Text Classification and Keyword Identification of the Sustainable Development Goals

Andres Gomez
University of Florida, College of
Electrical and Computer Engineering
Gainesville, FL
Andresggomez@ufl.edu

***Abstract* — We present a comparative study of four well-known machine learning algorithms, Multinomial Naïve Bayes (MNB), Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF), for text classification using the Open Software Development Governance - Code Dataset (OSDG-CD). We determined the key words for each class and evaluated the performance of each algorithm on the task of SDG classification. Our comparative study showed promising results, with SVM achieving the highest accuracy of 89.4%. We also compared our results with other models such as BERT and Label Powerset with SVM, which have achieved higher accuracies on this task, and discussed the advantages and disadvantages of these models.**

Keywords—Sustainable development goals, text classification, Keyword identification, OSDG-CD.

I. INTRODUCTION

The United Nations' Sustainable Development Goals (SDGs) are a set of 17 broad goals that seek to address some of the world's most urgent economic, social, and environmental issues. Achieving these goals requires a joint effort from governments, civil society organizations, and the private sector. With vast amounts of data being generated by the United Nations and outside organizations daily, automatic systems that utilize machine learning and natural language processing can dramatically reduce the time is takes to track the progress made in achieving the SDGs.

In this paper, we present a comparative study of four well-known machine learning algorithms to classify SDG-related documents. Additionally, we identify the keywords of each SDG. Our method utilizes Multinomial Naïve Bayes (MNB), Support Vector Machine (SVM), Logistic Regression (LR), and the Random Forest (RF) algorithm for text classification using the Open Software Development Governance - Code Dataset (OSDG-CD). The dataset contains over 30,000 text excerpts labeled with Sustainable Development Goals (SDGs) and reviewed by domain experts to provide a measure of agreement. We also utilize term frequency-inverse document frequency (TF-IDF) and to extract relevant features and identify keywords from each category.

Overall, this work underscores the potential of text classification and keyword identification techniques in accelerating the analysis of UN documents related to the SDGs. Our methods can effectively allow for improved monitoring of the progress made towards meeting the SDGs and contribute towards the goal of a more sustainable future.

II. DESCRIPTION

A. OSDG-CD

The Open Software Development Governance - Code Dataset (OSDG-CD) contains over 30,000 text excerpts gathered by thousands of volunteers from a majority of the UN Member States. The excerpts are of paragraph length and are derived from publicly available documents [1]. Roughly 3,000 of these documents are from UN-related sources. These documents are labeled (i.e. have associated SDGs) and are roughly 90 words in length.

The class label of each excerpt was reviewed by up to 9 domain experts and each expert was tasked with casting a vote on whether or not they agreed with the assigned class label. These votes are provided as well as an agreement score, which is calculated by taking the absolute difference between positive and negative votes divided by the sum of votes. So, the agreements are bounded between 0 and 1. Therefore, we are provided with a measure for how well the excerpt fits into its corresponding class. Additionally, we can see from Figure 1 that the classes are quite imbalanced.

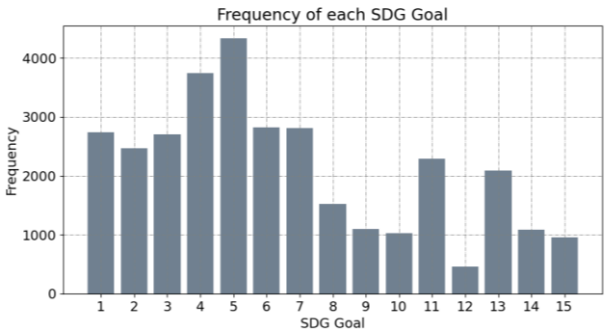

Figure 1: Frequency count of text excerpt by SDG.

In order to divide the data into train, test, and validation set, we created a categorical agreement of 5 bins, those between (0, 0.2), (0.2, 0.4), (0.4, 0.6), (0.6, 0.8), and (0.8, 1.1) – (a, b) corresponds to an inclusion of a and exclusion of b. We then used a stratified split method using both the class frequency and the categorical agreement score to ensure our various subsets of data had proportional distributions. The train, validation, and test size contained 70%, 15%, and 15% of the samples, respectively.

B. Text Classification

In this paper, we present a comparative study of four well-known machine learning algorithms, Multinomial Naïve Bayes (MNB), Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF), to classify text documents. We will compare their performances in terms of overall accuracies. For MNB, LR, and RF, we will examine

the top-3 accuracies, that is the percentage of times the true label is amongst the top 3 label predictions.

Additionally, we extract features using a Natural Language Processing (NLP) feature extraction technique, TF-IDF, and feed them into the previously mentioned classification algorithms and compare performances. Count vectorizer is another popular technique that was considered. The main difference between these two techniques is that count vectorizer technique considers the frequency of words in a corpus, while the TF-IDF technique accounts for how common or uncommon a word is. Words that occur in a greater frequency, like prepositions, are given lower weights and those occurring less frequently are given higher weights. In other words, the TF-IDF score can be thought of as a measure of importance. Ultimately, we decided to move forward with TF-IDF because the added complexity can provide more generalizable features.

B. Keyword Identification

The second task involved determining the key words for each class. In this dataset, we are provided documents for 15 different SDGs. Keyword extraction allows us to identify the most important words or phrases in a corpus. By separating the documents into their respective classes and implementing TF-IDF, we were able to identify the keywords for each class. We listed the top 20 words for each class below.

III. Related work

Several methods exist for text classification, including Bag of Words, Word2Vec embedding, fastest embedding, Convolutional Neural Networks, Long Short-term Memory, and BERT (Bidirectional encoder representations from transformers).

Previous works show success with applying the BERT model for multi-label classification tasks. Matsui T et al., tuned the Tohoku-BERT model on the SDGs corpus database and reconstructed the algorithm to classify texts [2]. Similarly, Guisiano J. et el. applied the BERT algorithm to the same dataset with an accuracy of 98% [3]. In earlier work, Guisiano J. et el. applied similar techniques to achieve an accuracy of 94.2% on the same task [4].

There are many advantages to using BERT for text classification. BERT can provide contextual representation or the context and meaning of words and phrases. In tasks where context is especially important, BERT will likely perform better than in traditional text classification methods that consider single words or n-grams. Additionally, BERT is a pretrained model, meaning it may perform better at tasks than other models trained from scratch, especially if the available dataset is relatively small. The downsides of using BERT are that it is computationally expensive, has a large memory footprint, and lacks interpretability.

Morales-Hernandez et al compared several models and found that a combination of Label Powerset with a Support Vector Machine achieved an accuracy of 91% [4]. A label powerset can be useful in cases where individual labels are not well-balanced and where there are a vast amount of label combinations. However, like BERT, the label powerset can be computationally expensive, especially if there are a large number of classes. SVMs are quite effective in high-dimensional spaces and can do well with small datasets. However, they can be computationally expensive, are sensitive to hyperparameter selection, and lack interpretability.

IV. Experiments and Evaluation

A. Text Classification

As previously mentioned, MNB SVM, LR, and RF were implemented alongside the TF-IDF feature extraction method. Several hyperparameters were tried for TF-IDF, including sublinear_df of [True, False], which alters the scale of the TF-IDF score. Additionally, there is a min_df parameter that ignores terms higher than the selected threshold. Min_df of [1, 2, 3, 4, 5] were tested. The maximum features were set to 72,400 because that was the maximum that we were able to fit on our 24GB-memory system while still being able to train models. Lastly, the stop words were set to "english" and our n_grams range was (1,2).

Additionally, three cases of training data were evaluated. The first involved training a model with equal sample weights and not utilizing the agreement feature. The second involved having equal sample weights and using the agreement as a feature. The third involved weighing the samples by agreement, that way, samples with high agreements contribute more to the model.

Next, we discuss hyperparameters specific to each model. For MNB, alphas of [1e-1, 1e-2, 1e-3, 0] were evaluated. For SVM and LR, we evaluated C values of [0.01, 0.1, 1]. Lastly, we experimented with the max depth, trying values of [300, 600] while keeping the number of estimators to the default (100).

Algorithm	**Validation Accuracy (%)**	**Top 3 Validation Set Accuracy (%)**
SVM	79.2	N/A
MNB	75.0	91.1
LR	78.3	93.2
RF	72.1	88.1

Table 1:Validation accuracy for SVM, MNB, LR, and RF. The top 3 accuracy is also shown for the latter 3 algorithms.

The optimal parameters for MNB were an alpha of 0.01, a sublinear_df of False, a min_df of 2 and the case where all samples are weighted equally with agreement being used as a feature. The optimal parameters for SVM were a C of 1, sublinear_df set to True, min_df of 2 and the case where all samples are weighted equally with agreement being used as a feature. The optimal parameters for LR were a C value of 1, sublinear_df set to False, min_df of 1, and the case where all samples are weighted equally with agreement being used as a feature. Lastly, the optimal hyperparameters for the RF were sublinear_df set to True, min_df of 3, and the case where all samples are weighted equally with agreement being used as a feature.

In Table 1, we see that SVM obtained the highest validation accuracy, followed closely by LR. We are unable to obtain a top 3 accuracy for SVM because it lacks a probabilistic interpretation. Of MNB, LR, and RF, LR had the highest top 3 accuracy on the validation set.

In Table 2, we can observe the generalization of each model to the test set. LR obtained both the highest overall

accuracy and top 3 accuracy on the test set. It is important to observe the top k accuracies because some of the SDGs are quite intertwined, as we will observe in the following section.

Algorithm	**Test Accuracy (%)**	**Top 3 Test set Accuracy (%)**
SVM	64	N/A
MNB	76.0	90.8
LR	78	93.4
RF	72	88.6

Table 2: Test set accuracies for SVM, MNB, LR, and RF. The top 3 accuracy is also shown for the latter 3 algorithms.

In Table 3, we show the Precision, Recall, and F1-score for the LR model evaluated on the test set. We see that the model very well on SDG 3 and quite poorly on SDG 8 and 10, as observed by the F1-score. The model obtains the lowest precision for SDG 8. In the following section, we show the top 20 keywords as ranked by the TF-IDF score, which may provide insights into the model's performance on these classes.

SDG	**Precision**	**Recall**	**F1-Score**
1	0.73	0.76	0.74
2	0.73	0.78	0.75
3	0.88	0.90	0.89
4	0.81	0.89	0.85
5	0.81	0.88	0.84
6	0.79	0.80	0.79
7	0.77	0.83	0.80
8	0.58	0.50	0.54
9	0.73	0.60	0.66
10	0.73	0.41	0.53
11	0.75	0.78	0.76
12	0.81	0.43	0.57
13	0.82	0.81	0.81
14	0.89	0.74	0.81
15	0.83	0.67	0.74
Accuracy			0.78

Table 3: Precision, Recall, F1-score, and overall accuracy for the LR model evaluated on the test set.

B. Keyword Identification

For the keyword extraction, we again used the TF-IDF method. We first separated the texts by their label. After applying the vectorizer, we ranked them in terms of their TF-IDF score and chose the top 20 words. The keywords for each of the 15 SDGs are shown below:

1. No Poverty

poverty, income, countries, children, social, poor, child, households, cent, household, deprivation, people, growth, rates, development, data, economic, population, employment, health

2. Zero Hunger

food, agricultural, countries, production, prices, land, price, farmers, agriculture, support, trade, development, policy, rural, market, policies, growth, use, farm, world

3. Good Health and Well-Being

health, care, countries, services, health care, The Organization for Economic Co-operation and Development (OECD), mental, quality, primary, data, population, patients, national, mental health, public, primary care, people, hospital, medical, average

4. Quality Education

education, school, students, schools, teachers, OECD, countries, skills, learning, development, teacher, level, training, teaching, higher, children, secondary, student, quality, average

5. Gender Equality

women, gender, countries, men, work, care, social, time, rights, equality, labour/labor, education, family, violence, female, children, employment, OECD, gender equality, economic

6. Clean Water and Sanitation

water, management, river, environmental, use, resources, groundwater, basin, countries, policy, national, irrigation, quality, level, costs, supply, public, development, economic, areas

7. Affordable and Clean Energy

energy, electricity, power, countries, costs, renewable, efficiency, capacity, sector, technologies, development, policy, investment, cost, energy efficiency, supply, market, use, emissions, demand

8. Decent Work and Economic Growth

employment, labour/labor, workers, work, countries, job, education, OECD, unemployment, growth, market, social, sector, economic, youth, policy, income, training, time, development

9. Industry, Innovation and Infrastructure

development, countries, innovation, infrastructure, research, new, policy, data, services, access, technology, trade, growth, digital, mobile, broadband, economic, public, developing, support

10. Reduced Inequality

income, countries, inequality, social, tax, OECD, growth, labour/labor, trade, employment, poverty, workers, market, distribution, policy, average, benefits, economic, wage, unemployment

11. Sustainable Cities and Communities

urban, development, transport, public, city, cities, local, land, housing, government, national, areas, use, countries, planning, policy, services, economic, new, road

12. Responsible Consumption and Production

waste, countries, environmental, management, economic, development, sustainable, policy, resource, use, products, collection, consumption, tourism, recycling, data, production, companies, materials, energy

13. Climate Action

climate, adaptation, finance, countries, development, change, climate change, information, national, climate finance, support, sector, policy, private, level, energy, emissions, country, global, reporting

14. Life Below Water

fisheries, fish, fishing, management, aquaculture, marine, production, species, ocean, value, economic, sea, fishery, vessels, resources, total, policy, stocks, countries, based

15. Life on Land

forest, biodiversity, areas, forests, species, national, land, management, use, protected, countries, development, area,

data, environmental, protected areas, services, resources, conservation, natural

If we again observe the Table 3, we see that the model performed poorly on SDG 8, as measured by the precision. It is worth reiterating that these SDGs are heavily intertwined. If we observe the keywords for SDG 8, we see that some of its highly ranked keywords are found in other SDGs, like in SDG 10 – Reduced Inequality and SDG 1 – No poverty. This may provide some insight into why the precision is so poor for this class.

V. Summary and Conclusion

In conclusion, the OSDG-CD provides a valuable resource for researchers interested in text classification and natural language processing. We trained and compared four well-known machine learning algorithms, Multinomial Naïve Bayes (MNB), Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF), to classify text documents using TF-IDF feature extraction technique. We also determined the key words for each class.

Our results show that the LR model obtained the second highest overall accuracy on the validation set and the highest top 3 accuracy. SVM obtained the highest overall accuracy on the validation set but generalized poorly on the test set. The LR model generalized the best to the test set, obtaining an overall accuracy of 78% and a top 3 accuracy of 93%.

While our comparative study showed promising results, there are other models like BERT and Label Powerset with SVM that have achieved higher accuracies on this task. We discussed the advantages and disadvantages of these models and highlighted the challenges of using them.

VI. References

[1] Osdg, et al. “OSDG Community Dataset (OSDG-CD).” Zenodo, 1 Oct. 2021, https://zenodo.org/record/5550238#.Y_AVNnbMKUl

[2]Matsui, Takanori, et al. “A Natural Language Processing Model for Supporting Sustainable Development Goals: Translating Semantics, Visualizing Nexus, and Connecting Stakeholders - Sustainability Science.” SpringerLink, Springer Japan, 4 Feb. 2022

[3] Guisiano, Jade Eva, et al. “SDG-Meter : a deep learning-based tool for automatic text classification of the Sustainable Development Goals.” HAL Open Source, 26 Jul 2022.

[4] Guisiano, Jade Eva, et al. “Automatic classification of multilabel texts related to Sustainable Development Goals (SDGs).” HAL Open Source, 27 Feb 2021.

[5] Morales-Hernandez, et al. “A Comparison of Multi-Label Text Classification Models in Research Articles Labeled with Sustainable Development Goals” IEEE Access, Vol 10. 17 Nov 2022