TCS_WITM_2022@FinSim4-ESG: Augmenting BERT with Linguistic and Semantic features for ESG data classification

Tushar Goel, Vipul Chauhan, Suyash Sangwan, Ishan Verma, Tirthankar Dasgupta, Lipika Dey
TCS Research
New Delhi India
(t.goel, chauhan.vipul, suyash.sangwan, ishan.verma, dasgupta.tirthankar, lipika.dey)@tcs.com

Abstract

Advanced neural network architectures have provided several opportunities to develop systems to automatically capture information from domain-specific unstructured text sources. The FinSim4-ESG shared task, collocated with the FinNLP workshop, proposed two sub-tasks. In sub-task1, the challenge was to design systems that could utilize contextual word embeddings along with sustainability resources to elaborate an ESG taxonomy. In the second sub-task, participants were asked to design a system that could classify sentences into sustainable or unsustainable sentences. In this paper, we utilize semantic similarity features along with BERT embeddings to segregate domain terms into a fixed number of class labels. The proposed model not only considers the contextual BERT embeddings but also incorporates Word2Vec, cosine, and Jaccard similarity which gives word-level importance to the model. For sentence classification, several linguistic elements along with BERT embeddings were used as classification features. We have shown a detailed ablation study for the proposed models.

1 Introduction

Sustainability disclosures have started gaining traction in Financial world. Sustainability disclosures reflects the procedures that an organization follow to fulfill its commitment towards Environment, Social and Governance (ESG) factors. Investors are increasingly considering the ESG commitments of organizations to aid their investments decisions. ESG information is not commonly part of financial reporting but organizations have started making ESG disclosures either as a part of their annual report or as a separate sustainability report altogether.

Recently, European Union has proposed new guidelines making ESG disclosures mandatory for companies providing ESG driven investment products. Companies in EU must comply with increasing obligations related to ESG in order to have their businesses qualify as sustainable and to be noticed by investors on the EU market. Several institutions, such as the Sustainability Accounting Standards Board (SASB), the Global Reporting Initiative (GRI), and the Task Force on Climate-related Financial Disclosures (TCFD) are also working to form standards and define standards to facilitate incorporation of ESG factors into the investment process. Organization need to align their disclosures to various ESG taxonomies available to facilitate effective utilization of the disclosures. Since, ESG domain and the allied taxonomies are still evolving, there is a need to have automated systems that can elaborate the existing taxonomies and also aid the organization to align their disclosures to the existing standards. The FinSim4-ESG shared task is one of the few attempts that primarily focused on automatically learning effective and precise semantic models adapted to the ESG domain. Specifically it addressed the task of automatic categorization of ESG terms into pre-defined categories to enhance the existing taxonomy and to create a language model that can understand the language of sustainability and segregate sustainable sentences from unsustainable ones. The specific details of the shared task is described in section 3.

In this paper, we propose to augment the transformer based BERT architecture with semantic similarity features to address the ESG term classification task. The base BERT model is first fine-tuned on a set of ESG document to facilitate capture of ESG language context. The embedding obtained from the resultant model is then augmented with semantic similarity features like Word2Vec, Cosine and Jaccard similarity to perform the classification task of segregating domain terms into fixed number of class labels. In a similar way, for sentence classification task, various lexical features are used along with BERT embeddings. Dimensionality reduction techniques are also incorporated into the experiments. For training the model, we have used the

Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP), pages 235 - 242

$BERT_{base}$ architecture. We have conducted multiple experiments with the model architecture such as taking combination of proposed features with different classifiers in both sub-tasks. Our experiments shows that a combination of ESG fine tuned BERT with Word2Vec, cosine and jaccard similarity with a Logistic Regression classifiers gives the best result for term classification. For sentence classification we exploited linguistic and NLP-based features like presence of specific named entities (like organisation, date), word-count in the document, etc. Our experimental results showed that when we combine these features along with fine-tuned BERT-base classifier, it lead to an increase of about 2% in system's accuracy.

2 Related Work

Rising sustainability awareness amongst consumers, investors and regulators is forcing organizations to pay attention towards Environment, Social or Governance (ESG) factors. Researchers around the world have conducted experiments to study the effect of ESG factors on financial performance of organizations. The last three editions of the FinSim proposed the challenge to automatically learn effective and precise semantic models for the financial domain (Mansar et al., 2021; El Maarouf et al., 2020; Kang et al., 2021). (Chersoni and Huang, 2021) solve the financial hypernym detection task by using the logistic regression classifier trained on a combination of word embeddings, semantic and string similarity metrics and won the challenge. (Nguyen et al., 2021) approached the task as a semantic textual similarity problem. They used a siamese network with pre-trained language model encoders to derive term embeddings and computed similarity scores between them while (Goel et al., 2021) leverage the use of given documents by extracting sentences corresponding to each term and then used a transformer based BERT architecture to perform a classification task. A number of approaches like use of publicly available knowledge graphs to generate explicit features (Portisch et al., 2021), customize word embeddings (Pei and Zhang, 2021), word ontology (Tian and Chen, 2021), pre-trained sentence embedding extracted from Universal Sentence Encoder along with cosine similarity (Anand et al., 2020) and use of static word embeddings (Fu et al., 2014; Nguyen et al., 2017) have also been proposed to automatically map financial concepts with its most relevant hypernym.

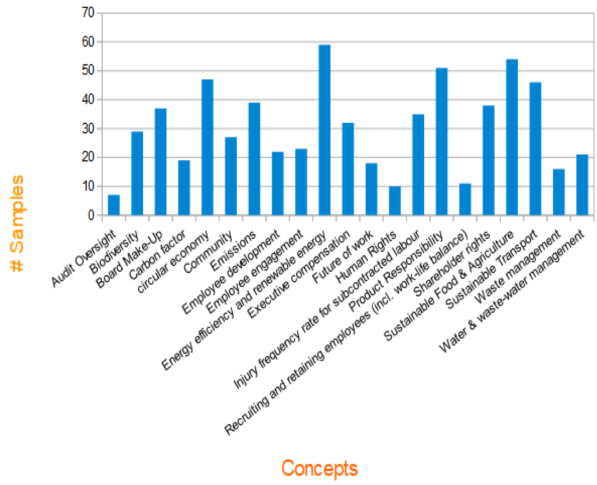

Figure 1: Sub-task1 Data Distribution

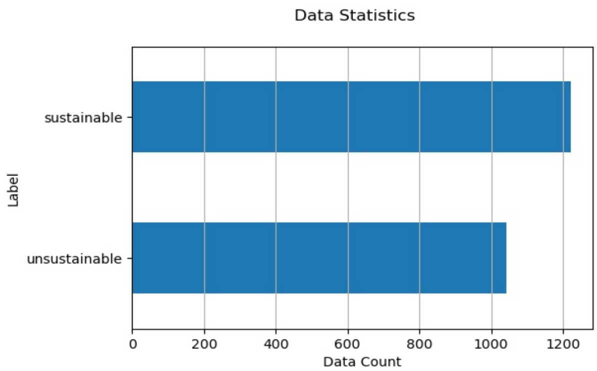

Figure 2: Sub-task2 Data Distribution

3 Shared Task Detail and Dataset Description

The FinSim4-ESG 2022 task focused on elaboration of an ESG taxonomy (ESG related concepts representations) based on the data like companies sustainability reports, annual reports, and environment reports, and make use of them to analyze how an economic activity complies with the taxonomy. The current edition proposed two sub-tasks :

- **Sub-task1** - Given a list of carefully selected terms from the sustainability domain such as "low-carbon", "Greenhouse gas emissions", the task was to design a system which can automatically classify them into the most relevant ESG-concept. For example, given the set of concepts "Future of Work", "Human Rights", "Biodiversity", "Community", "Waste Management", the most relevant concept of "Palm Oil" is "Biodiversity". The training data consisted of 647 unique terms with corresponding concepts. There were 25 unique concepts in the shared task but for few concepts data was insufficient like terms corresponding to concept "Diversity & Inclusion" were not present in the training data. Hence, we considered 641 unique terms which corresponded to 21 unique concepts for our experiments. Moreover, As shown in Figure 1, the training data was also imbalanced wherein concepts such as "Energy Efficiency and Renewable Energy" had 9.2% data and concept "Audit Oversight" had only 1.09% data. For test data, there were 145 terms to be classified into the correct concepts.

- **Sub-task2** - In this sub-task, the participants were asked to create an automated system which could classify sustainable and unsustainable sentences. Here, a sentence is considered sustainable if and only if it semantically mentions the Environmental or Social or Governance related factors. For example, a sentence "At Vauban Infrastructure Partners, we integrate in our daily work practices to avoid, reduce or offset our carbon emissions." is a sustainable sentence. For this purpose, the organisers provided the list of labeled sentences from the sustainability reports and other documents. The training data consisted of 2265 sentences, out of which 1223 sentences were sustainable and the remaining 1042 sentences were unsustainable. The dataset was balanced with a ratio of 0.852 as shown in Figure 2. There were 205 sentences in the test data which needs to be classified into their respective label categories.

4 Proposed Approach

The proposed approach can be divided into three main stages namely feature extraction stage, dimensionality reduction stage and classification stage as shown in Figure 3. For both sub-tasks, the input text is first subjected to feature extraction module where different NLP and text mining techniques are applied to obtain various Linguistic and Semantic features along with BERT embeddings. A combination of these features are then subjected to dimensionality reduction where a subset of features are selected as final input variables for classification module. The detailed description of the activities in each stage is explained next.

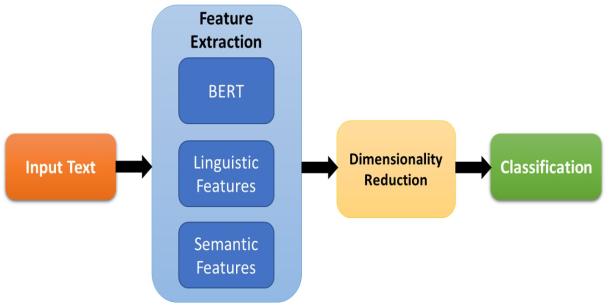

Figure 3: Proposed approach pipeline

4.1 Feature Extraction for Sub-task1

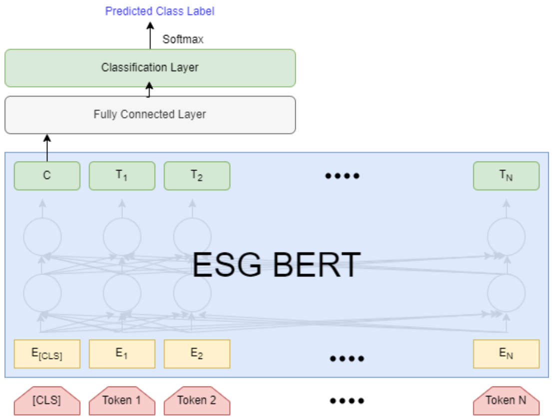

Figure 4: ESG BERT Architecture

- **ESG BERT** - BERT (Bidirection Encoder Representation from Transformers) is a state of the art model for language modeling (Devlin et al., 2018). The official BERT Repository[1] contains various pre-trained models that can be further used for the downstream task. In this work, ESG-BERT model[2] pre-trained on Sustainability corpora has been utilised to achieve better capability of understanding domain specific vocabulary. The model is fine-tuned further on the downstream task of concept classification as shown in Figure 4. Experiments show that the Fully Connected (FCN) layer added between the classification layer and BERT output layer benefited the model in learning the representation for the downstream task. The output of the FCN layer represents the term by an embedding vector.

[1] https://github.com/google-research/bert
[2] github.com/mukut03/ESG-BERT

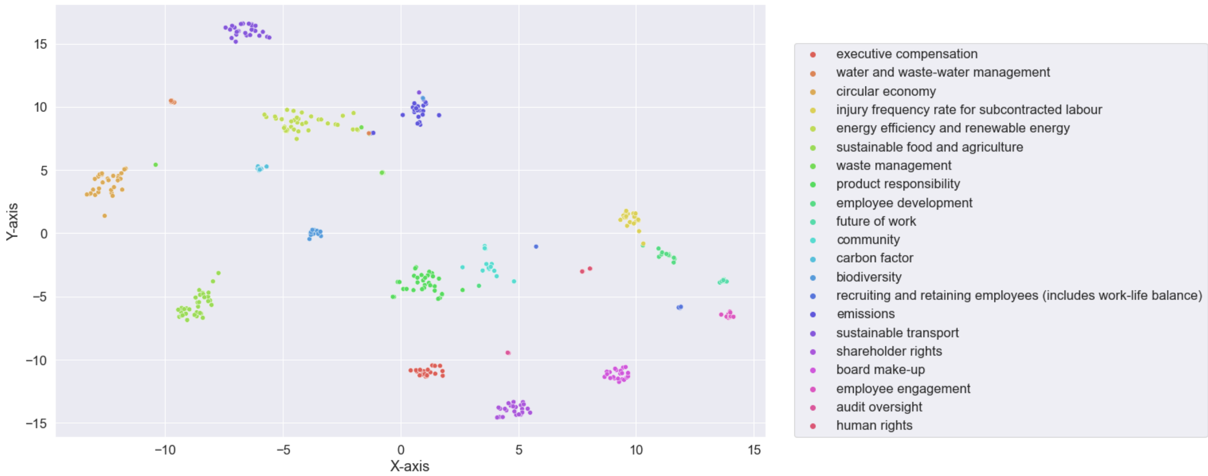

Figure 5: ESG BERT Embedding Plot

This vector has been used as one of the feature in our system for the classification. A tsne-plot of ESG BERT embedding of terms in training data is shown in Figure 5. The cluster segregation for different concepts can be seen clearly in the figure which shows the potential of ESG-BERT based embedding in classifying ESG concepts.

- **Cosine Similarity (CS)** - Semantic similarity between the terms and the concepts plays an important role in the classification. The ESG-BERT fine-tuned in the previous experiment is further used to obtain the cosine similarity between the terms and the pre-defined concepts. For each term, the embedding vector of the term is compared against the embedding vectors of each concept. Cosine similarity is used as the metric for comparing the vectors. This results into a N dimensional vector for each term where N is the number of concepts. Each value in this vector represents cosine similarity of a term with $concept_i$.

- **Word2Vec Features** - The organizers of the shared task have provided a set of 190 sustainability reports of various companies. The corpus extracted from these reports contains about 33 thousand unique tokens. These tokens are used to train the 100 dimensional domain-specific Word2Vec (Mikolov et al., 2013) word embeddings. For terms composed by multiple words, we simply represent them by summing the vectors of the individual words.

- **Jaccard Similarity** - It was observed that out of 641 samples, there were 326 samples with at least one word overlapping (excluding stop-words) between term and concept. Based on (Keswani et al., 2020) approach of pointing out "concept inclusion", we computed *Jaccard Similarity* as a N dimensional feature vector(one feature for each concept), where each values in this feature reflect the syntactic similarity between the terms and concepts.

Dimensionality Reduction with Principal Component Analysis (PCA) - The combination of embeddings and features used in our experiments results into large number of features as compared to number of samples. The performance of any machine learning model can suffer from the curse of dimensionality where the number of features becomes larger than the number of samples in the dataset. To tackle this issue, we have empirically reduce the dimensions of the features using Principal Component Anaysis (PCA) (Martinez and Kak, 2001) and found that reduction to 200 dimensions are giving the best results.

4.2 Feature Extraction for Sub-task2

Data pre-processing is a crucial first step before applying any text machine learning model. Firstly, we remove all the punctuations and convert the tokenized words into lower-case format. Then to clean the noise present in the sentences we removed stop-words. Subsequently, the cleaned data is subjected to feature extraction step. In this step, raw text data is transformed into feature vectors. We tried following different ideas to obtain relevant features from the dataset:

- **Count vector**- Count vector is the matrix notation of the dataset in which every row represents a sentence from the corpus, every column represents a term from the corpus, and every cell represents the frequency count of a particular term in a particular sentence.

- **Character level, N-gram level, Word level TF-IDF vectors as features**- This vector represents the relative importance of a character/n-gram/word to the given output class. TF-IDF stands for Term Frequency – Inverse Document Frequency. This weighting has been widely used for feature extraction in text data. Term frequency (TF) is equal to the frequency of a given word in the given sentence. Inverse Document Frequency (IDF) measures the information provided by the word. It is equal to the inverse of number of sentences containing the word. Hence, $tfidf(w, s) = tf(w, s) * idf(w, S)$ where w is the given word, s is the given sentence and S is the corpus containing all sentences. TF-IDF gives higher value to the words which are less frequent among sentences and vice versa.

- **Sentiment polarity as a feature**- We used TextBlob[3] to extract sentiment polarities of each input sentence.

- **Text based features**- A number of other text based features are extracted using NLTK[4] and SpaCy[5] to aid the text classification models. Some exmaples are-
  - **NLTK features**-
    - **Word count of the sentences**- Total number of words in the sentence.
    - **Punctuation count of the sentences**- Total number of punctuation marks in the sentence.
    - **Frequency distribution of POS tags**- Total number of nouns, verbs, adjectives, adverbs, and, pronouns in the sentence.
  - **SpaCy features**- We extracted Named Entities using SpaCy. Some of the extracted named entities are- Organisations, Places, Money, Date, and, Person.

[3]https://github.com/sloria/TextBlob
[4]https://github.com/nltk/nltk
[5]https://spacy.io/

Dimensionality reduction: We used correlation and p-values to select the final feature set. We compare the correlation between different pairs of features and remove one of the two features that have a correlation higher than 0.9. Now from the remaining set of features we remove different features randomly and measure the p-value in each case. These measured p-values are used to decide whether to keep a feature or not. Finally the feature set giving maximum p-value is selected.

Embedding models: With so many rampant advances taking place in NLP, it can sometimes become overwhelming to be able to objectively understand the differences between the different models. We experimented with different word and sentence level embeddings like Word2Vec (both CBOW and Skip-gram), GloVe (Pennington et al., 2014), FastText (Joulin et al., 2016), ELMo (Peters et al., 2018), InferSent[6], BERT, and ESG BERT. Finally fine-tuned ESG BERT embeddings performed best on our classification task. Therefore ESG BERT vectors are appended with above set of selected feature vectors which are then passed to the classification model.

4.3 Classification Model

Sub-task1 features and Sub-task2 features are used to train classifiers for their respective classification tasks. A classification model takes vectors generated from task dependent features corresponding to each task data as an input and generate the predicted results corresponding to each input. To find the best classifiers, we test several widely used classification methods including Logistic Regression, Gradient Boosting and XGBoost Classifier (all in the standard scikit-learn implementation). We gradually augmented the models by adding the features one by one and computed the scores. From the experimental studies, we found that linear classifier performed the best. Hence, we selected Logistic Regression as the classifier in our submitted systems. This observation is consistent with findings in the FinSim 2020 and FinSim 2021 shared tasks, that models learning linear boundaries perform better for these tasks (Mansar et al., 2021) (El Maarouf et al., 2020).

5 Metrics

The most important part of any system is to choose the most accurate model. The metrics used for

[6]https://github.com/facebookresearch/InferSent

S.No.	Model	Accuracy	MR
1	Baseline 1	0.4920	2.2146
2	Baseline 2	0.775	1.4728
3	BERT + CL	0.7596	1.5038
4	BERT FCN + CL	0.7751	1.4573
5	ESG BERT + CL	0.7905	1.4582
6	ESG BERT FCN + CL	0.7906	1.4573
7	ESG BERT FCN + LR	0.8139	1.341
8	ESG BERT FCN + Word2vec + LR	0.8217	1.34
9	ESG BERT FCN + CS + Jaccard + LR	0.8139	1.33
10	ESG BERT FCN + Word2vec+ CS + Jaccard +LR*	**0.8217**	**1.3255**
11	ESG BERT FCN + Word2vec+ CS + Jaccard + PCA 100 + LR	0.8154	1.379
12	ESG BERT FCN + Word2vec+ CS + Jaccard + PCA 150 + LR	**0.8217**	1.35
13	ESG BERT FCN + Word2vec+ CS + Jaccard + PCA 200 + LR**	**0.8217**	**1.3215**

Table 1: Sub-task1 Evaluation results on the validation data ('*' and '**' represent *TCSWITM_1* and *TCSWITM_2* submission respectively.)

S.No.	Model	Accuracy	Recall
1	TFIDF	0.61	0.59
2	Wod2Vec (CBOW)	0.61	0.59
3	Word2Vec (SkipGram)	0.79	0.78
4	GloVe	0.77	0.75
5	FastText	0.81	0.79
6	ELMo	0.83	0.82
7	InferSent	0.80	0.78
8	BERT	0.92	0.91
9	ESG BERT	0.93	0.92
10	ESG BERT+NLP based features (TCSWITM_1)	**0.95**	**0.94**

Table 2: Sub-task2-Evaluation results on Validation data

the evaluation is provided by the organizers. For Sub-task1, the evaluation metric used is Accuracy and Mean Rank. The model is required to predict the concepts in the ranking order (from the most probable to the least probable) for each term. The Accuracy and Mean Rank is defined as follows -

$$Accuracy = \frac{1}{n} * \sum_{i=1}^{n} I(y_i = y_i^l[0])$$

$$MeanRank = \frac{1}{n} * \sum_{i=1}^{n} rank_i$$

Note that $rank_i$ corresponds to the rank of the correct label.

For Sub-task2, the evaluation metric considered is Accuracy and Recall. Recall (also known as sensitivity) is the fraction of relevant (here sustainable) instances that were correctly retrieved.

6 Experiments and Results

Sub-task1 - We were provided with 641 data samples which included terms and their corresponding concepts for model development and validation. We had split the data in 80:20 train-validation split using 5 fold cross validation technique which results into 5 sets containing 512 data points for training and 129 data points for validation. All proposed models were trained and validated on the generated training and validation dataset respectively. The test set provided contained a list of 145 terms without their concepts in json format. Table 1 shows the performance on validation set for sub-task1 and Table 2 shows the performance on validation set for sub-task2 in terms of mean rank and accuracy whereas Table 3 shows the evaluation results of hidden test data for both the tasks.

Baseline 1 is a distance-based classifier using custom embeddings (Mikolov et al., 2013) whereas

Baseline 2 uses logistic regression classifier over these custom embeddings (Mikolov et al., 2013) given in FinSim4-ESG shared task. BERT (pre-trained on general English corpora[7]) based classification model is fine-tuned on the downstream task by adding a classification layer (CL). It can be seen from the Table 1 that the performance has been improved by BERT pre-trained on Sustainability corpora (ESG) after incorporating a Fully Connected Layer (FCN). In the further experiments, Embedding from the resultant BERT model is used as features in Logistic Regression (LR) classifier. The concatenation of BERT features with Word2vec, Cosine Similarity (CS) and Jaccard resulted into the best Mean Rank as shown in Table 1. The ablation study confirmed the gravity of these features. As the data was limited as compared to the features, employing dimension reduction technique (PCA) benefited the performance by improving feature selection. This can also be seen in the test data results in Table 3.

Model	Accuracy	MR
Sub-task1 Baseline1	0.4620	2.2758
Sub-task1 Baseline2	0.7448	1.5241
Sub-task1 TCSWITM_1	0.7724	1.4620
Sub-task1 TCSWITM_2	**0.7793**	**1.4482**
Sub-task2 Baseline1	0.4976	NA
Sub-task2 Baseline2	0.8195	NA
Sub-task2 TCSWITM_1	**0.8731**	NA

Table 3: Evaluation results on Hidden Test data

From Table 3, it can be observed that ESG-BERT combined with all other features along with PCA outperforms other models in terms of mean rank and accuracy. It is clear from the table that all our models performs better than the baseline models.

We have also observed misalignment amongst the terms and concepts in the given training data. For example, the alignment given in the training data is "CO2 Equivalent Emissions Indirect, Scope 3" - *Biodiversity*, "Accident spills" - *Energy efficiency and renewable energy*, "Gender diversity" - *Recruiting and retaining employees (incl. work-life balance)* etc. But as per our understanding the ideal assignment should be "CO2 Equivalent Emissions Indirect, Scope 3" - *Emissions*, "Accident spills" - *Injury frequency rate* and "Gender diversity" - *Diversity & Inclusion*.

Sub-task2- We split the data in 80:20 train-validation split using 5 fold cross validation technique which results into 5 sets. All proposed models were trained and validated on the generated training and validation dataset respectively. Sentence embeddings performed better than word based embeddings as shown in Table 2. One of the main reason is that word embeddings like Word2Vec and GloVe are unable to encode unknown or out-of-vocabulary words. Moreover word based embeddings don't take into consideration the order of words in which they appear which leads to loss of syntactic and semantic understanding of the sentence. Finally we used BERT-based features as our main feature set. BERT-based features when combined with above set of rule-based and NLP-based features (like presence of specific NERs (like organisation, date entity), and, word count of document, etc.) provided a significant improvement in performance of the classification system.

7 Conclusion

As part of FinSim4-ESG shared task on Learning Semantic Similarities for Financial Domain, we attempt to solve the problems of finding the most relevant ESG concept for each given domain term and classify a given sentence into sustainable or unsustainable based on provided training data. In sub-task1, in order to utilize contextual word embeddings along with sustainability resources to elaborate an ESG taxonomy, we proposed the use of semantic similarity features along with BERT embeddings to segregate domain terms into a fixed number of class labels. For sub-task2, several linguistic and NLP features along with BERT embeddings were used for classification. In our experiments we observed that linear classifier models like logistic regression performs the best. We have also compared our models with the given baseline accuracy and found that their performance is far superior to it. We have also shown a detailed ablation study for the proposed models. In future, we are planning to use better data augmentation techniques and explore the possibility of using the ontology hierarchy and definitions available as part of the EU taxonomy.

[7] https://huggingface.co/bert-base-uncased

References

Vivek Anand, Yash Agrawal, Aarti Pol, and Vasudeva Varma. 2020. Finsim20 at the finsim task: Making sense of text in financial domain. In *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, pages 104–107.

Emmanuele Chersoni and Chu-Ren Huang. 2021. Polyu-cbs at the finsim-2 task: combining distributional, string-based and transformers-based features for hypernymy detection in the financial domain. In *Companion Proceedings of the Web Conference 2021*, pages 316–319.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ismail El Maarouf, Youness Mansar, Virginie Mouilleron, and Dialekti Valsamou-Stanislawski. 2020. The finsim 2020 shared task: Learning semantic representations for the financial domain. In *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, pages 81–86.

Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning semantic hierarchies via word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1199–1209.

Tushar Goel, Vipul Chauhan, Ishan Verma, Tirthankar Dasgupta, and Lipika Dey. 2021. Tcs_witm_2021@ finsim-2: Transformer based models for automatic classification of financial terms. In *Companion Proceedings of the Web Conference 2021*, pages 311–315.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Juyeon Kang, Ismail El Maarouf, Sandra Bellato, and Mei Gan. 2021. Finsim-3: The 3rd shared task on learning semantic similarities for the financial domain. In *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*, pages 31–35.

Vishal Keswani, Sakshi Singh, and Ashutosh Modi. 2020. Iitk at the finsim task: Hypernym detection in financial domain via context-free and contextualized word embeddings. *arXiv preprint arXiv:2007.11201*.

Youness Mansar, Juyeon Kang, and Ismail El Maarouf. 2021. The finsim-2 2021 shared task: Learning semantic similarities for the financial domain. In *Companion Proceedings of the Web Conference 2021*, pages 288–292.

Aleix M Martinez and Avinash C Kak. 2001. Pca versus lda. *IEEE transactions on pattern analysis and machine intelligence*, 23(2):228–233.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Kim Anh Nguyen, Maximilian Köper, Sabine Schulte im Walde, and Ngoc Thang Vu. 2017. Hierarchical embeddings for hypernymy detection and directionality. *arXiv preprint arXiv:1707.07273*.

Nhu Khoa Nguyen, Emanuela Boros, Gaël Lejeune, Antoine Doucet, and Thierry Delahaut. 2021. L3i_lbpam at the finsim-2 task: Learning financial semantic similarities with siamese transformers. In *Companion Proceedings of the Web Conference 2021*, pages 302–306.

Yulong Pei and Qian Zhang. 2021. Goat at the finsim-2 task: Learning word representations of financial data with customized corpus. In *Companion Proceedings of the Web Conference 2021*, pages 307–310.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Jan Portisch, Michael Hladik, and Heiko Paulheim. 2021. Finmatcher at finsim-2: hypernym detection in the financial services domain using knowledge graphs. In *Companion Proceedings of the Web Conference 2021*, pages 293–297.

Ke Tian and Hua Chen. 2021. aiai at the finsim-2 task: Finance domain terms automatic classification via word ontology and embedding. In *Companion Proceedings of the Web Conference 2021*, pages 320–322.