ZERO-SHOT AND FEW-SHOT NAMED ENTITY RECOGNITION IN ENVIRONMENTAL SCIENCES DOMAIN

KEREM MERT DEMİRTAŞ

SEPTEMBER 2024

ÇANKAYA UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

DEPARTMENT OF COMPUTER ENGINEERING

M.Sc. Thesis in

COMPUTER ENGINEERING

ZERO-SHOT AND FEW-SHOT NAMED ENTITY RECOGNITION IN ENVIRONMENTAL SCIENCES DOMAIN

KEREM MERT DEMİRTAŞ

SEPTEMBER 2024

ABSTRACT

ZERO-SHOT AND FEW-SHOT NAMED ENTITY RECOGNITION IN ENVIRONMENTAL SCIENCES DOMAIN

DEMİRTAŞ, KEREM MERT

M.Sc. in Computer Engineering

Supervisor: Assist. Prof. Dr. Serdar Arslan

September 2024, 113 pages

Novel architectures in natural language processing enable to transfer knowledge of the model for specific tasks. For many downstream tasks, training the model from scratch has become unnecessary since transfer learning can be leveraged for such cases. This can be achieved by finetuning a pretrained Large Language Models (LLM). In this study, a lightweight version of BERT, DistilBERT which is pretrained to predict next sentence was fine-tuned to handle Named Entity Recognition, as one of the most important information extraction task in context of textual data. Transfer learning also enable to transfer knowledge of the model to unseen domains. In this context, we created a domain-specific dataset in the environmental sciences domain. Also, to recognize specific entities, custom NER labels for entities in environmental sciences domain have been defined. To evaluate transfer learning ability of the model, zero-shot, one-shot and ten-shots learning procedures have been conducted on created dataset. To improve transfer learning, we have pre-trained the model a generic Turkish dataset. Finally, artificially generated data that specific to environmental sciences domain have been combined with our created dataset to improve the prediction performance of the model in zero-shot and few-shot setups. In the study, pretraining the model with generic dataset and introducing artificially generated dataset evaluated individually and together. In addition, presence of semantically related entities in the dataset have been investigated

and improvements in prediction performance regardless of shot number are seen. The evaluation of tests demonstrates promising results and enlightens improvements in terms of transfer learning.

Keywords: Transfer Learning, Zero-shot Learning, Few-shot Learning, NER, BERT

ÖZET

ÇEVRE BİLİMLERİ ALANINDA SIFIR-ÖRNEKLİ VE AZ-ÖRNEKLİ ADLANDIRILMIŞ VARLIK TANIMA

DEMİRTAŞ, KEREM MERT
Bilgisayar Mühendisliği Yüksek Lisans

Danışman: Dr. Öğr. Üyesi Serdar ARSLAN
Eylül 2024, 113 sayfa

Doğal dil işlemede yeni mimariler, modelin bilgisini farklı görevlere aktarabilmeyi sağlar. Bu aktarımlı öğrenme sayesinde modeli bazı görevler için yeniden eğitme ihtiyacı ortadan kalkmıştır. Aktarımlı öğrenme, önceden eğitilmiş bir Büyük Dil Modeli'ni ince ayar yaparak sağlanabilir. Bu çalışmada sonraki cümleyi tahmin etmek için eğitilmiş bir model olan BERT'in daha sade bir versiyonu olan DistilBERT üzerinde ince ayar yapılarak, metinsel veriler üzerinde önemli bir bilgi erişim görevi olan Adlandırılmış Varlık Tanıma görevinin yapılması sağlanmıştır. Aktarımlı öğrenme, modelin daha önceden edindiği bilgileri daha önce görmediği alanlara aktarabilmeyi de sağlar. Bu bağlamda, çevre bilimleri alanına özgü bir veri kümesi oluşturduk. Ayrıca, belirli varlıkları tanımak için çevre bilimleri alanındaki varlıklar için özel varlık etiketleri tanımlanmıştır. Modelin transfer öğrenme yeteneğini değerlendirmek için oluşturulan veri kümesi üzerinde sıfır atışlı, bir atışlı ve on atışlı öğrenme prosedürleri gerçekleştirilmiştir. Transfer öğrenimini iyileştirmek için modeli genel bir Türkçe veri kümesi üzerinde önceden eğittik. Son olarak, modelin sıfır atışlı ve birkaç atışlı kurulumlardaki tahmin performansını iyileştirmek için bir büyük dil modeli kullanılarak oluşturulan çevre bilimleri alanına özgü veriler, oluşturduğumuz veri kümesiyle birleştirilmiştir. Çalışmada, modeli genel veri kümesiyle önceden eğitme işlemi ve yapay olarak oluşturulan veri kümesini tanıtma işlemi ayrı ayrı ve birlikte değerlendirilmiştir. Ayrıca, eğitim veri kümesinde anlamsal

olarak ilişkili varlıkların, modelin tahmin yeteneği üzerindeki etkisi incelenmiş olup, ilişkili varlıkların eğitim verisine eklenmesi sonucu tüm atış seçeneklerinde tahmin performansının iyileştiği görülmüştür. Testlerin değerlendirmesi umut verici sonuçlar göstermekte ve transfer öğrenimi açısından iyileştirmelere ışık tutmaktadır.

Anahtar Kelimeler: Aktarımlı Öğrenme, Sıfır-Atışlı Öğrenme, Az-Atışlı Öğrenme, İsimlendirilmiş Varlık Tanıma, BERT

ACKNOWLEDGEMENT

First and foremost, I would like to express my deepest gratitude to my thesis advisor, Assist. Prof. Dr. Serdar ARSLAN, for his unwavering support, guidance, and encouragement throughout this study. His invaluable insights, constructive feedback, and dedication to my success have been instrumental in shaping this thesis and my growth as a researcher. I am truly grateful for his mentorship and the hours he dedicated to helping me refine my ideas and overcome challenges.

To my family and friends, thank you for your strong support, encouragement, and patience throughout this challenging journey. Your belief in me has been an invaluable source of motivation, and your willingness to listen to my endless ramblings about research has been a lifeline. I am truly fortunate to have such a supportive network surrounding me and deeply grateful for your presence in my life and the countless sacrifices you have made to support my academic pursuits.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	: Artificial Intelligence
BERT	: Bidirectional Encoder Representations from Transformers
CLS	: Classification
SEP	: Separator
CoNLL	: Conference on Computational Natural Language Learning
GPT	: Generative Pretrained Transformer
IE	: Information Extraction
JSONL	: JavaScript Object Notation Line
MLM	: Masked Language Model
NER	: Named Entity Recognition
NLP	: Natural Language Processing

CHAPTER I
INTRODUCTION

1.1 INTRODUCTION AND OUTLINE OF THE STUDY

The concept and conversations surrounding the idea of computers to think and behave like humans are not unique to the modern period. Alan Turing is one of the first who brought this subject to light in his study Computing Machinery and Intelligence in the middle of the 20th century [1]. Ever since, this concept has been researched, and as computers have grown more capable and methodologies have been refined, new areas of artificial intelligence, such as robotics and computer vision, have evolved. Giving computers the ability to comprehend spoken and written natural language as a human could is the objective of the Artificial Intelligence (AI) field of study referred to as Natural Language Processing (NLP) [2]. Information extraction (IE) has become more significant as one of the main tasks of NLPs due to the increase in textual data available on the internet.

One of IE's key subtasks is "Named Entity Recognition and Classification (NERC)," which is the process of locating references to the predefined information units in text, such as individuals, organizations, and locations and numerical expressions such as those expressing time, date, money and percentages[3]. Beyond these that are referred to as generic entities, there is another category of entities that are designated as domain-specific entities, as in the biomedical domain, such as proteins, enzymes, and genes[4]. In this study, the field of environmental sciences took part as the specific domain, and entities were defined according to this field.

Over the past 10 years, deep learning has gained popularity because to advancements in computational capability, whereas earlier Named Entity Recognition studies required extensive feature engineering and well annotated data. As in other areas of artificial intelligence, deep learning made it possible to reduce labour-intensive tasks. Advancements in deep learning, like neural networks, led to more sophisticated methods, and with the abundance of textual data available, researchers were able to build enormous models to increase Natural Language Processing

capability. On the other hand, building such large models comes with the cost of training in economic and environmental terms, due to long running training sessions and high-power consumption hardware. One of the solutions to this challenge is to transfer knowledge gained throughout training for one specific task to another downstream task is called as transfer learning. This method lowers the requirement to retrain the model, reduces training expenses, and allows models to be utilized to a variety of applications. Small amounts of task-specific dataset can be used to fine-tune the pretrained models. One of the general-purpose language models DistilBERTurk the Turkish language model of the DistilBERT has been utilized in our work to apply its knowledge to our NER-specific goal by fine-tune it twice with task specific generalized dataset, and with our small environmental sciences domain-specific dataset.

Semantic connections between textual data are established by these language models. This relationship can be used to detect other semantically related categories without directly training on that category, as they were trained on a huge scale of data. This method, known as "zero-shot learning," may be useful in Named Entity Recognition tasks, especially if the entities required are domain-specific and producing the right labels of the dataset is a labour-intensive operation. Furthermore, by adding a tiny amount of data to the training dataset, the trained model can be enhanced when there is a limited supply of labelled data. This method is known as few-shot learning for greater quantities and one-shot learning for instances when there is only a single example of the unseen label. In this work, zero-shot learning, one-shot learning and ten-shots learning have been studied.

The Named Entity Recognition capability of transformer model BERT in zero-shot training setup is examined in this study, along with how this performance improves when a predetermined number of subject entities are introduced into the training data. Furthermore, the ability to connect semantically related entities of the model is examined, as well as how this performance improves when these related entities are present in the training data. Finally, the presence of generated and generic datasets in the training data is examined, along with their effects individually as well as together on the model's prediction performance.

There are three distinct datasets used in the study. A general Turkish Wiki NER dataset is the first one. The second dataset belongs to the environmental sciences domain and has been obtained from news sources. The final dataset was generated by

a LLM and focused on the same field, using straightforward prompts such as "Generate well-structured Turkish sentences for environmental news." The study has involved labelling these domain-specific datasets. The DistilBERTurk model was trained cumulatively, using a generalized dataset initially to optimize the model for NER-specific tasks and then introducing news datasets. Finally, the model trained with created data mixed with news dataset was used to observe the impact of generated data. To observe the impact of the general Turkish NER dataset, DistilBERTurk has also been trained directly using our domain-specific datasets. Zero-shot and few-shot training has been carried out for each label in our dataset by eliminating the subject class from the training data and, if necessary, by adding just one or ten examples. Lastly, hypernyms and/or hyponyms of the unknown class have been eliminated from the training data to further explore the semantic relationships. The F1 score was used to assess the prediction performance of each training checkpoint, and the T-Test was used to compare the evaluated configurations in pairs.

1.2 PERSONAL MOTIVATION

Natural language processing takes a crucial role in extracting information from structured and unstructured textual data. From statistical calculations to highly complex deep learning techniques, NLP can be applied by using a broad range of methods. Thanks to improving hardware capabilities and intense interest of researchers from computer science and other fields in the application of Natural Language Processing, these techniques have evolved into more advanced ones such as large language models.

Named Entity Recognition is one of the subtasks of NLP. It is used to extract generic information from textual data such as location, person, number. Also, it can be used to extract information belonging to specific domains such as biomedical and ecommerce. There are plenty of datasets for generic NER that contain annotated data. On the other hand, finding training datasets for specific domains is a hard task. Especially, if the domain is less studied, a new dataset may be created by the experts of the domain. Since this is a labor-intensive and expensive process, application of NER to resource scare domain specific datasets by reducing the cost of labelling data, a pretrained language model can be used for this specific purpose.

Large language models can be applied to different tasks such as language translation, semantic analysis, question answering and information retrieval. During

their self-supervised or unsupervised training sessions, they see high numbers of data. Since these models create semantic relationships within their hidden layers, they show high performance on common NLP tasks. Also, comprehensive definition of large language models offers researchers the opportunity to apply different architectures for different tasks to utilize their models for specific tasks. On the other hand, the need for high number of data and powerful hardware, training a large language model is an expensive task and forming a capable and comparable large language model is not easy. Despite that, flexible architecture of common language models gives chance to researchers to alter the model structure to apply on their tasks. This ability is also referred to as transfer learning that used to transfer the knowledge gained during training session to other tasks. In some cases, tasks that model has not seen can be done with transfer learning with application of zero-shot learning. Also, by introducing one or more examples to the model, its performance can be increased. This process is called few-shot learning.

As stated earlier, application of NER especially for a specific domain where annotated data is scare or not present, is an expensive process. By utilizing transfer learning ability of pretrained language models with zero-shot learning and few-shot learning, NER for unseen domains can be achieved. On the other hand, zero-shot and few-shot NER comes with low prediction performance compared to model trained with annotated data due to problems such as overfitting and underfitting.

Driven by a deep concern for the environment, I am inspired to explore the intersection of environmental science, natural language processing, zero-shot learning and few-shot learning. Since I have background from environmental engineering, combining these different disciplines together, I believe this study will shed light on multidisciplinary applications of NLP and NER. In this study, I did not only aim to leverage a language model for this task, at the same time I aimed to create a domain specific dataset that belongs to environmental sciences domain. Also, to improve the prediction result of the model, a large language model was leveraged to generate another dataset for same domain. Lastly, to get maximum from transfer learning, a generic Named Entity Recognition dataset has been used. These datasets and their combinations were used to gain different perspectives for this task and exceed limitations of the model. I believe all these improvements throughout the study could lead to innovative tools that empower stakeholders and decision makers to process and comprehend environmental information more effectively. Ultimately, this research

aims to contribute to a more sustainable planet by enabling informed decision-making and facilitating the development of adaptable and scalable solutions for environmental monitoring and analysis.

1.3 RESEARCH AIMS AND OBJECTIVES

There is the research on Named Entity Recognition, zero-shot learning and few-shot learning[5], which utilizes domain specific language models such as PubMedBERT[6] and trains the model on relatively large domain specific dataset. Apart from this approach, our study tries to shed light on zero-shot and few-shot NER task by using general-purpose encoder model DistilBERT's[7] Turkish version DistilBERTurk and a small dataset from Turkish environmental sciences domain.

This research has three contributions: first, it aims to determine whether a Turkish domain-specific dataset can be used for named entity recognition in zero-shot and few-shot learning, to achieve this a dataset specific to environmental sciences domain has been created; second, it seeks to determine how Turkish generic NER and AI-generated domain-specific datasets affect prediction performance in terms of F1 score; and third, it explores the relationship between semantically related classes in the dataset and by comparing F1 Score of available configurations using T-Test. To achieve this, semantically related relationships across NER classes has been identified as hypernyms and hyponyms. At last, by removing these related entities from the training data their effects on prediction performance have been investigated.

1.4 STRUCTURE OF THE THESIS

This section serves as a guide to the overall structure of the thesis, providing a roadmap to users to easily navigate between chapters.

Chapter I Introduction contains background information about natural language processing, named entity recognition, and zero-shot, few-shot learning, followed by the motivation behind this study and research aims and objectives.

In the Chapter II Literature Review previous studies that supports the idea of this work has been written.

As the third chapter, Chapter III Methodology includes detailed information about model's architecture, variants of the model, how named entity recognition can be applied using this model and zero-shot and few-shot named entity recognition. Also, this chapter contains information about named entities that have been defined to

conduct the experiments and their semantic relationships. Later, it gives information about datasets used in the study, labelling process and data augmentation. Finally, this chapter ends by referring to evaluation methods used in the study.

Chapter IV Experiments is the chapter where tools and software in the study have been used so far were mentioned. Also, it contains detailed information about training processes to complete the study.

Chapter V Results and Discussions, it gives detailed insights about results that obtained after the training and discusses performance differences between configurations, datasets and semantic relationships.

The thesis concludes with Chapter VI Conclusion, it summarizes the important implications of the study.

Additionally, in the Appendices chapter, figures of each semantic relationship can be seen.

CHAPTER II
LITERATURE REVIEW

Textual named entity identification and classification is known as named entity recognition, or NER. These entities might be individuals or groups of individuals, places, times, names of diseases, or other particular classifications. NER systems play a critical role in many Natural Language Processing (NLP) applications, such as relation extraction, information retrieval, and question answering. More complex language processing activities are made possible by machines' enhanced comprehension of textual meaning and context due to their capacity to precisely identify and classify named items[8].

The study titled “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer” Raffel, Shazeer, Robers, and Lee et al, gives detailed information about transfer learning and how it can be possible. Machine learning models must process text efficiently in order perform NLP tasks. This entails learning general language information, ranging from sophisticated global knowledge to basic word definitions. Traditionally, auxiliary activities that involve representing word meanings in a continuous space—where comparable words are positioned closer together—have been used to indirectly teach this knowledge. Models can acquire general skills that translate to different tasks because large language models are pretrained on large datasets. Additionally, most of the time unsupervised learning is employed since it is impossible to label a large enough amount of data to train a huge language model. The study shows that how a pretrained language model can be used to accomplish other downstream such as translating English to German with simple questions such as “translate English to German: this is good”, or summarizing text by asking “summarize the following text…”.[9]

Devlin, Chang, Lee and Toutanova of Google proposes a novel language model called BERT stands for Bidirectional Encoder Representations from Transformers.[10] The model considers both left and right context in every layer to pre-train deep bidirectional representations. This enables state-of-the-art performance

by quickly fine-tuning the pre-trained BERT model for different applications like question answering and language inference without requiring large architecture changes. In the paper, transfer learning applied for Named Entity Recognition by fine-tuning approach where it introduces the fewest task-specific parameters possible, and all pretrained parameters are just fine-tuned to train on the downstream tasks and with the feature-based method, the pre-trained representations are added as extra features to task-specific architectures.

In the study of Sainz, Ferrero, and Agerri et al., titled "Gollie: Annotation Guidelines Improve Zero-Shot Information-Extraction" a new model Gollie has been proposed. This large language model fine-tuned to learn from annotation guidelines to improve its prediction performance. According to the paper, Named Entity Recognition performance of the model has promising results that surpasses the other models such as GPT-3.5 and SOTA in domain-specific datasets in zero-shot learning fashion.[11]

The study titled "Application of BiLSTM-CRF model with different embeddings for product name extraction in unstructured Turkish text" proposes a Named Entity Recognition model as the combination of the architecture Bi-Directional Long Shor-Term Memory and the Layer Conditional Random Field. The study compares different embedding methods such as fastText, Glove, Word2Vec and BERT with proposed model and the model offers promising results in ecommerce domain-specific unstructured Turkish text.[12]

In the study titled "Learning Dense Representations of Phrases at Scale", a model for learning dense representations of phrases for open-domain quality assurance is introduced: DensePhrases. By learning phrase representations from reading comprehension tests, utilizing unique negative sampling and query-side fine-tuning, it achieves promising performance. The question answering ability of the model, allows to extract entities by querying the model with the prompts such as "{CLS} Name of the song {SEP} sung by."[13]

The paper of Wei, Bosma, Zhao et al, titled "Finetuned Language Models Are Zero-Shot Learners" combines the appealing aspects of fine-tuning and prompting methods for transfer acquired knowledge throughout training for different downstream tasks such as natural language inference, reading comprehension and closed-book QA with their model named as FLAN. To test zero shot learning in natural language inference, prompts like "<premise> mean that <hypothesis>?" fed into the model. The

study shows that the model outperforms models such as GPT-3, LamDA-PT, and GLaM in previously stated tasks in zero-shot learning[14].

Using the pretrained bi-directional encoding model PubMedBERT, the authors of the study “From Zero to Hero: Harnessing Transformers for Biomedical Named Entity Recognition in Zero- and Few-shot Contexts” by Kosprdic, Prodanovic, and Ljajic et al. investigated zero-shot and few-shot named entity recognition. Several domain-specific medical datasets were integrated into one big dataset for the study. Every input in the training set is prepared as "<Label Name>{SEP}<Sentence>," where the next sentence is the named entity, and the label name is the annotated class. Additionally, every input has an output vector with labels annotated with 1 and other tokens annotated with 0. The predictions of the encoder were evaluated throughout the two layers. The first one is Linear Layer, and the second one is SoftMax. Performance of the model were evaluated by F1 score. The study gives promising results especially for semantically related classes in zero-shot named entity recognition[5].

In the study with the title “Zero-Shot Learning in Named-Entity Recognition with External Knowledge”, authors combine the model LUKE (Language Understanding with Knowledge-based Embeddings) with external knowledge that originate from different word embeddings and creates a model called ZERO. This study applies zero-shot named entity recognition from music domain to science domain and vice versa. With the help of external knowledge, it outperforms LUKE especially in zero, one and five-shot NER settings[15].

The authors of study “Zshot: An Open-source Framework for Zero-Shot Named Entity Recognition and Relation Extraction” proposes a new framework to leverage models Named Entity Recognition ability by supplying simple definitions of entities that model was not trained to recognize. The framework is also used to extract relation of entities from the textual data as another subtask of information retrieval[16].

In the study titled “Self-Improving for Zero-Shot Named Entity Recognition with Large Language Models”, uses a large language model to annotate unseen dataset. Then they utilize different strategies to select correct annotations from the unsupervised labelled dataset. To label unannotated dataset, prompts that contain desired labels, simple task definition and text were introduced to the large language model. With the help of proposed method, the model shows promising results on zero-shot named entity recognition tasks on different datasets[17].

Authors of the study “Zero-shot evaluation of ChatGPT for food named-entity recognition and linking”, compare the zero-shot named entity recognition performance of ChatGPT-3.5 and its successor ChatGPT-4. In the study, prompts contain task description and unseen entities. Also, study shows both models can provide more detailed information other than entity name such as clinical Ids of entities that have been recognized[18].

The study “TEBNER: Domain Specific Named Entity Recognition with Type Expanded Boundary-aware Network” applies named entity recognition to domain-specific datasets. The proposed distantly supervised method TEBNER (Type Expanded Boundary-aware NER) processes a corpus to extend the boundaries of a named entity dictionary. Then, the authors post-process the dictionary to clean noisy data. Later, the extended dictionary is used to recognize named entity in the text. The proposed method outperforms the other methods on distant supervised named entity recognition such as AutoNER and HAMNER[19].

The paper of Sivarajkumar and Wang with the title “HealthPrompt: A Zero-shot Learning Paradigm for Clinical Natural Language Processing” describes a new framework to leverage models such as BERT, RoBERTa, BioBERT et cetera for zero-shot learning. In this study, instead of finetuning the models for specific tasks, authors have defined prompt functions to classify texts that belong to clinical domain. This approach allows authors to transfer masking ability of BERT based models to text classification task[20].

In the study with title “Zero-shot learning for requirements classification: An exploratory study”, authors have leveraged BERT based models to improve their zero-shot learning to classify text based on their functional and non-functional requirements, and security and non-security requirements. The study proposes a new zero-shot learning approach that does not require any labelled dataset or training. The approach shows promising results. In the study, a text that desired to be classified for its requirement is given into the model with following information contains unseen labels such as “usability”, and “security”. Then using cosine similarity, the output results are compared. The resulting output indicates one of the labels given in the input sequence[21].

CHAPTER III
METHODOLOGY

In the context of this study, the literature review conducted throughout demonstrates the potential to apply pretrained language models transfer learning abilities to other areas, particularly in Named Entity Recognition (NER) and domain specific datasets.

In this study a tiny dataset has been used which was created during the research process, as opposed to large domain-specific datasets which were used in similar studies. Afterwards, to test whether transfer learning can be applied in other models, DistilBERT, a smaller and faster variant of BERT[7], was utilized. Additionally, the model's few-shot, and zero-shot NER capabilities have been investigated. Finally, research has been done on how semantically related classes in the training data affect the accuracy of predictions made in few-shot and zero-shot configurations.

3.1 MODEL

3.1.1 Transformer Architecture

Transformer architecture was first proposed by Vaswani, Shazeer, and Paramer et al, in their study titled “Attention Is All You Need”. A transformer is an encoder-decoder architecture, where input sequences are mapped in encoder phase and output is generated by the decoder. Additionally, this architecture has a self-attention mechanism. The mapping of a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors, is known as an attention function, and after some intermediate processing output tokens have been generated[22]. Figure 1 demonstrates the architecture of the Transformer model.

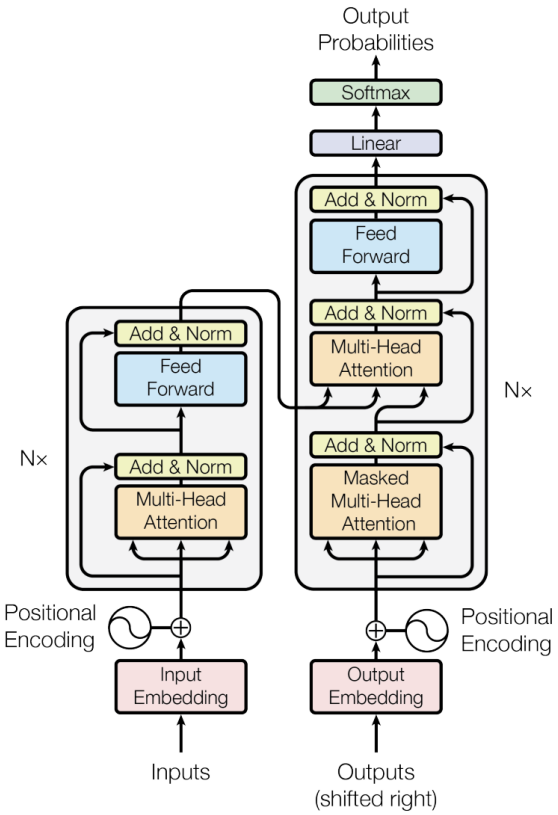

Figure 1: Transformer Model Architecture[22]

3.1.2 BERT

Simply defined, BERT (Bidirectional Encoder Representations from Transformers) is a transformer language model in which tokenizers are used to turn the original text input into tokens. In this case, the text is taken and split up into tokens by the tokenizer. A token can be a single letter, a word, or a word fragment. These tokens then mapped to distinct numbers. Apart from previously defined transformer architecture, BERT does not have a decoder part, and this makes it an Encoder Only Transformer Model. Since the token-based classification tasks such as Named Entity Recognition does not require decoding, this makes the BERT a good candidate for this task.

3.1.2.1 Pre-training BERT

3.1.2.1.1 MLM – Masked Language Model

The model has been pretrained on two tasks, the first one is Masked Language Model (MLM), which enables deep bi-directional training unlike previous models that can only learn left-to-right context or right-to-left context. In this task, a random

predefined percentage of input is masked. Later, these masked tokens are fed into the output vector from the vocabulary[10].

3.1.2.1.2 Next Sentence Prediction

The second task of the pretraining section is Next Sentence Prediction, where the model creates relationship between two sentences. Since the most of the Natural Language Processing task such as Question Answering where the first sentence is the question and the next one is the answer, or simply feeding a sequence of inputs and taking a sequence of outputs is the basic approach in such tasks, pretraining for next sentence prediction task allows the model's knowledge to transfer to other tasks without changing the output architecture as explained in detail in the section Application of Named Entity Recognition with DistilBERT.

3.1.3 DistilBERT

The study "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter" by Sanh, Debut, and Chaumond et al. states that while the model DistilBERT has the same architecture as BERT, it is a faster and smaller version. The study claims that the model maintains its 97% language understanding capacity while being 40% smaller and 60% faster than the BERT model[7]. These intriguing model parameters played a significant role in the model selection process for our Named Entity Recognition jobs.

3.1.4 Zero-Shot and Few-Shot Learning

Zero-shot learning in NLP is a configuration where a model's ability to achieve a downstream task is altered to carry out another downstream task, such as giving special prompts to analyse sentiment of the input text, to decide whether a hypothesis entails or not etc. by using the transfer learning ability of these pretrained language models[14].

Few-shot learning is a fine-tuning technique to improve performance of the pretrained language model for a downstream task that the model is not trained to carry out, by introducing one or more examples of unseen task. These tuning techniques create opportunities to do tasks such as language modelling, question answering, translation[23].

3.1.5 Application of Named Entity Recognition with DistilBERT

Next sentence prediction plays a part in the pretraining process, as was previously mentioned in the section Pre-training BERT. With this method, pretrained knowledge can be transferred without changing the architecture of the input layer. The study by Košprdić et al. was used as a guide, and the next sentence prediction task was transformed to the Named Entity Recognition (NER) task by providing the subject entity class in the first sentence and the subject entity-containing normal sentence input in the following sentence[5]. To mark which tokens are the entity in question, labels for each input token has been created in a way that all tokens in the first sentence and only the entity tokens in the second sentence marked as 1. On the other hand, all unrelated word tokens marked as 0 while special tokens such as [SEP], [CLS] and filling tokens marked as -100. Figure 2 shows how the input data were prepared for named entity recognition.

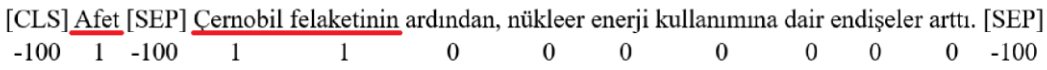

Figure 2: Input Data for Named Entity Recognition

3.1.6 Zero-Shot and Few-Shot Named Entity Recognition

To train the model to investigate its Zero-Shot Named Entity Recognition capability, unseen classes to be tested has been removed from the training and the validation data. For the Few-Shot NER, according to One-Shot or Ten-Shots configuration that number of examples of the subject class have been introduced into the training data.

3.2 NAMED ENTITIES

The study consists of three different datasets. This section covers labelling process, detailed description, obtaining and data distribution of the datasets, then continues with how the data labelled for Named Entity Recognition task and finally how the data prepared to train and test the DistilBERT model.

3.2.1 Defining Domain Specific Labels for Environmental Sciences

A total of 30 labels has been defined to examine the model's prediction behaviour in a domain-specific dataset, such the environmental sciences. These labels

are divided into categories: those that belong to a specific category, like Ecological Impact, and those that are generic and encompass a wide variety of things, like Ecosystem. The ability to augment the data and examine the impacts of semantically related entities in the training data was made possible by the labels' hierarchical structure in terms of semantic relationships. This relationship has been mentioned in the next section.

3.2.2 Semantic Relationship of Labels

Semantic relationships can be defined as those that exist between terms that are related to each other. For example, the terms "cheetah" and "lion" belong to the same category, "cat" where this category standing as the hypernym of "cheetah" and "lion." On the other hand, “cheetah”, and “lion” are the hyponym of “cat”. The same relationship can be observed in Turkish. As stated in previous section, hyponyms and hypernyms were considered while defining the labels in our domain-specific environmental sciences NER study.

Figure 3 illustrates how the semantic relationships have been constructed during the label definition phase. While the Environmental Impact is the hypernym of all the sub entities, these sub entities such as Well-Being Impact are the hyponyms of the Environmental Impact.

To simplify the study, labels were divided into three separate groups. The most general categories, or labels with just hyponyms, are considered top-level labels. However, specialized categories are regarded as bottom-level labels. Furthermore, mid-level labels are those that have both a hypernym and a hyponym. Finally, sibling labels are those that have the same hypernym and hierarchical level. Figure 3 shows that the top-level label is Environmental Impact, the bottom-level label is Economic Impact, and the mid-level label is Well-Being Impact. Other hierarchies can be seen in APPENDICES section.

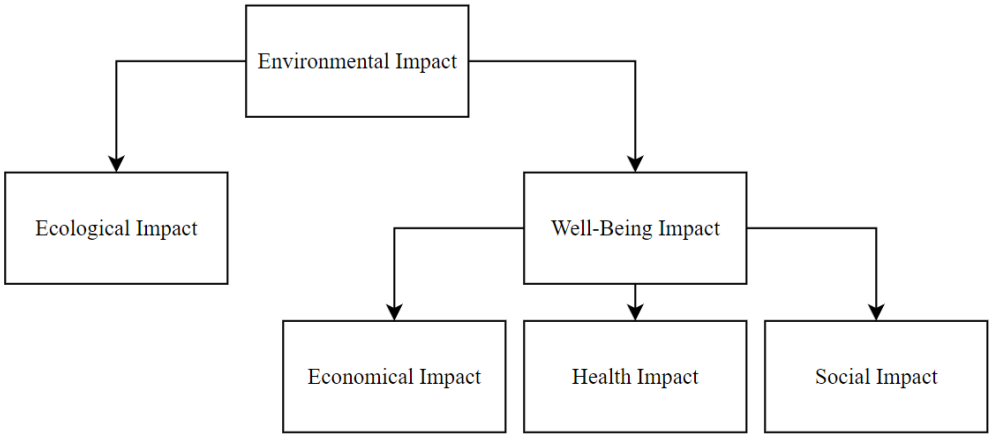

Figure 3 Semantic Relationships of Labels Related to Environmental Impact

Table 1: Label Distribution According to Semantic Hierarchy

Label Type	Count
Top-Level Labels	7
Mid-Level Labels	3
Bottom-Level Labels	20

Table 1 shows the label distribution according to semantic hierarchy with total of 30 labels.

3.3 DATASETS

3.3.1 Turkish Wiki Named Entity Recognition Dataset

Table 2: Turkish Wiki NER Dataset[24]

Tag	Count	Tag	Count
Cardinal	4295	Norp	4023
Date	6923	Ordinal	1711
Event	2392	Org	4583
Fac	944	Percent	182
Gpe	10368	Product	12787
Language	822	Quantity	990
Law	80	Time	131
Loc	1364	Title	2494
Money	100	Work of Art	2951

To further improve the DistilBERT model for our Named Entity Recognition tasks, we first employed the Turkish Wiki NER dataset. Furthermore, pretraining the model with a generic NER dataset could improve the prediction results because the

domain-specific news dataset is a small dataset. The Turkish Wiki NER dataset, which includes 20,000 annotated sentences with 357K words, was generated using Turkish Wikipedia entries[24]. The dataset's data distribution can be seen in Table 2. The obtained dataset is in CoNLL format.

3.3.2 News Dataset

The News Dataset has been created by obtaining Turkish news from the websites of news sources. The topics of these news are such as environmental pollution, global climate change, disasters including forest fires etc. Sentences have been selected considering the proper sentence structure, and occurrence of at least two entities in one sentence. Total of 175 sentences were added into this dataset. The details about labelling, augmentation processes, and label distribution for datasets can be found in next sections as in order Data Labelling, Data Augmentation and Label Distribution.

3.3.3 Artificially Generated Dataset

To study the effect of artificially generated dataset in training and test data, simple Turkish prompts has been asked to a large language model such as “Generate well-structured sentences for environmental news.”, “Generate news sentences about global climate change.”, “Generate news sentences about environmental disasters.”. The proper sentences have been selected from the answers of LLM. Total number of 118 sentences have been generated through this process.

3.3.4 Data Labelling

This was the most labour-intensive stage of the study out of all the steps. To label every sentence and entity in the news and AI-generated dataset, Doccano, an open-source software, was utilized. Following the labelling procedure, a JSONL (JSON Line) file was produced, and the study used this file as its data source.

3.3.5 Data Input Preparation

As described in section above, to label the entities in sentences, Doccano was used. The JSONL files that were produced converted into Dataset objects of Huggingface with proper formatting as shown in Figure 2: Input Data for Named

Entity Recognition. Filters were applied to this Dataset objects which are defined in detail in the section 4.2.2 Configurations.

3.3.6 Data Augmentation

As stated earlier, labelling data was not easy since the resources to do this job is scarce, to increase the number of examples in the dataset, data augmentation has been implemented. During the data labelling step, if possible, only the bottom-level entities has been labelled. Since a hyponym can be generalized as its hypernym in a way that an entity that represents Aquatic Ecosystem also represents Ecosystem. With this approach, number of inputs that will be fed into the model has been increased. Before the augmentation process, total input number was 1131, and after it become 2186.

3.3.7 Label Distribution

Table 3: Number of Bottom-Level Labels

Tag Name	**News Data**	**AI Data**	**Combined**
Aquatic Biota	20	11	31
Aquatic Ecosystem	60	48	108
Ecological Impact	74	53	127
Economic Impact	15	27	42
Energy	3	4	7
Gas Pollutant	12	4	16
Health Impact	50	23	73
Human	37	32	69
Human-Sourced Phenomenon	44	18	62
Human-Sourced Polluter	42	21	63
Liquid Pollutant	11	8	19
Mitigation	5	6	11
Natural Phenomenon	90	82	172
Natural Polluter	8	11	19
Regulation	11	2	13
Settlements	23	16	39
Social Impact	1	11	12
Solid Pollutant	13	3	16
Terrestrial Biota	14	29	43
Treatment	3	4	7
Total	536	413	949

Table 3 shows number of each bottom-level label in news dataset, AI dataset and combined. Since, augmentation takes place in this level, their number did not change after this process. In Figure 4, label distribution of these labels is shown.

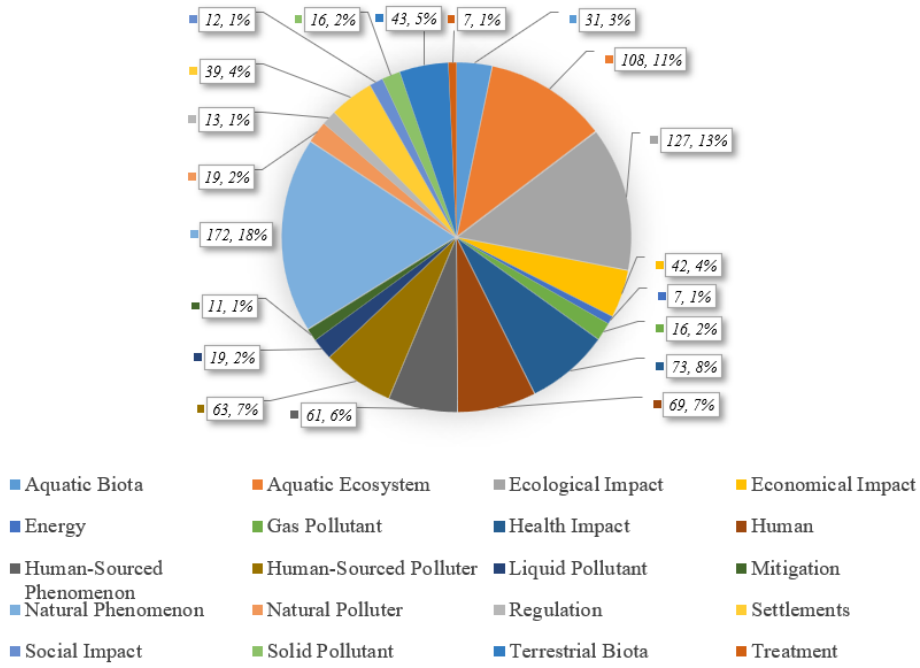

Figure 4: Bottom-Level Labels' Distribution (Count, Percent)

Table 4: Number of Top-Level Labels

Tag Name	Before Augmentation			After Augmentation		
	News Data	AI Data	Combined	News Data	AI Data	Combined
Biota	10	2	12	71	63	134
Ecosystem	30	14	44	100	79	179
Environmental Management	5	4	9	20	13	33
Environmental Impact	12	11	d	140	103	243
Phenomenon	1	0	1	127	99	226
Polluter	0	0	0	50	25	75
Pollutant	22	8	30	50	32	82
Total	80	39	119	558	414	972

As seen in Table 4, after the augmentation total number of top-level labels increased from 119 to 972. Since the main approach in labelling process is to label only bottom-level labels as much as possible, after the augmentation bottom-level labels defined as their top-level labels. Figure 5 shows label counts and their distribution of top-level labels after the augmentation process.

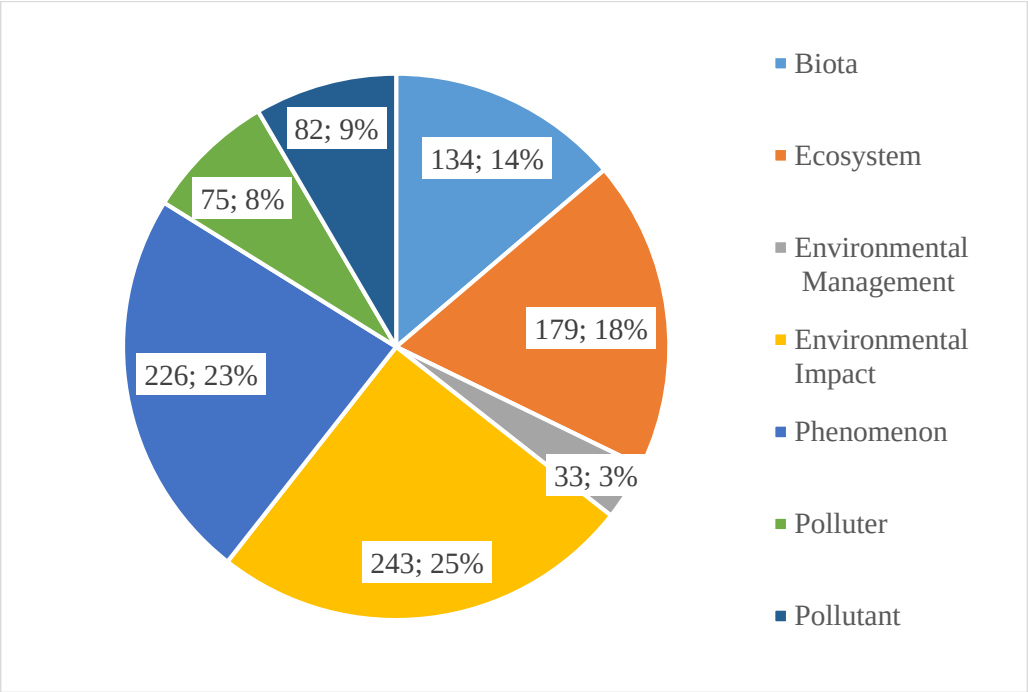

Figure 5: Top-Level Labels' Distribution (Count, Percent)

Table 5: Number of Mid-Level Labels

Tag Name	Before Augmentation			After Augmentation		
	News Data	AI Data	Combined	News Data	AI Data	Combined
Non-Anthropogenic Biota	3	0	3	35	39	74
Terrestrial Ecosystem	6	9	15	27	23	50
Well-Being Impact	27	18	45	78	63	141
Total	36	27	63	140	125	265

Table 5 shows mid-level labels and their counts before and after augmentation. These labels benefit from augmentation like top-level labels. Figure 6 below shows, their counts and distributions after the augmentation.

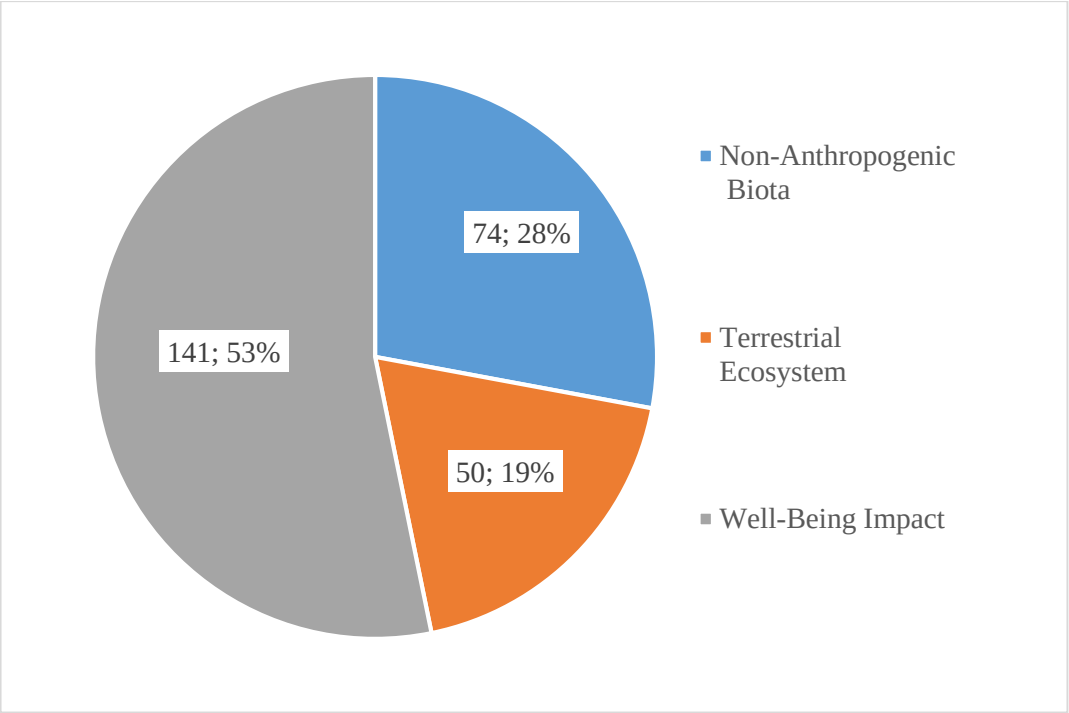

Figure 6: Mid-Level Labels' Distribution (Count, Percent)

Table 6: Numbers of Labels According to Levels

Label Type	**Before Augmentation**			**After Augmentation**		
	News Data	**AI Data**	**Combined**	**News Data**	**AI Data**	**Combined**
Bottom-Level Label	535	413	949	535	413	949
Mid-Level Label	36	27	63	140	125	265
Top-Level Label	80	39	119	558	414	972
Total	651	479	1131	1233	952	2186

Label counts of each level and their total numbers is shown in Table 6. As seen in results, after the augmentation almost size of the dataset is doubled. Figure 7 shows counts and distribution of these levels after the augmentation.

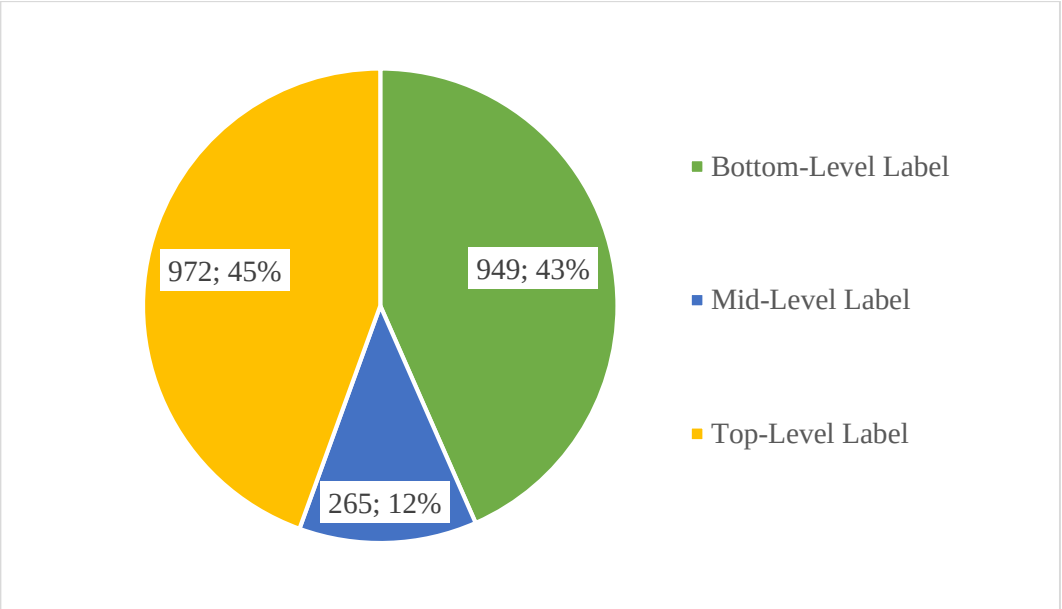

Figure 7: Labels' Level Distribution

3.3.8 Overlapping Entities

There are overlapping entities in the label definition, such as “Sea creatures,” which is aquatic biota, while "Sea" is an aquatic ecosystem. This circumstance is evident in various labels, such as Pollution and environmental management can coexist, for instance in the case of "Reducing carbon emissions,” where “carbon emissions” is a pollutant and "Reducing carbon emissions" is an environmental management procedure. Since each entity fed into the model in its own unique input, the architecture design of BERT allowed for the easy handling of this circumstance.

3.4 EVALUATION

There are many model trainings in the study, and the findings need to be analysed. Therefore, only the F1 Score metric was employed in the interpretation stage to keep it as straightforward as possible. Once more for the same reason, the T-Test score was utilized to compare the model's performances based on the datasets that it was trained on.

3.4.1 F1 Score

F1 Score, to put it simply, is a performance statistic that assesses the prediction capability of a model. The model's predictions are more accurate as it closer to the score of 1 or 100%. This score is harmonic mean of precision and recall, where precision is the proportion of correct number of tokens that defined as unseen class to

the total number of tokens that defined as unseen class by the model, and recall is on the other hand is the proportion of correct number of unseen class tokens that have been defined by model to the number of total number of unseen class tokens. As the essential units of widely used evaluation metrics such as Precision, Recall and F1-Score; details of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are given below.

- TP is the correct number of unseen class tokens predicted by the model.
- FP is the quantity of unseen class tokens that the model predicts but which are not actually unseen class tokens.
- TN is the correct number of tokens that predicted as tokens which do not belong to unseen class
- FN is the number of tokens that predicted as unseen class tokens but do not belong to unseen class

Formula of the metrics shown as follows:

$$Precision = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Positive\ (FP)} \quad (3.1)$$

$$Recall = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative\ (FN)} \quad (3.2)$$

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (3.3)$$

3.4.2 T-Test

This test is used to compare means of two different result set to test if change in the experiment -such as extending dataset or pretraining the model-- has affected the prediction results. To conduct a T-Test an alternative hypothesis must be constructed. In this study, since the focus is to analyse improvement of prediction performance of model as F1 score when dataset differs from base domain-specific news dataset, such as pretraining with a generic dataset or combining with artificially generated data, the alternative hypothesis is constructed as “These additions improve the prediction performance”. There are several T-Tests available, since the samples are paired, one-sided paired t-tests have been conducted. This test produces two outputs, T-Test score and P-Value. As the results p-value is lower than 0.05 or 0.5%, it means that our alternative hypothesis cannot be rejected and there is a significant difference in terms of prediction performance for these train configurations. On the

other hand, when the p-value greater than 0.05, it can be described as there is not a significant difference between two results.

To calculate the T-Test for each configuration pairs, we picked two configurations with one change in setup such as a model that was trained with same dataset, same relationship configuration but in one training session zero-shot setup and in the other one-shot setup have been used. With this approach, difference between zero-shot NER and one-shot NER was compared for the specific dataset combination. This test has been applied for different combinations such as comparison of two different datasets in training with same shot number and semantic relationship configurations. On the other hand, during the evaluation of these T-Test results, we were careful to not exceed the limitations of T-Test by comparing more than two configurations in the same test. Furthermore, we did not embrace statements without further tests such as if the configuration A outperforms configuration B, and configuration B outperforms the configuration C, then configuration A is better than configuration C. Also, as described above, number of classes in our dataset does not exceed 30, which is another aspect of our study that directed us to compare our configurations by using T-Test.

3.4.3 F1 Macro

As stated earlier, model performance is evaluated using F1 score. Addition to T-Test score, to compare the model's prediction performance for different dataset combinations, shot and semantic relationship configurations, F1 macro score is used. To calculate this score, summation of F1 scores for each class for the specific configuration has been divided into total class number. Simply, F1 macro score is arithmetic average of all produces F1 scores for current configuration. Also, there is F1 micro score that considers weight of all classes. Since, our dataset is not balanced where classes such as Environmental Impact has more than 200 examples to be tested, total number of examples for classes such as Treatment does not exceed 10, comparing configurations with F1 score where each class is treated equally by using F1 macro average became the better choice for us.

CHAPTER IV
EXPERIMENTS

4.1 TOOLS AND SOFTWARE

In this section, hardware and software that have been used to prepare the datasets, to train the models and to interpret the results are described.

To train the model and interpret the results, Google's paid product Google Colab Pro have been used in some cases with simultaneous two training sessions. This platform is a cloud-based solution where subscribers can allocate and utilize hardware in a period called sessions. To train the DistilBERTurk with generic Turkish Wiki NER dataset, session powered with Google Colab's V100 graphics processing units was run. Since this GPU had been replaced by A100 GPUs from the Google Colab, later training sessions with news dataset and AI combined datasets was completed with A100 GPUs.

To label the datasets, a free open-source software Doccano was setup on a local Docker server. This tool produced the dataset files in JSONL format. Dataset was labelled sentence by sentence.

To train the DistilBERTurk model, Huggingface's transformers library has been used. Since it has large scale library support and great community behind the models and training, this platform was selected. Also, Huggingface's Dataset library has been used to prepare and convert JSONL files to the datasets to be fed into transformers. PyTorch, NumPy and Pandas Dataframe libraries were used as intermediate numerical solutions on data structures.

To interpret the results with confusion matrix and F1 score, Sklearn's metrics library has been used. For visualization Matplotlib and to run T-tests, SciPy has been used. Output data of training results and interpretations were stored in a Google Drive folder.

4.2 TRAINING

In this section, steps to train the models are described such as pretraining with generic dataset, creating shot options and semantic relationship configurations, finally training the model with domain-specific dataset and training with combined datasets. Training with Generic Turkish NER Dataset

4.2.1 Training with Generic Dataset, Turkish Wiki NER

To train the model with this generic dataset and change the input architecture in a way that described in the section Application of Named Entity Recognition with DistilBERT, for each entity class in a sentence, the sentence is populated, and relevant tokens are tagged for this entity. Then, inputs prepared to be fed into the transformer model as follows "{The Entity}.{The Sentence Contains the Entity}". For training, 18.000 examples, and validation 1000 examples have been used. Trainer configurations are taken from the study of Kosprdic[5]. The trained model was stored in Google Drive, and by using Huggingface's interface, imported into the next sessions.

4.2.2 Configurations

Testing the model for different unseen classes, in absence and in presence of semantically related classes was not possible in one training. If the model trained with related classes, there is not a way to remove these classes from the training data of the trained model and revert the model weights to its previous state. To overcome this situation, specific configurations have been created for each scenario. These scenarios have been explained in next two sections. Total number of configurations that must be run individually are 199 for News Dataset, and 240 for News and AI generated data combined and total number of configurations is 439. Also, as stated earlier there are two start checkpoints the first one is DistilBERTurk, and the second one is DistilBERTurk pretrained with generic Turkish Wiki NER dataset. Hence, total number of configurations was doubled and became 878. All these sessions were run on Google Colab's A100 GPUs. Total time to complete all training sessions was about 12 hours.

4.2.2.1 Shot Configurations

As stated in previous sections, this study has been conducted to analyse the effect of zero-shot training, one-shot training and ten-shots training on prediction performance. To prepare this configuration, for each unseen class, zero-shot, one-shot and ten-shots training datasets have been prepared as follows: From the global dataset, data labelled with unseen classes, and the other data were split into two different datasets. The second dataset that does not have unseen classes was split into two with ratio of 85%, and 15% to create training and validation datasets. This ratio was taken into consideration instead of conventional ratios such as 80% training, 10%, and 10% test since there is not test data in this dataset as they were filtered out in previous step.

As stated earlier, ten-shots is the maximum number of examples to be introduced into training data from the data that have been labelled as unseen class in the context of this study. To test different shot options, we have created a simple strategy to keep the test dataset constant. In this strategy, test dataset has been divided into two without any further shuffling where the first half's length is ten. In zero-shot configuration, this dataset was not introduced into the training data. In one-shot configuration first data was taken and added into the training data. Finally, for ten-shot, all the data from the first half was introduced into the training data. On the other hand, the second half was used as test dataset and kept constant on all shot configurations.

This configuration can be described with the following statements: *Global* represents global dataset, *Unseen* represents dataset with unseen labels, *NotUnseen* represents dataset with labels that are not tagged as unseen labels. *Training* is 85% of the dataset *NotUnseen* as the training dataset, and *V*alidation is the 0.15 of the dataset *NotUnseen* as validation dataset. *ShotNumber* represents the current shot number that can be 0, 1 or 10. Here, the dataset *Unseen* is divided into two as *UnseenReserved* with constant length of 10 that represents unseen class examples that were reserved to be introduced into the training dataset, and *Test* represents remaining dataset from this split operation. If *ShotNumber* is other than 0, then this number of data were taken from *UnseenShot* and added into *Training*. In example, assume datasets have the following lengths, *Global* is 1200, *Unseen* is 200, then *NotUnseen*'s length becomes 1000, *Training* becomes 850 and *Validation* becomes 150. As stated earlier, *UnseenReserved*'s length is 10. Then the length of *Test* becomes 190. In zero-shot option none of reserved data introduced into the training data, so length of *Training*

dataset stays as 850. In one-shot, *Training* dataset becomes 851, and in ten-shot becomes 860 since that number of unseen class examples are introduced into the *Training* dataset.

4.2.2.2 Semantic Relationship Configurations

As explained in detail in the section Semantic Relationship of Labels, there are three different label groups as top-level, bottom-level and mid-level labels. To keep the number of training data in a reasonable limit, hyponyms for top-level labels, hypernyms for bottom-level labels are configured. This configuration was done with following steps, first each labels hyponyms and hypernyms are identified. If the label is described as top-level labels that only have hyponyms, then two configurations have been created, one of them where the hyponyms are in the training dataset, and the other one that hyponyms are removed from the dataset. For bottom-level labels, same two configurations were created for their hypernyms. Assume that *Phenomenon* is the current unseen class, which is a top-level label, so it only has hyponyms. Its hyponyms are *Natural Phenomenon* and *Human-Sourced Phenomenon.* To train the model in absence of hyponyms, these hyponyms are removed from the training and validation dataset. When the unseen class is *Natural Phenomenon* since its only hypernym is *Phenomenon*, to train the model in absence of hypernyms, all data tagged as *Phenomenon* were removed from the training dataset. Semantic relationships schema can be seen in the Appendix 1 Semantic Relationship Figures.

4.2.3 Training with News Dataset

To conduct the experiments, the model has been trained with domain-specific news dataset with the configurations described above. In the JSONL dataset obtained from labelling process with Doccano, each sentence is converted in a way that each entity class can be represented as "{The Entity}.{The Sentence Contains the Entity}". These structures were converted into Huggingface Dataset and fed into the Huggingface's trainer class. Train-validation-test split was discussed in the Shot Configurations section. The trainer configuration was taken from the study of Kosprdic[5].

4.2.4 News Data and Artificially Generated Data NER Model

To conduct the experiments with artificially extended dataset, JSONL dataset obtained from Doccano after AI dataset labelling process was concatenated with the JSONL file of news dataset. Then the same processes have been executed that explained in detail in the previous section Training with News Dataset.

CHAPTER V
RESULTS AND DISCUSSIONS

This section shows and discussed the results obtained during the testing step right after the training procedure. It divided into two subsections according to configurations where the first one shows shot comparions, prediction performance of zero-shot, one-shot and ten-shots results, all comparisons have been made in presence and absence of semantically related classes.

The second subsection's main focus is to compare the effect of semantically related classes when they are in the training data for subject entity. In this subsection prediction results of zero-shot, one-shot and ten-shots have been investigated separately.

5.1 SHOT COMPARISON

In the section, all models that are trained in this study are shown according to the number of unseen class examples introduced into the training data, in other terms shot number. At the end of this section, prediction performance of the model that was trained with four different dataset combination have been compared which are only news dataset, generic dataset and news dataset combined, news and ai generated datasets combined and as last one generic, news and ai generated datasets combined.

5.1.1 News Dataset

This section shows and compares the Named Entity Recognition prediction results of DistilBERTurk for zero-shot, one-shot and ten-shots setups. In this section, the model trained with only our news dataset.

Table 7: News Dataset T-Test Results According to Shot Numbers

	T-Test Score Without and With Relative		**P Value Without and With Relative**	
One shot vs Zero shot	7.07E-01	2.02E-01	2.43E-01	4.21E-01
Ten shot vs One shot	1.25E+00	2.98E+00	1.16E-01	5.00E-03
Ten shot vs Zero shot	1.24E+00	6.75E-01	1.17E-01	2.55E-01

Table 7 shows the T-Test results of F1 scores different shot configurations as seen in the first column. The T-Test compares the prediction performance of two different shot option. Also, presence and absence of semantically related classes have has been considered.

Except from the P value of Ten-Shot vs One-Shot and semantically related data are in the training dataset, which is smaller than 0, there is not clear evidence to support that introducing unseen classes into the training dataset improves the prediction performance of the model.

Table 8: News Dataset F1 Scores According to Shot Numbers

Unseen Class	**Zero Shot F1 Without and With Relative**		**One Shot F1 Without and With Relative**		**Ten Shot F1 Without and With Relative**	
Aquatic Biota	0.4675	0.6200	0.5778	0.6364		
Aquatic Ecosystem	0.2490	0.2439	0.2813	0.2155	0.2619	0.2531
Biota	0.1447	0.3760	0.1358	0.3137	0.1325	0.3306
Ecological Impact	0.7438	0.7234	0.7098	0.7218	0.7360	0.6960
Economic Impact	0.5949	0.6182	0.5854	0.6550		
Ecosystem	0.1530	0.2411	0.1791	0.2254	0.1515	0.2821
Environment Management	0.2727	0.4400	0.3102	0.4145		
Environmental Impact	0.1128	0.7732	0.1225	0.7612	0.1242	0.7616
Gas Pollutant	0.3500	0.4286	0.3721	0.4286		
Health Impact	0.8497	0.8289	0.8554	0.8454	0.8611	0.8376
Human	0.0588	0.0619	0.0632	0.0600	0.1031	0.0971
Human-Sourced Phenomenon	0.2972	0.3356	0.3000	0.3394	0.3656	0.3469
Human-Sourced Polluter	0.2723	0.2527	0.2396	0.2857	0.2360	0.2959
Liquid Pollutant	0.2000	0.1905	0.1600	0.2642		
Mitigation	0.0769	0.0769	0.0769	0.0714		
Natural Phenomenon	0.3910	0.5881	0.3946	0.3462	0.3551	0.4200
Natural Polluter	0.2727	0.2727	0.3636	0.3200		
Non-Anthropogenic Biota	0.3043	0.4842	0.3182	0.4835	0.2796	0.5714
Phenomenon	0.1014	0.4133	0.0789	0.4069	0.1050	0.5098
Pollutant	0.1274	0.1911	0.0964	0.1600	0.1739	0.2370
Polluter	0.1389	0.2667	0.1531	0.2973	0.1754	0.2692
Regulation	0.5439	0.5138	0.4909	0.5310		
Settlements	0.1148	0.2800	0.1094	0.3000		
Solid Pollutant	0.2222	0.1395	0.1923	0.2051		
Terrestrial Biota	0.3415	0.3556	0.4783	0.4444		
Terrestrial Ecosystem	0.2143	0.3146	0.1860	0.3505		
Well-Being Impact	0.2111	0.8059	0.1643	0.8150	0.6505	0.8408

F1 scores of each label in presence and in absence of relatives for zero-shot, one-shot and ten-shots can be in seen in Table 8. There are empty cells in the table since these labels do not have enough data to create ten-shots training configurations when only news dataset have been used.

In Figure 8, F1 score results difference of model's prediction in absence of relative classes can be seen. In the graph, the class named Well-Being Impact most benefits from increased number of shots. On the other hand, one-shot and ten-shots prediction results of classes such as Biota, Ecological Impact and Human-Sourced Pollutant is lower than the zero-shot training predictions.

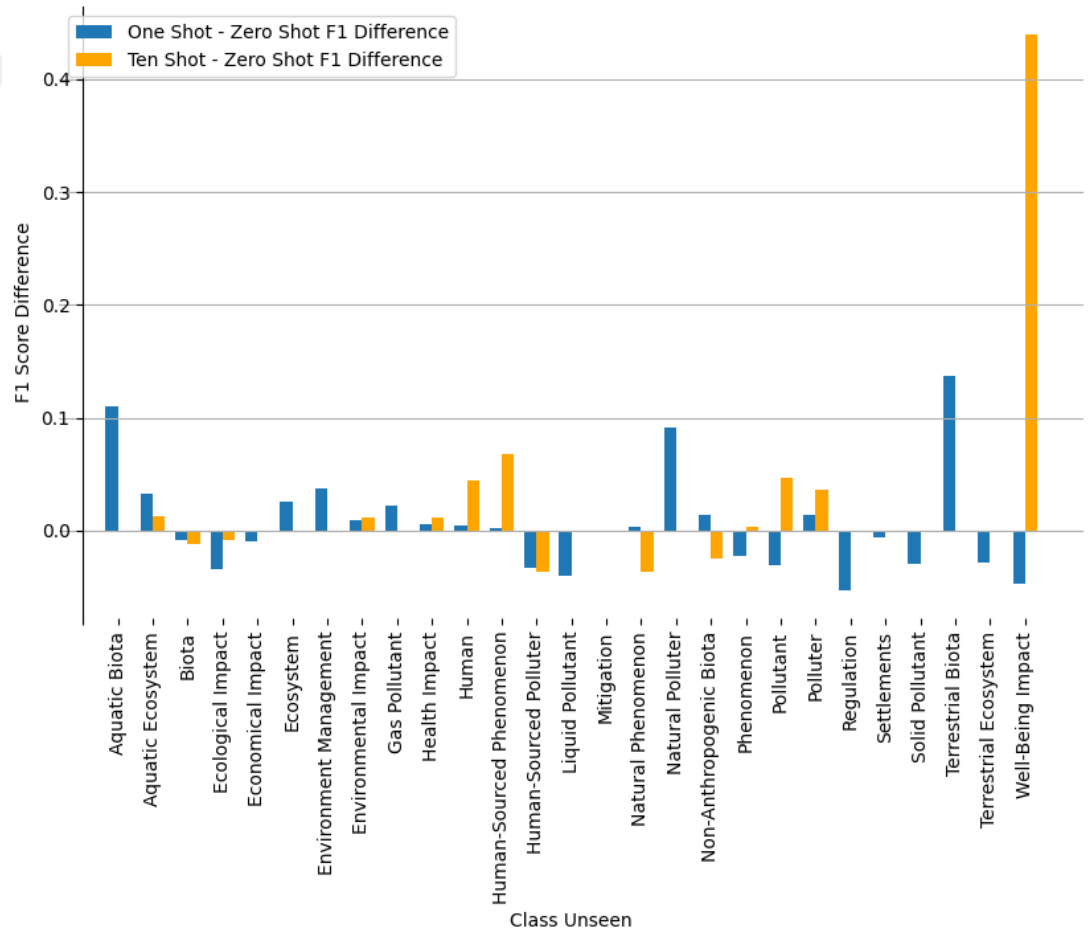

Figure 8: News Dataset F1 Scores Differences According to Shot Numbers Without Relatives

Figure 9 shows F1 scores differences for labels in the predictions according to the shot numbers in presence of related classes, where the blue line shows difference between one-shot and zero-shot training predictions, the yellow line shows difference between ten-shots training predictions and zero-shot training predictions. As seen in

the graph, there is an increase in 14 of 27 classes' one-shot predictions and some of them do not change. Also, there is a general trend in ten-shots predictions where almost all of them outperform the zero and one-shot predictions. On the other hand, there are some decreases in unseen class prediction such as, Biota, Environmental Impact, and by far the most decrement in Natural Phenomenon. It may be caused by that the model cannot create proper relationships with additional examples and these examples may create an overfitting problem.

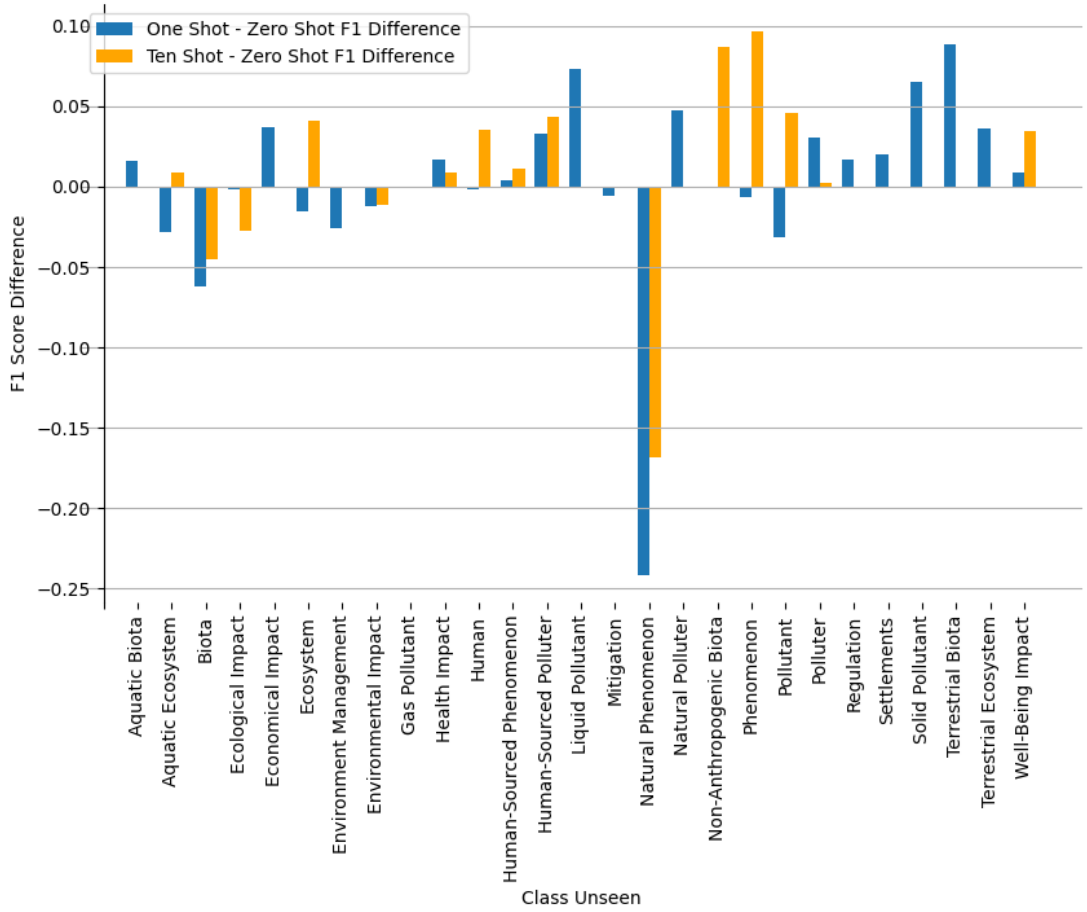

Figure 9: News Dataset F1 Scores Differences According to Shot Numbers With Relatives

5.1.2 Turkish Wiki NER and News Datasets

In this section, NER prediction result for DistilBERTurk that has been trained with Turkish Wiki NER dataset and then later with News Dataset is shown.

Table 9: Turkish Wiki NER and News Datasets T-Test Results According to Shot Numbers

	T-Test Score Without and With Relative		**P Value Without and With Relative**	
One shot vs Zero shot	2.50E+00	2.43E+00	9.43E-03	1.11E-02
Ten shot vs One shot	1.70E+00	1.98E+00	5.53E-02	3.38E-02
Ten shot vs Zero shot	2.23E+00	2.46E+00	2.13E-02	1.39E-02

As seen in T-Test results that shown in Table 9, when there are not semantically related classes in the training dataset introducing one or more unseen classes improves the prediction performance of the model. On the other hand, there is not enough evidence to support the claim ten-shots has higher prediction performance than one-shot training in absence of related classes. The rightmost column of the table shows prediction performance comparison in presence of related classes, and it can be said that ten-shots prediction performance is better than one-shot and zero-shot, and one-shot prediction performance is better than zero-shot when the model is pretrained with a generic dataset, trained and tested with news dataset.

Table 10: Turkish Wiki NER and News Datasets F1 Scores According to Shot Numbers

Unseen Class	**Zero Shot F1 Without and With Relative**		**One Shot F1 Without and With Relative**		**Ten Shot F1 Without and With Relative**	
Aquatic Biota	0.5612	0.6372	0.6774	0.8224		
Aquatic Ecosystem	0.1683	0.1856	0.3417	0.5217	0.6491	0.5799
Biota	0.1117	0.4313	0.1557	0.4748	0.1053	0.4615
Ecological Impact	0.4974	0.6536	0.4823	0.6299	0.7121	0.7081
Economic Impact	0.5969	0.6136	0.6023	0.6136		
Ecosystem	0.2246	0.4487	0.2093	0.3972	0.2082	0.5490
Environment Management	0.4645	0.6008	0.2577	0.4550		
Environmental Impact	0.2345	0.7225	0.1825	0.7353	0.2956	0.7376
Gas Pollutant	0.4091	0.3810	0.4583	0.4500		
Health Impact	0.8107	0.8054	0.8203	0.7770	0.7606	0.7925
Human	0.2000	0.2701	0.2136	0.1695	0.1270	0.1967
Human-Sourced Phenomenon	0.2613	0.2941	0.3772	0.4149	0.4222	0.4633
Human-Sourced Polluter	0.5214	0.5176	0.5425	0.5856	0.5347	0.4479
Liquid Pollutant	0.4286	0.4783	0.4444	0.4490		
Mitigation	0.2581	0.0714	0.4000	0.2581		
Natural Phenomenon	0.5359	0.5084	0.5408	0.5734	0.5444	0.5870
Natural Polluter	0.5000	0.4444	0.5217	0.5217		
Non-Anthropogenic Biota	0.4891	0.7521	0.4970	0.7778	0.4516	0.7759
Phenomenon	0.1427	0.4766	0.0935	0.5194	0.1753	0.6045
Pollutant	0.2116	0.3737	0.1951	0.3687	0.2525	0.4094
Polluter	0.0585	0.3200	0.2162	0.3458	0.2691	0.5174
Regulation	0.3960	0.3030	0.5299	0.5781		
Settlements	0.2182	0.2824	0.2182	0.3173		
Solid Pollutant	0.4231	0.3913	0.5200	0.4727		
Terrestrial Biota	0.7619	0.8077	0.8000	0.7586		
Terrestrial Ecosystem	0.0845	0.2368	0.1818	0.3210		
Well-Being Impact	0.3039	0.7707	0.5082	0.7965	0.5776	0.8114

Table 10 shows F1 scores of each class. Since there aren't enough examples of some labels in the news dataset to train for ten-shots configuration, there are empty cells for ten-shots F1 scores in the table.

Figure 10, shows F1 score differences for one-shot and zero-shot training predictions with zero-shot training predictions of the model pretrained with Turkish Wiki NER and trained with news data. As seen in the graph, most of the classes benefit from introduced examples in the training data. On the other hand, there are some classes such as Health Impact, Human and Non-Anthropogenic data which introducing ten examples reduces prediction performance of the model which may be caused by overfitting.

Figure 11 shows F1 scores when semantically related classes are in the training dataset, and it shows similar differences compared to Figure 10. The effect of relative classes in the training data is discussed in detail in the section 5.2.

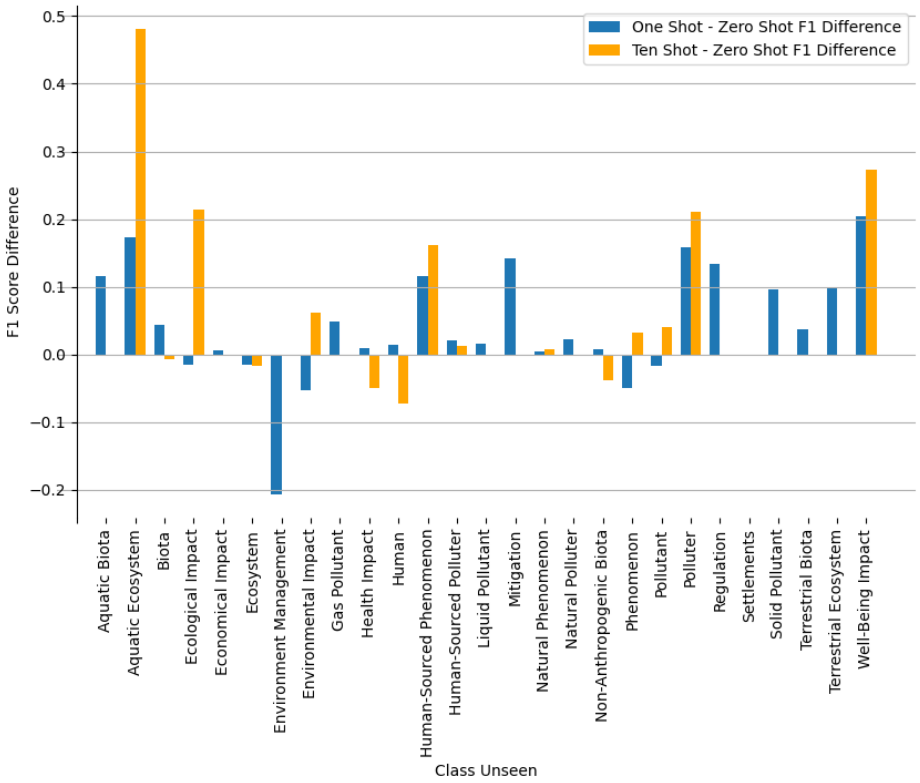

Figure 10: Turkish Wiki NER and News Dataset F1 Scores Differences According to Shot Numbers without Relatives

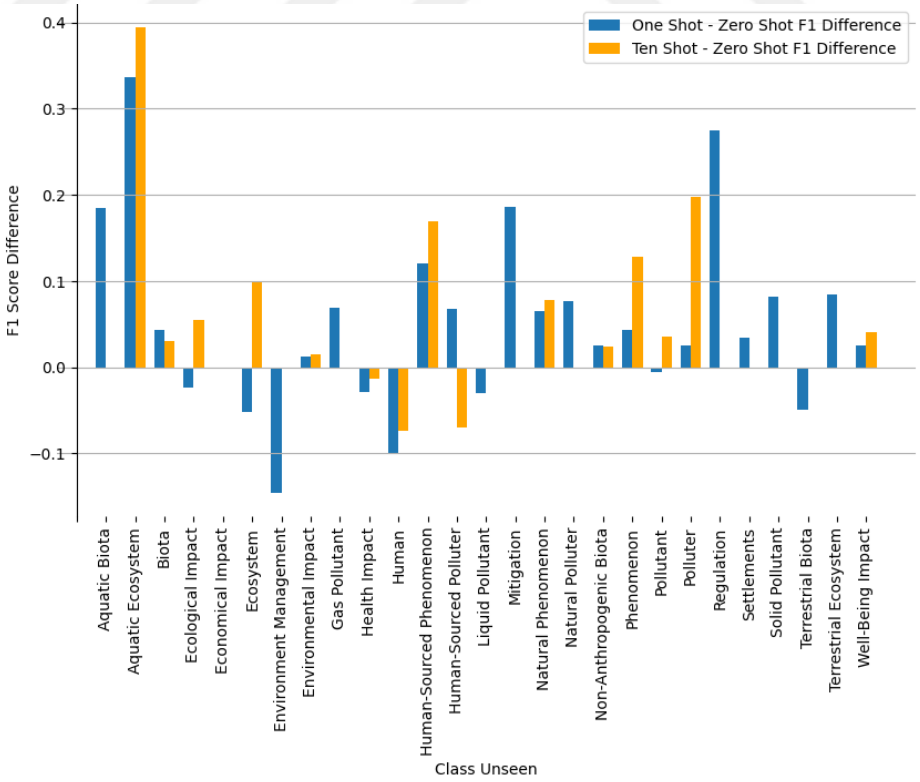

Figure 11: Turkish Wiki NER and News Dataset F1 Scores Differences According to Shot Numbers with Relatives

5.1.3 News and AI Datasets

In this section, F1 results for Named Entity Recognition predictions of the models that has been trained with news and AI datasets are shown.

Table 11: News and AI Datasets T-Test Results According to Shot Numbers

	T-Test Score Without and With Relative		**P Value Without and With Relative**	
One shot vs Zero shot	1.30E+00	1.47E+00	1.03E-01	7.66E-02
Ten shot vs One shot	1.22E+00	3.46E+00	1.18E-01	1.25E-03
Ten shot vs Zero shot	2.34E+00	2.89E+00	1.48E-02	4.52E-03

Table 11 demonstrates that when one example of an unseen class—neither the relative classes present nor absent—is added to the model, the prediction performance of the model trained on news and AI datasets does not change in a meaningful manner. However, ten-shots training prediction performance differs significantly from one-shot training prediction when related classes are included in the training data. The model performs much better in ten-shots training than in zero-shot training in both scenarios where related classes are present in the training data and those that are not.

Table 12: News and AI Datasets F1 Scores According to Shot Numbers

Unseen Class	**Zero Shot F1 Without and With Relative**		**One Shot F1 Without and With Relative**		**Ten Shot F1 Without and With Relative**	
Aquatic Biota	0.7568	0.7586	0.8257	0.7925	0.6538	0.7500
Aquatic Ecosystem	0.6716	0.7232	0.7034	0.6389	0.6926	0.7096
Biota	0.0876	0.5405	0.1038	0.5570	0.0989	0.6383
Ecological Impact	0.7936	0.7754	0.8033	0.7667	0.7823	0.7616
Economic Impact	0.7289	0.7284	0.7207	0.7688	0.8057	0.7602
Ecosystem	0.1399	0.4795	0.1732	0.4489	0.1416	0.6035
Energy	0.4000	0.3158	0.3158	0.5385		
Environment Management	0.1793	0.3851	0.2517	0.4162	0.3222	0.5368
Environmental Impact	0.1739	0.8071	0.1682	0.8196	0.3814	0.8202
Gas Pollutant	0.4746	0.4231	0.4231	0.4231		
Health Impact	0.8428	0.8288	0.8454	0.8433	0.8441	0.8371
Human	0.1517	0.2115	0.2618	0.1875	0.3602	0.1630
Human-Sourced Phenomenon	0.4224	0.5989	0.5760	0.5836	0.5682	0.6766
Human-Sourced Polluter	0.5592	0.4840	0.5215	0.5248	0.5738	0.6039
Liquid Pollutant	0.3721	0.2466	0.2439	0.3797		
Mitigation	0.2353	0.2222	0.1923	0.2182		
Natural Phenomenon	0.7502	0.8054	0.8201	0.8571	0.7729	0.8379
Natural Polluter	0.5476	0.6027	0.6410	0.6105		
Non-Anthropogenic Biota	0.4724	0.7510	0.4351	0.7295	0.5296	0.7368
Phenomenon	0.0888	0.7734	0.0789	0.7610	0.0957	0.7903
Pollutant	0.1586	0.2689	0.1742	0.2490	0.1793	0.2867
Polluter	0.2479	0.4246	0.2284	0.4324	0.3018	0.4936
Regulation	0.5854	0.6142	0.5470	0.6567		
Settlements	0.1217	0.2314	0.1452	0.2680	0.1284	0.3486
Social Impact	0.6582	0.6377	0.6133	0.6032		
Solid Pollutant	0.4783	0.4390	0.5000	0.6364		
Terrestrial Biota	0.7402	0.7639	0.8456	0.7681	0.7606	0.8511
Terrestrial Ecosystem	0.5912	0.6875	0.5625	0.5752	0.5180	0.6289
Treatment	0.4848	0.4390	0.5000	0.4571		
Well-Being Impact	0.3034	0.8369	0.4778	0.8360	0.7600	0.8309

Table 12 has some empty results since the number of examples of classes with empty results are not enough to prepare the model for ten-shots training and prediction.

Figure 12 shows F1 score differences of one-shot and ten-shots training predictions with zero-shot training predictions. In the graph, most of the unseen classes benefit from introduced examples especially in ten-shots training configuration. Figure 13 shows F1 score results of the model in presence of relative classes.

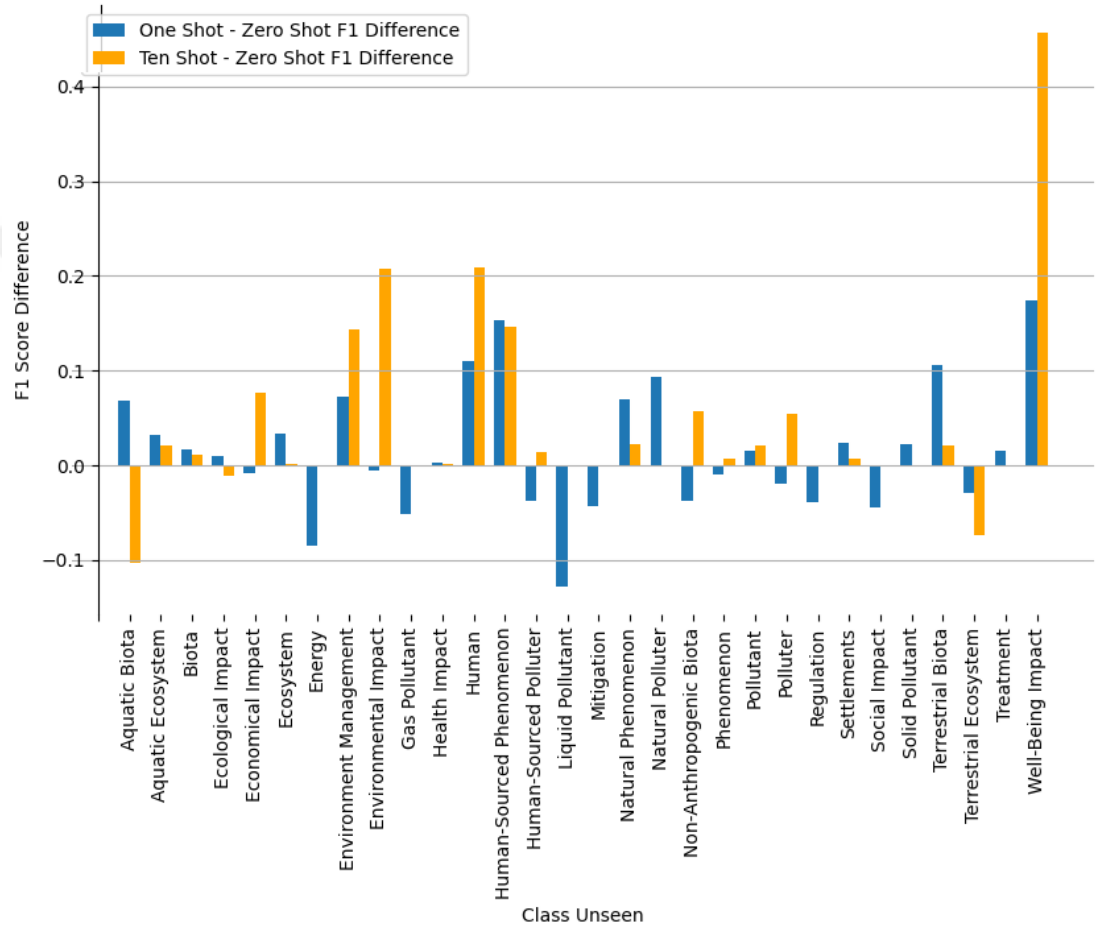

Figure 12: News and AI Dataset F1 Scores Differences According to Shot Numbers without Relatives

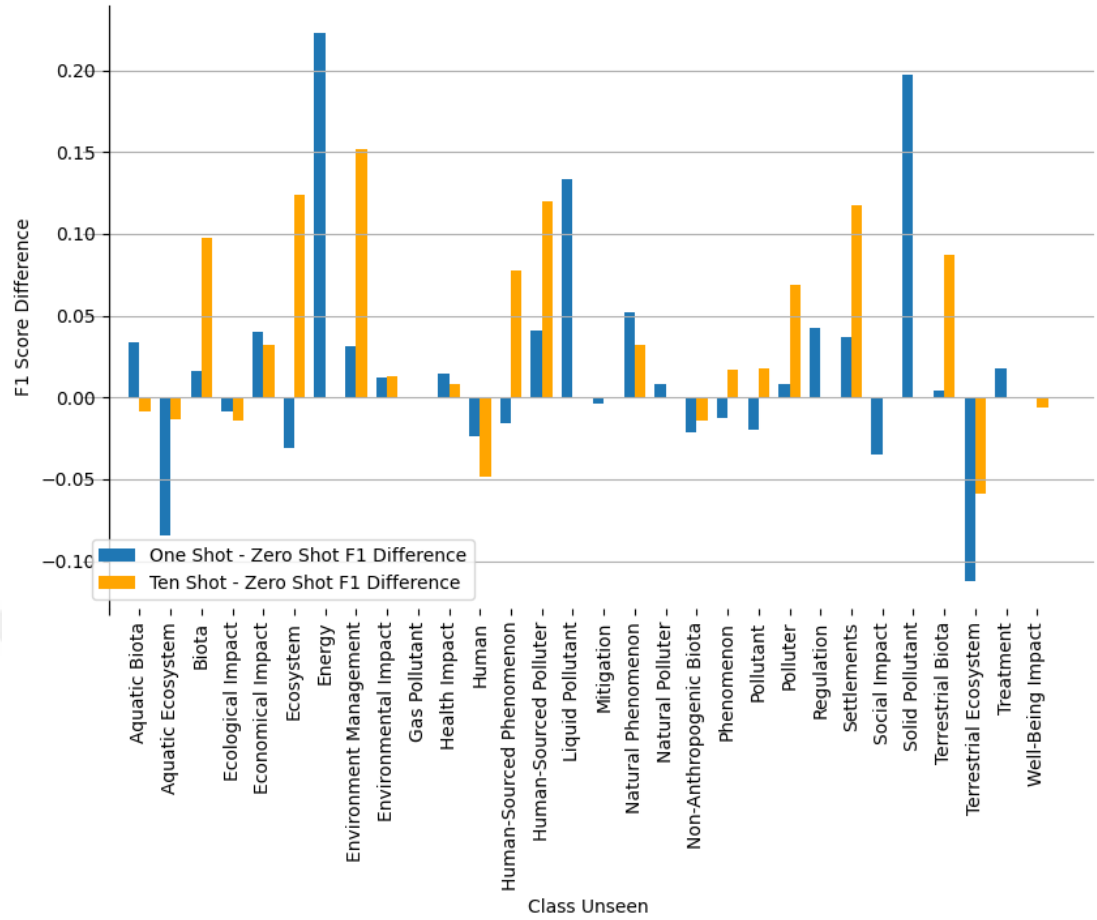

Figure 13: News and AI Dataset F1 Scores Differences According to Shot Numbers with Relatives

5.1.4 Turkish Wiki NER, News and AI Datasets

In this section, F1 scores for Named Entity Recognition prediction results of DistilBERTurk, that has been trained initially with Turkish Wiki NER, then News and AI Datasets concatenated are shown.

Table 13: Turkish Wiki NER, News, AI Datasets T-Test Results According to Shot Numbers

	T-Test Score Without and With Relative		**P Value Without and With Relative**	
One shot vs Zero shot	1.94E+00	1.69E+00	3.14E-02	5.05E-02
Ten shot vs One shot	4.41E+00	2.75E+00	1.36E-04	6.17E-03
Ten shot vs Zero shot	4.17E+00	3.60E+00	2.36E-04	9.02E-04

T-Test results shown in Table 13 indicate that Turkish Wiki NER dataset combining with artificial intelligence and news data displays an enhancement in the model's prediction ability when one or more unseen classes are included to the training

set. This tendency is comparable to the performance of the same model trained with only Turkish Wiki NER and news data. Since, zero-shot training benefits greater than few-shot training predictions when there are related classes in the training data, the only failed test is one-shot training predictions over zero-shot training predictions in presence of semantically related classes. More details about effect of related classes in training data are available in the next section 5.2.

Table 14: Turkish Wiki NER, News and AI Datasets F1 Scores According to Shot Numbers

Unseen Class	**Zero Shot F1 Without and With Relative**		**One Shot F1 Without and With Relative**		**Ten Shot F1 Without and With Relative**	
Aquatic Biota	0.8793	0.8504	0.7736	0.8739	0.8544	0.7475
Aquatic Ecosystem	0.1857	0.5533	0.5899	0.7140	0.7401	0.7528
Biota	0.1596	0.5328	0.2582	0.6414	0.3280	0.6489
Ecological Impact	0.3966	0.5062	0.7347	0.7349	0.7898	0.7742
Economic Impact	0.7219	0.7614	0.7701	0.7718	0.7919	0.7884
Ecosystem	0.2423	0.5886	0.2308	0.5879	0.2853	0.7295
Energy	0.2000	0.2105	0.1905	0.3077		
Environment Management	0.3478	0.4857	0.4892	0.5929	0.3818	0.4729
Environmental Impact	0.2301	0.8144	0.2187	0.8190	0.3297	0.8141
Gas Pollutant	0.6563	0.5660	0.5417	0.5306		
Health Impact	0.8393	0.8212	0.7945	0.8113	0.8240	0.8182
Human	0.2462	0.2326	0.2269	0.3110	0.4239	0.3980
Human-Sourced Phenomenon	0.3492	0.6369	0.3583	0.5848	0.6308	0.7042
Human-Sourced Polluter	0.6667	0.6146	0.6734	0.5685	0.6938	0.6822
Liquid Pollutant	0.5500	0.2593	0.4762	0.3467		
Mitigation	0.2642	0.3582	0.3492	0.3279		
Natural Phenomenon	0.8037	0.8560	0.8743	0.8409	0.8681	0.8663
Natural Polluter	0.7677	0.6667	0.7692	0.7816		
Non-Anthropogenic Biota	0.5079	0.6777	0.5907	0.7368	0.6107	0.8507
Phenomenon	0.0594	0.8072	0.0532	0.8074	0.2202	0.7980
Pollutant	0.1754	0.3742	0.2157	0.3583	0.3077	0.4219
Polluter	0.1071	0.3835	0.1155	0.2120	0.2507	0.5149
Regulation	0.2941	0.4870	0.6412	0.5378		
Settlements	0.1767	0.2482	0.2115	0.2623	0.4060	0.5667
Social Impact	0.6329	0.6173	0.4848	0.6024		
Solid Pollutant	0.5965	0.5556	0.5484	0.5926		
Terrestrial Biota	0.7559	0.7846	0.7727	0.7669	0.8252	0.8095
Terrestrial Ecosystem	0.5468	0.5522	0.4604	0.4959	0.4559	0.6438
Treatment	0.5455	0.5532	0.6842	0.6111		
Well-Being Impact	0.2101	0.7824	0.4372	0.7293	0.6508	0.7681

Same with the Table 13, since both dataset configurations use the same domain-specific datasets, Table 14 has some empty values for ten-shots training prediction results where the number of examples for that class is not enough to create proper setup.

Figure 14 shows promising results where almost all the unseen classes benefit from more examples in the training data. On the other hand, in Figure 15, number of classes that have lower prediction for one-shot training prediction increases.

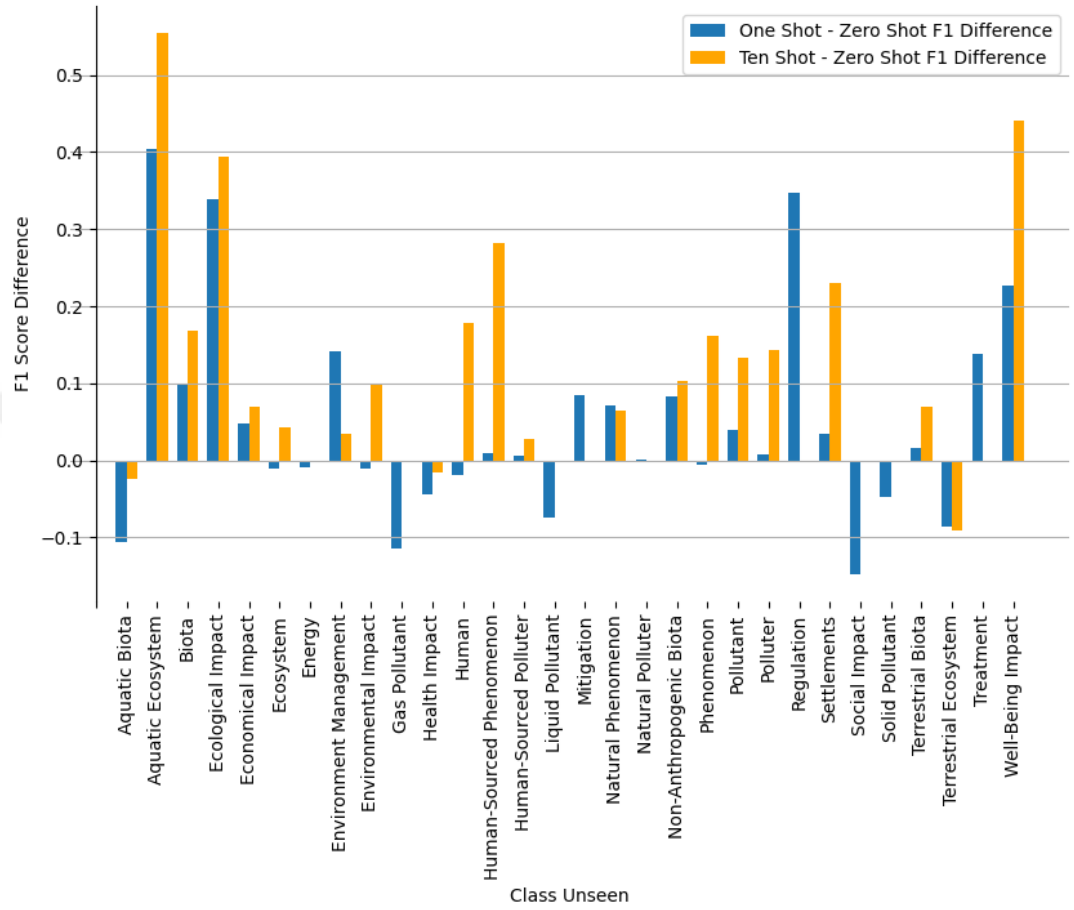

Figure 14: Turkish Wiki NER, News and AI Dataset F1 Scores Differences According to Shot Numbers with Relatives

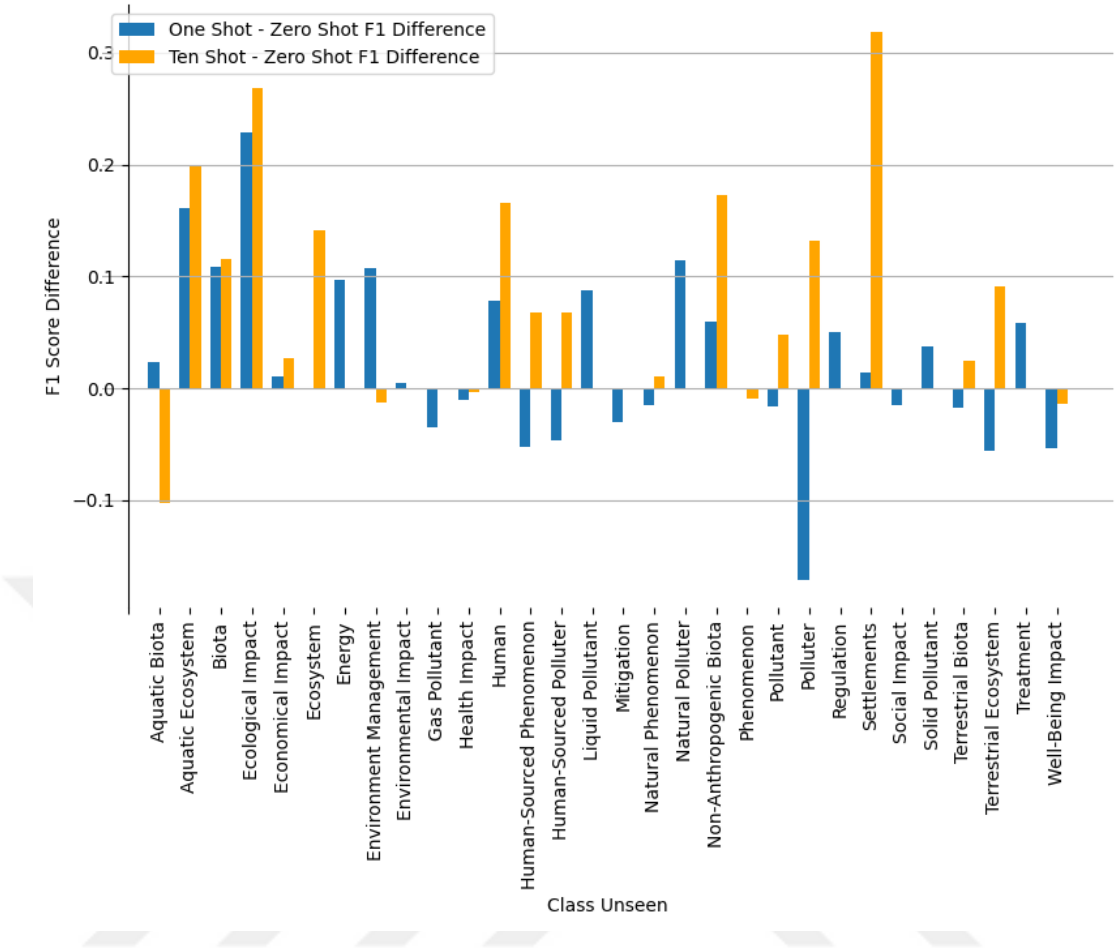

Figure 15: Turkish Wiki NER, News and AI Dataset F1 Scores Differences According to Shot Numbers without Relatives

5.1.5 Datasets Comparison

In this section, all configurations that has been discussed so far in Shot Configurations are compared. Since, the pivot dataset is our news dataset and it exists in all combinations, the other two datasets that are generic dataset and artificially generated dataset have been toggled in all configurations. The first subsection considers pretraining the model with generic Turkish Wiki NER dataset. All T-Tests in this subsection have been conducted to see if there is a significant difference in prediction performance when model is pretrained with a generic dataset. Second subsection considers presence of artificially generated dataset. All T-Tests in this subsection have been conducted to see if there is a significant difference in prediction performance when the news dataset is combined with artificially generated dataset.

5.1.5.1 Effect of Turkish Wiki NER Dataset with News Dataset

In this subsection two models are compared as the first one DistilBERTurk without any further pretraining procedure, and DistilBERTurk pretrained with generic Turkish Wiki NER dataset. As stated earlier, T-Tests are conducted to test to see pretraining the model improves its prediction performance.

5.1.5.1.1 Without AI Generated Data

Table 15: T-Test Scores of Turkish Wiki NER and News Datasets compared to News Dataset

Results	T-Test Score	P-Value
Zero-Shot Results	3.79E+00	1.50E-04
One-Shot Results	6.54E+00	3.05E-09
Ten-Shots Results	5.00E+00	5.31E-06

In this test that the results are shown in Table 15, all trainings have been conducted with only news dataset without combining artificially generated data. As seen in the table, P-Value of each shot option is lower than 0.05 and it means all tests shows there are significant improvement in terms of prediction performance when the model is pretrained with generic Turkish Wiki NER dataset.

5.1.5.1.2 Combined with AI Generated Data

Table 16: T-Test Scores of Turkish Wiki NER, News Combined With AI Datasets compared to News Combined with AI Dataset

Results	T-Test Score	P-Value
Zero-Shot Results	-1.27E-01	5.50E-01
One-Shot Results	5.73E-01	2.84E-01
Ten-Shots Results	2.93E+00	2.43E-03

In this test that the results are shown in Table 15, all trainings have been conducted with news dataset combined with artificially generated data. As seen in Table 16, pretraining the model with a generic dataset does not change significantly the prediction results of the model apart from ten-shots training.

5.1.5.2 Effect of AI Generated Dataset

In this section, F1 score T-Tests of the models that have been trained with and without artificially generated data are shown. Since, the model pretrained with Turkish Wiki NER and News combined with AI generated datasets was shown in previous section as seen in Table 16, this comparison is skipped in this section. The tests are conducted with following hypothesis; combining artificially generated dataset with news dataset improves the performance of the model in terms of F1 score.

5.1.5.2.1 Without Turkish Wiki NER Dataset

Table 17: T-Test Scores of News Dataset compared to News Combined with AI Datasets

Results	T-Test Score	P-Value
Zero-Shot Results	9.13E+00	2.96E-14
One-Shot Results	9.90E+00	1.00E-15
Ten-Shots Results	7.88E+00	4.17E-10

T-Test results in the Table 17 shows, model trained with artificially generated dataset in addition to news dataset outperforms the model that have only trained with news dataset regardless of the introduced number of unseen class examples.

5.1.5.2.2 Pretrained with Turkish Wiki NER Dataset

Table 18: T-Test Scores of Pretrained, News Dataset compared to Pretrained News Combined with AI Datasets

Results	T-Test Score	P-Value
Zero-Shot Results	4.94E+00	2.19E-06
One-Shot Results	5.20E+00	8.07E-07
Ten-Shots Results	7.38E+00	2.09E-09

As seen in Table 18, training the model with news data combined with artificially generated dataset that is pretrained with a generic dataset such as Turkish Wiki NER has better outcomes in the scope of prediction performance and F1 score.

5.1.5.3 F1 Macro

In this section the model's prediction results for different shot and dataset combinations have been compared by using F1 macro scores.

Table 19: F1 Macro Scores of Dataset Combinations

Dataset Combination	Zero Shot	One Shot	Ten Shot
News Dataset	0.4013	0.4036	0.4499
Turkish Wiki NER and News Dataset	0.4733	0.5224	0.5761
News and AI Datasets	0.5601	0.5782	0.6508
Turkish Wiki NER, News and AI Datasets	0.5713	0.5953	0.6939

As seen in the Table 19, the highest F1 macro for zero shot result has been achieved with model pretrained with Turkish Wiki NER and trained with News and AI Datasets. It can be seen in the Figure 16, each complexity added to the dataset and each number of examples introduced into the model during training increases F1 macro score.

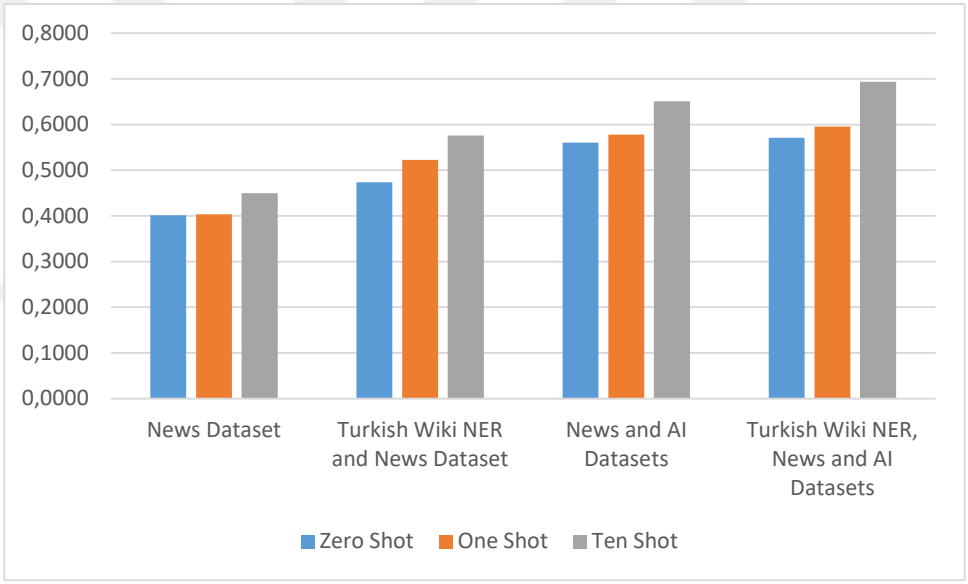

Figure 16: F1 Macro Scores of Dataset Combinations

5.1.6 Discussion

In the previous sections above, DistilBERTurk's prediction performance in terms of F1 score have been shown and later compared in four different dataset combinations which the first one is news dataset for training, the second one is generic dataset for pretraining and news dataset for training, the third one news dataset and artificially generated dataset for training and finally the fourth one is generic dataset for pretraining and news and artificially generated for training.

Individual combination results show that, if the model is trained with only news dataset only ten-shot training shows significant improvement over one-shot training.

On the other hand, neither one-shot training nor ten-shot training have superiority over zero-shot training. This situation changes when the model is pretrained with Turkish Wiki NER generic dataset. In this case, model benefits from increased number unseen class examples in the training dataset and its prediction performance improves in almost all configurations. When the model is not pretrained with generic dataset but on the other hand artificially generated data are introduced into the dataset, ten-shot learning outperforms almost all of the other shot options. On the other hand, one-shot learning does not create significant difference over zero-shot learning. Finally, if the model is pretrained with generic dataset and artificially generated data are introduced into the dataset, ten-shot learning outperforms other options and one-shot learning outperforms zero-shot learning. It means that, the most complicated dataset configuration benefits from the increased number of unseen class examples in the training data.

These comparisons have been made without considering absence of semantically related classes. It means all dataset and shot combinations have been evaluated when there are related classes of subject class in the training and validation data. The results of T-Test can be interpreted as follows; if the model is to be trained with domain specific dataset, pretraining the model with generic dataset improves its prediction performance regardless of number of unseen class examples in the training dataset. On the other hand, if the dataset consists of artificially generated data, the model only benefits in ten-shot configuration from the pretraining with generic dataset procedure. This may be because the presence of artificially generated data has already improved prediction performance to its limit for zero-shot and one-shot configurations. On the other hand, introducing AI data into the dataset significantly improves prediction performance regardless of shot configuration and pretraining with generic dataset.

According to the F1-macro scores of each dataset combination and shot number configuration, pretraining the model with Turkish Wiki NER or introducing artificially generated data into training data, or doing both operations together, prediction performance of the model increases. On the other hand, introducing artificially generated data into the training dataset outperforms pretraining the model with generic dataset. Best prediction results achieved when the model is pretrained and AI data introcuded into the training data.

5.2 EFFECTS OF SEMANTICALLY RELATED CLASSES

In this section, the presence of semantically hierarchical entities in the training dataset and their effects on the prediction performance are discussed. The tests are conducted with the following hypothesis: Introducing semantically related classes into the training dataset, prediction performance of the model in terms of F1 score increases.

5.2.1 Hyponyms

This section only discusses top-level classes, presence and absence of the classes as hyponyms in other terms that are more-specific classes below the subject unseen class in the semantical hierarchy. All T-Tests have been conducted to see significant improvements in F1 scores when hyponyms are introduced into the dataset.

5.2.1.1 News Dataset

In this subsection, model has been trained with news dataset. There is not a such as pretraining with generic dataset and combination of artificially generated data.

Table 20: Presence of Hyponyms, News Dataset Comparison

Results	**T-Test Score**	**P-Value**
Zero-Shot Results	3.03E+00	1.15E-02
One-Shot Results	2.72E+00	1.74E-02
Ten-Shots Results	2.79E+00	1.93E-02

Table 20 shows the T-Test score of the comparison of the predictions in presence and in absence of hyponyms in the training data for the model that is trained with news dataset. As seen in the results, the model benefits from presence of more specialized class definitions in the dataset.

Table 21: Presence of Hyponyms, News Dataset F1 Scores

Unseen Class	Shot Number	Without Hyponym	With Hyponym
Biota	0	0.1447	0.3760
	1	0.1358	0.3137
	10	0.1325	0.3306
Ecosystem	0	0.1530	0.2411
	1	0.1791	0.2254
	10	0.1515	0.2821
Environment Management	0	0.2727	0.4400
	1	0.3102	0.4145
Environmental Impact	0	0.1128	0.7732
	1	0.1225	0.7612
	10	0.1242	0.7616
Phenomenon	0	0.1014	0.4133
	1	0.0789	0.4069
	10	0.1050	0.5098
Pollutant	0	0.1274	0.1911
	1	0.0964	0.1600
	10	0.1739	0.2370
Polluter	0	0.1389	0.2667
	1	0.1531	0.2973
	10	0.1754	0.2692

Table 21 shows F1 scores of the model trained with news dataset. As seen in table, all classes' predictions have been increased when their hyponyms are introduced into the training dataset. There are some drastic improvements in unseen classes predictions such as Environmental Impact where its zero-shot training performance increases from 11% to 77%.

Figure 17, Figure 18, and Figure 19 shows how the prediction performance of the model changes when there are hyponyms in the training data. As seen in figures, all shot options show the same trend.

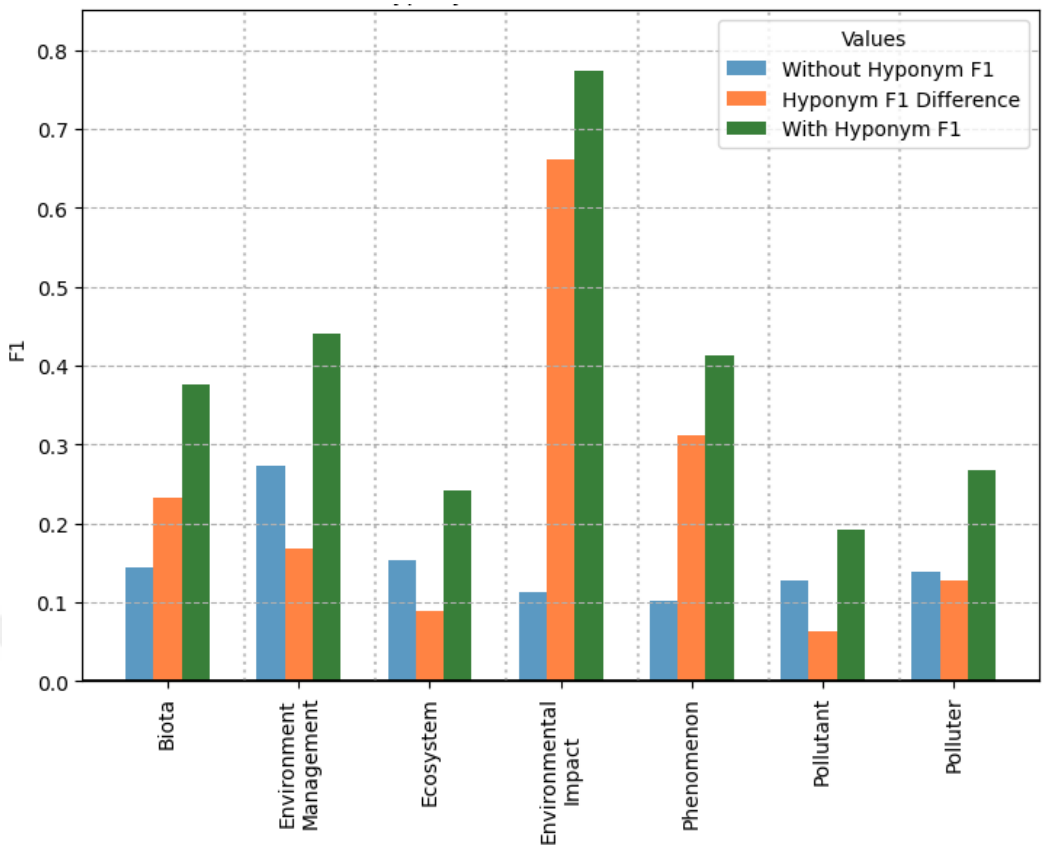

Figure 17: Presence of Hyponyms, News Data Zero-Shot F1 Scores

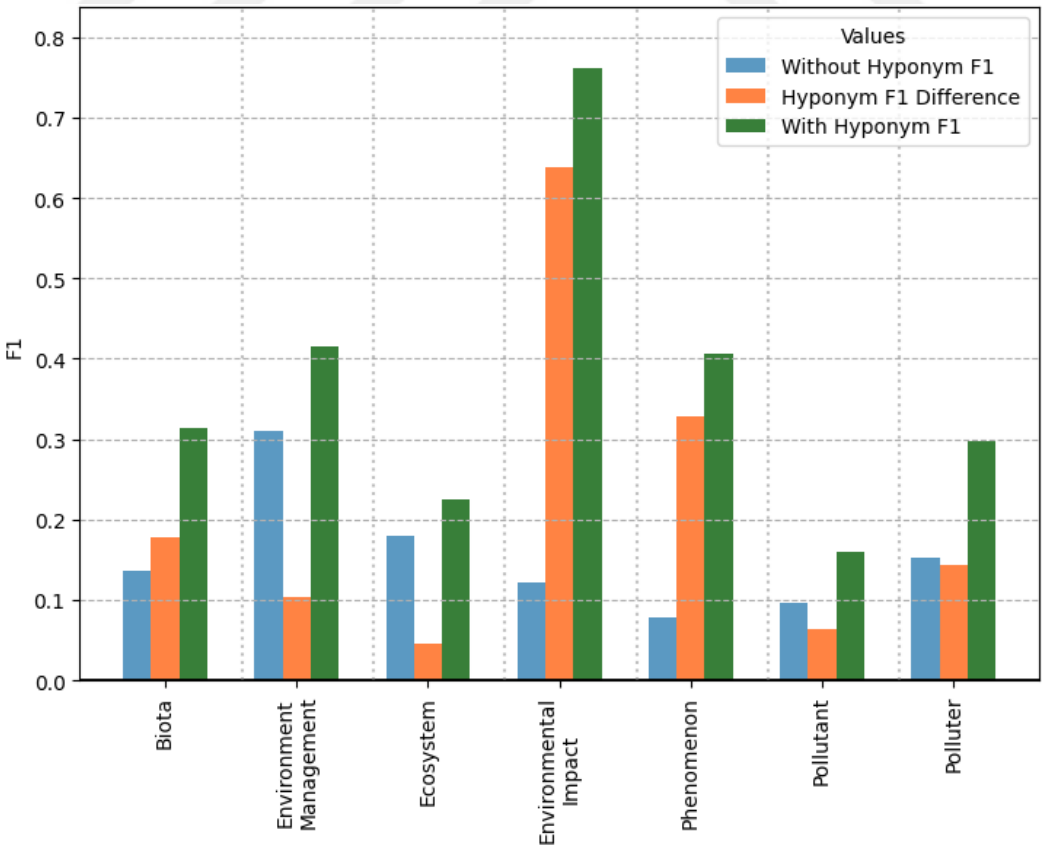

Figure 18: Presence of Hyponyms, News Data One-Shot F1 Scores

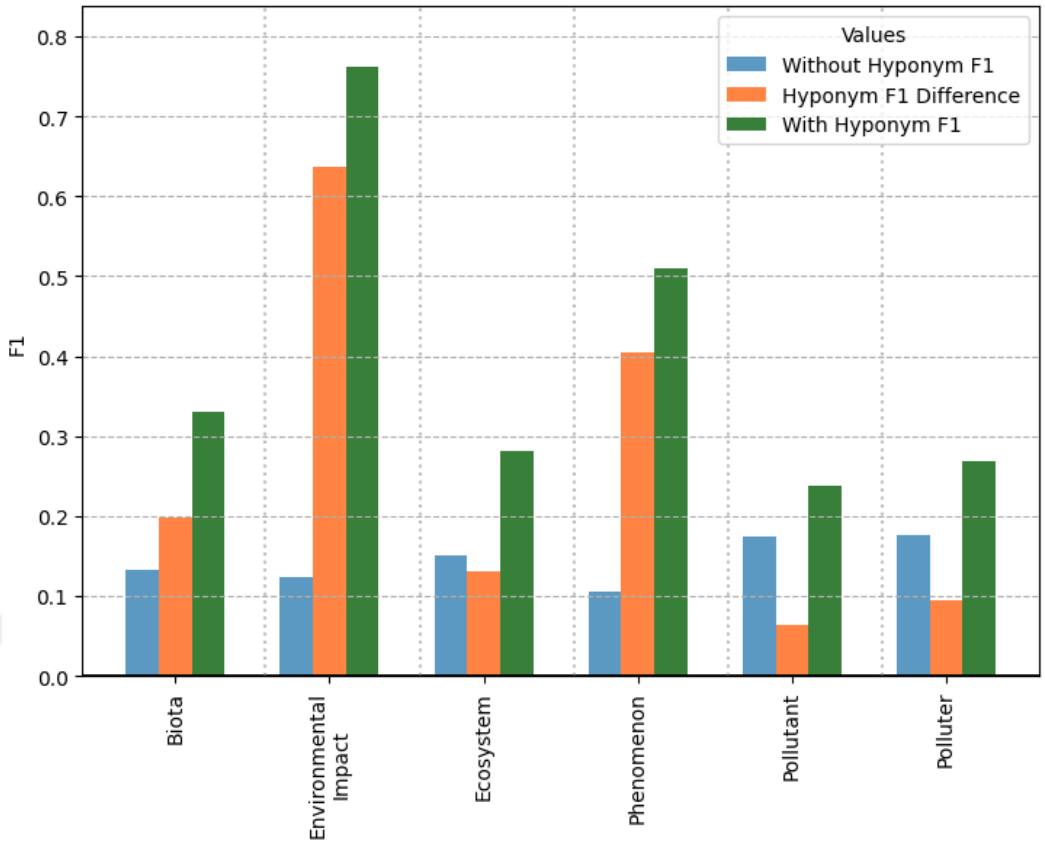

Figure 19: Presence of Hyponyms, News Data Ten-Shots F1 Scores

5.2.1.2 Turkish Wiki NER and News Datasets

This section contains results for the model that was gone through the pretraining with generic dataset process. The training dataset for downstream task does not contain artificially generated data.

Table 22: Presence of Hyponyms, Turkish Wiki NER and News Datasets Comparison

Results	**T-Test Score**	**P-Value**
Zero-Shot Results	6.09E+00	4.44E-04
One-Shot Results	4.81E+00	1.49E-03
Ten-Shots Results	7.36E+00	3.63E-04

As seen in Table 22, presence of hyponyms increases the prediction results of the model that is pretrained with a generic dataset and later with domain specific dataset regardless the number of unseen classes in the training data.

Table 23: Presence of Hyponyms, Turkish Wiki NER and News Datasets F1 Scores

Unseen Class	Shot Number	Without Hyponym	With Hyponym
Biota	0	0.1117	0.4313
	1	0.1557	0.4748
	10	0.1053	0.4615
Ecosystem	0	0.2246	0.4487
	1	0.2093	0.3972
	10	0.2082	0.5490
Environment Management	0	0.4645	0.6008
	1	0.2577	0.4550
Environmental Impact	0	0.2345	0.7225
	1	0.1825	0.7353
	10	0.2956	0.7376
Phenomenon	0	0.1427	0.4766
	1	0.0935	0.5194
	10	0.1753	0.6045
Pollutant	0	0.2116	0.3737
	1	0.1951	0.3687
	10	0.2525	0.4094
Polluter	0	0.0585	0.3200
	1	0.2162	0.3458
	10	0.2691	0.5174

Table 23, shows prediction results of all configurations increase when there are hyponyms in the training dataset.

Figure 20, Figure 21, and Figure 22 shows how F1 scores change when hyponyms are introduced into the training dataset.

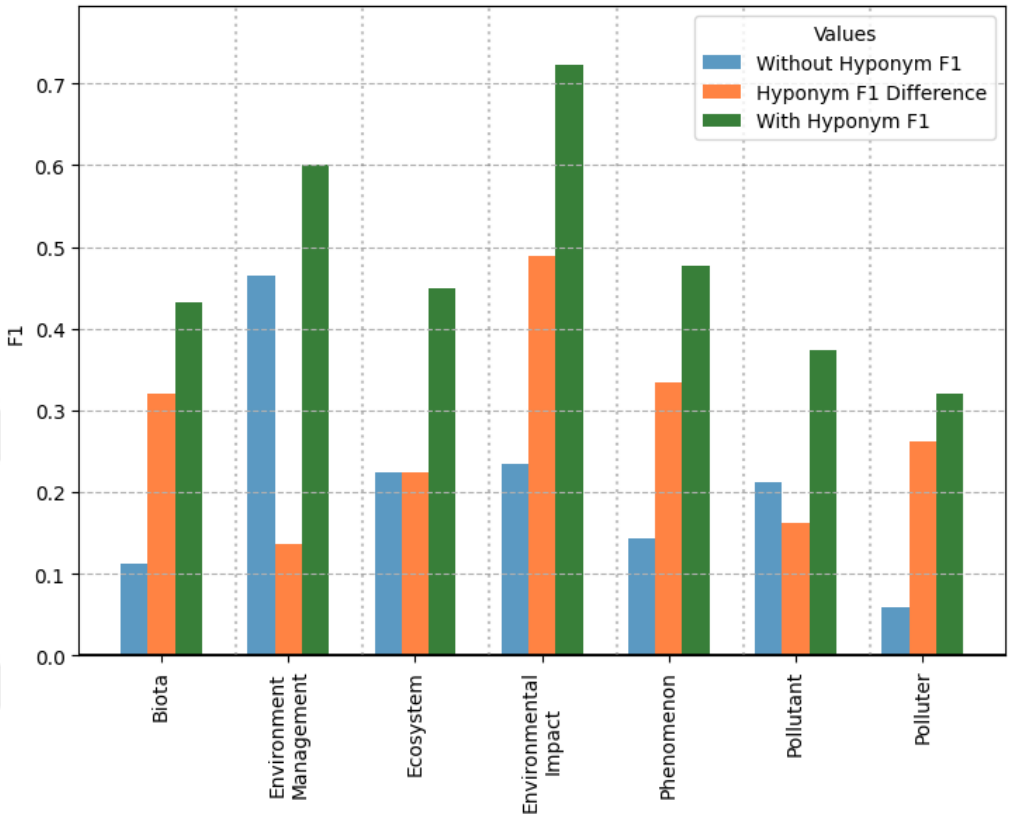

Figure 20: Presence of Hyponyms, Turkish Wiki NER and News Datasets, Zero-Shot F1 Scores

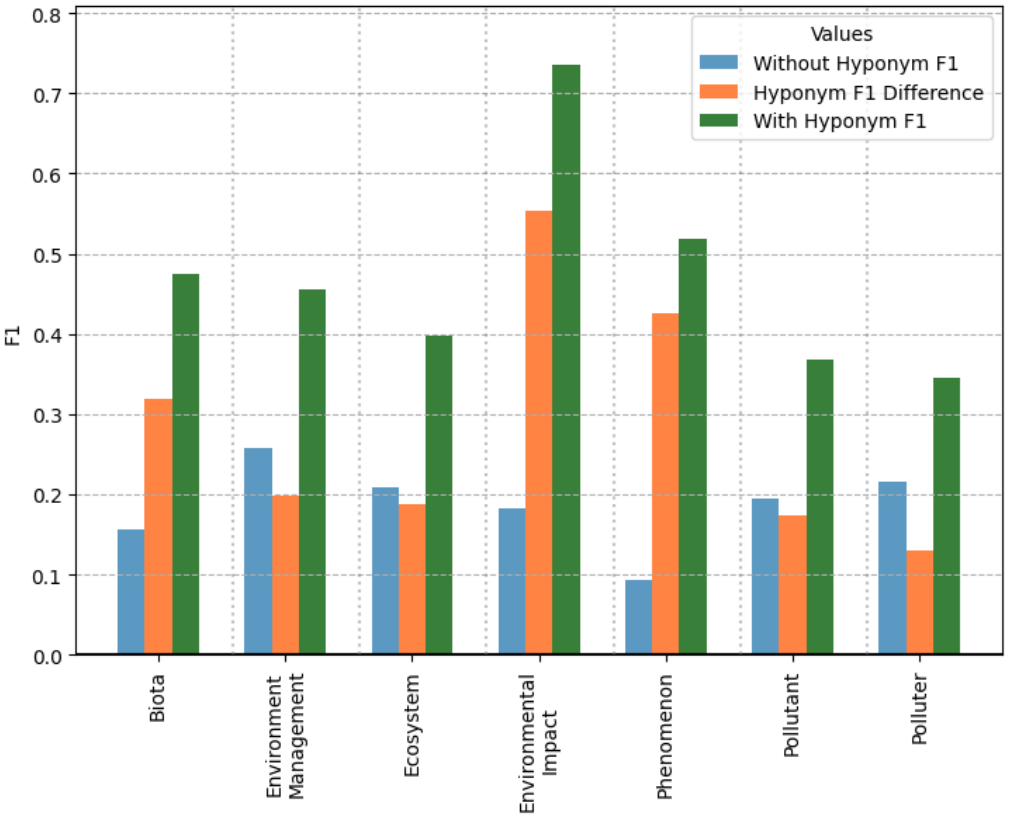

Figure 21: Presence of Hyponyms, Turkish Wiki NER and News Datasets, One-Shot F1 Scores

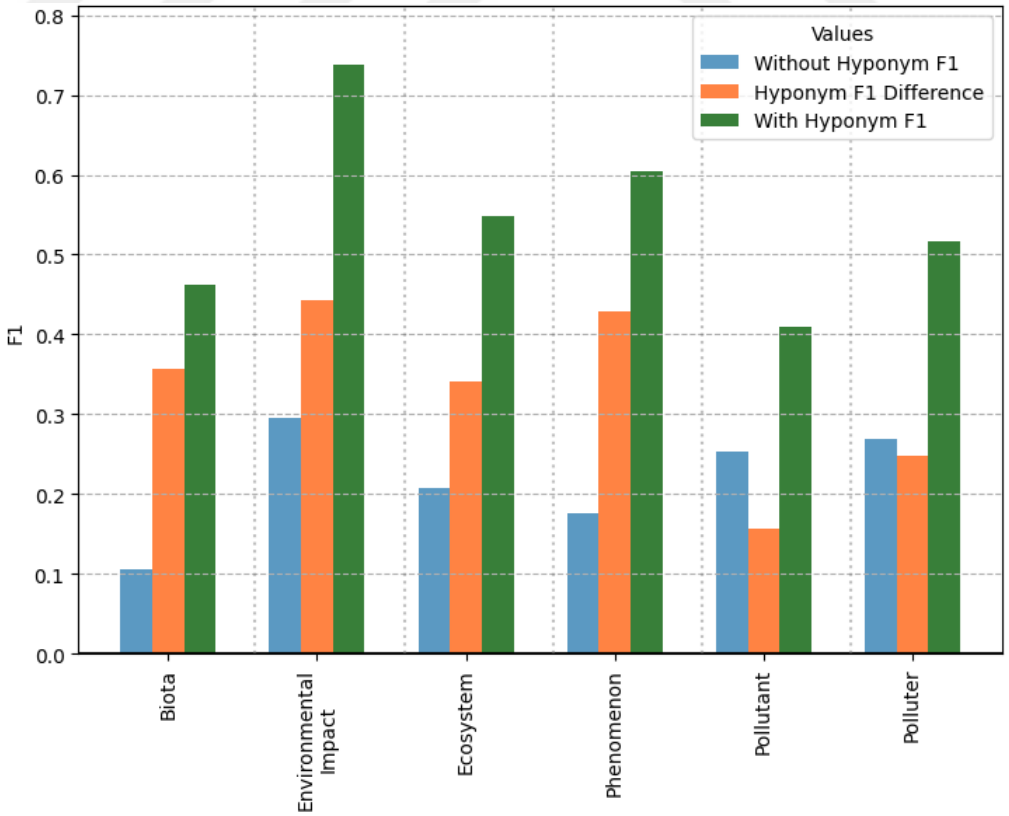

Figure 22: Presence of Hyponyms, Turkish Wiki NER and News Datasets, Ten-Shots F1 Scores

5.2.1.3 News and AI Datasets

Table 24: Presence of Hyponyms, News and AI Datasets Comparison

Results	**T-Test Score**	**P-Value**
Zero-Shot Results	4.34E+00	2.43E-03
One-Shot Results	3.93E+00	3.85E-03
Ten-Shots Results	4.71E+00	1.65E-03

Table 24 shows that, presence of hyponyms in the training dataset for the model is trained with news and artificially generated dataset increase the prediction performance of the model for zero-shot and few-shot training configurations.

Table 25: Presence of Hyponyms, News, and AI Datasets F1 Scores

Unseen Class	**Shot Number**	**Without Hyponym**	**With Hyponym**
Biota	0	0.0876	0.5405
	1	0.1038	0.5570
	10	0.0989	0.6383
Ecosystem	0	0.1399	0.4795
	1	0.1732	0.4489
	10	0.1416	0.6035
Environment Management	0	0.1793	0.3851
	1	0.2517	0.4162
	10	0.3222	0.5368
Environmental Impact	0	0.1739	0.8071
	1	0.1682	0.8196
	10	0.3814	0.8202
Phenomenon	0	0.0888	0.7734
	1	0.0789	0.7610
	10	0.0957	0.7903
Pollutant	0	0.1586	0.2689
	1	0.1742	0.2490
	10	0.1793	0.2867
Polluter	0	0.2479	0.4246
	1	0.2284	0.4324
	10	0.3018	0.4936

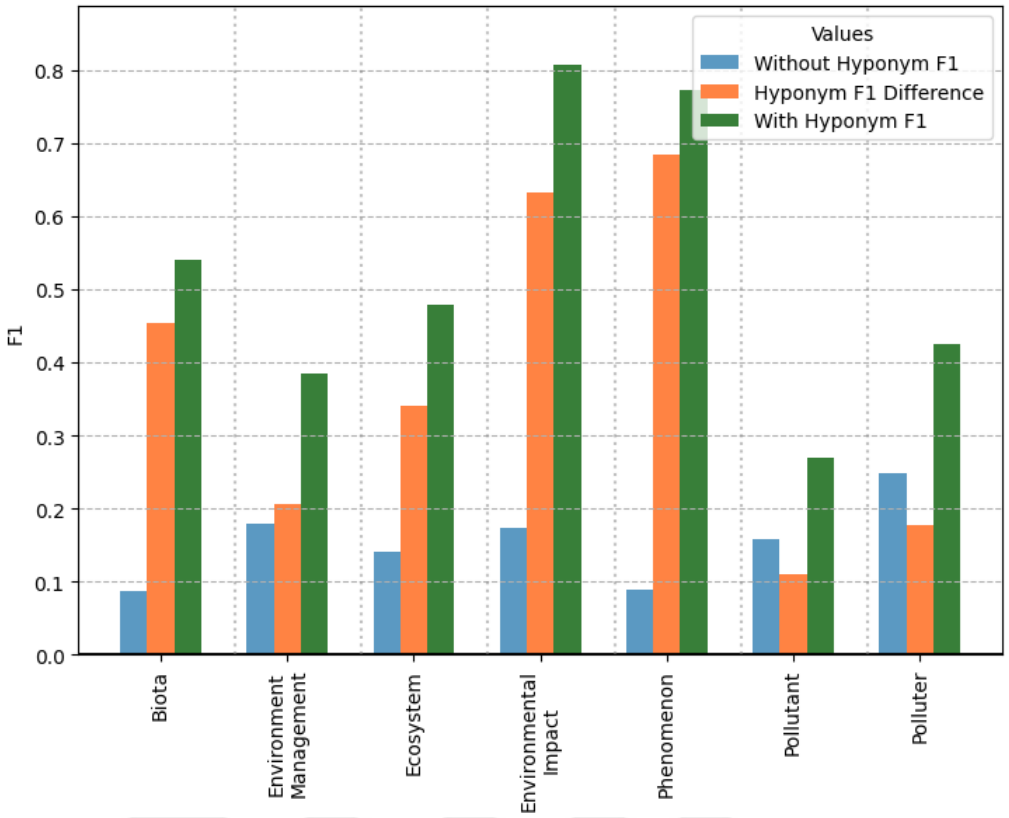

Figure 23: Presence of Hyponyms, News and AI Datasets, Zero-Shot F1 Scores

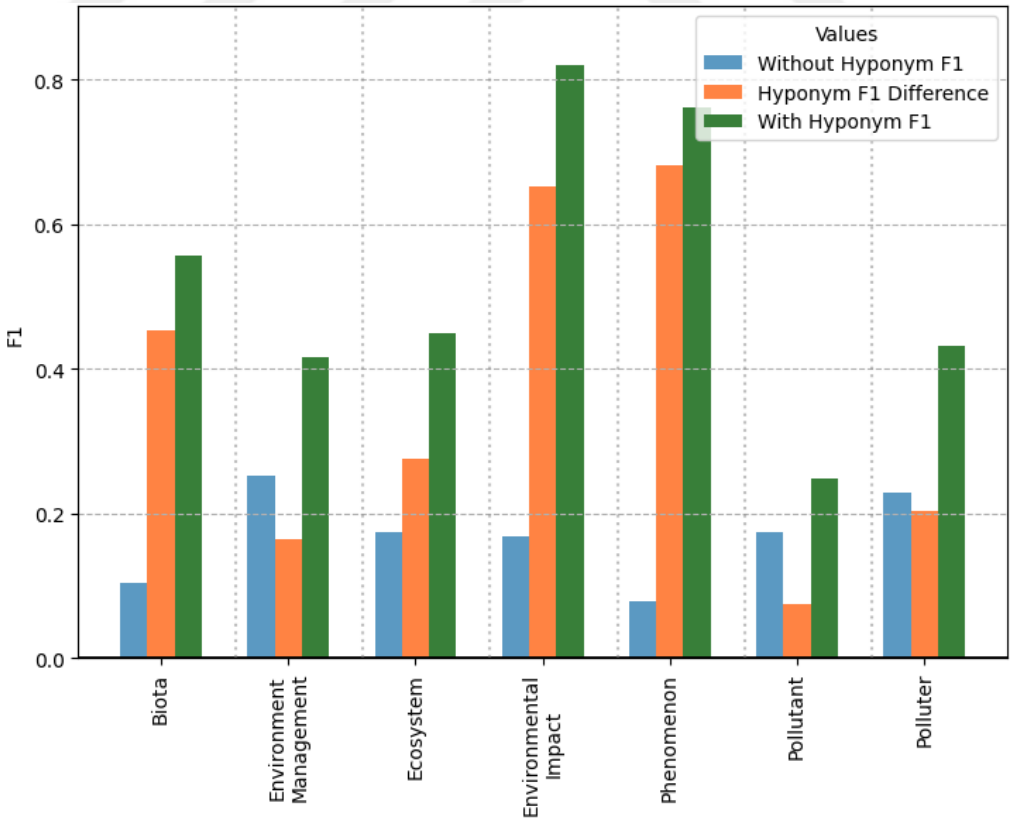

Figure 24: Presence of Hyponyms, News and AI Datasets, One-Shot F1 Scores

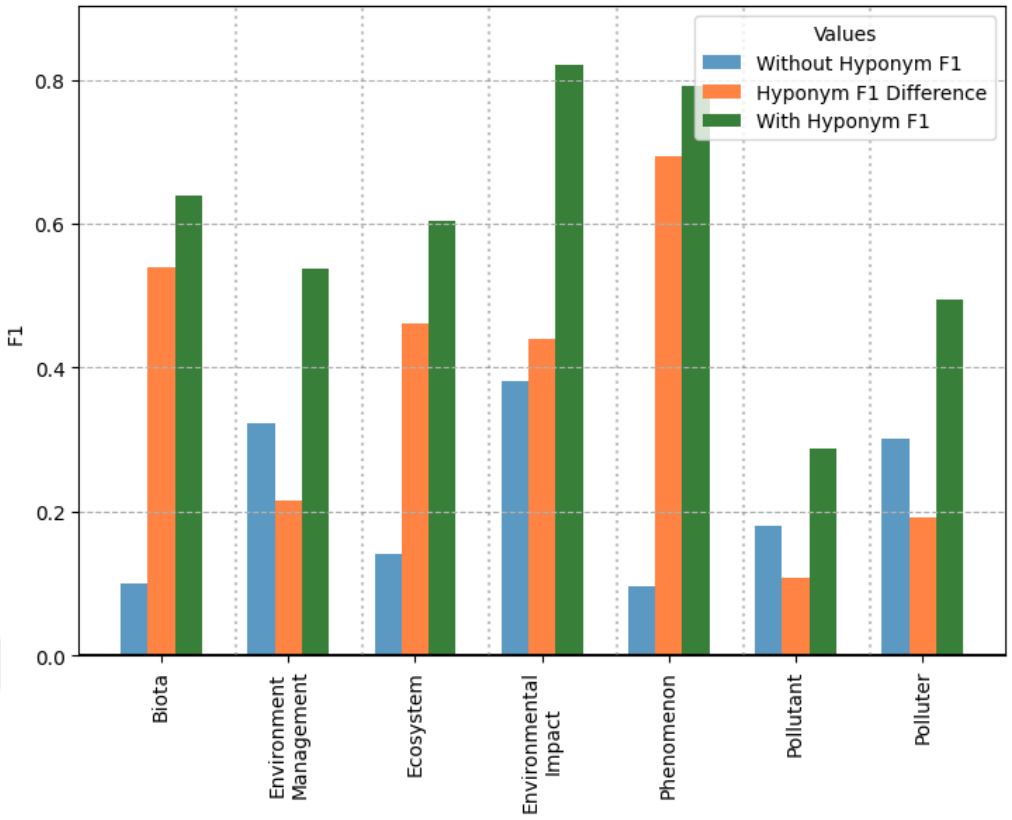

Figure 25: Presence of Hyponyms, News and AI Datasets, Ten-Shots F1 Scores

5.2.1.4 Turkish Wiki NER, News and AI Datasets

In this section, changes in prediction performance of the model first pretrained with Turkish Wiki NER, then trained with news and artificially generated datasets combined have been discussed.

Table 26: Presence of Hyponyms, Turkish Wiki NER, News and AI Datasets Comparison

Results	**T-Test Score**	**P-Value**
Zero-Shot Results	4.66E+00	1.74E-03
One-Shot Results	3.59E+00	5.75E-03
Ten-Shots Results	4.68E+00	1.70E-03

According to the Table 26, for each shot option, performance of the model when there are hyponyms in the training data is better when the model pretrained with generic dataset and trained with news and AI datasets combined.

Table 27: Presence of Hyponyms, Turkish Wiki NER, News and AI Datasets F1 Scores

Unseen Class	Shot Number	Without Hyponym	With Hyponym
Biota	0	0.0876	0.5405
	1	0.1038	0.5570
	10	0.0989	0.6383
Ecosystem	0	0.1399	0.4795
	1	0.1732	0.4489
	10	0.1416	0.6035
Environment Management	0	0.1793	0.3851
	1	0.2517	0.4162
	10	0.3222	0.5368
Environmental Impact	0	0.1739	0.8071
	1	0.1682	0.8196
	10	0.3814	0.8202
Phenomenon	0	0.0888	0.7734
	1	0.0789	0.7610
	10	0.0957	0.7903
Pollutant	0	0.1586	0.2689
	1	0.1742	0.2490
	10	0.1793	0.2867
Polluter	0	0.2479	0.4246
	1	0.2284	0.4324
	10	0.3018	0.4936

As seen in the Table 27, F1 scores increase for all classes and shot options when there are hyponyms in the training data.

Figure 26, Figure 27, and Figure 28 shows differences in F1 scores for each predicted in the test dataset. As seen in light blue lines in the graph, all of them are positive that means F1 scores increase for all classes.

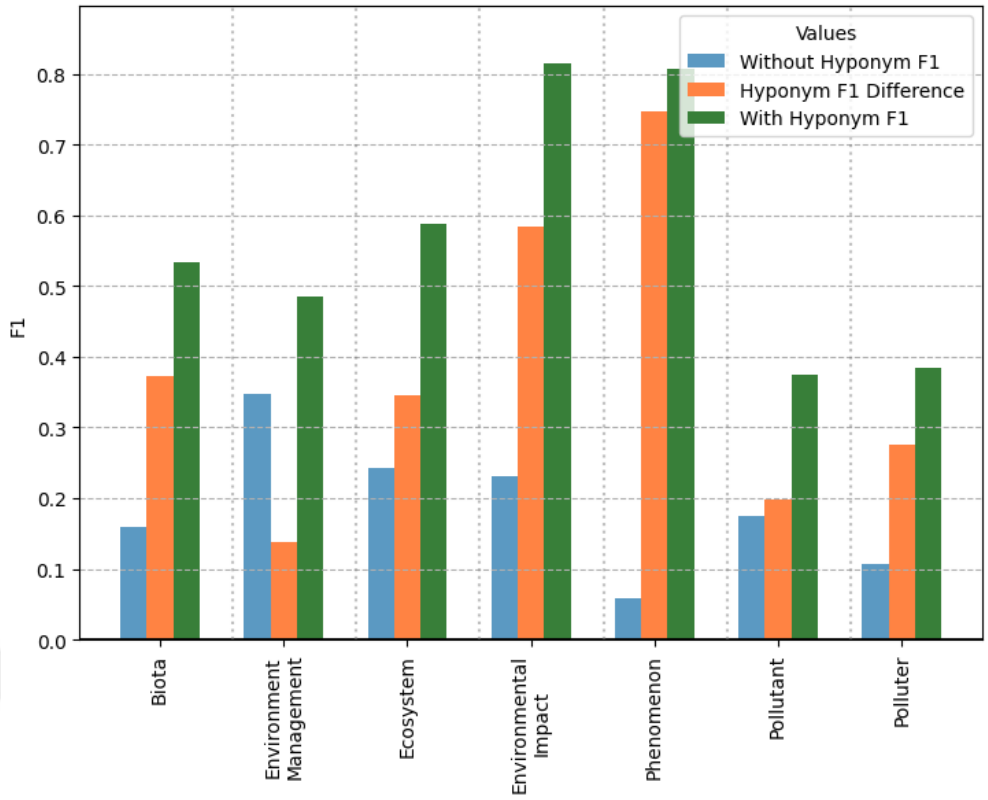

Figure 26: Presence of Hyponyms, Turkish Wiki NER, News and AI Datasets, Zero-Shot F1 Scores

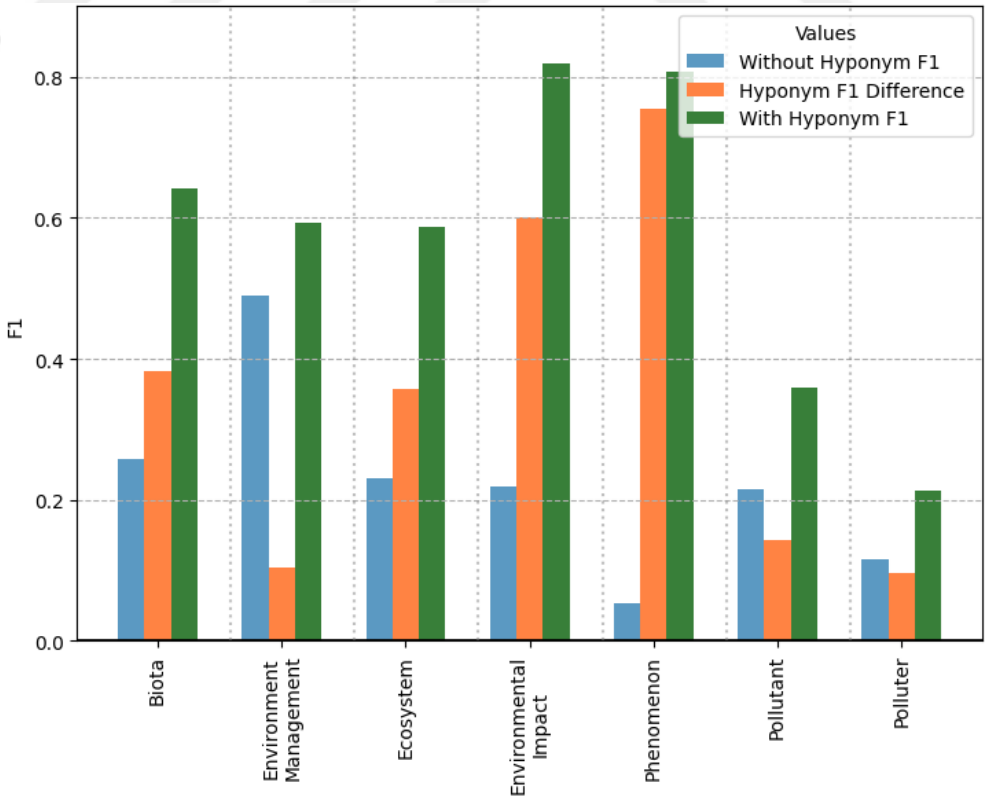

Figure 27: Presence of Hyponyms, Turkish Wiki NER, News and AI Datasets, One-Shot F1 Scores

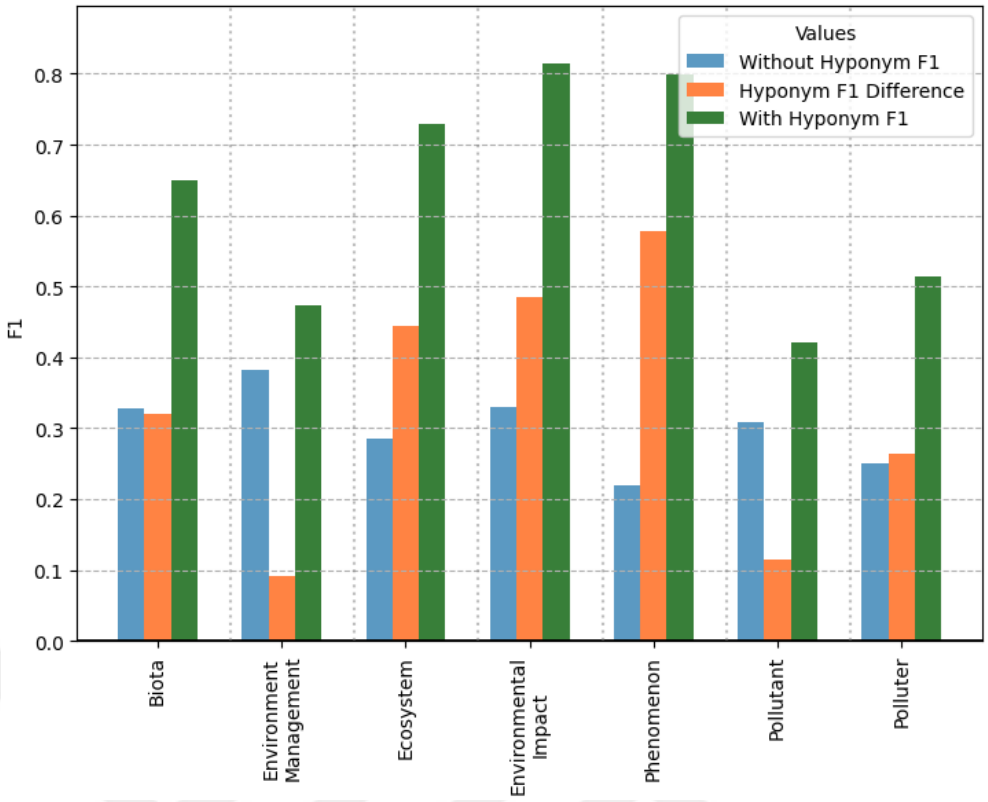

Figure 28: Presence of Hyponyms, Turkish Wiki NER, News and AI Datasets, Ten-Shots F1 Scores

5.2.1.5 Datasets Comparison

In this section, effect of datasets for the predictions in presence and in absence of hyponyms has been discussed.

5.2.1.5.1 Effect of Turkish Wiki NER Dataset on News Dataset

This section discusses F1 scores of the model that has been pretrained with generic dataset, Turkish Wiki NER.

5.2.1.5.1.1 Without AI Generated Data

Table 28: Presence of Hyponyms, T-Test Score of Turkish Wiki NER and News Datasets

Results	**T-Test Score**		**P-Value**	
	Without Hyponyms	**With Hyponyms**	**Without Hyponyms**	**With Hyponyms**
Zero-Shot Results	1.64E+00	2.78E+00	7.65E-02	1.60E-02
One-Shot Results	1.85E+00	3.21E+00	5.73E-02	9.22E-03
Ten-Shots Results	2.83E+00	3.39E+00	1.83E-02	9.78E-03

In Table 28, almost all setups have better prediction performance when the model has been trained with generic dataset apart from zero-shot and one-shot results in absence of hyponyms.

5.2.1.5.1.2 Combined with AI Generated Data

Table 29: Presence of Hyponyms, T-Test Score of Turkish Wiki NER, News and AI Datasets

Results	T-Test Score		P-Value	
	Without Hyponyms	With Hyponyms	Without Hyponyms	With Hyponyms
Zero-Shot Results	1.34E+00	9.60E-01	1.15E-01	1.87E-01
One-Shot Results	2.11E+00	1.21E+00	3.98E-02	1.36E-01
Ten-Shots Results	9.32E-01	1.89E+00	1.94E-01	5.35E-02

As seen in Table 29, when AI generated data have been introduced into the training data, pretraining the model with generic dataset starts to lose its meaning since only one-shot training has significantly higher F1 scores in absence of hyponyms apart from all shot results combined.

5.2.1.5.2 Effect of AI Generated Dataset

This section discusses effect of AI generated dataset combined with news dataset.

5.2.1.5.2.1 Without Turkish Wiki NER Dataset

Table 30: Presence of Hyponyms, T-Test Score of News and AI Datasets

Results	T-Test Score		P-Value	
	Without Hyponyms	With Hyponyms	Without Hyponyms	With Hyponyms
Zero-Shot Results	7.31E-01	3.42E+00	2.46E-01	7.10E-03
One-Shot Results	1.21E+00	4.12E+00	1.41E-01	4.58E-03
Ten-Shots Results	1.38E-01	2.70E+00	4.47E-01	1.77E-02

The Table 30 shows that when hyponyms are not in the training dataset, there is not significant difference between the prediction performance of the model trained and not trained with AI generated data. On the other hand, when hyponyms are introduced into the training data, difference become significant, meaning that model

trained with artificially generated dataset combined with news dataset has better outcomes.

5.2.1.5.2.2 Pretrained with Turkish Wiki NER Dataset

Table 31: Presence of Hyponyms, T-Test Score of News and AI Datasets Pretrained With Turkish Wiki NER

Results	T-Test Score		P-Value	
	Without Hyponyms	With Hyponyms	Without Hyponyms	With Hyponyms
Zero-Shot Results	9.67E-01	1.96E+00	1.86E-01	4.88E-02
One-Shot Results	2.08E+00	2.91E+00	4.61E-02	1.66E-02
Ten-Shots Results	-7.45E-01	1.70E+00	7.58E-01	6.99E-02

In Table 31, when there are not hyponyms in the training data, training the model with AI generated data does not make significant difference apart from the one-shot predictions. On the other hand, introduction of hyponyms into the training data shows better result, except for ten-shots prediction results.

5.2.2 Hypernyms

This section only discusses bottom-level classes and presence and absence of the classes as hypernym that are more-generalized classes above them in semantic hierarchy. Since the number of bottom-level classes is high and showing all results in tables in this section would make the study hard to read, only T-Test scores and graphs will be shown. F1-Scores tables can be seen in APPENDICES section. T-Tests have been conducted to determine there is improvement in predictions when hypernyms are introduced into the dataset.

5.2.2.1 News Dataset

Table 32: Presence of Hypernyms, News Dataset Comparison

Results	T-Test Score	P-Value
Zero-Shot Results	4.31E+00	2.72E-04
One-Shot Results	4.65E+00	1.33E-04
Ten-Shots Results	3.79E+00	4.51E-03

As seen in Table 32, presence of hypernyms in the dataset when the model is trained with domain-specific news dataset increase the prediction results. Figure 29, Figure 30, and Figure 31 shows for all entities, introducing hypernyms has positive impact on the prediction of unseen entity. Since, the number of examples of some entities such as Aquatic Biota, Liquid Pollutant etc. to prepare a ten-shots training setup, these entities have been removed from the ten-shots training prediction results.

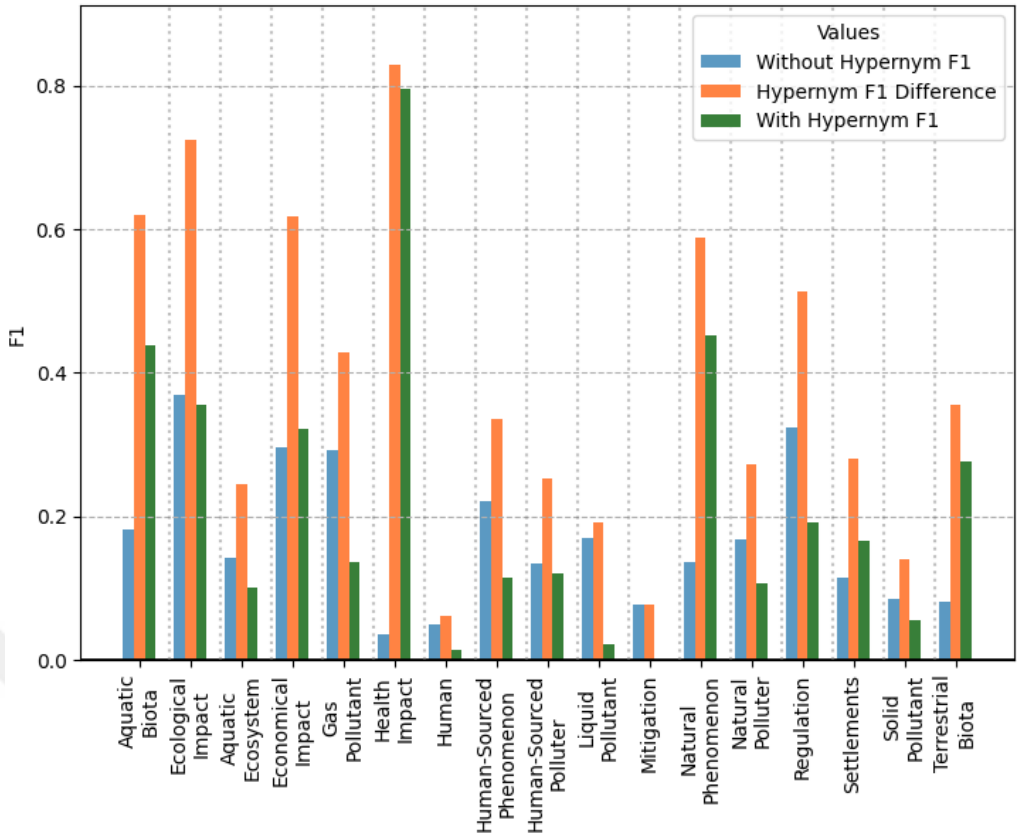

Figure 29: Presence of Hypernyms, News Data Zero-Shot F1 Scores

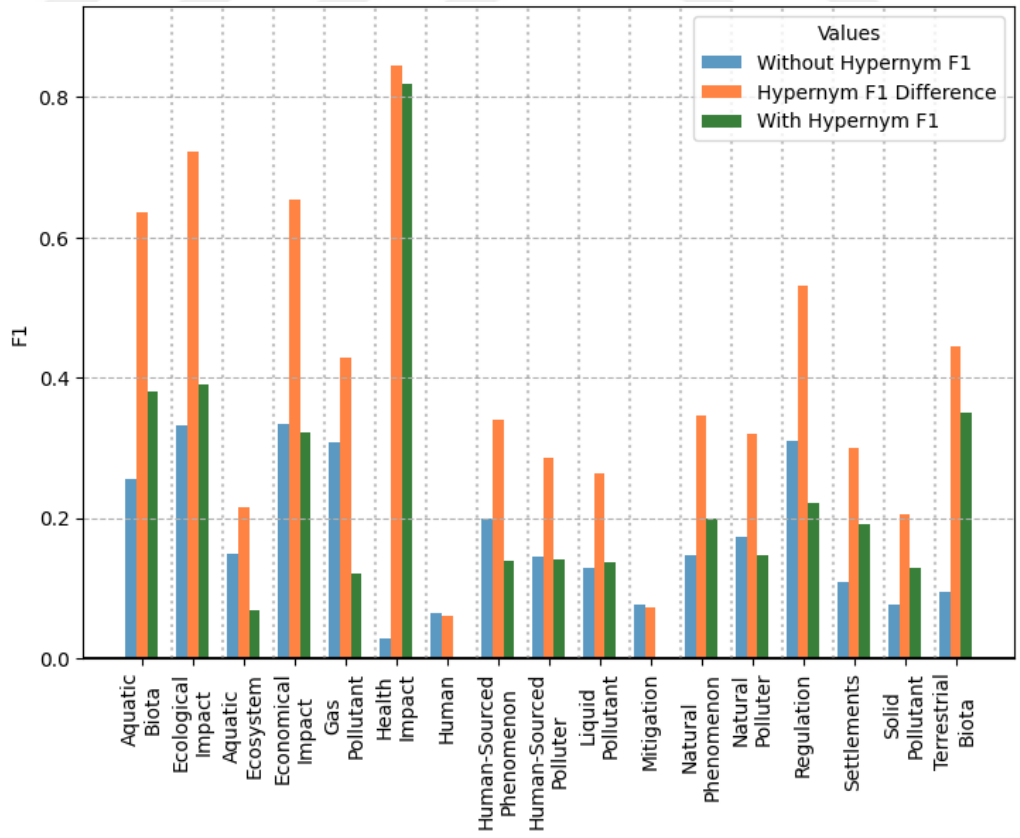

Figure 30: Presence of Hypernyms, News Data One-Shot F1 Scores

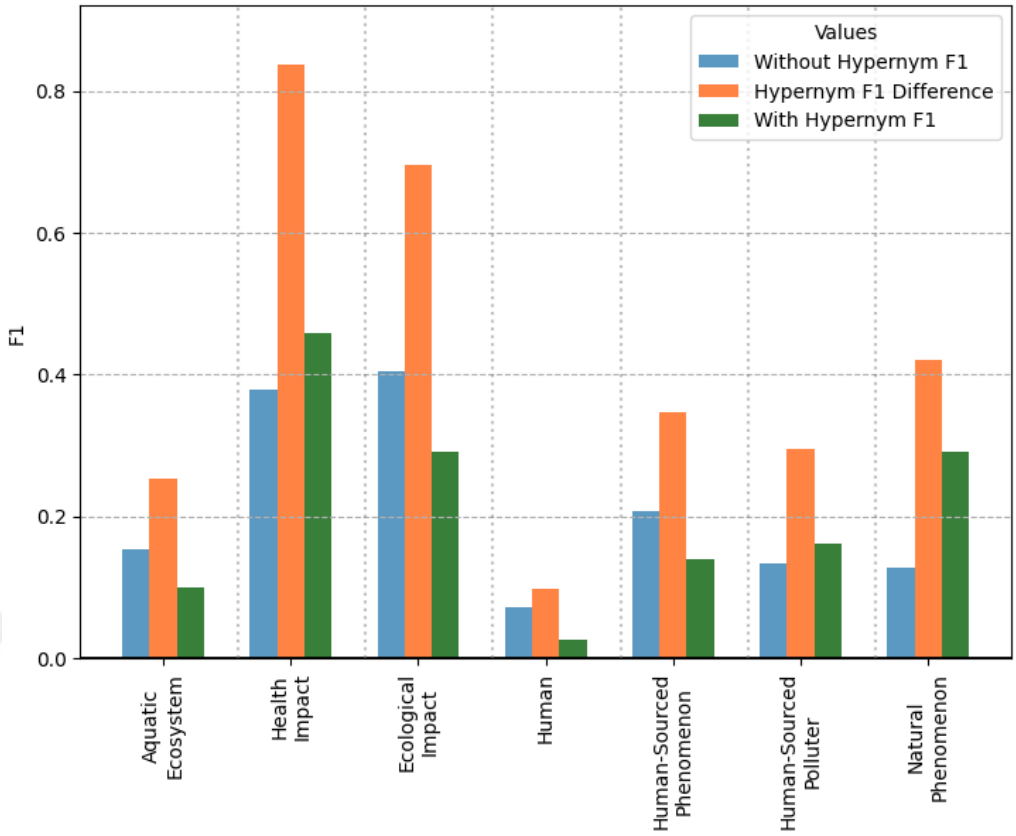

Figure 31: Presence of Hypernyms, News Data Ten-Shots F1 Scores

5.2.2.2 Turkish Wiki NER and News Datasets

Table 33: Presence of Hypernyms, Turkish Wiki NER and News Datasets Comparison

Results	**T-Test Score**	**P-Value**
Zero-Shot Results	5.45E+00	2.66E-05
One-Shot Results	7.83E+00	3.64E-07
Ten-Shots Results	5.57E+00	7.13E-04

As seen in Table 33, introducing hypernym into the training dataset to the model that is pretrained with a generic dataset and then trained with news dataset, prediction performance of the model increases. Figure 32, Figure 33, and Figure 34 show that this statement is valid for every unseen class regardless of the shot number in training data.

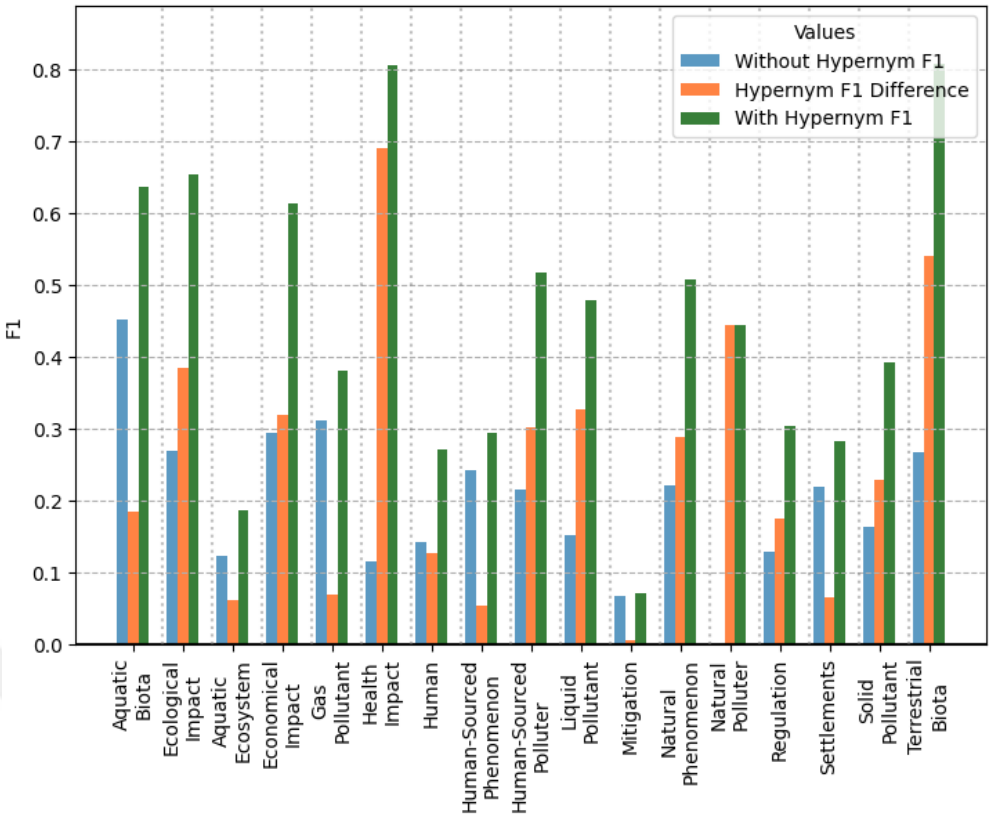

Figure 32: Presence of Hypernyms, Turkish Wiki NER and News Datasets, Zero-Shot F1 Scores

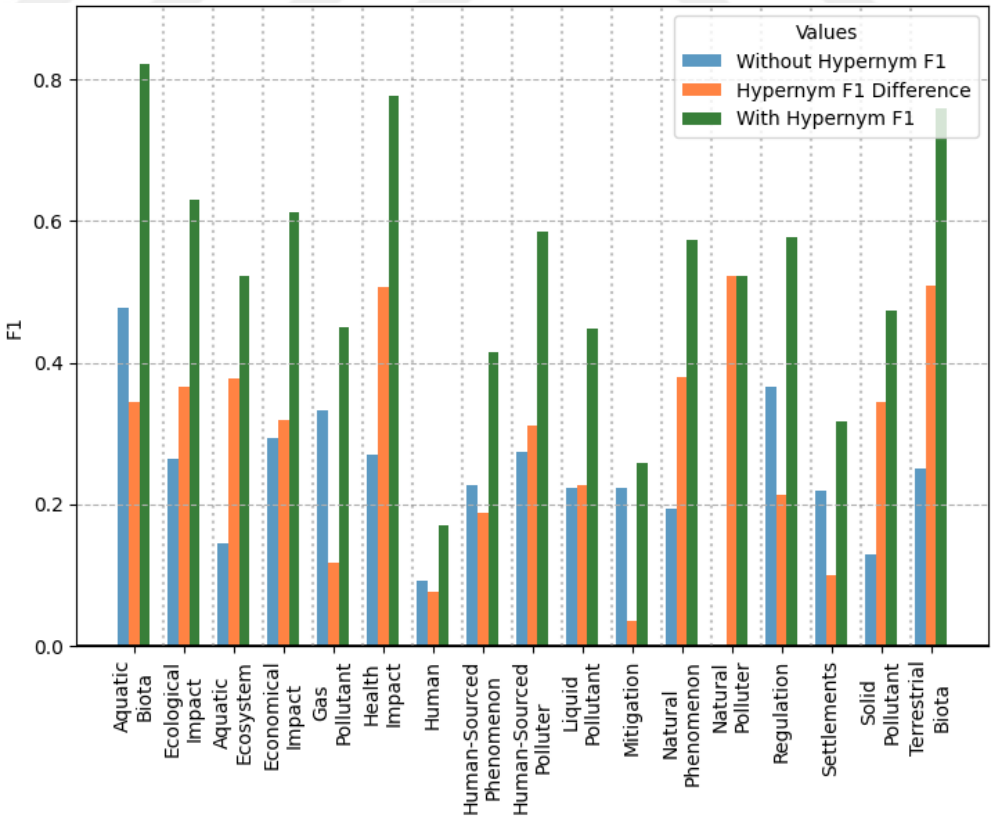

Figure 33: Presence of Hypernyms, Turkish Wiki NER and News Datasets, One-Shot F1 Scores

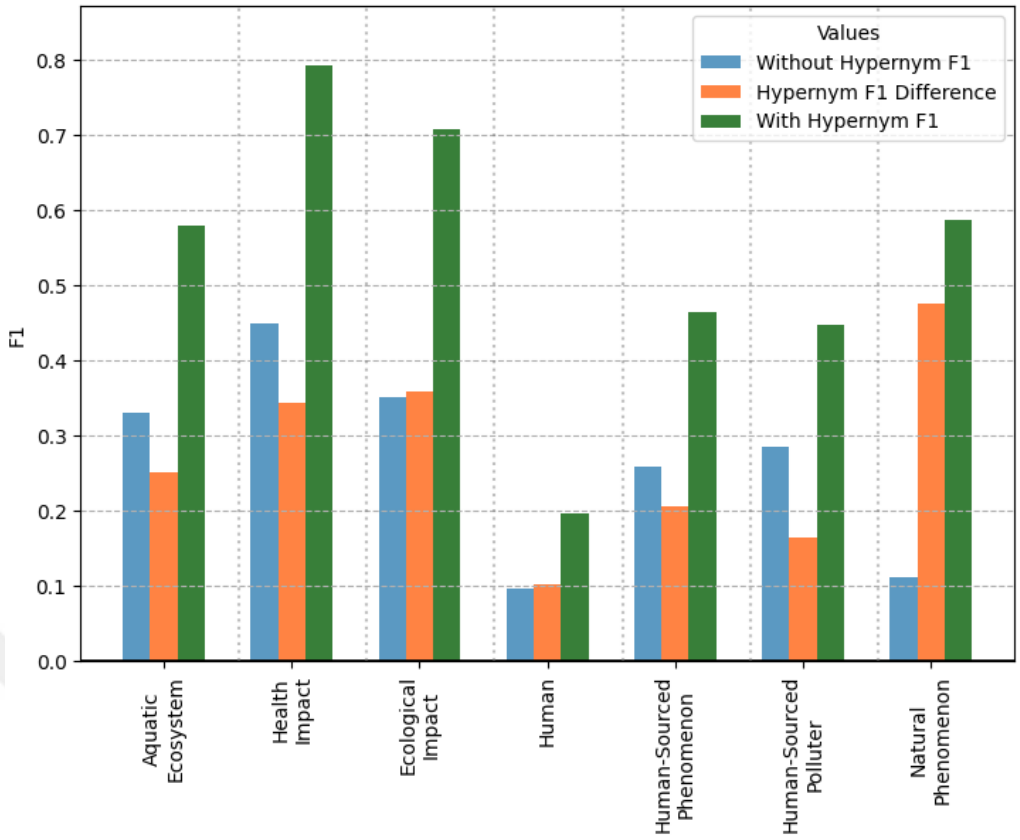

Figure 34: Presence of Hypernyms, Turkish Wiki NER and News Datasets, Ten-Shots F1 Scores

5.2.2.3 News and AI Datasets

Table 34: Presence of Hypernyms, News and AI Datasets Comparison

Results	**T-Test Score**	**P-Value**
Zero-Shot Results	4.34E+00	2.43E-03
One-Shot Results	3.93E+00	3.85E-03
Ten-Shots Results	4.71E+00	1.65E-03

Table 34 shows there is significant improvement in F1 scores for prediction results of unseen classes when the hypernyms are introduced into the dataset that trained with news and artificial generated datasets combined. Figure 35, Figure 36, and Figure 37 supports this claim. As seen in graphs, F1 scores of all unseen classes are increased with the introduction of hypernyms.

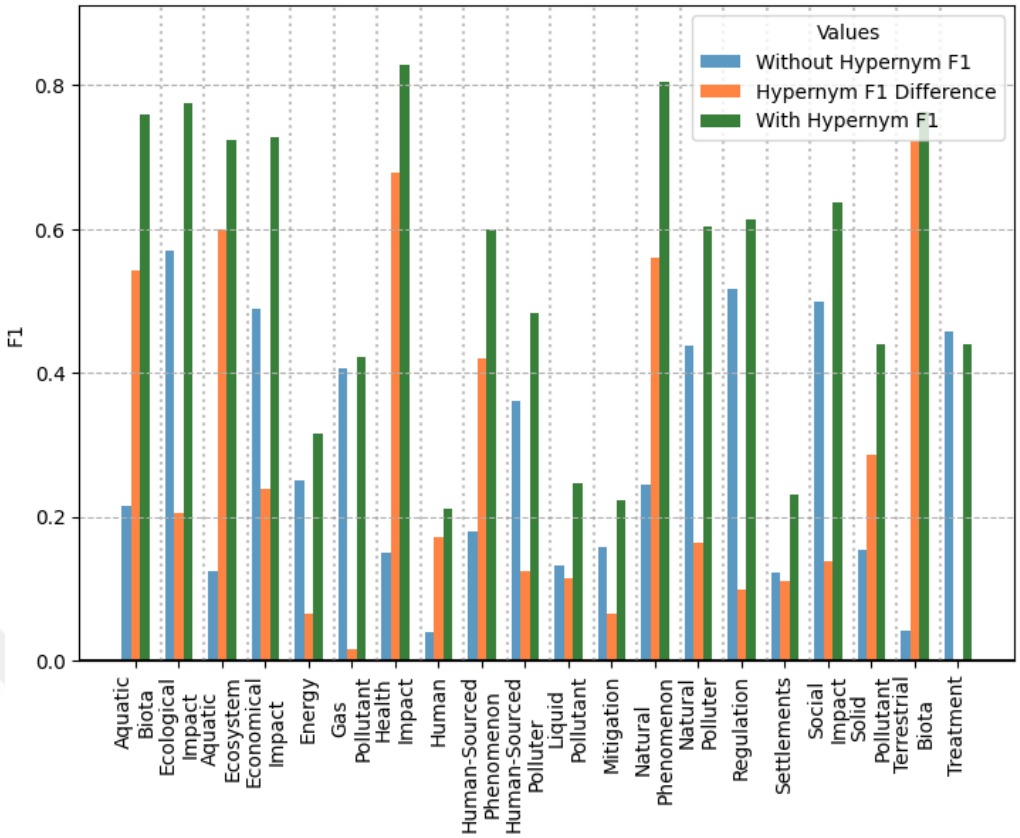

Figure 35: Presence of Hypernyms, News and AI Datasets, Zero-Shot F1 Scores

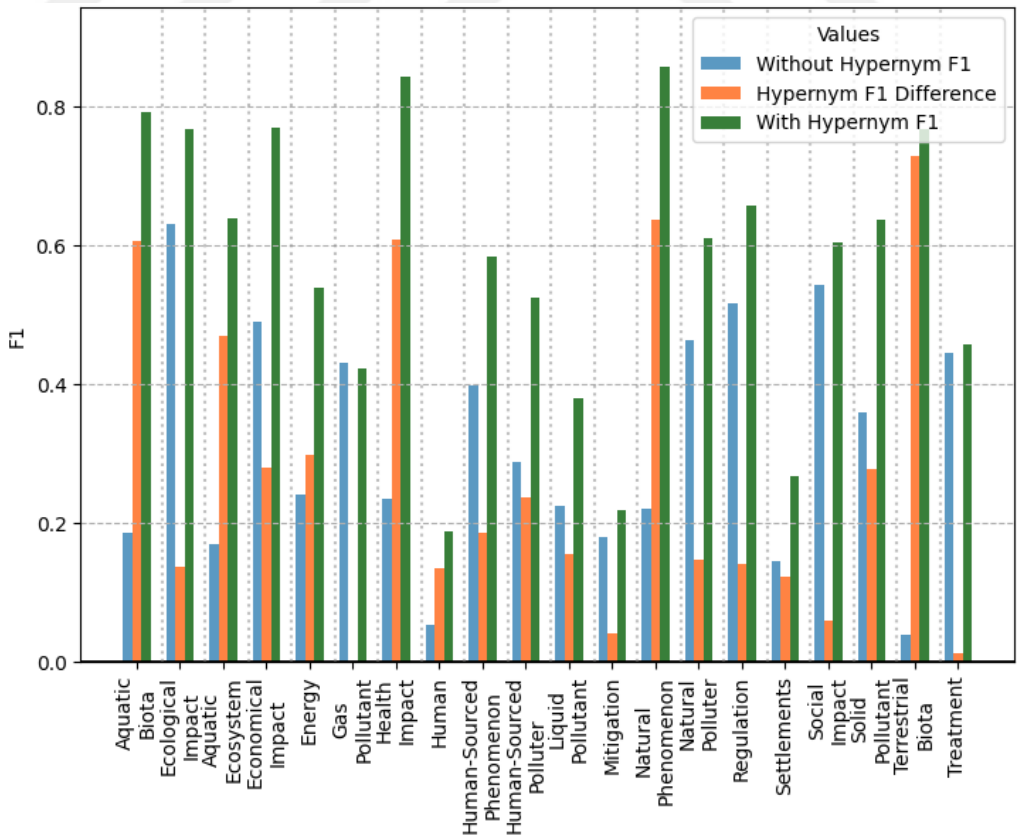

Figure 36: Presence of Hypernyms, News and AI Datasets, One-Shot F1 Scores

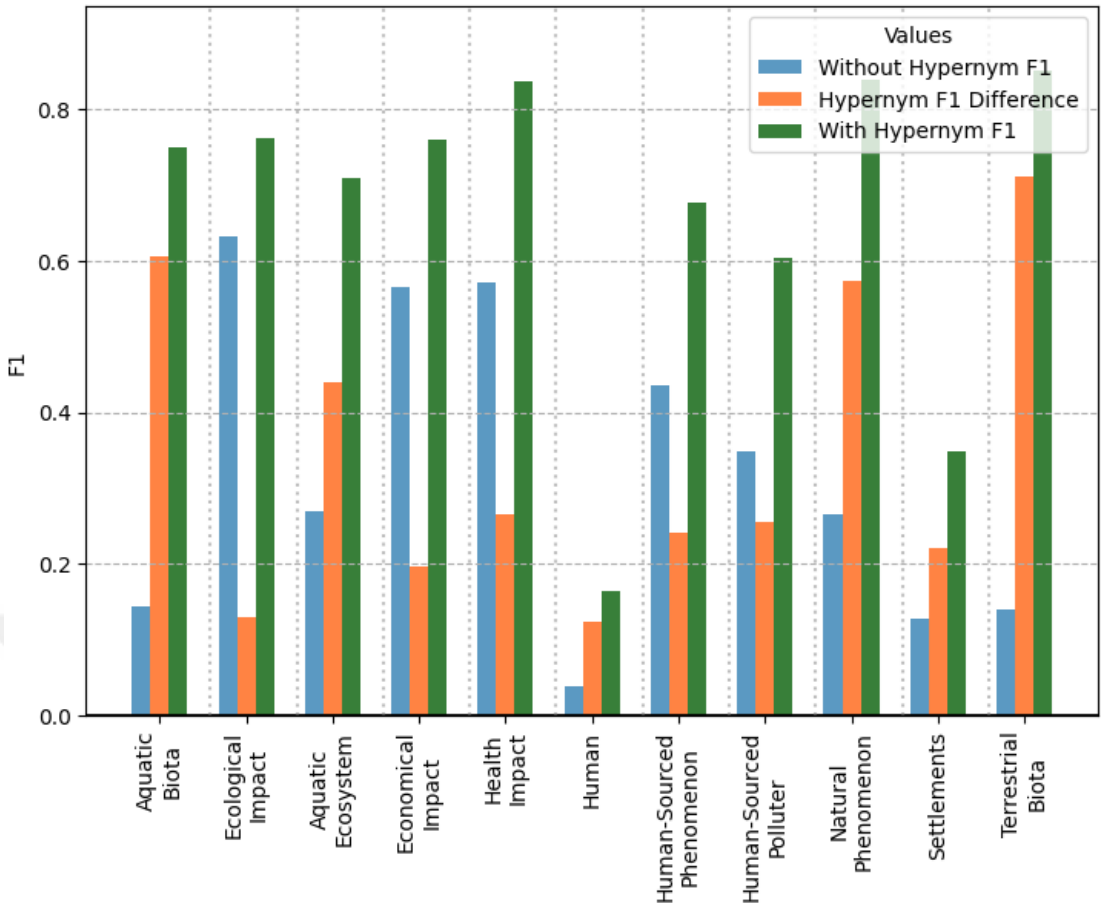

Figure 37: Presence of Hypernyms, News and AI Datasets, Ten-Shots F1 Scores

5.2.2.4 Turkish Wiki NER, News and AI Datasets

Table 35: Presence of Hypernyms, Turkish Wiki NER, News and AI Datasets Comparison

Results	**T-Test Score**	**P-Value**
Zero-Shot Results	6.27E+00	2.53E-06
One-Shot Results	6.11E+00	3.52E-06
Ten-Shots Results	8.90E+00	2.28E-06

Table 35 shows that all introducing hypernyms into the training data for the model that is pretrained with generic dataset and trained with news and artificially generated dataset combined, F1-scores for all shot configurations increase. Figure 38, Figure 39, and Figure 40 shows, model benefits from introduction of hypernyms regardless of the unseen class and shot number.

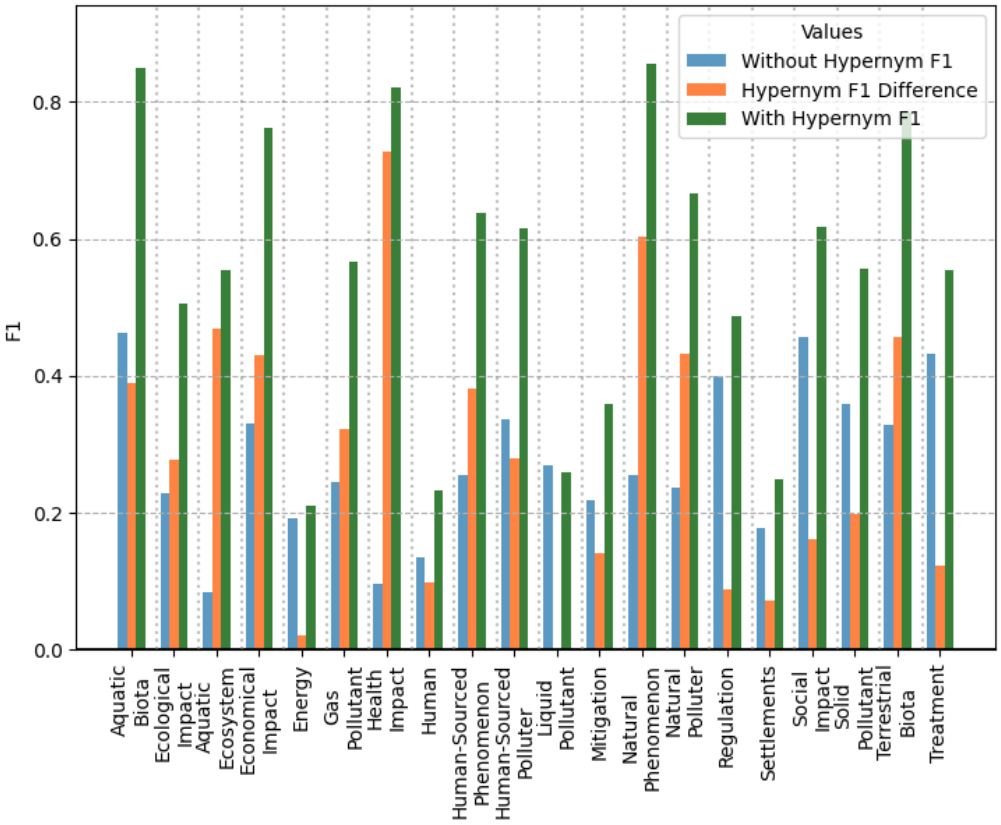

Figure 38: Presence of Hypernyms, Turkish Wiki NER and News Datasets, Zero-Shot F1 Scores

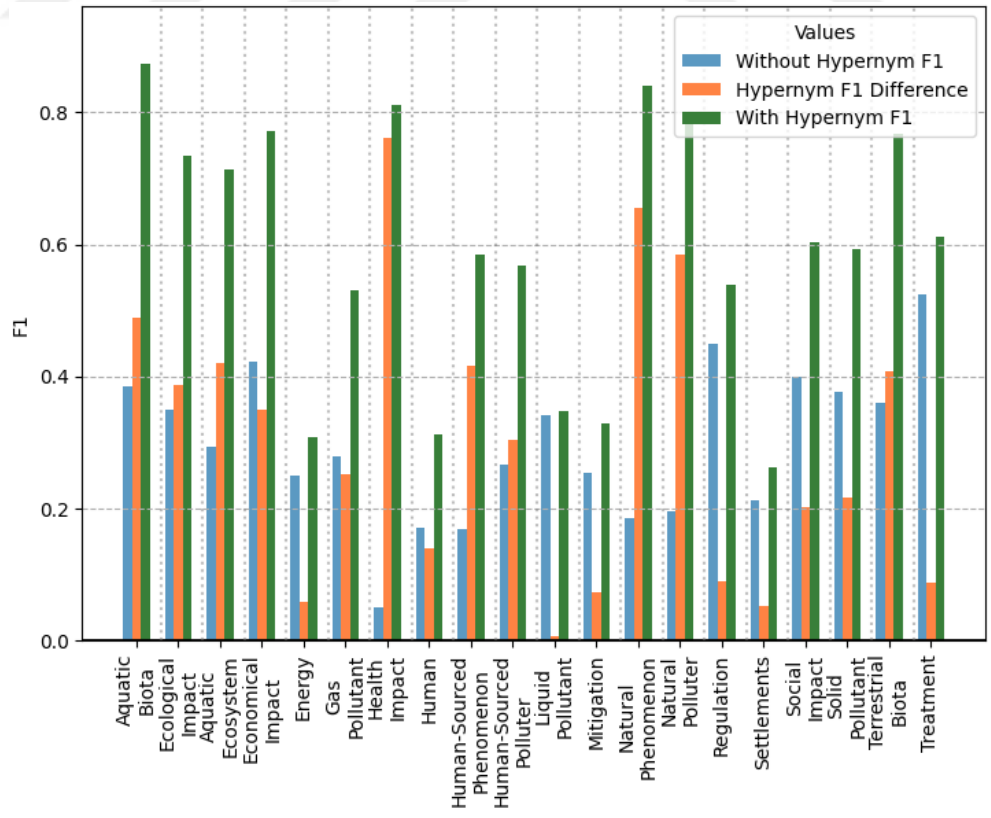

Figure 39: Presence of Hypernyms, Turkish Wiki NER and News Datasets, One-Shot F1 Scores

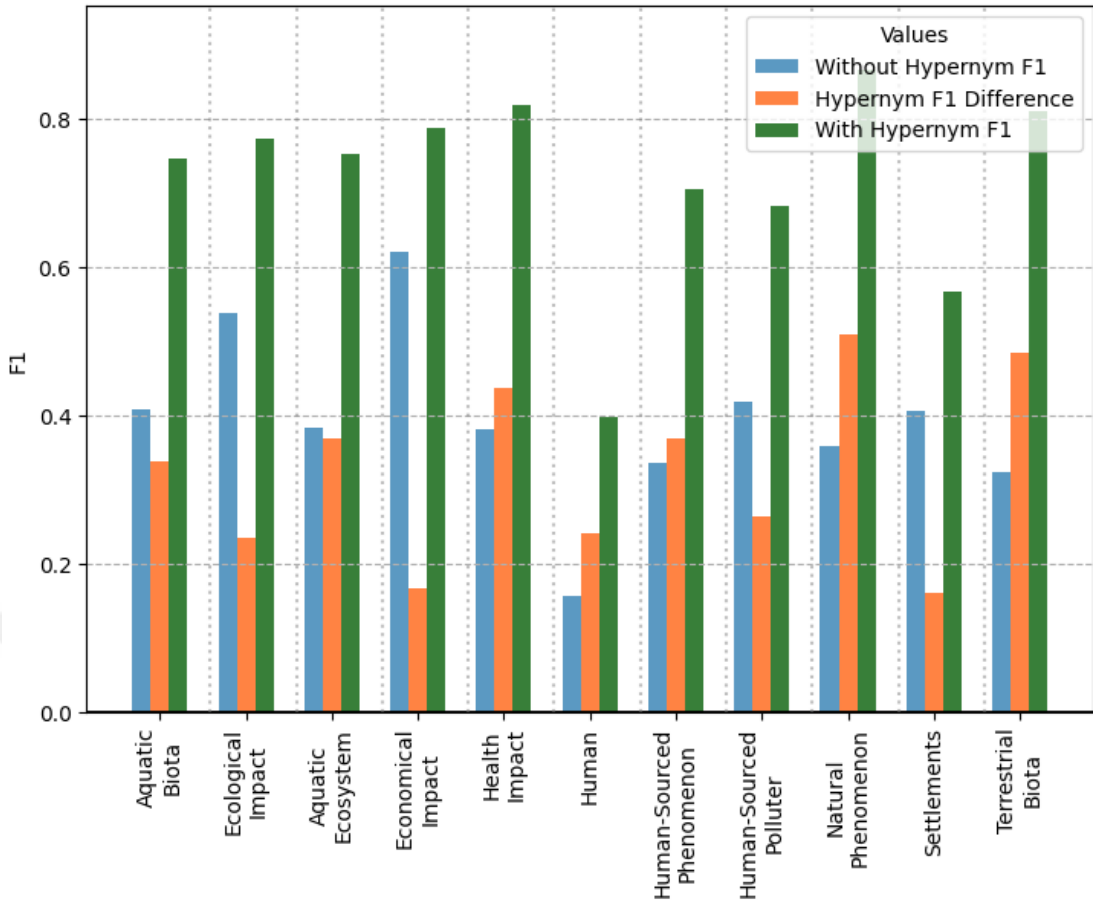

Figure 40: Presence of Hypernyms, Turkish Wiki NER and News Datasets, Ten-Shots F1 Scores

5.2.2.5 Dataset Comparison

In this section, effect of datasets for the predictions in presence and in absence of hypernyms has been discussed.

5.2.2.5.1 Effect of Turkish Wiki NER Dataset on News Dataset

This section discusses F1 scores of the model that has been pretrained with generic dataset, Turkish Wiki NER.

5.2.2.5.1.1 Without AI Generated Data

Table 36: Presence of Hypernyms, T-Test Score of Turkish Wiki NER and News Datasets

Results	T-Test Score		P-Value	
	Without Hypernyms	**With Hypernyms**	**Without Hypernyms**	**With Hypernyms**
Zero-Shot Results	1.57E+00	1.55E+00	6.90E-02	6.99E-02
One-Shot Results	3.50E+00	4.05E+00	1.61E-03	4.68E-04
Ten-Shots Results	1.81E+00	2.63E+00	6.00E-02	1.96E-02

Table 36 shows that model pretrained with generic dataset outperforms its not-pretrained version, where Turkish Wiki NER is not used in pretraining phase, only on one-shot training prediction in absence and in presence of hypernyms. On the other hand, pretrained model benefits more from introducing hypernyms into the training data during ten-shots training.

5.2.2.5.1.2 Combined with AI Generated Data

Table 37: Presence of Hypernyms, T-Test Score of Turkish Wiki NER, News and AI Datasets

Results	T-Test Score		P-Value	
	Without Hypernyms	**With Hypernyms**	**Without Hypernyms**	**With Hypernyms**
Zero-Shot Results	-8.34E-02	5.92E-01	5.33E-01	2.80E-01
One-Shot Results	-4.55E-01	8.36E-01	6.73E-01	2.07E-01
Ten-Shots Results	1.61E+00	2.03E+00	6.91E-02	3.49E-02

Table 37 shows pretraining the model with Turkish Wiki NER dataset does not create significant difference in terms of prediction performance and F1-Score when artificially generated data is combined with news dataset during training, apart from ten-shots training where hypernyms are introduced into the dataset.

5.2.2.5.2Effect of AI Generated Dataset

This section discusses effect of AI generated dataset combined with news dataset.

5.2.2.5.2.1 Without Turkish Wiki NER Dataset

Table 38: Presence of Hypernyms, T-Test Score of News and AI Datasets

Results	T-Test Score		P-Value	
	Without Hypernyms	**With Hypernyms**	**Without Hypernyms**	**With Hypernyms**
Zero-Shot Results	3.46E+00	4.77E+00	1.62E-03	1.04E-04
One-Shot Results	4.40E+00	4.89E+00	2.24E-04	8.19E-05
Ten-Shots Results	4.36E+00	3.32E+00	2.39E-03	7.97E-03

As seen in Table 38, when the model training dataset consists of AI generated data along with news dataset, performance of the model in terms of F1-scores improves regardless of presence of hypernyms and number of examples of unseen class introduced into the training data.

5.2.2.5.2.2 Pretrained with Turkish Wiki NER Dataset

Table 39: Presence of Hyponyms, T-Test Score of News and AI Datasets Pretrained With Turkish Wiki NER

Results	T-Test Score		P-Value	
	Without Hypernyms	**With Hypernyms**	**Without Hypernyms**	**With Hypernyms**
Zero-Shot Results	2.06E+00	2.94E+00	2.88E-02	4.84E-03
One-Shot Results	1.30E+00	3.27E+00	1.07E-01	2.40E-03
Ten-Shots Results	2.56E+00	4.88E+00	2.14E-02	1.38E-03

Table 39 shows that apart from one-shot training in absence of hypernyms in the training dataset, training the model with news and AI datasets combined improves the prediction ability when the model pretrained with Turkish Wiki NER dataset.

5.2.3 Discussion

In this section how the prediction performance of the model changes when semantically related entities are introduced into the training dataset is discussed according to the results were shown in the sections above. This section discusses results in two different contexts. First each dataset combination's performance in presence and absence of related entities were discussed individually. Later, dataset

combinations have been compared in presence and absence of related entities as conducted similar in previous section where shot options were discussed.

The results show regardless of dataset combination and shot configuration; introducing hyponyms and hypernyms into the training data improves the prediction performance of the model.

Dataset combination comparisons show different results, and a generic conclusion is hard to make. According to hyponym presence comparisons, when there are not hyponyms in the training dataset; pretraining with generic dataset or introducing artificially generated data into the training dataset does not affect model's prediction performance significantly. On the other hand, when hyponyms are in the training dataset, pretraining with generic dataset or combining training dataset with artificially generated dataset or pretraining the model with generic dataset to be trained with AI and news data combined improves prediction performance of the model on almost all shot setup.

As the hypernyms results show, apart from hyponyms; introducing artificially generated data in the training dataset improves prediction performance of the model regardless of the shot-option, regardless of pretraining with generic dataset procedure and regardless of presence of hypernyms.

To summarize, having both artificially generated data and related classes in the training dataset improves the prediction performance of the model. On the other hand, presence of hypernyms -or in other terms entities above the semantical hierarchy-, has slightly more benefits.

CHAPTER VI
CONCLUSION

The study presents multiple applications of zero-shot, one-shot and ten shots Named Entity Recognition (NER) for a dataset specific to environmental sciences domain, to investigate improvements in terms of prediction performance of the DistilBERTurk model which is the lightweight version of the BERT[7] as the first aspect of the study. During training and testing, Huggingface's Transformers and Dataset libraries have been utilized.

In this study, NER labels specific to environmental sciences domain have been defined with total number of 30. Also, as the second aspect of this study, semantical hierarchies of these entities have been constructed. These classes will be mentioned as related classes in this section.

The study consisted of three different datasets as the first two datasets belong to environmental sciences domain and they are used in training and testing processes. Texts for these datasets are obtained from news websites and generated by large language models. As the third aspect of this study, combinations of these datasets have been created. Data cleansing, labelling and augmentation processes have been conducted throughout the study. The third dataset is a generic dataset named Turkish Wiki NER[24], and this dataset has been used as intermediate pretraining step to improve prediction performance of the DistilBERTurk.

The study is based on three different hypotheses to investigate the prediction performance of the model. These are pretraining the model with generic dataset and or introducing artificially generated data improves performance, increasing shot-numbers improves performance, introducing related classes improves performance in terms of prediction ability. To test these hypotheses, different configurations have been created. First, to see effect of increased shot numbers zero-shot, one-shot and ten-shots trainings have been conducted only for domain specific news dataset. Second, these training procedures have been repeated for different dataset combinations. Third,

related classes had been removed from the domain-specific datasets and all trainings were repeated. To evaluate the prediction results, F1 score have been calculated. Also, to compare these results, one-sided paired T-Tests have been conducted. Comparisons have been made according to the aspects and hypotheses described previously.

The study shows promising results in terms of zero-shot and few-shot learning. Training the model with news dataset without AI data and intermediate pretraining procedure, increased number of unseen class examples did not create significant difference in terms of F1 score. On the other hand, pretraining the model just before the training, or introducing AI generated data into the training dataset, or doing these both operations together have improved the model's ability to recognize unseen classes on one-shot and ten-shots configurations. Especially, combining pretraining with generic dataset and introducing AI data into news dataset showed the best improvement. Later the dataset combinations had been compared and following conclusions have been drawn; introducing artificially generated data into the dataset improves prediction performance of the model regardless of the shot-number. On the other hand, pretraining the model with generic dataset shows similar improvements except for artificially generated data is in the dataset.

Presence of semantically related classes in the training data improves the performance of the model as excepted regardless of the shot number and dataset combination. Especially, combination of artificially generated data and related classes in the dataset creates most beneficial option. These results prove the ability of BERT to create semantic relationships between entities. Also, introducing well-structured data such as AI data into the dataset help the model to predict unseen classes.

The number of combinations that can be conducted in context of this study is not limited to the presented above. Instead of using a generic dataset, another domain specific dataset may be used to pretrained the model. On the other hand, in this study, AI data was introduced into test dataset if AI dataset is in consideration, instead of this, AI data can only be in the training dataset. Since, obtaining data, cleaning and labelling is a labour-intensive task, this study stayed limited to ten-shots learning. This limit can be extended with further additions into dataset for hundred-shots learning. Also, in this study, news data and AI data combined to create a dataset combination, there was not a combination to train the model with AI generated dataset and test with real-life -in context of this study our news data—dataset.

As the need of information retrieval increases through the age of technology, importance of named entity recognition for low resource domain-specific datasets increases. This study touches on important points in this context, especially including but not limited to environmental sciences domain to enlighten different options to extract specific information from the texts.

REFERENCES

[1] TURING Alan Mathison (1950), "Computing Machinery and Intelligence", *Mind*, Vol. 59, No. 236, pp. 433-60, DOI: 10.1093/mind/LIX.236.433.

[2] GOODFELLOW Ian, BENGIO Yoshua and COURVILLE Aaron (2016), *Deep Learning*, p. 461, Mit Press, USA.

[3] NADEAU David and SEKINE Satoshi (2007), "A survey of named entity recognition and classification", *Lingvisticae Investigationes*, Vol. 30, No. 1, pp. 3-26, DOI: 10.1075/li.30.1.03nad.

[4] LI Jing, SUN Aixin, HAN Jianglei and LI Chenliang (2022), "A Survey on Deep Learning for Named Entity Recognition", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 34, No. 1, pp. 50-70, DOI: 10.1109/TKDE.2020.2981314.

[5] KOŠPRDIĆ Miloš, PRODANOVIĆ Nikola, LJAJIĆ Adela, BAŠARAGIN Bojana and MILOŠEVIĆ Nikola (2024), "From Zero to Hero: Harnessing Transformers for Biomedical Named Entity Recognition in Zero- and Few-shot Contexts", *Artificial Intelligence in Medicine*, Vol. 156, No. 1, DOI: 10.1016/j.artmed.2024.102970.

[6] GU Yu, TINN Robert, CHENG Hao, LUCAS Michael, USUYAMA Naoto, LIU Xiaodong, NAUMANN Tristan, GAO Jianfeng and POON Hoifung (2022), "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing", *ACM Transactions on Computing for Healthcare*, Vol. 3, No. 1, pp. 1-23, DOI: 10.1145/3458754.

[7] SANH Victor, DEBUT Lysandre, CHAUMOND Julien and WOLF Thomas (2019), "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter", *Arxiv Preprint*, DOI:10.48550/arXiv.1910.01108.

[8] YADAV Vikas and BETHARD Steven (2018), “A Survey on Recent Advances in Named Entity Recognition from Deep Learning models”, *In, Proceedings of the 27th International Conference on Computational Linguistics,* Eds. Emily M. Bender, Leon Derczynski, Pierre Isabelle, pp.2145-2158,Association for Computational Linguistics, USA.

[9] RAFFEL Colin, SHAZEER Noam, ROBERTS Adam, LEE Katherine, NARANG Sharan, MATENA Micheal, ZHOU Yanqi, LI Wei and LIU Peter (2019), “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”, *The Journal of Machine Learning Research*, Vol. 21, No. 1, pp. 5485-5551.

[10] DEVLIN Jacob, CHANG Ming-Wei, LEE Kenton and TOUTANOVA Kristina (2019), “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, *In, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Eds. Jill Burstein, Christy Doran, Thamar Solorio, pp. 4171-4186, Association for Computational Linguistics, Minesota, DOI:10.18653/v1/N19-1423

[11] SAINZ Oscar, GARCIA-FERRERO Iker, AGERRI Rodrigo, LOPEZ DE LACALLE Oier, RIGAU German, AGIRRE Eneko (2023), “GoLLIE: Annotation Guidelines improve Zero-Shot Information-Extraction”, *ICLR 2024*, Vienna, Austria.

[12] ARSLAN Serdar (2024), “Application of BiLSTM-CRF model with different embeddings for product name extraction in unstructured Turkish text”, *Neural Computing and Applications*, vol. 36, No. 1, pp. 8371-8382, DOI: 10.1007/s00521-024-09532-1.

[13] LEE Jinhyuk, SUNG Mujeen, KANG Jaewo and CHEN Danqi (2020), “Learning Dense Representations of Phrases at Scale”, *In, Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Eds. Chengqing Zong, Fei Xia, Wenjie Li, Roberto Navigli, pp.6634–6647, Association for Computational Linguistics, Online, DOI:10.18653/v1/2021.acl-long.518.

[14] WEI Jason, BOSMA Maarten, ZHAO Vincent, GUU Kelvin, YU Adams Wei, LESTER Brian, DU Nan, DAI Andrew and LE Quoc (2022), "Finetuned Language Models Are Zero-Shot Learners", Virtual Conference, *ICLR 2022*.

[15] VAN HOANG Nguyen, MULVAD Soereen Hougaard, DEXTER Neo and YUE Yang (2021), "Zero-Shot Learning in Named-Entity Recognition with External Knowledge", *Arxiv Preprint*, DOI: 10.48550/arXiv.2111.07734.

[16] PICCO Gabriele, GALINDO Marcos Martinez, PURPURA Alberto, FUCHS Leopold, LOPEZ Vanessa and LAM Thanh (2023), "Zshot: An Open-source Framework for Zero-Shot Named Entity Recognition and Relation Extraction", *In, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Volume 3 (System Demonstrations)*, pp. 357-368, Association for Computational Linguistics, Toronto.

[17] XIE Tingyu, LI Qi, ZHANG Yan, LIU Zuozhu, WANG Hongwei (2024), "Self-Improving for Zero-Shot Named Entity Recognition with Large Language Models", *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 583-593, Association for Computational Linguistics, Mexico City.

[18] MATEVŽ Ogrinc, SELJAK Barbara Korousic, EFTIMOV Tome (2024), "Zero-shot evaluation of ChatGPT for food named-entity recognition and linking", *Frontiers in Nutrition*, Vol. 11, Article Number 1429259, DOI:10.3389/fnut.2024.1429259.

[19] FANG Zheng, CAO Yanan, LI Tai, JIA Ruipeng, FANG Fang, SHANG Yanmin and LU Yuhai (2021), "TEBNER: Domain Specific Named Entity Recognition with Type Expanded Boundary-aware Network", *In, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Eds. Marie-Francine Moens, Xuanjing Huang, Lucia Specia, Scott Wen-tau Yih, pp.198-207, Association for Computational Linguistics, Online and Punta Cana.

[20] SIVARAJKUMAR Sonish and WANG Yanshan (2022), "HealthPrompt: A Zero-shot Learning Paradigm for Clinical Natural Language Processing", *Arxiv Preprint*, DOI:10.48550/arXiv.2203.05061.

[21] ALHOSHAN Waad, FERRARI Alessio and ZHAO Liping (2023), “Zero-shot learning for requirements classification: An exploratory study”, *Information and Software Technology*, vol. 159, pp. 107-202, DOI: 10.1016/j.infsof.2023.107202.

[22] VASWANI Ashish, SHAZEER Noam, PARMAR Niki, USZKOREIT Jakob, JONES Llion, GOMEZ Aidan, KAISER Lukasz and POLOSUKHIN Illia (2017), “Attention Is All You Need”, *NIPS’17*, California, USA.

[23] BROWN Tom, MANN Benjamin, RYDER Nick, SUBBIAH Melanie and KAPLAN Jared, DHARIWAL Prafulla, NEELAKANTAN Arvind, SHYAM Pranav, SASTRY Girish, ASKELL Amanda, AGARWAL Sandhini, HERBERT-VOSS Ariel, KRUEGER Gretchen, HENIGHAN Tom, CHILD Rewon, RAMESH Aditya, ZIEGLER Daniel, WU Jeffrey, WINTER Clemens, HESSE Chris, CHEN Mark, SIGLER Eric, LITWIN Mateusz, GRAY Scott, CHESS Benjamin, CLARK Jack, BERNER Christopher, MCCANDLISH Sam, RADFORD Alec, SUTSKEVER ILYA and AMODEI Dario (2020), “Language Models are Few-Shot Learners”, *In, Advances in Neural Information Processing Systems*, Volume 33, Eds. H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan and H. Lin, pp. 1877-1901, Curran Associates Inc., New York.

[24] ALTINOK Duygu (2023), “A Diverse Set of Freely Available Linguistic Resources for Turkish”, *In, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Eds. Anna Rogers, Jordan Boyd-Graber, Naoaki Okazaki, pp.13739–13750, Association for Computational Linguistics, Canada, DOI:10.18653/v1/2023.acl-long.768.

APPENDICES

APPENDIX 1: SEMANTIC RELATIONSHIP FIGURES

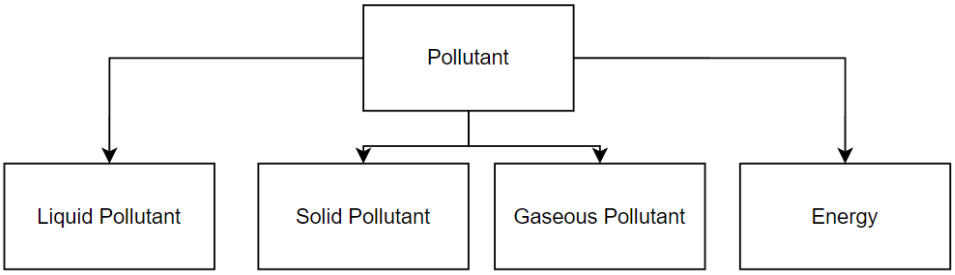

Figure 41: Semantic Relationships of Labels Related to Pollutant

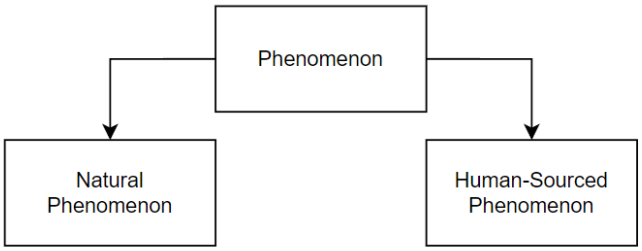

Figure 42: Semantic Relationships of Labels Related to Phenomenon

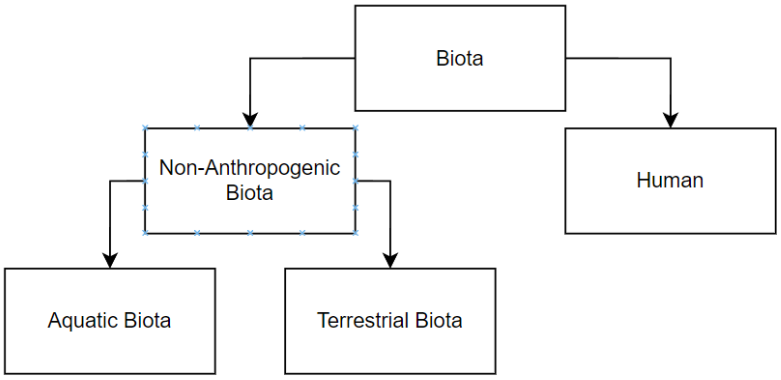

Figure 43: Semantic Relationships of Labels Related to Biota

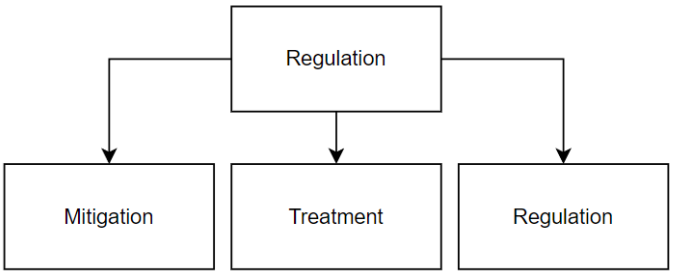

Figure 44: Semantic Relationships of Labels Related to Regulation

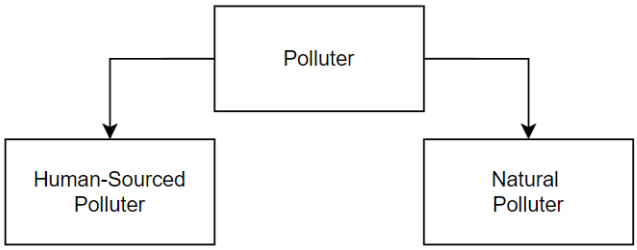

Figure 45: Semantic Relationships of Labels Related to Polluter

APPENDIX 2: HYPERNYMS PREDICTION F1 SCORES

News Dataset

Table 40: Presence of Hypernyms, News Dataset F1 Scores

Unseen Class	Shot Number	Without Hypernym	With Hypernym
Aquatic Biota	0	0.1818	0.6200
	1	0.2553	0.6364
Aquatic Ecosystem	0	0.1429	0.2439
	1	0.1481	0.2155
	10	0.1533	0.2531
Ecological Impact	0	0.3683	0.7234
	1	0.3317	0.7218
	10	0.4041	0.6960
Economic Impact	0	0.2968	0.6182
	1	0.3333	0.6550
Gas Pollutant	0	0.2927	0.4286
	1	0.3077	0.4286
Health Impact	0	0.0346	0.8289
	1	0.0270	0.8454
	10	0.3795	0.8376
Human	0	0.0484	0.0619
	1	0.0645	0.0600
	10	0.0719	0.0971
Human-Sourced Phenomenon	0	0.2205	0.3356
	1	0.2000	0.3394
	10	0.2066	0.3469
Human-Sourced Polluter	0	0.1333	0.2527
	1	0.1444	0.2857
	10	0.1341	0.2959
Liquid Pollutant	0	0.1695	0.1905
	1	0.1277	0.2642
Mitigation	0	0.0769	0.0769
	1	0.0769	0.0714

Table 40 continued

Natural Phenomenon	0	0.1357	0.5881
	1	0.1471	0.3462
	10	0.1283	0.4200
Natural Polluter	0	0.1667	0.2727
	1	0.1739	0.3200
Regulation	0	0.3232	0.5138
	1	0.3093	0.5310
Settlements	0	0.1148	0.2800
	1	0.1094	0.3000
Solid Pollutant	0	0.0851	0.1395
	1	0.0769	0.2051
Terrestrial Biota	0	0.0800	0.3556
	1	0.0952	0.4444

Turkish Wiki NER and News Datasets

Table 41: Presence of Hypernyms, Turkish Wiki NER and News Datasets F1 Scores

Unseen Class	**Shot Number**	**Without Hypernym**	**With Hypernym**
Aquatic Biota	0	0.4522	0.6372
	1	0.4779	0.8224
Aquatic Ecosystem	0	0.1235	0.1856
	1	0.1440	0.5217
	10	0.3299	0.5799
Ecological Impact	0	0.2700	0.6536
	1	0.2638	0.6299
	10	0.3505	0.7081
Economic Impact	0	0.2938	0.6136
	1	0.2941	0.6136
Gas Pollutant	0	0.3111	0.3810
	1	0.3333	0.4500
Health Impact	0	0.1155	0.8054
	1	0.2704	0.7770
	10	0.4491	0.7925
Human	0	0.1429	0.2701
	1	0.0924	0.1695
	10	0.0952	0.1967

Table 41 continued

Human-Sourced Phenomenon	0	0.2414	0.2941
	1	0.2264	0.4149
	10	0.2581	0.4633
Human-Sourced Polluter	0	0.2150	0.5176
	1	0.2747	0.5856
	10	0.2844	0.4479
Liquid Pollutant	0	0.1509	0.4783
	1	0.2222	0.4490
Mitigation	0	0.0667	0.0714
	1	0.2222	0.2581
Natural Phenomenon	0	0.2206	0.5084
	1	0.1932	0.5734
	10	0.1116	0.5870
Natural Polluter	0	0.0000	0.4444
	1	0.0000	0.5217
Regulation	0	0.1290	0.3030
	1	0.3652	0.5781
Settlements	0	0.2182	0.2824
	1	0.2182	0.3173
Solid Pollutant	0	0.1633	0.3913
	1	0.1290	0.4727
Terrestrial Biota	0	0.2667	0.8077
	1	0.2500	0.7586

News and AI Datasets

Table 42: Presence of Hypernyms, News and AI Datasets Comparison

Unseen Class	Shot Number	Without Hypernym	With Hypernym
Aquatic Biota	0	0.2157	0.7586
	1	0.1860	0.7925
	10	0.1443	0.7500
Aquatic Ecosystem	0	0.1239	0.7232
	1	0.1699	0.6389
	10	0.2697	0.7096
Ecological Impact	0	0.5703	0.7754
	1	0.6307	0.7667
	10	0.6313	0.7616
Economic Impact	0	0.4901	0.7284
	1	0.4903	0.7688
	10	0.5643	0.7602
Energy	0	0.2500	0.3158
	1	0.2400	0.5385
Gas Pollutant	0	0.4068	0.4231
	1	0.4308	0.4231
Health Impact	0	0.1497	0.8288
	1	0.2344	0.8433
	10	0.5714	0.8371
Human	0	0.0391	0.2115
	1	0.0526	0.1875
	10	0.0386	0.1630
Human-Sourced Phenomenon	0	0.1785	0.5989
	1	0.3978	0.5836
	10	0.4348	0.6766
Human-Sourced Polluter	0	0.3603	0.4840
	1	0.2883	0.5248
	10	0.3490	0.6039
Liquid Pollutant	0	0.1319	0.2466
	1	0.2247	0.3797
Mitigation	0	0.1569	0.2222
	1	0.1786	0.2182

Table 42 continued

Natural Phenomenon	0	0.2450	0.8054
	1	0.2211	0.8571
	10	0.2645	0.8379
Natural Polluter	0	0.4384	0.6027
	1	0.4638	0.6105
Regulation	0	0.5161	0.6142
	1	0.5167	0.6567
Settlements	0	0.1217	0.2314
	1	0.1452	0.2680
	10	0.1284	0.3486
Social Impact	0	0.5000	0.6377
	1	0.5435	0.6032
Solid Pollutant	0	0.1538	0.4390
	1	0.3600	0.6364
Terrestrial Biota	0	0.0412	0.7639
	1	0.0388	0.7681
	10	0.1391	0.8511
Treatment	0	0.4571	0.4390
	1	0.4444	0.4571

Turkish Wiki NER, News and AI Datasets

Table 43: Presence of Hypernyms, Turkish Wiki NER, News and AI Datasets F1 Scores

Unseen Class	Shot Number	Without Hypernym	With Hypernym
Aquatic Biota	0	0.46153846	0.8503937
	1	0.38554217	0.87394958
	10	0.40909091	0.74747475
Aquatic Ecosystem	0	0.08403361	0.55328798
	1	0.29390681	0.71403813
	10	0.38379531	0.75276753
Ecological Impact	0	0.22871665	0.50618673
	1	0.34862385	0.73490814
	10	0.53880464	0.77424613
Economic Impact	0	0.33082707	0.76136364
	1	0.42176871	0.77183099
	10	0.62089552	0.7884058
Energy	0	0.19047619	0.21052632
	1	0.25	0.30769231
Gas Pollutant	0	0.24489796	0.56603774
	1	0.27906977	0.53061224
Health Impact	0	0.09459459	0.82122905
	1	0.05	0.81134752
	10	0.38045375	0.81818182
Human	0	0.1352657	0.23255814
	1	0.17142857	0.31099196
	10	0.15730337	0.39800995
Human-Sourced Phenomenon	0	0.25503356	0.63687151
	1	0.16858238	0.58479532
	10	0.33623188	0.70422535
Human-Sourced Polluter	0	0.336	0.61463415
	1	0.26573427	0.56852792
	10	0.41758242	0.68221574
Liquid Pollutant	0	0.26966292	0.25925926
	1	0.34090909	0.34666667
Mitigation	0	0.21818182	0.35820896
	1	0.25454545	0.32786885

Table 43 continued

Natural Phenomenon	0	0.25373134	0.85601578
	1	0.18507891	0.84090909
	10	0.35778175	0.86630037
Natural Polluter	0	0.23529412	0.66666667
	1	0.19607843	0.7816092
Regulation	0	0.4	0.48695652
	1	0.44827586	0.53781513
Settlements	0	0.17667845	0.24817518
	1	0.21145374	0.26229508
	10	0.40601504	0.56666667
Social Impact	0	0.4556962	0.61728395
	1	0.4	0.60240964
Solid Pollutant	0	0.35820896	0.55555556
	1	0.37681159	0.59259259
Terrestrial Biota	0	0.328125	0.78461538
	1	0.359375	0.76691729
	10	0.32432432	0.80952381
Treatment	0	0.43137255	0.55319149
	1	0.52380952	0.61111111