LUND UNIVERSITY

SCHOOL OF ECONOMICS AND MANAGEMENT

Multi-Label Classification of Sustainability Topics in Media Coverage: A Comparative Study of Transformer Models

James McGoldrick

Department of Economics

Lund University School of Economics and Management

Supervisor: Muhammad Qasim

DABN01 - Master Thesis 15 ECTS

May 2025

Abstract

This study examines the use of transformer-based language models for multi-label classification of Environmental, Social, and Governance (ESG) topics in media coverage of DAX-listed companies. Using a dataset of approximately 11,500 third-party media texts, thirty ESG topics were assigned to documents through a pipeline combining lemmatised keyword matching and semantic similarity.

Four models were compared: BERT, FinBERT, RoBERTa, and DistilBERT. All of them were fine-tuned on the same conditions on the same ESG taxonomy and evaluation framework. FinBERT performed the best overall, which may reflect the benefits of domain-specific pretraining on financial texts. DistilBERT performed well too despite its smaller size, which showed that smaller models can compete with proper fine-tuning. RoBERTa and BERT both performed well but had lower results compared to the other models.

The study also highlighted the importance of threshold tuning since each model reached its best performance at a value lower than the standard of 0.50. The tuned thresholds significantly improved both micro and macro F1 scores.

These findings shows the successful application of transformer-based models to ESG classification within external media outlets. The results are a contribution to applied NLP research on sustainability and provide a reproducible approach to large-scale ESG text classification.
Keywords: ESG classification, multi-label learning, BERT, FinBERT, NLP

Acknowledgements

I would like to thank my supervisor, Muhammad Qasim, for the steady guidance, constructive feedback, and encouragement throughout this thesis.

I would like to express my deepest gratitude to my family for their unwavering support, encouragement, and belief in me throughout this journey. This thesis is as much yours as it is mine.

To my classmates and friends, Fynn, Paolo, Viktorija, Elide, Oskar, Nikolas, Diego, Analiz, Dennis, Dagny, Sophie, Dilay, Ingi, Arnar, Denisse, Alina, thank you for the conversations, the laughs, and the shared frustrations that made this process feel a lot less solitary. Your support, whether academic or simply knowing when to take a break, made a real difference.

A special thanks to Viktorija Pasichnyk for being such a steady presence in my life. Your calm nature, your quiet kindness, and the way you are always there have meant more than I can put into words. I feel incredibly lucky to have had you around this past year.

I am especially grateful to Fynn Reinders, it is hard to imagine this year without you. I am grateful for all the time we spent together. Whether we were out and about or just doing nothing in particular, you made each day something I looked forward to. I am really glad that we got to share so much of this time together and hope we get to share more good times in the future.

And finally, to Paolo Totaro, thank you for being a constant through it all. We have worked on so much together over the past year, and I honestly don't think I would have enjoyed any of it nearly so much without you. Even when nothing was happening, it never was boring with you there. I really appreciate your company and it has been lovely to have all of this to experience with you.

CONTENTS

1 Introduction

1.1 Background and Context

Environmental, Social, and Governance (ESG) issues have gained significant attention in recent years (Khan et al., 2016; Dyck et al., 2019; Liang and Renneboog, 2020), not only from investors and regulators but also from consumers and the general public. With growing global attention on sustainability, companies are under increasing pressure to demonstrate and report their ESG performance and approach. Such issues are not only discussed in official reports but also in the broader public domain including media coverage and investor commentary.

Despite the growing volume of information related to ESG, the lack of a standardised framework for discussing these topics remains a major issue. Public discourse around ESG varies widely: some sources offer systematic reporting, while others embed ESG information in broader financial or general social commentary. Terminology can also differ significantly between media outlets, industries, and authors and is generally difficult to extract and classify information consistently. As a result, manual analysis of ESG-related texts is time consuming and prone to inconsistency (Berg et al., 2022), whereas rule-based or keyword-driven methods tend not to capture the subtleties and implicit references that are common in sustainability discourse (Loughran and McDonald, 2011).

Recent advances in natural language processing (NLP) have enabled us to tackle these challenges. Transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers) have demonstrated strong capabilities in understanding context, handling long text-form unstructured text, and high-accuracy classification. These types of models are also well-suited to deal with multi-label classification problems because it is not unusual for a single document to cover multiple overlapping ESG themes. This is especially relevant in public discourse related to ESG, where a single article or commentary can simultaneously refer to issues such as climate change, human rights, diversity, and governance.

In recent years, BERT-based models have been employed in various NLP applications within the sustainability domain (Chung and Latifi, 2024), such as sentiment analysis (Saxena et al., 2024), taxonomy mapping (Ong et al., 2025), and topic classification (Linhares, 2023). However, the majority of studies either utilise domain-specific corpora such

as regulatory filings or curated sustainability reports and use narrowly defined taxonomies. Comparatively little work has assessed and compared the performance of different BERT variants on unstructured, external ESG discourse under consistent evaluation conditions. Given the complexity and diversity of ESG text, it remains unclear which models offer the most reliable performance in practice.

This thesis contributes to that conversation by applying deep learning methods to ESG topic classification. It builds on recent progress in natural language processing to address the challenges of identifying multiple overlapping sustainability themes within context-rich, unstructured media texts.

1.2 Aim of the Thesis

The aim of this thesis is to develop and evaluate a deep learning framework to automatically classify ESG-related texts into multiple relevant topics. This problem is framed as a multi-label classification problem that reflects the structure of sustainability reports where environmental, social, and governance issues often arise in a single text.

To achieve this, the thesis fine-tunes four pre-trained transformer-based language models: BERT, FinBERT, RoBERTa, and DistilBERT. The models are trained on a semi-automatically labelled dataset of ESG-related documents using a customised taxonomy of sustainability topics. The dataset is drawn from publicly available corporate communications and third-party media content, with the vast majority consisting of external news articles and commentary, and only a small subset comprising official company reports. All models are trained using a consistent pipeline, with separate threshold optimisation and hyperparameter tuning to ensure fair and robust comparison across architectures.

The objective is to examine whether domain-specific pretraining, architectural simplification, or general-purpose design has a meaningful impact on ESG topic classification performance. In particular, this analysis focuses on the extent to which models can handle imbalanced data, identify context-sensitive topics, and produce reliable multi-label predictions.

The broader goal is to contribute to the growing literature on applying natural language processing to ESG analysis and to provide a reproducible foundation for future work in automated sustainability classification of sustainability-related coporate communication.

1.3 Outline of the Thesis

The rest of the thesis is organised as follows: Chapter 2 reviews the relevant literature on ESG reporting, multi-label text classification, and transformer-based models. Chapter 3 describes the dataset used, the ESG labelling schema, and preprocessing steps applied. Chapter 4 outlines the modelling approach, including the selected BERT variants, activation functions, training setup, and optimisation strategy. Chapter 5 presents the empirical results, including performance comparisons across models and hyperparameter tuning outcomes. Finally, Chapter 6 summarises the findings and offers suggestions for future research directions.

2 Literature Review

In this section, I summarise the main theoretical concepts and provide an overview of existing literature on ESG reporting, multi-label text classification, and the application of BERT-based models in natural language processing.

2.1 ESG Reporting and Unstructured Text Data

Information on a company's governance, social responsibility and environmental impact can appear in a variety of formats, including corporate sustainability reports, press releases, and third-party news media. Although investors and regulators increasingly rely on such data to evaluate non-financial performance (Linhares, 2023), much of it is presented as lengthy, qualitative, and unstructured text.

The presentation of ESG information can vary widely between organisations and document types, from structured reports to loosely framed new articles, with no single universal format (Amel-Zadeh and Serafeim, 2018) . Some reports can accordingly make clear references to ESG issues, while others may unintentionally or intentionally raise them in a broader discussion. Such inconsistency has been a challenge in conducting any structured analysis. Approaches using hard rules or simple keyword scans often end up looking beyond the subtlety or various ways the same topic can be phrased (Mehra et al., 2022).

At the same time, there is greater volume and complexity in ESG information, driven by cross-border regulation and investor pressure to disclose sustainability performance. As a result, there has been increased interest in applying NLP techniques to extract structured meaning from unstructured ESG text (Sahu et al., 2025). Pre-trained transformer-based models such as BERT (Devlin et al., 2019), FinBERT (Araci, 2019), RoBERTa (Liu et al., 2019) and DistilBERT (Sanh et al., 2019) have achieved impressive performance in different language understanding tasks and are now being used in ESG settings, for example, issue classification, sentiment extraction, and sustainability scoring.

2.2 Multi-Label Text Classification

Multi-label text classification (MLTC) is a supervised learning task in which an input can be assigned multiple labels simultaneously. It differs from single-label classification,

where the labels are exclusive of each other. MLTC is especially suitable for use in complex and overlapping theme applications, such as legal, biomedical, and financial text, where documents will cover several themes in one story (Zhang and Zhou, 2014).

Traditional approaches to MLTC include binary relevance and classifier chains, which breaks down the problem into multiple single-label classification problems. Although straightforward, these methods neglect correlations between labels, and easily suffer from label imbalance and few positive instances (Tsoumakas and Katakis, 2009). Recent attempts have incorporated neural architectures, such as convolutional and recurrent networks, to learn semantic features from text directly (Liu et al., 2016). These approaches improved representation learning, although they were not that effective at handling long-range dependencies or word meaning that is highly contextual.

Transformer model development has greatly improved MLTC performance on domain and general testing. BERT and its extensions, when with sigmoid output layers and binary cross-entropy loss adapted, have demonstrated strong performance for multi-label classification on long-form documents and highly imbalanced label distributions (Yin et al., 2019). Studies such as (Ruberg, 2021) and (Linhares, 2023) expanded such models to multi-label classification for ESG and sustainability cases, using taxonomies such as GRI (Global Reporting Initiative) and MSCI (Morgan Stanley Capital International) for labelling corporate and media texts.

Transformer-based models also support token-level context encoding, which is beneficial for representing overlapping or implicit labels in multi-label settings. They can also be generalised to hierarchical label spaces, where certain labels are nested under broader categories. However, challenges persist in the handling of infrequent labels, tuning thresholds, and measuring the performance on imbalanced data. Measurements such as micro- and macro-averaged F1 score are frequently used in the literature to solve these issues (Zhang and Zhou, 2014; Yin et al., 2019).

2.3 Applications of BERT-Based Models in ESG Classification

Existing ESG classification research has evolved from simple experimentation with the effectiveness of BERT-based models to considering how the models can be adjusted to address the special needs of the ESG text. Existing research focuses on multilingual settings, label standardisation, document organisation, and ensemble learning. The re-

search yields significant insights on increasing classification accuracy and robustness for real-world ESG data. (Linhares, 2023) addressed multilingual ESG labelling in the Multi-Lingual ESG Issue Identification (ML-ESG) shared task, where texts were labelled with the MSCI ESG taxonomy. Their RoBERTa-based classifier performed best on English data and highlighted the impact of training corpora and tokenisation variation across languages. Interestingly, more advanced semantic similarity approaches underperformed relative to standard fine-tuning, demonstrating the strength of direct transformer-based classification.

Mehra et al. (2022) were interested in prediction adjustments to environment risk scores from 10-Q financial statements. Their pipeline went beyond finetuning and involved selecting contextually comparable segments through sentence similarity prior to classification based on FinBERT. This assisted in addressing the sparsity and scattering of ESG-related data in lengthy documents, particularly where the same was limited to only a part of it being relevant for environmental outcomes.

In yet another line of work, ESG-labeled corpora that are structured on a framework similar to that of the European Union's Sustainable Finance Taxonomy and the United Nations Sustainable Development Goals (SDGs) have been utilised by researchers. Overlapping or hierarchical label structures in the datasets used are common in most cases, so researchers have applied transformer-based models such as RoBERTa where classification heads have been optimised for multi-label prediction. For example, Angin et al. (2022) pre-trained RoBERTa on Open-Source SDG Community Dataset, a large human-annotated text dataset designed for supervised classification towards the 17 UN SDGs. The authors' model performed excellent classification, with an F1 score of 0.92 for multi-label. The research further highlighted the value of label curation and agreement-based filtering in reducing annotation noise and increasing semantic consistency within the overlapping SDG categories.

Other researchers have explored ensemble methods to improve model generalisation. Veeramani et al. (2023) experimented with early and late fusion methods using multilingual models such as mBERT, FlauBERT, and ALBERT. While macro-F1 improvements were limited, their results emphasised the need for robustness when handling noisy and weakly labelled ESG text, particularly across languages and data sources.

Finally, model interpretability has started to make appearances. Model explanation

architectures and attention visualisation are now ever-present in research in areas like finance where there is an interest in regulation that requires models to be transparent. Most work is still in its infancy and most ESG NLP pipelines have been focusing on metrics over explainability.

Overall, the literature shows an improvement from demonstrating the feasibility of using BERT-based models to ESG text to building more sophisticated pipelines with a feature to meet the challenges posed by real corpora. These include label sparsity, multilingual variation, to needing a fine-grained level of control over thresholding and classification behaviour. While performance continues to improve across experiments, open problems persist in the areas of interpretability, cross-domain generalisation, and the standardisation of ESG taxonomies.

Applications to broader public ESG discourse, such as media narratives or third-party commentary, remain limited in the literature. Most work has focused on structured, company-authored disclosures.

2.4 ESG Taxonomies and Labelling Strategies in NLP

One of the most fundamental challenges of ESG text classification is the absence of a shared taxonomy to describe ESG themes. Studies differ in how they categorise sets of ESG labels, depending on the kind of data and the intended application. Others adapt straight from formal frameworks such as MSCI ESG Ratings (Linhares, 2023), the United Nations Sustainable Development Goals (Angin et al., 2022), or the GRI, while others formalise personal taxonomies suitable for the corpus at hand (Ruberg, 2021).

Most of the literature uses a pre-defined set of ESG labels, which are typically employed in deterministic or rule-based labelling methods such as keyword matching or phrase detection. For example, Ruberg (2021) created a GRI-based labelling scheme for sustainability reports and used exact string matching for labelling.

Although these methods are transparent and reproducible, they will inevitably fail to capture indirect references or soft language and therefore have low recall.

In order to overcome this constraint, there have been some recent suggestions of embedded and semantic solutions. Mehra et al. (2022) used a sentence similarity filtering step to exclude ESG-irrelevant portions of long 10-Q reports before applying FinBERT classification. Ong et al. (2025) proposed using ESGSenticNet, which uses a semantic-

structured framework with a concept ontology along with transformer-based reasoning in automatically tagging ESG concerns in disclosure reports. These methods are less versatile but introduce noise, are uninterpretable, and need high-quality training sets to function.

Label imbalance is also a common issue. Labels such as climate change, diversity, or corporate governance are usually the most common in datasets, whereas labels such as biodiversity or water use appear sporadically. This imbalance affects both model training and evaluation. To counter this, most papers present both micro- and macro-averaged F1 scores so that the performance is not overwhelmed by common labels (Huang et al., 2021; Veeramani et al., 2023).

There is no consensus in the literature on the best label granularity. Some models use broad categories (for example E/S/G), while others use fine-grained hierarchies with 20–50 different topics. Fine-grained taxonomies give more informative supervision but result in sparsity and potentially redundant labels. This makes it not only difficult for model learning, but also for human interpretation of classification outputs (Ruberg, 2021).

In general, the majority of studies favour predefined taxonomies with rule-based labelling due to their scalability and transparency. At the same time, recent work continues to explore semantic methods to address flexibility and coverage, although standardisation among ESG NLP systems remains limited to date, particularly in applications involving third-party media sources.

2.5 Summary and Research Gap

The literature reviewed illustrates a broad diversity of BERT-based model solutions to ESG text classification. Studies differ greatly in label schemes, model architectures, and evaluation strategies, such that results are neither comparable nor generalisable. While transformer models have shown good performance for multi-label ESG use cases, there is little agreement on the definition of ESG categories, the labelling of instances, or the consistent comparison of model outputs.

There is an evident research gap in the form of a shortage of benchmarking among baseline transformer models on a fixed, interpretable ESG taxonomy. Most existing work either suggests complex labelling pipelines, uses semantic methods without transparency, or evaluates on corpora with limited reproducibility. Comparatively few studies system-

atically contrast model behaviour under identical labelling and evaluation conditions.

This thesis fills that gap by benchmarking BERT, FinBERT, RoBERTa, and DistilBERT on ESG-related disclosures from the DAX ESG Media Dataset. Unlike previous studies, this dataset consists mainly of external narratives rather than formal corporate disclosures. A controlled taxonomy developed for this study is applied and all models are evaluated under identical training, tuning, and validation settings. The goal is to quantify baseline model performance on third-party ESG-related media texts and provide a reproducible framework for future ESG classification research in public corporate discourse.

3 Data

3.1 Source of Data

The data used in this thesis is the DAX ESG Media Dataset which is made available publicly on Kaggle at: https://www.kaggle.com/datasets/equintel/dax-esg-media-dataset/data. It includes approximately 11,500 English-language documents published between 2018 and 2021, all related to companies listed on Germany's DAX 30 stock market index. The overall focus of the dataset is corporate sustainability and communication related to ESG.

The documents are drawn from two sources of information. The first source involves firm-authored documents such as sustainability reports, annual reports, and press releases. These capture the way firms determine to report their ESG intentions, plans, and activities. The second source involves third-party material such as news stories and external commentary offering an objective viewpoint of corporate ESG activities. While both types are technically included, the dataset is overwhelmingly composed of external material (approximately 99%), with only a small minority (approximately 1%) consisting of official company disclosures.

Each document includes contextual information such as company name, publication date, document title, and domain retrieved from. Most importantly, the dataset includes the full-text content of each document, which is the primary input to all the classification tasks in this thesis.

3.2 Variable Overview

Each row in the data is one document related to ESG topics published by a DAX 30 company. In addition to the main text content, each record includes several metadata fields providing contextual information about the source and nature of the document. The key variables used within this thesis are described below:

Table 1: Original Variables in the ESG Dataset

Variable	Description
`company`	The name of the organisation that is linked with the document.
`title`	The document title or name, which can be linked to a sustainability report, news article, or press release.
`content`	The full text of the document. This is the main input for the classification task and typically includes discussions of ESG policies, disclosures, risks, and performance.
`datatype`	A categorical variable stating if the document is an internal corporate report, external media, or general disclosure.
`date`	The publication date of the document.
`domain`	The originating domain where the document was obtained, i.e., a corporate site or a newspaper publisher.
`symbol`	The stock ticker symbol of the associated company.
`url`	The original URL of the document, when available.

Table 2: Additional Variables Created During Preprocessing

Variable	Description
`text`	A combined version of the `title` and `content` fields. This was used as the input text for classification, providing more context than the body text alone.
`esg_labels`	A list of ESG topics automatically assigned to each document based on the taxonomy described in Section 3.2.
`Topic columns`	A set of 30 binary columns corresponding to each ESG topic. A value of 1 indicates the topic was assigned to the document; 0 indicates it was not. These columns form the multi-label classification targets used during model training and evaluation.

Additional variables were created during preprocessing to support the classification

task. These variables together provide the inputs and targets needed for multi-label classification of ESG topics.

3.3 ESG Taxonomy Description

This thesis uses a 30-topic ESG taxonomy to enable supervised multi-label classification. The labels were obtained from ESG frameworks like GRI, MSCI, and the UN Sustainable Development Goals and narrowed down through analysing common themes within the frameworks and in the DAX ESG Media Dataset.

Each label in the taxonomy is accompanied by a brief definition and some sample phrases, which help define its scope. This structure ensures that themes are clear, relevant, and grounded in actual ESG discussions. The taxonomy was crafted to take a middle-ground approach between frequency and specificity. Problems were made specific enough to identify important ESG problems (e.g labour rights and fair wages or board composition and independence) without being too detailed and resulting in sparse label distributions and decreased classification performance.

Labels were assigned automatically using a combined matching pipeline. First, the representative phrases for each label were lemmatised and compared against the document text to identify direct or partial matches. In parallel, semantic similarity was calculated between ESG topic descriptions and document segments using Sentence-BERT embeddings. The final label set for each document was constructed by taking the union of matches from both methods. This approach improved topic coverage, particularly in cases where ESG issues were discussed using indirect or varied language.

The result of this process was a binary label matrix in which each of the 30 ESG topics is represented by a dedicated target column, indicating whether the topic was assigned to the document. This served as the target label set for all model training and evaluation tasks throughout the thesis.

3.4 Preprocessing

The original dataset was also run through a series of preprocessing tasks towards preparing input text and label targets for the multi-label classification problem. The operations included the combination of text fields, label mapping creation in alignment with the ESG taxonomy, and preparing data for model training.

The first step was to combine the title and content fields for each document into a new variable called text. This allowed both the headline and the full body of the document to be included in the model input, giving better context for identifying ESG-related content. Documents with missing or empty text were removed at this stage.

Labelling was performed using a two-stage pipeline against Section 3.2's 30-topic ESG taxonomy. The first stage comprised lemmatised keyword matching. Every ESG topic contained some defining phrases, which were converted to their base form using spaCy's lemmatiser. A document was labelled if it contained at least two of the matching lemmas for that label. This stage was used to eliminate straightforward and overt mentions of ESG issues.

The second half of the pipeline was designed to use semantic similarity to pick up on less direct mentions. Documents were divided into overlapping windows by a sliding window of 2,000 characters with a stride of 1,000 characters. Every chunk was encoded with the Sentence-BERT model all-MiniLM-L6-v2, and cosine similarity was computed between the chunk encodings and precomputed encodings of the ESG topic descriptions. The top 5 most similar topics were retrieved per document, and every topic with an average similarity score greater than 0.45 was selected as a fallback label. The values for the similarity score and most similar topics were selected based on a small grid search across threshold values (0.30-0.45) and top candidates (3-8). Each combination was evaluated on a 100-document sample by computing average Jaccard similarity between assigned and reference labels. The combination of 5 and 0.45 yielded consistent coverage and high overlap, so it was chosen for all semantic fallback labelling.

The final set of labels per document was formed by taking the union of the keyword matches and top semantic matches. There was no additional ranking or confidence weighting applied; any topic appearing in either the keyword match or embedding match set was included in the final labels. This was to achieve a balance between coverage and precision, merging more confident direct matches with a broader semantic fallback to pick up varied ESG language.

Once labels were applied, the data was transformed to a binary format. All 30 ESG topics were provided in individual columns. When the label existed within the document, 1 was used, and when the label did not exist, 0 was used. This binary label matrix served as the target during model training and testing for all models.

Following preprocessing, the dataset was randomly split into an 80/20 training and validation set for model development and evaluation.

4 Methods

4.1 BERT-Based Architecture

Transformer-based language models have become the standard approach for a wide range of natural language processing tasks, particularly in text classification. One of the most influential among these is BERT (Bidirectional Encoder Representations from Transformers), which introduced a shift in how contextual information is learned by training on large corpora such as English Wikipedia and BooksCorpus. Its bidirectional attention mechanism allows the model to take into account both the left and right context of a word at the same time. This makes it well-suited to understanding subtle or context-dependent language.

BERT-based models are particularly well-suited to multi-label classification, in which documents can be assigned to multiple overlapping classes. In contrast to conventional approaches based on bag-of-words features or fixed-length n-grams, transformer models do capture semantic relationships and long-range dependencies in the input. This is particularly crucial for the classification of lengthy unstructured documents like corporate sustainability reports and third-party media, where information relevant to any particular class may be scattered unevenly throughout the text.

In this thesis, four transformer models are compared for ESG topic classification: BERT, FinBERT, RoBERTa, and DistilBERT. Both domain-specific and general-purpose models are used. All of the models are fine-tuned on the same labelled dataset to ensure consistent evaluation of classification performances.

4.1.1 BERT

BERT, developed by Devlin et al. (2019), marked a significant advancement in NLP by leveraging bidirectional self-attention mechanisms within transformer architectures. Compared to traditional language models that read text from left-to-right or right-to-left, BERT conditionally employs both directions at the same time so that it can discover intricate interdependencies within a sentence. It is therefore effective for tasks such as document classification, where the meaning tends to be context-dependent with respect to the surrounding environment.

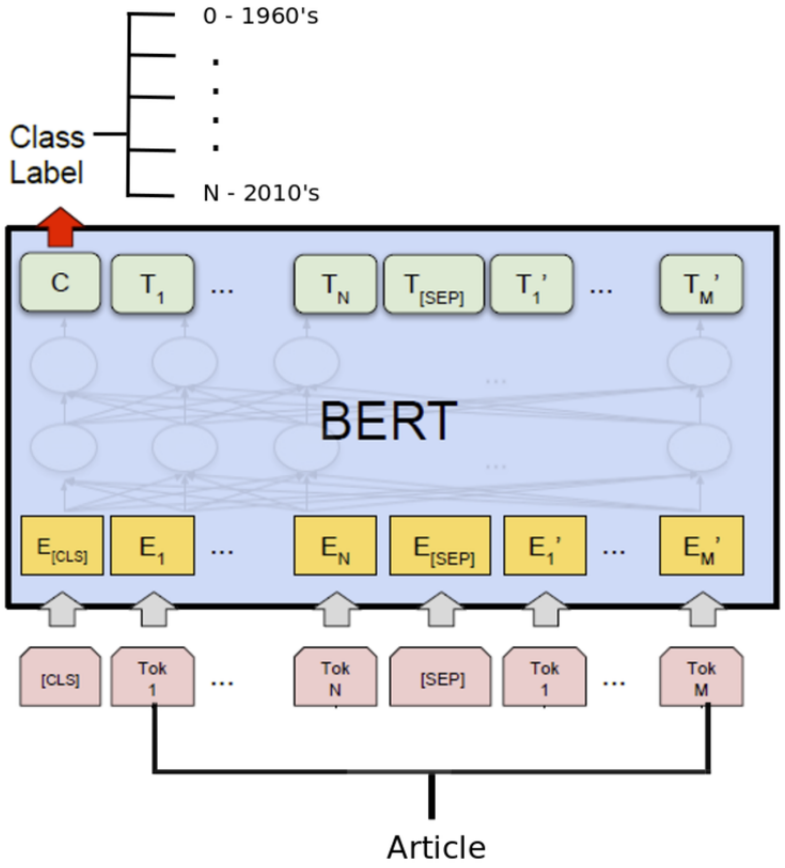

Figure 1: Overview of the BERT model architecture used. The model consists of embedding and positional encoding layers followed by twelve transformer encoder blocks. A sigmoid-activated output layer is applied for multi-label ESG classification.

The general architecture of the BERT model is shown in Figure 1 (Devlin et al., 2019). The specific model used in this paper is the bert-base-uncased, which is a 12-layer pre-trained model on BooksCorpus and English Wikipedia doing two unsupervised tasks. Firstly, masked language modelling (MLM) which involves randomly masking tokens and training the model to predict them in the context in which they are used, and next sentence prediction (NSP) which allows the model to learn relations between the next sentences.

For multi-label ESG classification, BERT is trained with a classification head that consists of a linear layer and a dropout layer and sigmoid-activated linear layer. In contrast to softmax used in single-label scenarios, sigmoid allows single prediction for each label to support multiple assignments per document at once.

Since BERT has a maximum input length of 512 tokens, and sustainability documents often exceed this limit, a sliding window strategy is applied. Each document is split into overlapping chunks, which are passed through the model independently. The resulting

predictions are aggregated to form the final label set for the entire document.

BERT serves as the baseline model in this thesis. Its general-purpose design and strong performance on a wide range of NLP tasks make it a useful reference point when comparing against more specialised architectures.

4.1.2 FinBERT

Although general-purpose models such as BERT perform well on a broad range of language tasks, domain-specific variants have been shown to increase accuracy in specific environments. FinBERT is one such model and is specifically designed for financial text classification. Introduced by Araci (2019), it has the same structure as bert-base-uncased but also undergoes additional pretraining on financial texts such as earnings reports, analyst reports, and financial news. This extended pretraining allows the model to capture more strongly the vocabulary and semantic patterns of finance language.

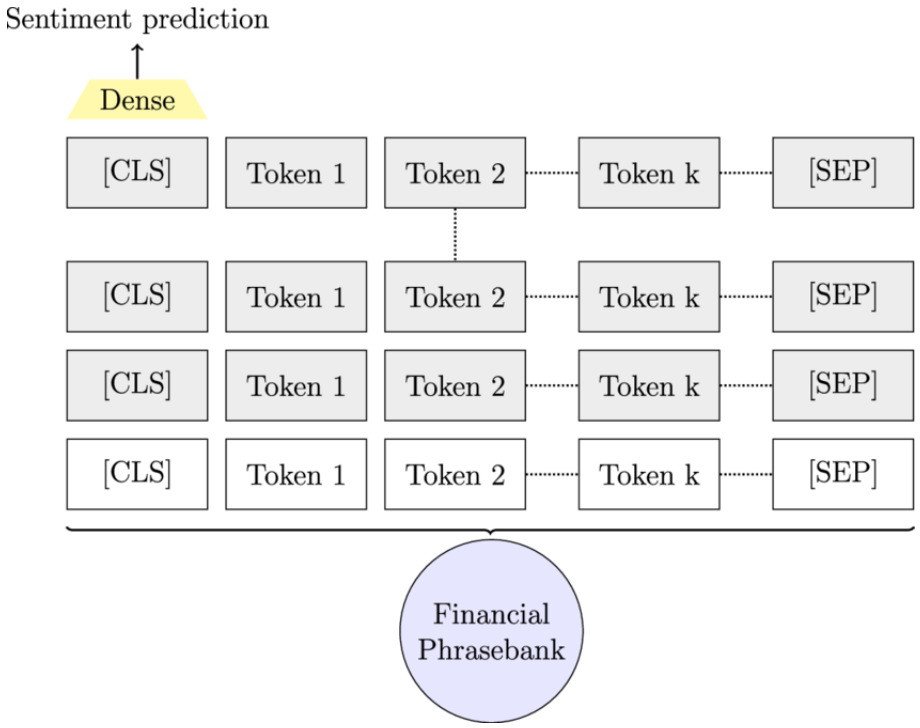

Figure 2: Overview of the FinBERT model architecture used. The model is based on the BERT encoder structure, with domain-specific pretraining on financial text. A dense classification layer with sigmoid activation is applied to support multi-label ESG topic prediction.

The architecture of FinBERT is shown in Figure 2 (Araci, 2019). Even though the figure is an example of sentiment classification, the underlying structure, e.g., the use of

the classification token, transformer encoder layers, and the final dense layer, is the same as used in this thesis.

FinBERT is included in this thesis because a majority of ESG documents are of a financial and regulatory nature. Sustainability reports of corporations are likely to use language related to capital distribution, regulatory compliance, risk of investments, and economic performance. Phrases such as "carbon risk exposure," "capital expenditure alignment," or "decarbonisation roadmap" may have more nuanced financial connotations that general-purpose models could be less effective at detecting. FinBERT, which has been trained in a domain-adapted environment, is better equipped to recognise and interpret such financially nuanced terminology.

The model was sourced from the HuggingFace Transformers library and fine-tuned on the same ESG-labelled dataset as the rest of the architectures. Fine-tuning is comparable to BERT's, using a dropout layer and sigmoid-activated linear classifier to enable multi-label output. This offers a controlled comparison to FinBERT with non-specialised baselines.

Due to the length of many ESG disclosures, the sliding window approach described in Section 4.1.1 was again used. Documents were segmented into overlapping chunks, each independently processed, with the resulting predictions aggregated to produce a final label assignment for the full document. This allows the model to have access to long-distance context without violating the input token limit.

The comparison of FinBERT with BERT allows domain adaptation to be measured directly within the ESG context. Even though BERT has extensive linguistic coverage, industry-specific training of FinBERT may make it more capable of finding ESG-concerned topics buried in high-level financial technicalities.

4.1.3 RoBERTa

RoBERTa (A Robustly Optimised BERT Pretraining Approach), introduced by Liu et al. (2019), is yet another modification of the baseline BERT architecture that focuses on improving pretraining efficiency as well as downstream task performance. Although RoBERTa retains BERT's core transformer design, it modifies the pretraining process in several ways: it drops the next sentence prediction (NSP) task, trains on much larger corpora such as Common Crawl, increases the maximum input sequence length, and employs

larger batch sizes and longer training durations. These changes result in a more stable optimisation and improved generalisation across a range of natural language understanding tasks.

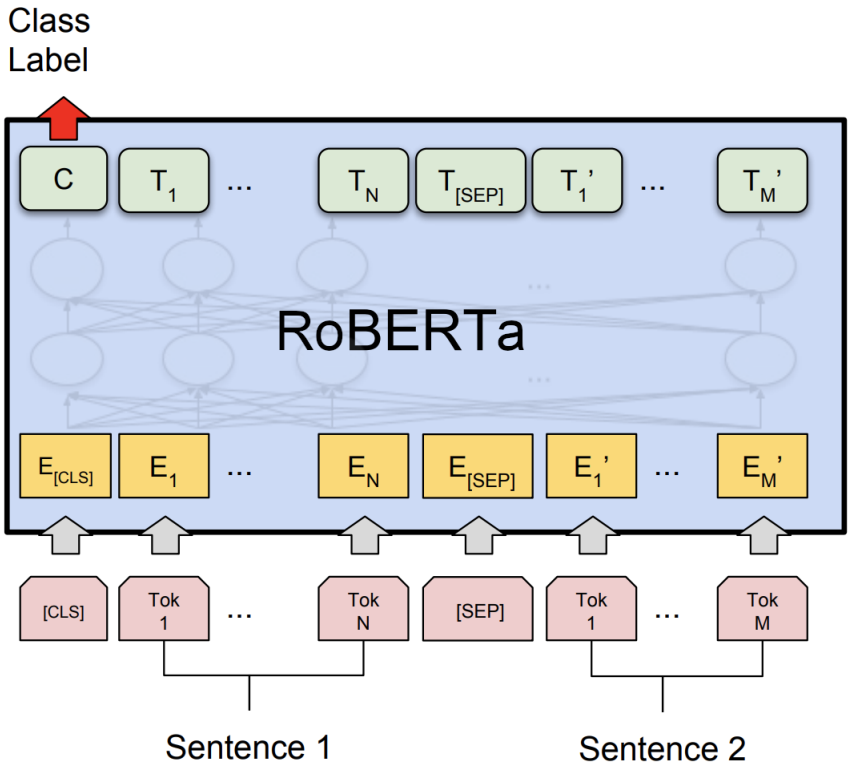

Figure 3: Overview of the RoBERTa model architecture. The input text is tokenised and passed through embedding and positional encoding layers, followed by multiple transformer encoder layers. The final hidden states are used by a sigmoid-activated classification layer to produce multi-label predictions for ESG topics.

An overview of the RoBERTa architecture is provided in Figure 3 (Khusuma et al., 2023), illustrating the sequence of components used in the implementation described below.

RoBERTa is included in this thesis to evaluate whether architectural and training-scale improvements made to a general-purpose model can enhance ESG topic classification performance. Corporate sustainability documents vary widely in structure, tone, and terminology, and BERT has been shown to exhibit instability when handling longer inputs. RoBERTa's more aggressive pretraining schedule may yield more robust resistance in these cases, particularly in dealing with subtle or scattered ESG-related language.

The model used here is roberta-base, which can be obtained from the HuggingFace Transformers library. It is trained on the same ESG-tagged dataset used on BERT and

FinBERT with an identical classification setup being a dropout layer followed by a linear layer with sigmoid activation. This setup allows for individual prediction of each ESG tag in a multi-label scenario. For the limit on input length (up to 512 tokens), the same sliding window approach is applied to partition documents into overlapping chunks, and predictions aggregated over chunks to yield final label assignments.

RoBERTa provides a useful baseline in this thesis for evaluating the impact of more extensive pretraining in general-purpose models. By comparing its performance with both BERT and FinBERT, the analysis investigates whether architectural refinements and large-scale pretraining alone can match the benefits of domain adaptation in specialised classification tasks such as ESG topic detection.

4.1.4 DistilBERT

DistilBERT, introduced by Sanh et al. (2019), is a compressed version of the original BERT architecture that retains most of its performance while significantly reducing computational cost. It is created through a process called knowledge distillation, in which a smaller model is trained to replicate the behaviour of a larger pretrained model. DistilBERT contains around 40% fewer parameters and 60% quicker inference compared to BERT, maintaining more than 95% of its language understanding across typical NLP benchmarks.

An illustration of the model architecture is provided below in Figure 4 (Izadi et al., 2021). Although the figure shows a general topic classification scenario, the structure with embedding layers, encoder blocks, and a dense output layer with sigmoid activation is equivalent to the implementation used in this thesis. The fully connected layer (FCL) takes the final hidden state from the encoder and maps it to a set of 30 outputs, each corresponding to one ESG topic in the classification schema. These outputs are then passed through a sigmoid function to generate probabilities for each topic so that the model can assign multiple labels to a single document. This approach makes the model suitable for ESG documents, as several themes can be discussed. A simplified view of the fully connected layer used for this purpose is shown in Figure 5. In this layer, each node from the input representation is connected to every node in the output layer. This allows the model to combine information across all features and produce a probability for each ESG topic.

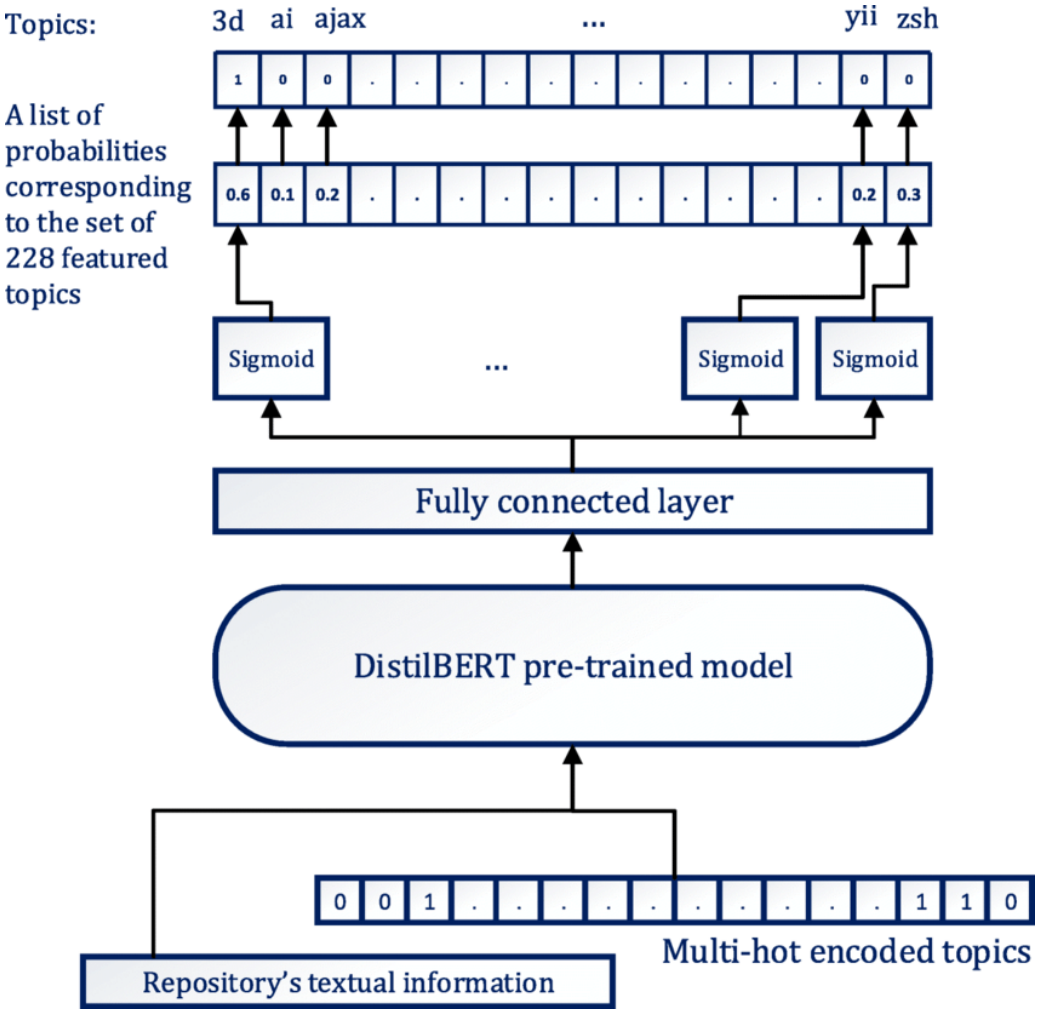

Figure 4: Overview of the DistilBERT model architecture used in this thesis. The model is a compressed version of BERT, retaining the encoder stack while reducing depth. A fully connected layer with sigmoid activation supports multi-label ESG topic prediction.

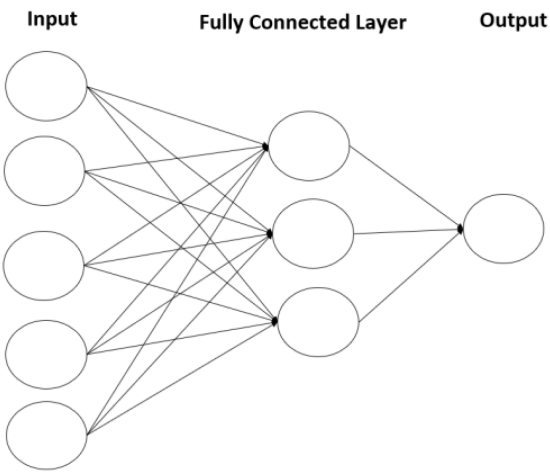

Figure 5: Structure of a Fully Connected Layer (FCL).

The thesis includes DistilBERT to examine the trade-off between model efficiency and classification performance when used to classify ESG documents. Although models like RoBERTa and FinBERT offer performance advantages via increased scale or domain adaptation pretraining, they are computationally more demanding. DistilBERT offers a lighter alternative that may be more suitable in hardware-constrained situations or real-time inference.

In this study, the distilbert-base-uncased model is obtained from the HuggingFace Transformers library and fine-tuned on the same ESG-labelled dataset used for the other architectures. The classification head follows the same structure: a dropout layer followed by a linear output layer with sigmoid activation. This allows for independent label predictions in a multi-label classification setting.

Due to the same 512-token input limitation present in other BERT-based models, the sliding window approach is applied here as well.

Including DistilBERT alongside BERT, FinBERT, and RoBERTa provides performance comparisons for model sizes and pretraining strategies. In particular, it investigates whether a smaller general-purpose model can match the performance of larger models on complex classification tasks on long specialised documents.

Together, these four transformer models represent a diverse set of pretraining strategies and architectural designs, ranging from general-purpose to domain-adapted, and from resource-heavy to lightweight implementations. By fine-tuning each model on the same dataset using an identical training setup, the analysis aims to isolate the impact of model choice on ESG classification performance. The following sections describe the technical configuration used during fine-tuning, including the activation function, optimisation strategy, and training procedure.

4.2 Activation Function

In multi-label classification tasks, each label is treated as an independent binary classification problem. This is distinct from single-label classification, where one label is selected from a mutually exclusive set. To support this structure, the final layer of each model in this thesis uses the sigmoid activation function.

The sigmoid function maps raw model outputs (logits) to values between zero and one. For a given input $z \in \mathbb{R}$, the sigmoid function $\sigma(z)$ is defined as:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{1}$$

This function is applied independently to each of the L output logits, where $L = 30$, the number of ESG labels in the classification task. The output is interpreted as the probability that a given ESG topic is present in the document.

The derivative of the sigmoid function, which is used during backpropagation for gradient-based optimisation, is given by:

$$\sigma'(z) = \sigma(z)(1 - \sigma(z)) \tag{2}$$

This derivative is computationally efficient in the sense that it can be expressed in terms of the sigmoid function itself, without the need for any more complex gradient computation. Since $\sigma(z)$ is being computed during the forward pass, the derivative may simply be reused during the backward pass, saving overall computational cost and allowing training to be performed more efficiently.

In combination with the sigmoid activation, the models are optimised using the binary cross-entropy loss function, which is well-suited to scenarios where multiple labels may be active simultaneously. For a single prediction $\hat{y}$ and true label $y \in \{0, 1\}$, the binary cross-entropy loss is defined as:

$$\mathcal{L}(y, \hat{y}) = -\left[y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y})\right] \tag{3}$$

In the case of multi-label, this loss is computed independently for each label and then averaged across all labels for each document.

The binary cross-entropy loss and sigmoid function usage allow the model to acquire label-specific decision boundaries but maintain the flexibility needed to assign more than one topic to a single document. This design reflects the nature of ESG disclosures, where documents routinely reference multiple overlapping themes such as climate risk, labour rights, and governance practices.

4.3 Training

All models were fine-tuned using the same multi-label ESG dataset and a consistent classification setup. Each document was tokenised and segmented using a sliding window

to comply with input length constraints. The model outputs were passed through sigmoid activation to generate per-label probabilities, and predictions were evaluated using binary cross entropy loss.

Although the training framework was consistent across models, several hyperparameters such as learning rate, dropout, and classification threshold were tuned individually. The following sections provide details on the optimisation strategy and the use of the Adam optimiser.

4.3.1 Optimisation

The training objective for all models in this thesis is to minimise binary cross entropy loss across 30 ESG topic labels, formulated as a multi-label classification task. Each document may be associated with several topics simultaneously, and each topic is treated as an independent binary classification problem. To enable this structure, the models output a probability for each label with sigmoid activation, and loss is computed per label before averaging over the set of labels.

Given a true label vector $\mathbf{y} = [y_1, y_2, \ldots, y_L]$ and the corresponding predicted probabilities $\hat{\mathbf{y}} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_L]$, the binary cross entropy loss function is defined as:

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{L} \sum_{i=1}^{L} [-y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i)] \tag{4}$$

This formulation is widely used for multi-label classification tasks where each label is conditionally independent (Goodfellow et al., 2016).This enables the model to learn distinct decision boundaries for each ESG topic and allow overlapping label assignments that are common in sustainability reporting.

During training, a starting threshold of 0.5 was used for converting model predictions to binary labels. This threshold was then optimised on the validation set following training to improve the balance between precision and recall. A grid search was performed over threshold values ranging from 0.1 to 0.9 in increments of 0.02. For each candidate threshold, F1 scores were computed, and the threshold that maximised the F1 micro score was selected for final evaluation. This step was done to improve the predictive balance in a label space that exhibits significant frequency variation and class imbalance.

Each model was trained for five epochs. While the batch size and number of training epochs were held constant to maintain comparability, key hyperparameters such as

learning rate and dropout were optimised individually. Three candidate learning rates and five dropout values were tested for each model. The best configuration was selected based on validation set performance, as measured by F1 micro and F1 macro scores. This tuning process ensured that each model was given the opportunity to learn under optimal regularisation and convergence conditions, without introducing confounding differences in model comparison.

By isolating model architecture as the only experimental variable and applying a controlled, performance-driven tuning procedure, this optimisation setup provides a reliable foundation for comparing general-purpose and domain-specific transformer models in the context of ESG text classification.

4.3.2 Adam

All models in this thesis were trained using the AdamW algorithm, an improved version of the original Adam optimiser that splits weight decay from the gradient update. This is a known solution to an issue in Adam where using L_2 regularisation in the gradient update causes less-than-ideal behaviour. AdamW resolves this by decaying weights directly before the gradient step, resulting in more stable and generalisable updates (Loshchilov and Hutter, 2019).

The optimiser was used with its standard configuration, including $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. Weight decay was enabled in the default form provided by the HuggingFace Trainer and PyTorch AdamW implementation. No changes were made to the underlying optimiser structure.

A linear learning rate scheduler without warm-up steps was applied during training. The learning rate decreased steadily from its initial value over the course of each model's five training epochs. This scheduling strategy helps improve training stability, especially towards the later stages of convergence.

The Adam algorithm itself maintains moving averages of both the gradients and their squared values. At each time step t, given the gradient of the loss with respect to parameter θ, denoted g_t, the update proceeds as follows:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{5}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \tag{6}$$

Here, m_t and v_t represent the first and second moment estimates, respectively. To correct their initial bias towards zero, the algorithm computes:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} \tag{7}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t} \tag{8}$$

The model parameters θ_t, which represent the trainable weights of the transformer model at time step t, are then updated according to:

$$\theta_{t+1} = \theta_t - \alpha \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{9}$$

Here, α is the learning rate that adjusts the size of the update step, and ϵ is a small constant added to ensure numerical stability. This formulation allows it to adjust the learning rate for each weight individually based on the size and direction of historical gradients. It is also effective with noisy or sparse updates, which is particularly convenient when fine-tuning a large-scale language model on noisy, multi-label ESG data.

AdamW was chosen for its widespread adoption and strong empirical performance in transformer-based language models. Its efficiency, adaptive learning rates, and improved handling of weight regularisation made it a suitable choice for fine-tuning BERT-based architectures on the ESG classification task.

4.4 Evaluation Metrics

To evaluate model performance on the ESG classification task, this thesis makes use of two standard metrics for multi-label classification which are F1 micro and F1 macro. These both are based on the F1 score, which balances precision and recall by considering both false positives and false negatives. However, they differ in how performance is averaged across labels.

Let L denote the number of ESG topic labels, and for each label $i \in \{1, 2, \ldots, L\}$, define:

- TP_i: number of true positives
- FP_i: number of false positives
- FN_i: number of false negatives

The *precision* P_i and *recall* R_i for each label i are defined as:

$$P_i = \frac{TP_i}{TP_i + FP_i}, \qquad R_i = \frac{TP_i}{TP_i + FN_i} \tag{10}$$

The corresponding *F1 score* for each label is:

$$F1_i = \frac{2 \cdot P_i \cdot R_i}{P_i + R_i} \tag{11}$$

F1 micro aggregates true positives, false positives, and false negatives across all classes prior to calculating the score:

$$F1_{\text{micro}} = \frac{2 \cdot \sum_{i=1}^{L} TP_i}{2 \cdot \sum_{i=1}^{L} TP_i + \sum_{i=1}^{L} FP_i + \sum_{i=1}^{L} FN_i} \tag{12}$$

It therefore gives equal weight to all instances and is therefore suitable for measuring overall model performance with class imbalance. This is particularly important in this thesis, where common ESG themes such as climate change or diversity occur far more frequently than less-discussed themes such as water use or tax governance.

F1 macro, on the other hand, calculates the F1 score for each individual label and then goes on to calculate an average:

$$F1_{\text{macro}} = \frac{1}{L} \sum_{i=1}^{L} F1_i \tag{13}$$

Each label is treated equally, regardless of how often it appears in the dataset. It is therefore sensitive to the model's ability to correctly classify rare or underrepresented ESG topics, and is complementary to the F1 micro in that it evaluates the degree to which the model covers label space in total.

Both metrics are reported during the model selection and final evaluation stages. During threshold tuning, they guide the selection of the optimal threshold by determining

the value that yields the optimal balance between precision and recall. In final model comparison, they help analyse the trade-off between general performance and performance on low-frequency labels.

This dual-metric approach ensures that the evaluation reflects both general predictive accuracy and the model's ability to detect underrepresented ESG concerns.

4.5 Overview of Analysis Pipeline

The following table summarises the main steps of the ESG topic classification pipeline, highlighting how each component was applied consistently across all four transformer-based models used in this thesis.

Table 3: Overview of the ESG Text Classification Pipeline Across All Models

Step	**Description**
Data Collection	ESG-related documents collected from the DAX ESG Media Dataset (11,547 entries).
Label Assignment	Soft multi-label annotation using lemmatised keyword matching and semantic similarity (Sentence-BERT).
Preprocessing	Text tokenised and chunked using a 512-token sliding window to handle input length limits.
Model Architectures	Four transformer-based models: BERT, FinBERT (finance-specific), RoBERTa (no NSP, dynamic masking), and DistilBERT (compressed).
Activation Function	Sigmoid function applied to the final layer to enable multi-label output.
Optimisation	Models fine-tuned using the AdamW optimiser with default configuration.
Evaluation	Model outputs evaluated using F1-micro and F1-macro metrics to balance frequent and rare ESG topics.

5 Results

5.1 Overview

This section presents the evaluation results of the four BERT-based models after applying the best-performing configurations identified during training. All the models were assessed using their tuned threshold values that were optimised and present both F1 micro and F1 macro scores. The following sections provide a breakdown of model-level performance and the effects of hyperparameter tuning.

5.2 Model Performance Summary

Table 4 presents the best-performing configuration for each model after threshold tuning and hyperparameter optimisation. All models were evaluated using the same ESG-labelled dataset and identical training conditions, ensuring that the comparison reflects architectural differences rather than external factors.

Table 4: Final model performance scores based on best configuration

Model	**LR**	**Dropout**	**Threshold**	**F1 Micro**	**F1 Macro**	**Notes**
BERT	5e-5	0.1	0.32	0.7587	0.7362	Baseline
FinBERT	5e-5	0.1	0.36	0.7749	0.7511	Financial domain
RoBERTa	3e-5	0.1	0.40	0.7578	0.7349	Extended pretraining
DistilBERT	5e-5	0.2	0.38	0.7703	0.7469	Lightweight model

Note: Best configurations were selected via manual grid search over learning rate, dropout, and threshold, using F1-micro and F1-macro as selection criteria.

FinBERT recorded the highest performance among all models, with an F1 micro score of 0.7749 and an F1 macro score of 0.7511. This is consistent with previous research such as Araci (2019) and Mehra et al. (2022) and this result supports the expectation that domain-specific pretraining on financial text can improve performance on ESG classification tasks, which often include economic, regulatory, and risk-related language. DistilBERT achieved nearly equivalent performance (F1 micro: 0.7703, F1 macro: 0.7469) while being significantly smaller in size, suggesting that lightweight models can perform

equally well with correct fine-tuning. BERT, the overall baseline , was slightly below these models, with RoBERTa being comparable due to its large-scale pretraining.

The next sections detail the full tuning results for each model, including dropout and threshold experiments, label-wise performance, and error patterns.

5.3 BERT

The BERT model was trained on the ESG-labelled corpus using three different learning rates (2e-5, 3e-5, 5e-5) and five different dropout values (0.1 to 0.5). Training was carried out for five epochs per configuration and F1 scores were computed on the validation set. Initial predictions were thresholded at 0.5, and following training, threshold tuning was performed using a grid search from 0.1 to 0.9 (step size 0.02) to optimise F1 micro. The best threshold obtained was 0.32, and the F1 micro score of 0.7438 and F1 macro was 0.7193.

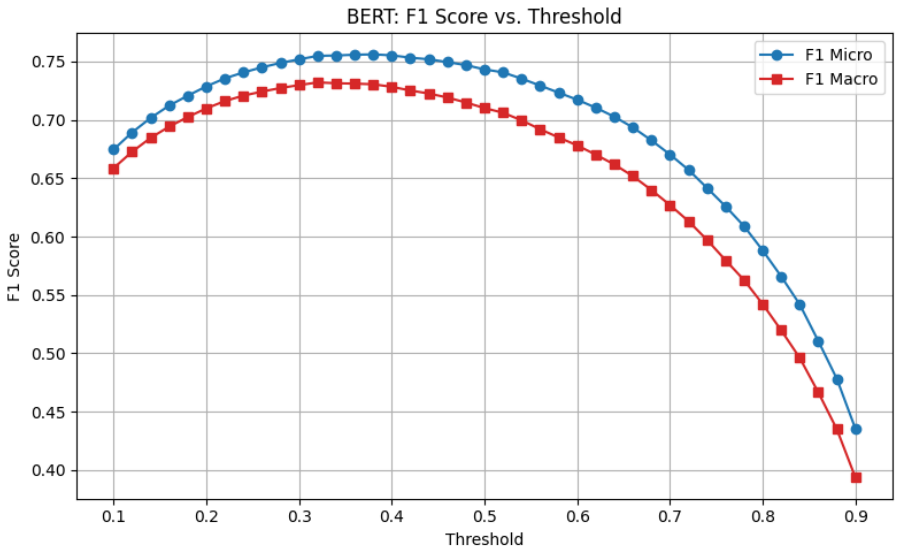

Figure 6: Threshold tuning results for the BERT model. The F1 Micro and F1 Macro scores are plotted across thresholds from 0.1 to 0.9. The optimal threshold was found at 0.32, maximising the F1 Micro score.

The model with the best overall performance was achieved with 5e-5 as the learning rate and a dropout of 0.1. At epoch 5, this had an average validation loss of 0.3012, with F1 micro of 0.7587 and F1 macro of 0.7362.

Table 5: Best configuration and final validation performance for BERT

Parameter	Value
Learning Rate	5e-5
Dropout	0.1
Threshold	0.32
F1 Micro	0.7587
F1 Macro	0.7362
Validation Loss	0.3012

The other learning rates performed less well, with 3e-5 achieving an F1 micro of 0.7440 and macro of 0.7213, and 2e-5 reaching at 0.7316 (micro) and 0.7079 (macro).

Table 6: BERT performance comparison across learning rates (dropout fixed at 0.3)

Learning Rate	F1 Micro	F1 Macro
2e-5	0.7316	0.7079
3e-5	0.7440	0.7213
5e-5	0.7583	0.7369

Dropout rates between 0.2 and 0.5 resulted in somewhat poorer performance, confirming that less dropout (0.1) performed optimally for this model.

Table 7: Effect of dropout on BERT performance (LR = 5e-5)

Dropout	F1 Micro	F1 Macro
0.1	0.7587	0.7362
0.2	0.7557	0.7335
0.3	0.7583	0.7369
0.4	0.7549	0.7337
0.5	0.7547	0.7320

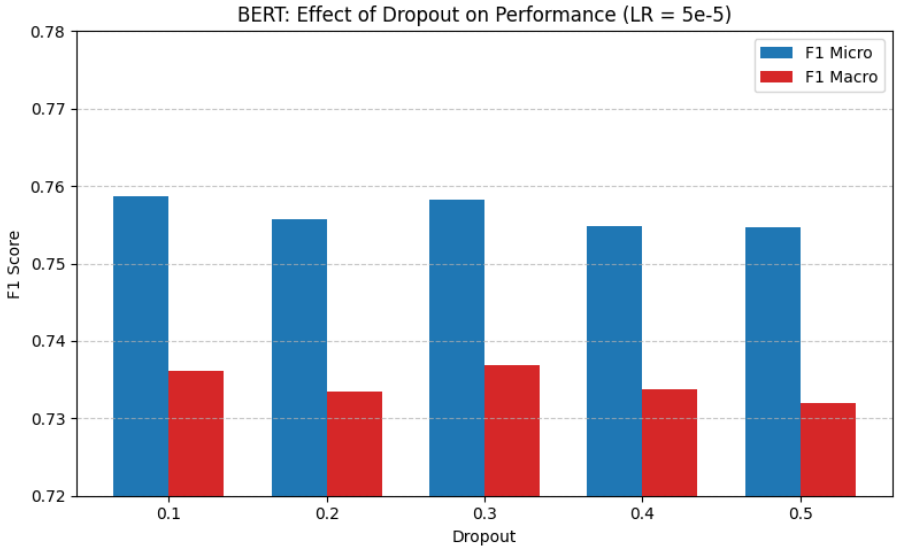

Figure 7: Dropout tuning results for the BERT model with learning rate 5e-5. The highest F1 scores were observed at dropout 0.1. Performance declined slightly as dropout increased.

These results confirm that BERT, being a general language model, is competitive on ESG topic classification. It is, however, narrowly beaten by its domain-specific counterparts in later sections. The persistent performance improvement through threshold tuning also highlights the importance of calibrating decision boundaries in imbalanced multi-label domains.

5.4 FinBERT

The evaluation of FinBERT followed the same procedure as BERT, using three learning rates (2e-5, 3e-5, and 5e-5) and tuning dropout values between 0.1 and 0.5. Training was conducted over five epochs for each configuration, and validation scores were recorded after every run. Like the other models, the predictions were initially thresholded at 0.5. After training, threshold tuning was performed using a grid search from 0.1 up to 0.9 in steps of 0.02. The optimal threshold for FinBERT was found to be 0.36 and had an F1 micro score of 0.7737 and an F1 macro score of 0.7520. These were the highest values obtained by any of the models experimented with in this thesis.

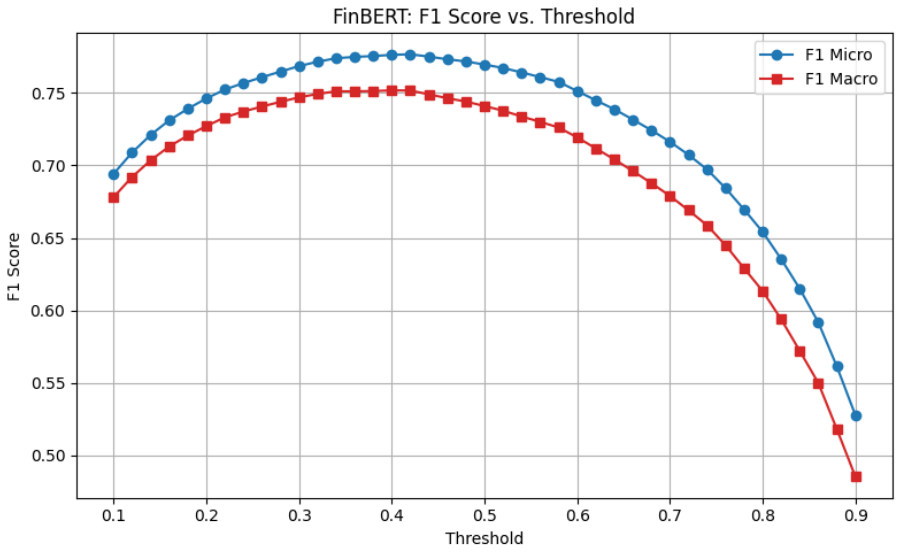

Figure 8: Threshold tuning results for the FinBERT model. F1 Micro and F1 Macro scores are plotted across thresholds from 0.1 to 0.9. The best-performing threshold was 0.36.

The best overall performance was achieved with a learning rate of 5e-5 and a dropout of 0.1. At epoch five, this configuration resulted in a validation loss of 0.2588 and delivered F1 micro and macro scores of 0.7749 and 0.7511 respectively.

Table 8: Best configuration and final validation performance for FinBERT

Parameter	**Value**
Learning Rate	5e-5
Dropout	0.1
Threshold	0.36
F1 Micro	0.7749
F1 Macro	0.7511
Validation Loss	0.2588

The other two learning rates led to slightly weaker performance, with 3e-5 reaching 0.7568 (micro) and 0.7323 (macro), and 2e-5 producing 0.7460 and 0.7196 respectively.

Table 9: FinBERT performance comparison across learning rates (dropout fixed at 0.3)

Learning Rate	F1 Micro	F1 Macro
2e-5	0.7460	0.7196
3e-5	0.7568	0.7323
5e-5	0.7749	0.7511

Dropout values from 0.2 to 0.5 resulted in slightly reduced performance compared to 0.1. This confirms that FinBERT benefited from minimal regularisation in this context.

Table 10: Effect of dropout on FinBERT performance (LR = 5e-5)

Dropout	F1 Micro	F1 Macro
0.1	0.7749	0.7511
0.2	0.7714	0.7483
0.3	0.7711	0.7481
0.4	0.7682	0.7449
0.5	0.7693	0.7452

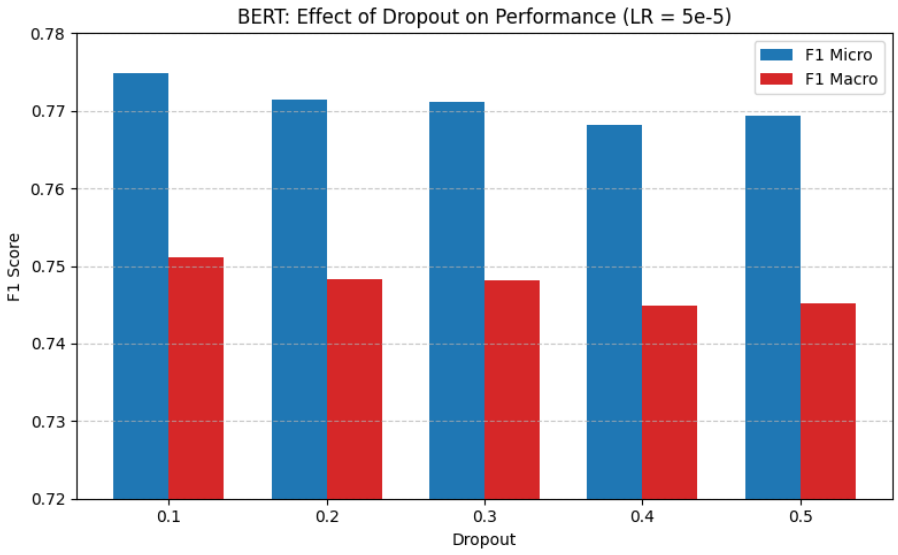

Figure 9: Dropout tuning results for the FinBERT model with learning rate 5e-5. The highest F1 scores were observed at dropout 0.1, with a gradual decline as dropout increased.

These results indicate that FinBERT, a domain-specific variant of BERT, is particularly well suited for ESG classification. Its advantage is most evident when working with

disclosures that contain financial or regulatory language. Compared to general-purpose models, FinBERT showed more consistent gains during threshold tuning and was less affected by regularisation.

5.5 RoBERTa

RoBERTa was trained with the same data and training procedure as the other models. Three learning rates were tried (2e-5, 3e-5, and 5e-5), and each setup was trained for five epochs. The model, similar to BERT and FinBERT, used a dropout layer and sigmoid-activated output to facilitate multi-label classification, with longer documents processed using a sliding window approach.

Threshold tuning was performed after training to improve classification performance. As before, with a grid search between 0.1 and 0.9, the optimal threshold value was found to be 0.40. In this configuration, RoBERTa achieved an F1 micro value of 0.7564 and an F1 macro value of 0.7349, showing great overall performance.

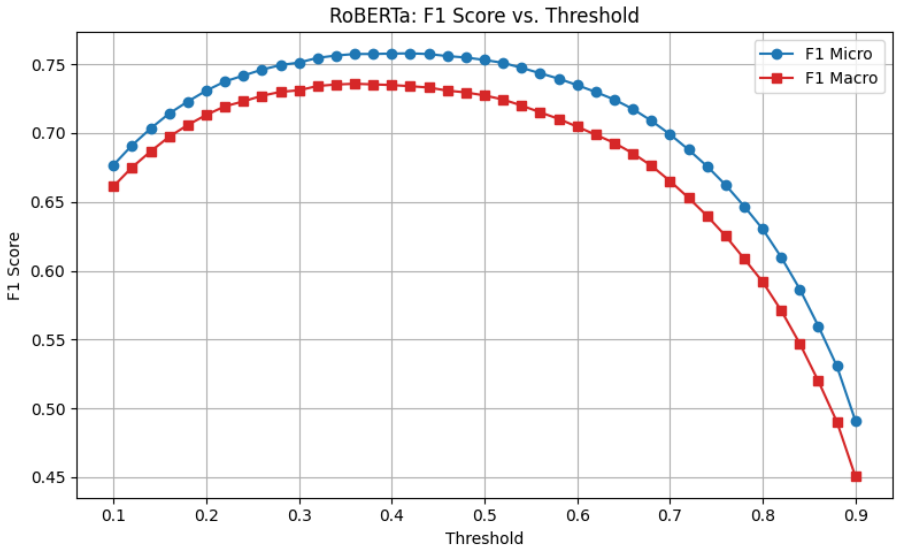

Figure 10: Threshold tuning results for the RoBERTa model. The F1 Micro and F1 Macro scores are plotted across thresholds from 0.1 to 0.9. The best-performing threshold was 0.40.

The best overall configuration occurred at a learning rate of 3e-5 and dropout at 0.1. At epoch five, this setup yielded a validation loss of 0.3446 and delivered the highest F1 scores in every run.

Table 11: Best configuration and final validation performance for RoBERTa

Parameter	Value
Learning Rate	3e-5
Dropout	0.1
Threshold	0.40
F1 Micro	0.7578
F1 Macro	0.7349
Validation Loss	0.3446

Learning rates of 2e-5 and 5e-5 were slightly suboptimal. The former recorded an F1 micro of 0.7472, while the latter recorded 0.7538, with both models falling behind the performance recorded at 3e-5.

Table 12: RoBERTa performance comparison across learning rates (dropout fixed at 0.3)

Learning Rate	F1 Micro	F1 Macro
2e-5	0.7472	0.7235
3e-5	0.7559	0.7320
5e-5	0.7559	0.7320

Dropout tuning also showed that 0.1 gave the most stable outcomes, with the optimal learning rate. Dropout rates 0.2 and 0.3 provided acceptable results but with a minimal drop in performance from 0.1.

Table 13: Effect of dropout on RoBERTa performance (LR = 3e-5)

Dropout	F1 Micro	F1 Macro
0.1	0.7578	0.7349
0.2	0.7571	0.7335
0.3	0.7559	0.7320
0.4	0.7540	0.7302
0.5	0.7470	0.7232

RoBERTa proved to be a strong performer, consistently scoring highly scores across dropout levels. While it lacks domain-specific knowledge like FinBERT, its scale and

pretraining improvements gave it consistent results. Its strong performance across tuning configurations suggests that general-purpose models may be highly competitive on ESG classification tasks, especially if tuned very well.

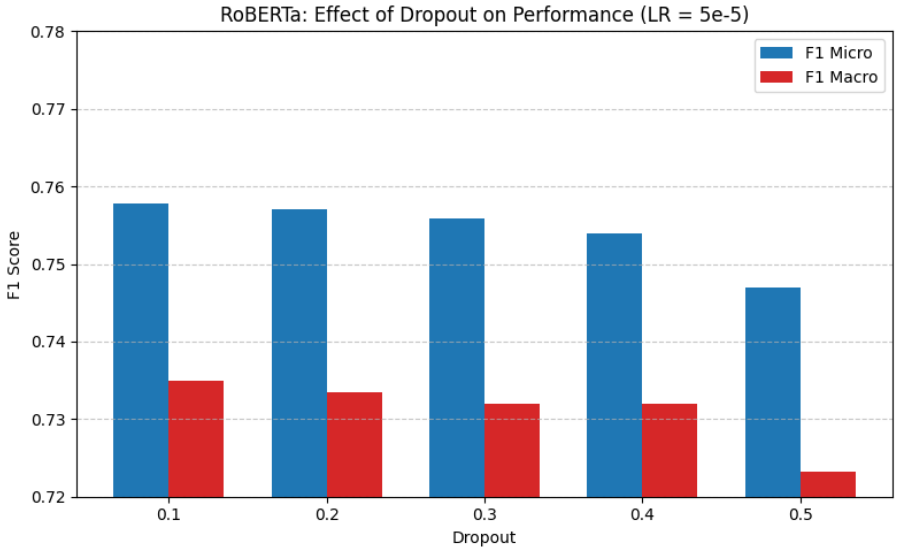

Figure 11: Dropout tuning results for the RoBERTa model with learning rate 3e-5. The highest F1 scores were observed at dropout 0.1, with a gradual decline as dropout increased.

5.6 DistilBERT

As with the other models, DistilBERT was fine-tuned on the same ESG-labelled dataset using a classification head with dropout and sigmoid activation to allow for multi-label output. Predictions were initially thresholded at 0.5 and then refined through grid search. The best threshold was found to be 0.38.

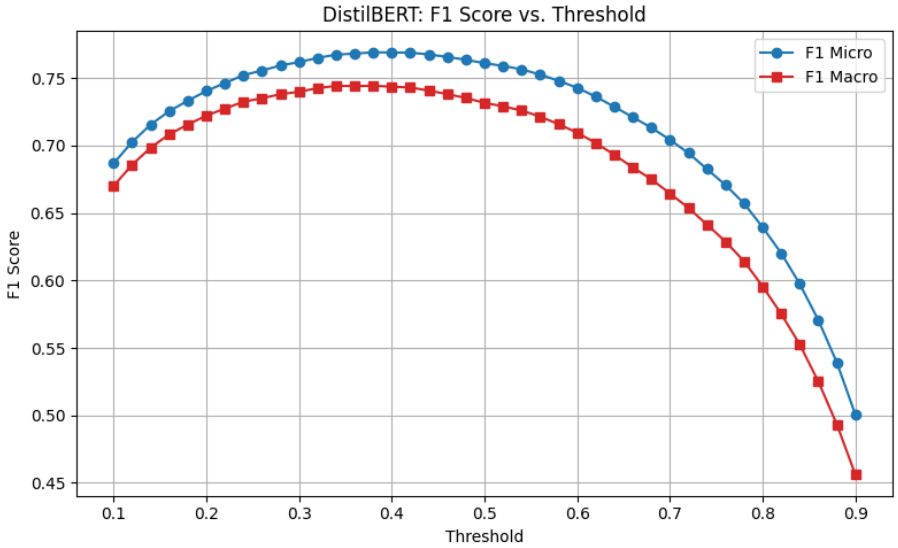

Figure 12: Threshold tuning results for the DistilBERT model. The optimal threshold was 0.38, maximising the F1 Micro score.

The best learning rate was 5e-5, which was then used for dropout tuning. The model performed best with a dropout of 0.2, where it achieved its highest F1 scores. As dropout increased, performance dropped slightly, consistent with patterns observed in earlier models.

Table 14: Best configuration and final validation performance for DistilBERT

Parameter	Value
Learning Rate	5e-5
Dropout	0.2
Threshold	0.38
F1 Micro	0.7703
F1 Macro	0.7469
Validation Loss	0.2770

Table 15: DistilBERT performance comparison across learning rates (dropout fixed at 0.3)

Learning Rate	F1 Micro	F1 Macro
2e-5	0.7361	0.7071
3e-5	0.7515	0.7256
5e-5	0.7688	0.7452

Dropout rates between 0.3 and 0.5 resulted in slightly reduced performance, with 0.2 achieving the highest macro F1. This suggests that minimal regularisation was optimal for DistilBERT under the given training setup.

Table 16: Effect of dropout on DistilBERT performance (LR = 5e-5)

Dropout	F1 Micro	F1 Macro
0.1	0.7691	0.7443
0.2	0.7703	0.7469
0.3	0.7668	0.7427
0.4	0.7623	0.7368
0.5	0.7688	0.7452

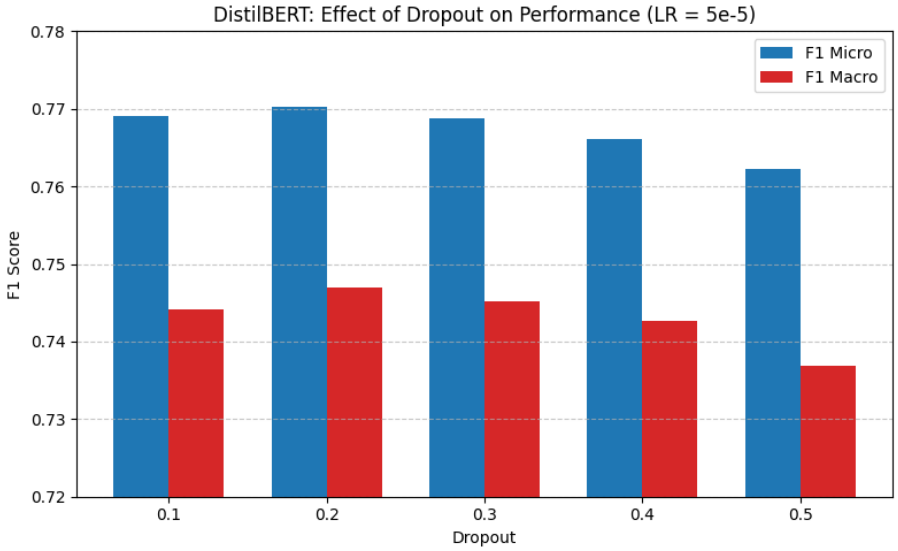

Figure 13: Dropout tuning results for the DistilBERT model with learning rate 5e-5. Performance peaked at 0.1 and declined at higher dropout values.

DistilBERT performed strongly overall, nearly matching FinBERT despite having fewer parameters. Its results suggest that smaller models can still perform well on complex classification tasks when properly tuned. This makes it a promising option in contexts where model size or runtime is a constraint.

6 Conclusion

6.1 Key Findings

This study compared the performance of four transformer-based models in classifying ESG topics on a large dataset of third-party media texts related to companies listed on the DAX index. These models were BERT, FinBERT, RoBERTa, and DistilBERT, which were all trained on the same dataset using the same taxonomy of thirty ESG topics. Experimental conditions were kept constant across models to ensure that performance differences were due to the architecture or pretraining strategy rather than external factors.

Best overall performance was achieved by FinBERT, with the highest F1 micro and macro scores. Pretraining on financial texts gave it a clear advantage when applied to ESG content that reflects business, economic, and regulatory language. DistilBERT, a smaller and lighter model, also performed well. Its result show that lighter models can deliver competitive performance when fine-tuned carefully. RoBERTa performed well but underperformed compared to the baseline BERT model, which could be due to its larger and more diverse pretraining corpus. However, the performance difference between FinBERT and DistilBERT was relatively small, which raises questions about the additional value of larger architectures in this setting.

Threshold selection was also an important finding. Rather than using the common default threshold of 0.5, each model performed best at a different, lower value. BERT achieved its highest performance at 0.32, FinBERT at 0.36, RoBERTa at 0.40, and DistilBERT at 0.38. These tuned thresholds improved both the micro and macro F1 scores. This highlights the importance of calibration when working with imbalanced multi-label classification tasks.

Overall, the results demonstrate that transformer models can be used effectively to classify ESG themes in external media content. Both the choice of model architecture and the tuning of classification thresholds were both important to achieve reliable performance across a wide range of ESG topics. While FinBERT was the strongest overall, all four models showed solid performance when trained and calibrated properly.

6.2 Connection to Aim

The aim of this thesis was to explore whether transformer-based language models can be used to classify ESG-related topics in external media coverage of publicly listed firms. The research also sought to examine whether domain-specific pretraining or model efficiency could provide practical benefits in this context.

The results address this purpose directly. They confirm that models trained on financial language, such as FinBERT, are better suited to the classification of ESG topics that appear in business and regulatory media texts. At the same time, the performance of DistilBERT shows that smaller models can still be strong options for ESG classification tasks when properly fine-tuned.

By comparing multiple models under identical training conditions and evaluating performance using a fixed ESG taxonomy, the study met its objective of providing a fair and structured benchmark. The threshold tuning analysis further contributed to this goal by showing that performance depends not only on model choice but also on calibration decisions. These findings support the original purpose of the study by demonstrating how model architecture and threshold optimisation influence the ability to extract ESG information from unstructured third-party text sources.

6.3 Theoretical and Methodological Contributions

This thesis contributes to the existing literature in the classification of ESG topics as follows. Firstly, it provides a direct comparison between four different transformer models under a controlled experiment. By keeping the dataset, label structure, and training parameters consistent, this study offers a clear view of how different model architectures and pretraining methods could impact performance in text classification when it comes to ESG.

From a methodological perspective, the thesis demonstrates the value of combining lemmatised keyword matching with semantic similarity for ESG label assignment. This approach improved label coverage without relying on manually annotated data and allowed for a scalable way to construct a multi-label training set. The method offers a practical solution for researchers working with unstructured ESG texts where labelled data is limited or inconsistent.

The findings also add to the understanding of model calibration in multi-label classifi-

cation. The impact of threshold selection on both micro and macro F1 scores shows that classification accuracy cannot be separated from post-processing decisions. This highlights the need to from default settings for employing large language models to tasks with imbalanced labeled distributions, such as ESG.

Through the examination of both architectural and procedural components, this thesis provides a foundation for future research on improving ESG topic classification from text using language models.

6.4 Implications for Policy and Practice

The results of this study have practical relevance for organisations and institutions seeking to analyse ESG-related content at scale. The finding that transformer-based models, particularly FinBERT and DistilBERT, can classify ESG topics with relatively high accuracy suggests that these models could be used to support automated monitoring of sustainability themes in external media.

For companies, this can be a way of tracking how ESG issues are being discussed in public narratives and what trends there are in the framing of particular issues, such as climate risk or governance practices, are presented in the media. This enables corporations to benchmark their ESG footprint against competitors or industry standards based on external sources.

From a policymaker's perspective, the ability to systematically classify ESG content across large volumes of media texts could be useful regulators or researchers looking to gauge the public perception of issues related to sustainability.Models like FinBERT could be used to identify gaps between corporate ESG reporting and the external discourse, offering a way to flag areas where regulatory oversight, public concern, or corporate transparency might be misaligned.

Overall, the study shows that transformer models offer a scalable and effective way to analyse ESG themes beyond formal disclosures, supporting more dynamic approaches to sustainability analysis in both business and policy contexts.

6.5 Future Research

There are several different directions for future research based on the findings and limitations of this study. One clear direction is to extend the analysis to a larger set of ESG

topics or to explore hierarchical taxonomies that reflect how ESG themes are structured in practice. More precise labels might more accurately classify topics and allow for more sophisticated assessments of ESG performance. Future studies could also employ classification models to multilingual data, particularly in regions where ESG language might be found in multiple languages. This would test the generalisability of current models and raise questions about whether cross-linguality pretraining or translation-based preprocessing is necessary.

Another area of interest is interpretability. While this study focused on classification performance, it did not investigate how or why certain models identified specific ESG topics. Incorporating explainability methods such as attention visualisation, gradient-based attribution, or concept-based explanations could help researchers and practitioners better understand model output.

Lastly, future work could examine the performance of these models over time by assessing them on media coverage from different reporting periods or major events. This could give perspective on how ESG-related storylines develop over time and can be used to track the development of public interest in specific sustainability issues over a period time.

References

Amel-Zadeh, A. and Serafeim, G. (2018). Why and how investors use esg information: Evidence from a global survey. *Financial Analysts Journal*, 74(3):87–103.

Angin, M., Taşdemir, B., Yılmaz, C. A., Demiralp, G., Atay, M., Angın, P., and Dikmener, G. (2022). A roberta approach for automated processing of sustainability reports. *Sustainability (Switzerland)*, 14(23).

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Berg, F., Kölbel, J., and Rigobon, R. (2022). Aggregate confusion: The divergence of esg ratings. *Review of Finance*, 26(6):1315–1344.

Chung, J. and Latifi, E. (2024). Nlp models for esg disclosure analysis. *Journal of Sustainable Finance and Investment*. Forthcoming.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 1:4171–4186.

Dyck, A., Lins, K., Roth, L., and Wagner, H. (2019). Do institutional investors drive corporate social responsibility? international evidence. *Journal of Financial Economics*, 131(3):693–714.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Huang, Y., Giledereli, B., Köksal, A., Özgür, A., and Ozkirimli, E. (2021). Balancing methods for multi-label text classification with long-tailed class distribution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8153–8161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Izadi, M., Heydarnoori, A., and Gousios, G. (2021). Topic recommendation for software repositories using multi-label classification algorithms. *Empirical Software Engineering*, 26.

Khan, M., Serafeim, G., and Yoon, A. (2016). Corporate sustainability: First evidence on materiality. *The Accounting Review*, 91(6):1697–1724.

Khusuma, R., Maharani, W., and Gani, P. (2023). Personality detection on twitter user with roberta. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, 7:542.

Liang, H. and Renneboog, L. (2020). On the foundations of corporate social responsibility. *Finance Research Letters*, 36:101360.

Linhares, R. (2023). Topic modelling for esg in financial disclosures: A bert-based approach. *Sustainability Analytics*, 5(2):113–130.

Liu, P., Qiu, X., and Huang, X. (2016). Deep multi-task learning with shared memory for text classification. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 118–127.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65.

Mehra, S., Louka, R., and Zhang, Y. (2022). Esgbert: Language model to help with classification tasks related to companies' environmental, social, and governance practices. In *Embedded Systems and Applications (EMSA 2022)*, EMSA 2022, pages 183–190. Academy and Industry Research Collaboration Center (AIRCC).

Ong, D., Morais, A., and Jensen, L. (2025). Concept-aware esg classification with esgsenticnet. In *ACL 2025*.

Ruberg, N. (2021). Bert goes sustainable: An nlp approach to esg financing. Master's thesis, Università di Bologna, Department of Computer Science and Engineering, Bologna.

Sahu, A. K., Debata, B., and Khanna, G. (2025). Unveiling the nexus between esg performance, climate policy uncertainty and corporate innovation: Evidence from textual analysis. *Social Responsibility Journal*, 21(4):893–921.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Saxena, A., Narayanan, A., and Sharma, R. (2024). Applying transformer models to esg news classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Tsoumakas, G. and Katakis, I. (2009). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3:1–13.

Veeramani, H., Thapa, S., and Naseem, U. (2023). Enhancing esg impact type identification through early fusion and multilingual models. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*, pages 84–90, Bali, Indonesia. Association for Computational Linguistics.

Yin, W., Hay, J., and Roth, D. (2019). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*.

Zhang, M.-L. and Zhou, Z.-H. (2014). A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837.

Appendix A: Tables and Figures

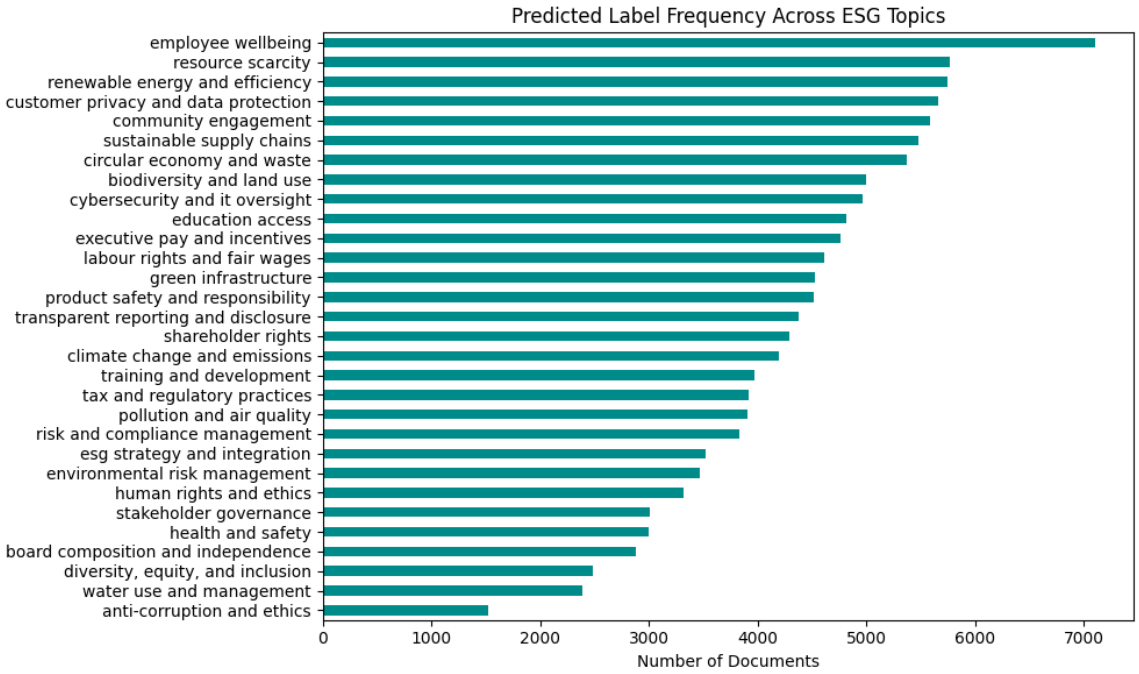

Figure 14: Frequency distribution of predicted ESG topic labels across the dataset by DistilBERT. The most frequently assigned topics include "employee wellbeing," "resource scarcity," and "renewable energy and efficiency," while less common topics include "anti-corruption and ethics" and "water use and management.

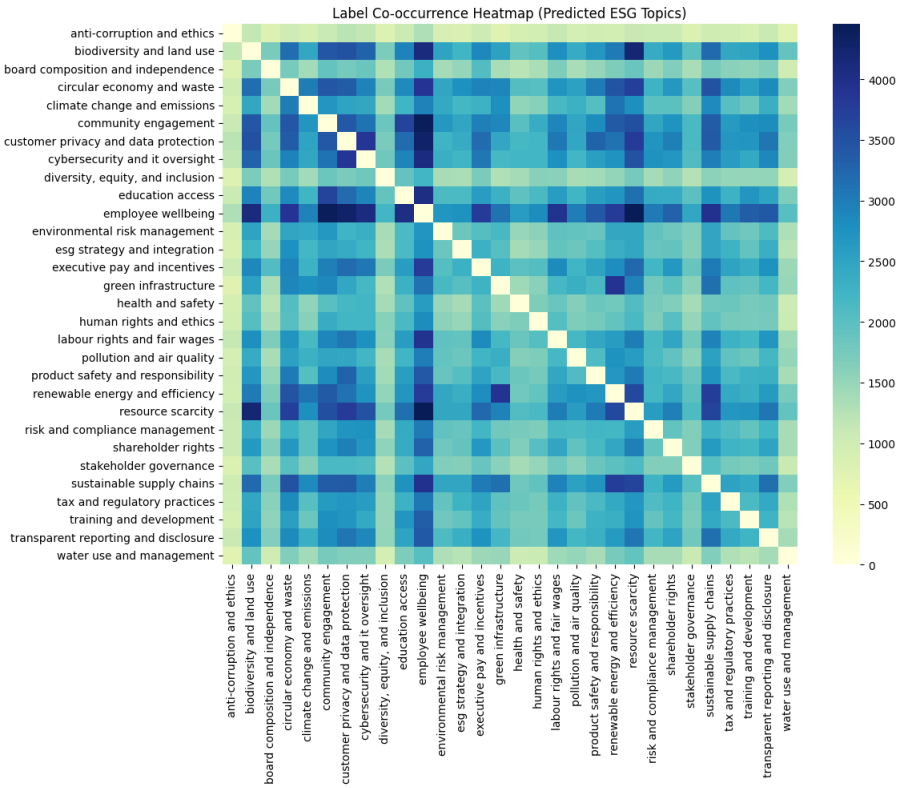

Figure 15: Co-occurrence heatmap of ESG topic labels predicted by DistilBERT. Each cell represents the number of documents where a pair of ESG topics were assigned together. Higher values indicate stronger overlap between topics

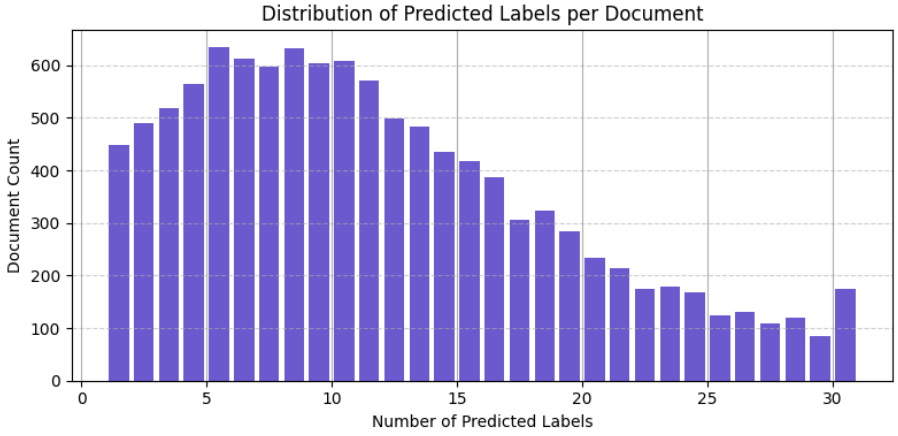

Figure 16: Distribution of label counts per document after classification.

Source Type	Count	Percentage (%)
External (Third-party)	11456	99.2
Internal (Reports)	92	0.8

Table 17: Distribution of the number of ESG labels predicted per document using DistilBERT. Most documents received between 5 and 15 labels, reflecting the multi-topic nature of ESG media content and the overlapping structure of the classification scheme.

ID	SIM_THRESHOLD	TOP_K	Avg Jaccard	Avg Labels
15	0.45	8	1.0000	11.12
14	0.45	6	1.0000	11.12
13	0.45	5	1.0000	11.12
12	0.45	3	1.0000	11.12
11	0.40	8	1.0000	11.12
10	0.40	6	1.0000	11.12
9	0.40	5	1.0000	11.12
8	0.40	3	1.0000	11.12
4	0.35	3	0.9986	11.13
5	0.35	5	0.9986	11.13
6	0.35	6	0.9986	11.13
7	0.35	8	0.9986	11.13
0	0.30	3	0.9886	11.14
2	0.30	6	0.9879	11.15
3	0.30	8	0.9879	11.15
1	0.30	5	0.9879	11.15

Table 18: Results from varying similarity thresholds and Top-K settings for ESG tag assignment. Each row represents a specific configuration used during semantic matching, where `SIM_THRESHOLD` defines the cosine similarity cutoff and `TOP_K` indicates the number of top-matching tags retained. `Avg Jaccard` shows the average Jaccard similarity between assigned and reference labels, and `Avg Labels` refers to the average number of labels assigned per document under each configuration.

Appendix B: AI Statement

I used artificial intelligence while writing and structuring this thesis. I used ChatGPT to help refine sections of text that I wrote by improving the academic phrasing, and restructuring content when necessary for better flow. AI assistance was also used in Overleaf formatting, as I had no prior experience with LaTeX. Additionally, ChatGPT was also used for coding support, mainly debugging, comprehending model behavior, and implementing functions. I also used ChatGPT to help explain concepts and different metrics that I was not familiar with.

I decided on all content, structure, and final wording of the thesis. The analysis, interpretation of results, and overall argument are my own word. No material was automatically produced and all the material in this thesis was reviewed, tested, and confirmed by me.