








# Comparison Between SVM and DistilBERT for Multi-label Text Classification of Scientific Papers Aligned with Sustainable Development Goals

Roberto Carlos Morales-Hernández<sup>1</sup> , David Becerra-Alonso<sup>2</sup> ,  
Eduardo Romero Vivas<sup>1</sup> , and Joaquín Gutiérrez<sup>1</sup>  

<sup>1</sup> Centro de Investigaciones Biológicas del Noroeste, S.C., Av. Instituto Politécnico Nacional 195, Playa Palo de Santa Rita, 23096 La Paz, B.C.S, México  
joaquin04@cibnor.mx

<sup>2</sup> Department of Quantitative Methods, Universidad Loyola Andalucía, 41704 Seville, Spain

**Abstract.** The scientific articles identification with the 17 sustainable development goals of the UN 2030 Agenda is a valuable task for research and educational institutions. Finding an efficient and practical multi-label classification model using machine or deep learning remains relevant. This work refers to the performance comparison of a text classification model that combines Label Powerset (LP) and Support Vector Machine (SVM) against a transfer learning language model such as DistilBERT in 5 different imbalanced and balanced dataset scenarios of scientific papers. A proposed classification process was implemented with performance metrics, which have confirmed that the combination LP-SVM continues to be an option with remarkable results in multi-label text classification.

**Keywords:** Multi-label text classification · Label powerset · Support vector machine · Transfer learning · DistilBERT · Sustainable development goals

## 1 Introduction

For research centers and universities, identifying their scientific production with sustainable goals or policies becomes crucial to assess their contribution and influence. In this context, Natural Language Processing (NLP) through Machine or Deep Learning enables large-scale data handling for text classification. Text classification is a technique of text analysis to categorize data into different types, forms, or any other distinct predefined class [1]. According the number of classes, classification problems can be grouped in three types: Binary, Multi-class, and Multi-label.

In supervised learning, Multi-label Text Classification (MLTC) refers to models that learn from training data, to classify new instances by assigning a correct class label to each of them [2]. Binary classification algorithms, such as Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF) or Logistic Regression (LR), need methods to transform the multi-label instances into a set of binary or multi-class datasets [3].

Problem transformation methods like Binary Relevance (BR), Label Powerset (LP), or Classifier Chains (CC) remain convenient to this day to help binary algorithms for MLTC [4–8].

While in machine learning a combination of problem transformation method with a classification algorithm is a traditional model, in deep learning, transfer learning models such as BERT are pre-trained methods with the state-of-art performance in classification [9]. However, this pre-trained model could have the problem of consuming high computational resources, making it difficult to adopt. Nevertheless, more and more methods develop to create models to consume less resources based on BERT (DistilBERT [10], DocBERT [11], LegalDB [12], or TinyBERT [13]).

This paper aims to evaluate two MLTC models. One model, more traditional by combining LP with SVM, and the second one, by implementing a light and small pre-trained model, DistilBERT [10]. The database is a collection of scientific articles from the domain of knowledge in Organic Agriculture 3.0 aligned to the 17 Sustainable Development Goals (SDG) of the United Nations 2030 Agenda [14]. Likewise, the performance of the classification models is tested by proposing five scenarios with different balances and imbalances of the dataset.

The contributions of this project are as follows:

- Dataset creation with 31,434 scientific papers from year 2018 with title and abstract from organic agriculture 3.0 domain, labeled with the 17 SDG classification.
- LP-SVM results a competitive traditional model under the five dataset scenarios with balanced and imbalanced SDG labels.
- DistilBERT with a *minimum* configuration, evaluated under the five dataset scenarios of scientific papers with sustainable developments labels.

This research quantifies the performance of two MLTC models, comparing the LP-SVM and DistilBERT models to classify scientific articles (title and abstract) under the domain of organic agriculture.

## 2 Related Work

MLTC have different ways of being implemented. This section shows two widely used ways, a traditional machine learning combining a problem transformation method with a single-label classification algorithm, and a transfer learning model with a pre-trained method: DistilBERT.

### 2.1 Problem Transformation Method and Classification Algorithm

Unlike binary or multiclass text classification, MLTC presents more challenges because each text document can have multiple labels. It can find solutions through so-called multi-label learning methods such as Problem transformation, Problem Adaptation, and Ensemble [15, 16]. Problem transformation techniques change to one (or more binary) or multi-class datasets to be managed by single-label or multi-class classification algorithms [17].

LP is a problem transformation method. LP [18] transforms multi-labels from each instance into one single-label. This approach converts the multi-label problem in a multi-class classification. With this transformation, a single-label classifier such as SVM can perform the needed classification. LP has the advantage of taking label correlations into account albeit increasing the number of label classes, where most of them represent few or very few instances.

SVM is a linear classification model that maximizes the margin between data instances and a hyperplane, acting as a division boundary [19]. Some studies maintain the experimentation and performance evaluation with acceptable results of SVM as a multi-label classifier [20–22].

## 2.2 Transfer Learning Model

For NLP, another relevant area of study with an influence and paradigm change has been transfer learning where different types of word embeddings and pre-trained language models are proposed. Transfer learning refers more specifically to pre-trained language representations [1]. NLP has two types of pre-trained languages representations: feature-based and fine-tuning models. The first are often used to initialize the first layer of a neural network, the latter are fine-tuned as an entire model for a specific downstream task [9].

To create a lighter version of BERT, in DistilBERT the token-type embeddings and the pooler are removed from the architecture, thus reducing the number of layers by a factor of 2 [23]. Knowledge distillation is a compression technique in which a compact model - the student - is trained to reproduce the behavior of a larger model - the teacher - or an ensemble of models [10].

Comparison studies between SVM and a pre-trained model for classification continue to be carried out, which is why SVM performance evaluations remain current in a wide variety of scenarios, beyond the performance level that pre-trained models generally present [21, 24–27].

## 3 Methodology

### 3.1 Framework for Multi-label Text Classification

The proposed framework to classify scientific articles into 17 SDG multi-label data classes is described in Fig. 1. It shows a typical MLTC pipeline to apply classification methods in four condensed phases: Information Retrieval and Dataset Creation, Data Analysis and Preprocessing, Model Building, and Model Evaluation.

**Information Retrieval and Dataset Creation.** A first phase where data collection is attained from bibliographic resources. The dataset used here was obtained from Dimensions, a bibliographic database produced by the company Digital Science, who offers a feasible categorization scheme in scientific papers for the seventeen SDG [28]. Scientific articles had three features: Title, Abstract and labels of the 17 SDGs of the UN 2030 Agenda (Table 1). For this study, year 2018 was selected to create dataset from the dominion organic agriculture 3.0.

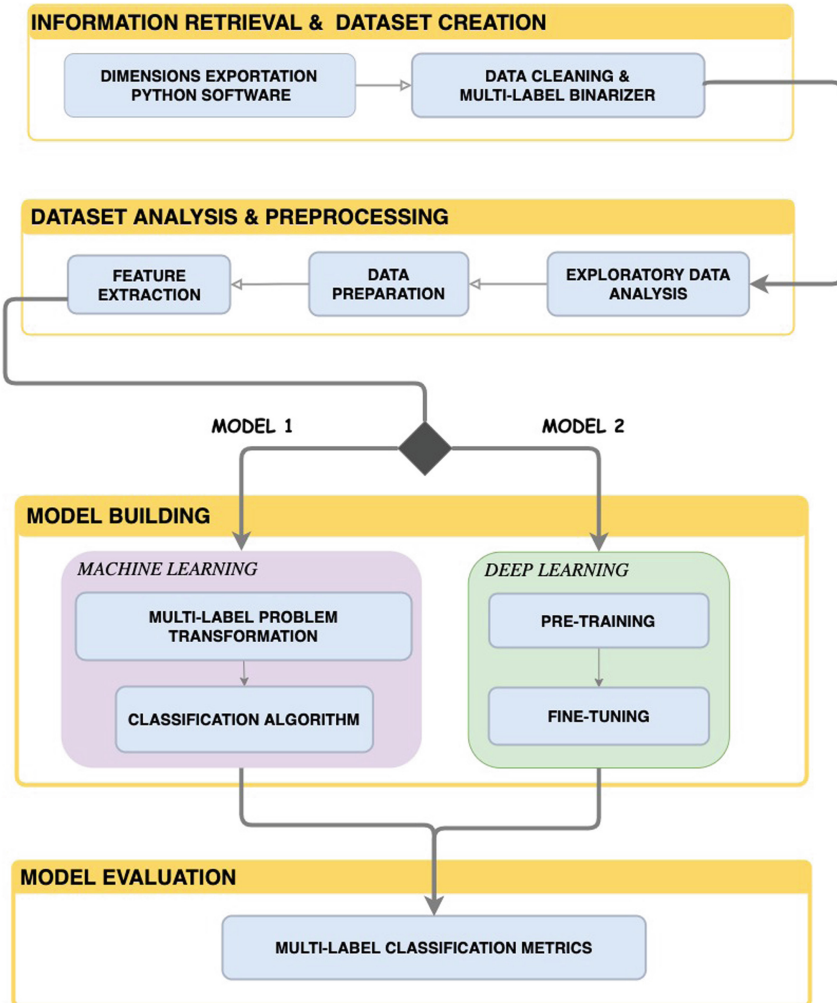


Fig. 1. Framework for the comparison experiment of text classification models

**Data Analysis and Preprocessing.** In this second phase, a relevant dataset preprocessing involves the extraction of text to create clean word sequences. Also, in this stage, datasets are created and adjusted in label distribution for classification model evaluation according to experimental requirements. Stop-word elimination, stemming and tokenization are pre-processing tools. Term Frequency-Inverse Document Frequency (TF-IDF) is used as a feature extraction method. TF-IDF can describe how important the word is in a text and is applied as a weighting factor in text mining [29].

**Table 1.** Sustainable development goals from United Nations Agenda 2030.

Sustainable Development Goals (SDG) from UN Agenda 2030		
<b>SDG1</b> No Poverty	<b>SDG7</b> Affordable and Clean Energy	<b>SDG13</b> Climate Action
<b>SDG2</b> Zero Hunger	<b>SDG8</b> Decent Work and Economic Growth	<b>SDG14</b> Life Below Water
<b>SDG3</b> Good Health and Well-being	<b>SDG9</b> Industry, Innovation, and Infrastructure	<b>SDG15</b> Life on Land
<b>SDG4</b> Quality Education	<b>SDG10</b> Reducing Inequality	<b>SDG16</b> Peace, Justice, and Strong Institutions
<b>SDG5</b> Gender Equality	<b>SDG11</b> Sustainable Cities and Communities	<b>SDG17</b> Partnerships for the Goals
<b>SDG6</b> Clean Water and Sanitation	<b>SDG12</b> Responsible Consumption and Production	

**Model Building.** This is a stage where classification models are established, configured, and run. According to Fig. 1, Model 1 is constructed with LP as the problem transformation method to convert the multi-label to multi-class classification from SVMV as classification algorithm. Scikit-multilearn is a multi-label classification software module that builds on top of the scikit-learn python framework with transformation methods such as LP. SVM algorithm is implemented through scikit-learn, a tool for predictive data analysis.

In Model 2, the pre-training phase refers to distilbert-base-uncased model as a distilled version of the BERT [9] base model to tokenize the data (distilbert-base-uncased has 66 million parameters against 110 million for BERT-base). For the model construction, maximum length Bert tokenizer, learning rate, batch size, epochs, its loss, and an optimizer are parameters to be defined in this step. For this project, in the fine-tuning stage, a training function is defined to train the neural network on the training dataset via pytorch. Parameters are default values, and none of their respective hyperparameters are optimized in both models. This criterion enables a fair comparison among the methods.

**Model Evaluation.** Three multi-label classification metrics are selected to evaluate the experiment multi-label classification models: Accuracy, F1-Score (micro), and Hamming loss. Accuracy is defined as the ratio of observations predicted correctly to the total number of observations. Hamming loss refers to an average binary classification error [30] represented by the fraction of labels that are incorrectly predicted. F1-Score (micro) is the harmonic mean (weighted) of Recall (the ratio of true positives to the sum of true positives and false negatives across all labels) and Precision (refers to the percentage of predicted labels that are relevant) [30].

## 4 Model Experiments

This section presents several experiments with different dataset scenarios to find performance for the two multi-label text classification models, according to Fig. 1.

### 4.1 Dataset

For this experiment, dataset creation was produced from Dimensions with organic agriculture 3.0 as a knowledge domain from 2018. Total instances collected with SDG labels: 31,434. This study proposed five different dataset scenarios described in Table 2. These dataset scenarios let discover the performance for both proposed MLTC models. For the SC2, SC3, and SC4 scenarios, six SDG tags were discarded for having less than 1,000 instances and being considered noisy tags (SDG 1, 5, 8, 9, 10, and 17). Scenario 4 (SC4) involved creating, in turn, 11 datasets. In each one, a label has the number of instances equal to the sum of the remaining 10 labels. In SC4, classification models are applied to each dataset and the average is the result presented.

**Table 2.** 2018 Dataset scenarios for classification models performance evaluation.

Five Dataset Scenarios	Instances		
	Total	Train 66%	Test 33%
<b>SC1.</b> Imbalanced with all 17 SDG labels	31,434	20,745	10,687
<b>SC2.</b> Imbalanced with 11 SDG labels greater than 1,000 examples	30,480	20,098	10,353
<b>SC3.</b> Balanced with equal number of instances in 11 SDG labels	13,623	7,791	4,014
<b>SC4.</b> <i>Extreme</i> imbalanced (10 to 1) from one label vs other 10 labels	5,310*	3,540	1,770
<b>SC5.</b> Instances with only one SDG label (multi-class)	27,400	18,084	9,316
*Average			

### 4.2 Data Preprocessing

The dataset undergoes feature selection implemented with libraries, such as: Re (for symbol filtering), NLTK (for stop words removal), NumPy (for rows randomization), and scikit-learn (for tokenization). The feature extraction was made vectorizing with TF-IDF from scikit-learn. Databases were split for training and test with a 2:1 ratio.

### 4.3 Models Building

In Model 1, the LP-SVM model is configured with default values and none of its hyperparameters are optimized, both for the problem transformation method LP and the classification algorithm SVM.

For Model 2, the transfer learning model, DistilBERT pre-trained features are selected using the distilbert-base-uncased model, i.e., the distilled version of the BERT base model [9] to tokenize the data. For the training/fine-tuning stage, the PyTorch library defines a series of minimal standard parametrizations that shape and control the data pre-processing and its passage to the neural network: batch size (4), maximum length (128), optimizer (Adam), learning rate (1-e-15), and epochs (3).

### 4.4 Model Evaluation

Scikit-learn library in Python offers a series of reports for the quantitative model evaluation for classification. Accuracy, F1-Score (micro) and Hamming loss are the performance metrics relevant for this experiment.

## 5 Results and Discussion

This section discusses in detail the results obtained for the individual models. Table 3 presents the accuracy, F1 (micro), and Hamming loss results for both classification models under five dataset scenarios.

In SC1, in accuracy, LP-SVM had a better performance with respect to DistilBERT by less than 2%. It is remarkable that LP-SVM achieves 81% in accuracy with standard parameters provided by the scikit-learn library. SC1 features the closest performance between both models.

Eliminating noisy labels in SC2, SC3, and SC4, yields slight improvements on both models. However, LP-SVM had a higher enhancement than DistilBERT (Fig. 2). For instance, in F1 (micro), while in DistilBERT the improvement is around 2%, for LP-SVM the improvement was 5%.

**Table 3.** Comparison of Accuracy, F1 and Hamming loss scores of DistilBERT and LP-SVM on the five Dataset Scenarios

Metric	Model	Dataset Scenarios				
		SC1	SC2	SC3	SC4	SC5
Accuracy	DistilBERT	0.791	0.813	0.787	0.736	0.875
	LP-SVM	<b>0.809</b>	<b>0.868</b>	<b>0.833</b>	<b>0.756</b>	<b>0.893</b>
F1-score (micro)	DistilBERT	0.855	0.874	0.858	0.826	0.892
	LP-SVM	0.855	<b>0.901</b>	<b>0.870</b>	<b>0.834</b>	<b>0.893</b>
Hamming loss	DistilBERT	0.018	0.024	0.028	<b>0.032</b>	0.019
	LP-SVM	0.018	<b>0.022</b>	<b>0.025</b>	0.034	<b>0.013</b>

In SC3, with balanced instances, both models had an acceptable performance with few differences between them.

SC4 has the worst performance for both models due to the low number of instances.

Finally, in SC5 (multi-class dataset) both models had the best performance in accuracy, LP-SVM with 89% and DistilBert with 88%.

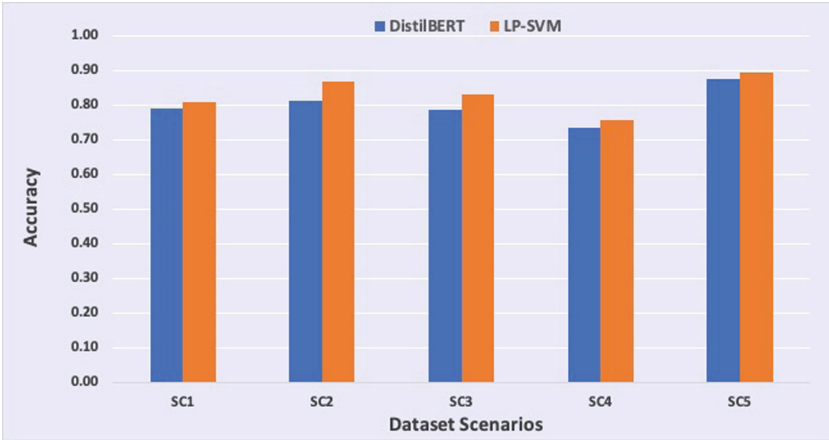


Fig. 2. Accuracy comparison of DistilBERT and LP-SVM in five dataset scenarios.

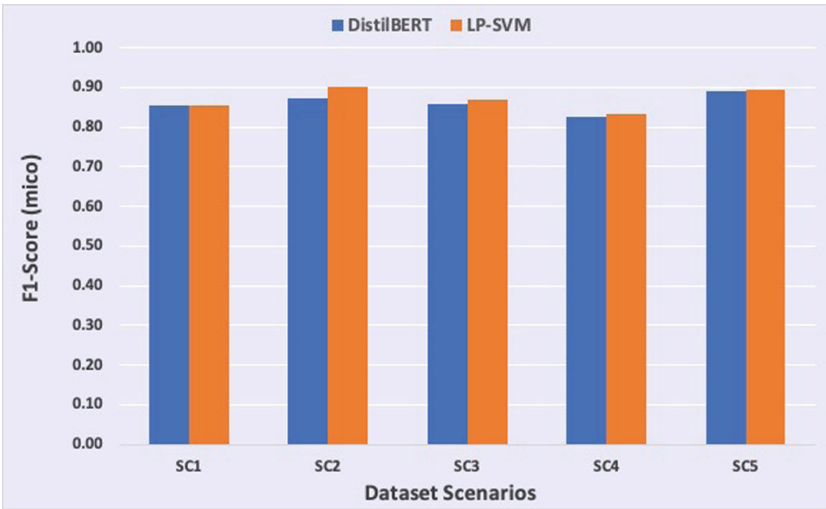
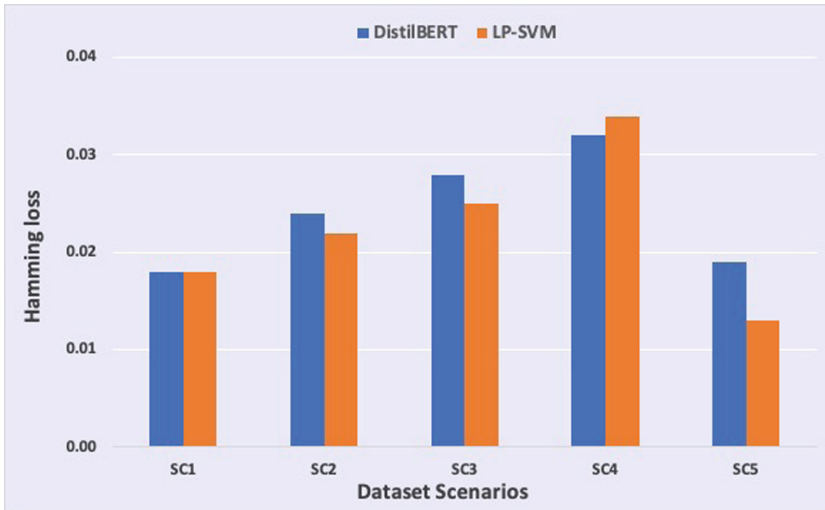


Fig. 3. F1-Score (micro) comparison of DistilBERT and LP-SVM in five dataset scenarios.





**Fig. 4.** Hamming loss comparison of DistilBERT and LP-SVM in five dataset scenarios.

## 6 Conclusion and Future Work

This study presented a comparison review of multi-label text classification models based on their performance. The results support the framework that implemented a combination of transformation methods with classification algorithms and a pre-trained model with acceptable classification performance.

LP-SVM, even with default parameters, had a remarkable result from almost all scenarios.

DistilBert, with similar results compared to the other model, has the disadvantage of requiring more computer resources and this is a disadvantage for some institutions that wish to implement the recognition of their academic products aligned with the SDGs. Thus, this study confirms the complexity of pre-trained models and the need to deepen the tuning of the model. Future work includes defining adjustments to hyperparameters in both models and quantifying performance improvements.

Institutions with little computing resource capacity can implement LP-SVM to classify their scientific production with respect to the SDG.

For future work, the dataset with organic agriculture 3.0 as a dominant theme could be a promising source of information. Topic modeling and recognition of emerging trends could bring opportunities for data mining and knowledge management applying artificial intelligence.

## References

1. Medina, S.R., Niamir, A., Dadvar, M.: Multi-Label Text Classification with Transfer Learning for Policy Documents The Case of the Sustainable Development Goals. Uppsala University (2019)

2. Aggarwal, C. :Data Classification: Algorithms and Applications. CRC press (2014)
3. Rivolli, A., Read, J., Soares, C., Pfahringer, B., de Carvalho, A.C.P.L.F.: An empirical analysis of binary transformation strategies and base algorithms for multi-label learning. *Mach. Learn.* **109**(8), 1509–1563 (2020). <https://doi.org/10.1007/s10994-020-05879-3>
4. Dudzik, W., Nalepa, J., Kawulok, M.: Evolving data-adaptive support vector machines for binary classification. *Knowl.-Based Syst.* **227**, 107221 (2021). <https://doi.org/10.1016/j.knsys.2021.107221>
5. Shah, K., Patel, H., Sanghvi, D., Shah, M.: A comparative analysis of logistic regression, random forest and KNN models for the text classification. *Augmented Hum. Res.* **5**(1), 1–16 (2020). <https://doi.org/10.1007/s41133-020-00032-0>
6. Xu, S.: Bayesian Naïve Bayes classifiers to text classification. *J. Inf. Sci.* **44**(1), 48–59 (2018). <https://doi.org/10.1177/0165551516677946>
7. Wu, X., Gao, Y., Jiao, D.: Multi-Label classification based on random forest algorithm for non-intrusive load monitoring system. *Processes* **7**(6), 337 (2019). <https://doi.org/10.3390/pr7060337>
8. Abdullahi, A., Samsudin, N.A., Khalid, S.K.A., Othman, Z.A.: An improved multi-label classifier chain method for automated text classification. *Int. J. Adv. Comput. Sci. Appl.* **12**(3), 442–449 (2021). <https://doi.org/10.14569/IJACSA.2021.0120352>
9. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186 (2019). Available: <http://arxiv.org/abs/1810.04805>
10. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In: Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing (EMC2) co-located with the Thirty-third Conference on Neural Information Processing Systems (NeurIPS 2019), pp. 1–5 (2019). Available: <http://arxiv.org/abs/1910.01108>
11. Adhikari, A., Ram, A., Tang, R., Lin, J.: DocBERT: BERT for document classification. In: Proceedings of the 5th Workshop on Representation Learning for NLP, pp. 72–77 (2020). Accessed: 26 Jun 2022. [Online]. Available: <https://aclanthology.org/2020.repl4nlp-1.10.pdf>
12. Bambroo P., Awasthi, A.: LegalDB: long distilbert for legal document classification. In: Proceedings of the 2021 1st International Conference on Advances in Electrical, Computing, Communications and Sustainable Technologies, ICAECT 2021 (2021). <https://doi.org/10.1109/ICAECT49130.2021.9392558>
13. Jiao, X., Hui, K., Sun, L., Sun, Y.: TinyBERT: distilling BERT for natural language understanding. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 4163–4174 (2020). Accessed: 26 May 2022 [Online]. Available: <https://aclanthology.org/2020.findings-emnlp.372.pdf>
14. United-Nations, “Resolution 70/1. Transforming our world: the 2030 Agenda for Sustainable Development,” United Nations (2015)
15. Madjarov, G., Kocev, D., Gjorgjevič, D., Džeroski, S.: An extensive experimental comparison of methods for multi-label learning. *Pattern Recogn.* **45**(9), 3084–3104 (2012). <https://doi.org/10.1016/j.patcog.2012.03.004>
16. Tsoumakas, G., Katakis, I.: Multi-Label classification: an overview. *Int. J. Data Warehouse. Min.* **3**(3), 1–13 (2007). <https://doi.org/10.4018/jdwm.2007070101>
17. Read, J.: *Advances in Multi-label Classification* (2011)
18. Tsoumakas, G., Vlahavas, I.: Random k-labelsets: an ensemble method for multilabel classification. In: Kok, J.N., Koronacki, J., Mantaras, R.L.D., Matwin, S., Mladenič, D., Skowron, A. (eds.) *Machine Learning: ECML 2007. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence)*, vol. 4701, pp. 406–417. Springer, Heidelberg (2007). [https://doi.org/10.1007/978-3-540-74958-5\\_38](https://doi.org/10.1007/978-3-540-74958-5_38)

19. Cortes, C., Vapnik, V.: Support-vector networks. *Mach. Learn.* **20**(3), 273–297 (1995)
20. Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., Lopez, A.: A comprehensive survey on support vector machine classification: applications, challenges and trends. *Neurocomputing* **408**, 189–215 (2020). <https://doi.org/10.1016/j.neucom.2019.10.118>
21. Hana, K.M., Adiwijaya, S., Faraby, A., Bramantoro, A.: Multi-label classification of Indonesian hate speech on Twitter using support vector machines. In: 2020 International Conference on Data Science and Its Applications (ICoDSA), pp. 1–7 (2020). <https://doi.org/10.1109/ICoDSA50139.2020.9212992>
22. Saeed, S., Ong, H.C.: Performance of SVM with multiple kernel learning for classification tasks of imbalanced datasets. *Pertanika J. Sci. Technol.* **27**(1), 527–545 (2019)
23. Büyüköz, B., Hürriyetoğlu, A., Özgür, A.: Analyzing ELMo and DistilBERT on socio-political news classification. In: Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020, pp. 9–18 (2020). Available: <https://www.aclweb.org/anthology/2020.aespen-1.4>
24. Clavié, B., Alphonso, M.: The unreasonable effectiveness of the baseline: discussing SVMs in legal text classification. *Front. Artif. Intell. Appl.* **346**, 58–61 (2021). <https://doi.org/10.3233/FAIA210317>
25. Menger, V., Scheepers, F., Spruit, M.: Comparing deep learning and classical machine learning approaches for predicting inpatient violence incidents from clinical text. *Appl. Sci. (Switzerland)* **8**(6), (2018). <https://doi.org/10.3390/app8060981>
26. Alammery, A.S.: BERT models for Arabic text classification: a systematic review. *Appl. Sci.* **12**(11), 5720 (2022). <https://doi.org/10.3390/app12115720>
27. Lagutina, K.: Topical text classification of Russian news: a comparison of BERT and standard models. In: 2022 31st Conference of Open Innovations Association (FRUCT), pp. 160–166 (2022). <https://doi.org/10.23919/FRUCT54823.2022.9770920>
28. Wastl, J., Porter, S., Draux, H., Fane, B., Hook, D.: Contextualizing sustainable development research. *Digit. Sci.* (2020). Available: <https://doi.org/10.6084/m9.figshare.12200081>
29. Mishra, A., Vishwakarma, S.: Analysis of TF-IDF model and its variant for document retrieval. In: 2015 International Conference on Computational Intelligence and Communication Networks (CICN), pp. 772–776 (2015). <https://doi.org/10.1109/CICN.2015.157>
30. Nasierding, G., Kouzani, A.Z.: Comparative evaluation of multi-label classification methods. In: Proceedings - 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2012, pp. 679–683 (2012). <https://doi.org/10.1109/FSKD.2012.6234347>