22] C. Cortes and V. Vapnik, "Support-vector networks," *Mach Learn*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.
[23] MIT Sloan Management Review, *Lessons from Becoming a Data-Driven Organization*. EY – Ernst & Young, 2016.
[24] E. Brynjolfsson and K. McElheran, "The rapid adoption of data-driven decision-making," in *American Economic Review*, American Economic Association, May 2016, pp. 133–139. doi: 10.1257/aer.p20161016.