Zero-Shot Learning for Requirement Classification using the CrowdRE Dataset

Project Report

ECS412: BS Project

Submitted by

Rohan Mehra
BS Student
Department of Electrical Engineering and Computer Science
Indian Institute of Science Education and Research Bhopal
E-mail: rohan21@iiserb.ac.in

Under the guidance of

Dr. Arpit Sharma
Associate Professor
Department of Electrical Engineering and Computer Science Indian
Institute of Science Education and Research Bhopal
Bhopal, India

January - April 2025

Contents

Abstract

This project addresses a significant challenge in Requirements Engineering (RE): the need for effective requirement classification without extensive labeled training data. Traditional machine learning approaches for requirements classification typically require large amounts of labeled data, which is often expensive and time-consuming to obtain. We propose using Zero-Shot Learning (ZSL) with the CrowdRE dataset, containing approximately 3,000 smart home requirements classified into five categories: Health, Entertainment, Safety, Energy, and Others.

Our approach leverages pre-trained language models (Sentence-BERT, All-MiniLM-L12-v2, BERTOverflow, and BERT4RE) and explores various label configurations to enable classification without task-specific training data. We implement and evaluate multiple classification scenarios: One-vs-Rest, One-vs-One, and multi-class classifications with different class combinations. The study also investigates the effectiveness of different label representations, including original labels, expert-curated labels, and word-embedding-based labels.

Results demonstrate the viability of ZSL for requirements classification, with performance varying across different classification tasks and label configurations. The approach shows particular promise in One-vs-One classification scenarios, achieving competitive results without the need for labeled training data. This work contributes to addressing the data scarcity problem in Requirements Engineering and offers insights into the application of zero-shot learning for requirements classification tasks.

Chapter 1

Introduction

Requirements engineering (RE) researchers have been increasingly adopting machine learning (ML) approaches for various RE tasks, particularly in requirements classification [8]. However, most current approaches rely heavily on supervised learning techniques, which require substantial amounts of labeled training data. This dependency on labeled data poses a significant challenge in RE, where creating high-quality labeled datasets is expensive, time-consuming, and requires domain expertise [2].

The emergence of crowdsourced requirements datasets, such as the CrowdRE dataset [3], has provided valuable resources for requirements analysis in specific domains. The CrowdRE dataset, focusing on smart home requirements, presents a structured collection of requirements classified into five distinct categories: Health, Entertainment, Safety, Energy, and Others. While this dataset offers rich insights into user requirements for smart home systems, traditional supervised learning approaches would require extensive labeled data for each new classification task or domain.

1.0.1 Problem Statement

The key challenges in requirements classification are:

- The need for large amounts of labeled training data for each new classification task
- Limited generalizability of supervised models to new domains or classification schemes
- High cost and time investment in creating labeled datasets
- Dependency on domain expertise for accurate labeling

1.0.2 Proposed Approach

To address these challenges, we propose using Zero-Shot Learning (ZSL) [1], an emerging paradigm that enables classification without task-specific training data. Our approach leverages pre-trained language models (Sentence-BERT [4], All-MiniLM-L12-v2, BERTOverflow, and BERT4RE) and explores various label configurations to enable effective classification without the need for labeled training data.

The ZSL approach offers several advantages:

- Eliminates the need for task-specific labeled training data

- Provides flexibility in handling new classification schemes
- Enables rapid adaptation to different domains
- Reduces dependency on domain experts for labeling

1.0.3 Objectives

The primary objectives of this study are:

- Implement and evaluate ZSL for requirements classification using the CrowdRE dataset
- Compare different label configurations:
  - Original labels
  - Expert-curated labels
  - Word-embedding-based labels (top 20) for pre-trained Word2Vec and GloVe
  - Word-embedding-based labels (top 50) for pre-trained Word2Vec and GloVe
  - Combined Orginal, Expert-curated, and Word-embedding-based labels
- Assess performance across various classification scenarios:
  - One vs Rest (OvR) classification
  - One vs One (OvO) classification
  - Multi-class classification with different combinations (3-class, 4-class, and 5-class)
- Analyze the effectiveness of different pre-trained language models for ZSL in requirements classification

This work contributes to addressing the data scarcity problem in requirements engineering while demonstrating the practical applicability of zero-shot learning for requirements classification tasks. The findings have implications for both research and practice, particularly in scenarios where labeled training data is scarce or expensive to obtain.

Chapter 2

Background and Literature Review

This chapter provides the foundational concepts of our zero-shot learning approach for requirements classification.

2.1 The CrowdRE Dataset

The CrowdRE dataset [3] provides a structured collection of smart home requirements, containing approximately 3,000 requirements categorized into five classes: Health, Entertainment, Safety, Energy, and Other. This dataset represents real-world requirements gathered through crowdsourcing, making it particularly suitable for evaluating automated classification approaches. The class distribution and percentages in the dataset are shown in Table 2.1.

Class	**Number of Requirements**	**Percentage (%)**
Safety	892	30.07
Energy	626	21.11
Health	593	19.99
Entertainment	471	15.88
Other	384	12.95
Total	2966	100.00

Table 2.1: Class distribution and percentages in the CrowdRE dataset.

2.2 Zero-Shot Learning

Zero-shot learning (ZSL) enables classification without task-specific training data [1]. Unlike traditional supervised approaches that require extensive labeled datasets, ZSL leverages pre-trained language models and semantic similarity to classify requirements into predefined categories without additional training. This work replicates and builds upon the experiments conducted by Waad Alhoshan, Alessio Ferrari, and Liping Zhao [1] on a different dataset, applying their framework to a new context. By doing so, we demonstrate the adaptability and generalizability of ZSL methodologies to diverse requirements engineering scenarios.

2.3 Language Models

Our implementation utilizes four pre-trained language models, divided into generic and domain-specific categories:

2.3.1 Generic Language Models

These models are trained on general-purpose text and are designed for broad applicability:

- **Sentence-BERT (Sbert)** [4]:
  - Based on BERT architecture
  - Optimized for sentence-level embeddings
  - Specifically designed for semantic similarity tasks
  - Pre-trained on large-scale general text corpora
- **All-MiniLM-L12-v2 (AllMini)** [7]:
  - Lightweight version of BERT
  - Optimized for efficiency while maintaining performance
  - Suitable for resource-constrained environments
  - Effective for semantic similarity computations

2.3.2 Domain-Specific Models

These models are specifically trained or fine-tuned for technical and requirements-related content:

- **BERTOverflow (SObert)** [6]:
  - Trained on Stack Overflow data
  - Specialized for technical and software engineering text
  - Better understanding of technical terminology
  - Pre-trained on 152 million technical sentences
- **BERT4RE (Bert4RE)** [5]:
  - Fine-tuned specifically for requirements engineering
  - Trained on requirements documentation
  - Optimized for requirements-specific terminology
  - Enhanced understanding of requirements context
  - Some weights of RobertaModel were not initialized from the model checkpoint at thearod5/bert4re and are newly initialized: ['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']

2.4 Label Configuration Overview

The effectiveness of ZSL depends significantly on how class labels are represented. Our approach explores five label configuration strategies namely Original Labels, Expert-Curated Label, Word-embedding-based labels (top 20 & top 50), and Combined Original, Expert-curated, and Word-embedding-based labels

The detailed implementation and comparison of these configurations are discussed in Chapter 3 (Methodology).

Chapter 3

Methodology

This chapter details our systematic approach to implementing zero-shot learning for requirements classification, focusing on label configurations, classification processes, and experimental setup.

3.1 Label Configurations

Our approach utilizes eight distinct label configurations, each designed to capture different aspects of requirement classes. These configurations progressively increase in complexity and semantic richness.

3.1.1 Original Labels (Configuration A)

The baseline configuration uses simple, direct class names. This configuration serves as a control group, using only the primary class identifier (e.g., "Health" for health-related requirements). While simple, this configuration helps evaluate whether basic class names are sufficient for zero-shot classification.

3.1.2 Expert-Curated Labels (Configuration B)

This configuration incorporates domain expertise in requirements engineering and smart home systems. For each class, relevant terms are manually curated that capture different aspects of the category. For example, the health category includes terms related to wellness, medical monitoring, and fitness tracking. These terms were selected based on:

- Common terminology in smart home requirements
- Domain-specific vocabulary
- Synonyms and related concepts
- Functional aspects of each category

3.1.3 Word Embedding Labels (Configuration C, D, E, F)

Word2Vec embeddings and GloVe embeddings are utilized to generate word embedding-based labels. Specifically:

- **Model:** Pre-trained Word2Vec model (from Google) and GloVe pretrained model (glove-wiki-gigaword-300)
- **Similarity Metric:** Cosine similarity between word vectors

The process for generating these labels involved:

- Starting with the base class term (e.g., "health")
- Computing semantic similarity with domain-specific vocabulary
- Ranking terms by similarity scores
- Selecting top-20 terms for Configuration C
- Extending to top-50 terms for Configuration D
- Hyphenating compound terms for consistency (for Word2Vec)

3.1.4 Combined Labels (Configuration G and H)

This configuration merges the strengths (for both Word2Vec and GloVe) of manual expertise and automated generation by combining:

- Original class labels
- Expert-curated terms
- Terms from word embeddings

3.2 Classification Tasks

We implement three types of classification scenarios for each label configuration and each model:

3.2.1 One vs Rest (OvR)

Binary classification comparing each class against all others:

- Health vs Not-Health
- Entertainment vs Not-Entertainment
- Safety vs Not-Safety
- Energy vs Not-Energy
- Other vs Not-Other

3.2.2 One vs One (OvO)

Pairwise classification between classes, resulting in 10 combinations:

- Health vs Entertainment
- Health vs Safety
- Health vs Energy

- Health vs Other
- Entertainment vs Safety
- Entertainment vs Energy
- Entertainment vs Other
- Safety vs Energy
- Safety vs Other
- Energy vs Other

3.2.3 Multi-class Classification

Three different multi-class scenarios:

- Three-class combinations (10 possible combinations)
- Four-class combinations (5 possible combinations)
- Five-class classification (1 combination)

3.3 Experimental Setup

3.3.1 Hardware Configuration

- Google Colab environment (Jupyter)
- GPU: NVIDIA T4
- CUDA acceleration enabled

3.3.2 Implementation Details

- Batch size: 32 (optimized for GPU memory)
- Maximum sequence length: 512 tokens
- PyTorch framework for model implementation
- Transformers library (version 4.x) for language models

3.3.3 Classification Process

For each classification task:

- Generate embeddings for requirements and labels
- Compute semantic similarities
- Determine classifications based on highest similarity scores
- Record predictions

3.4 Evaluation Framework

3.4.1 Performance Metrics

We evaluate performance using:

- **Precision**: Accuracy of positive predictions.

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

- **Recall**: Completeness of positive predictions.

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

- **F1-Score**: Harmonic mean of precision and recall.

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Class-specific performance metrics**: These metrics (Precision, Recall, F1-score) are computed for each class individually, providing a detailed view of performance across different categories.

- **Overall classification accuracy**: Proportion of correctly classified instances out of all instances.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

3.4.2 Comparative Analysis

Our analysis includes:

- Performance comparison across label configurations
- Model effectiveness for different classification scenarios
- Impact of label configuration on classification accuracy
- Scenario-specific performance patterns

Chapter 4

Results

The performance of each model under each configuration was evaluated using the following metrics:

- **Precision**
- **Recall**
- **F1-Score**

4.1 Summary of Experiment Cases

The table below summarizes the total number of experimental cases across models, classification strategies, and label configurations.

Classification Type	Cases / Label Configuration	Total Cases (All Models & Config.)
One-vs-Rest (OvR)	5	160
One-vs-One (OvO)	10	320
3-class Classification	10	320
4-class Classification	5	160
5-class Classification	1	32
Total Cases	31	992

Table 4.1: Summary of Experiment Cases

This comprehensive evaluation across 620 cases allows us to compare model performance under varying class configurations and classification strategies, offering insights into the suitability of different Zero-Shot Learning models for requirement classification. In the tables, the following short forms are used for brevity and clarity:

- **P1**: Precision for Class1
- **R1**: Recall for Class1
- **P2**: Precision for Class2
- **R2**: Recall for Class2
- **Macro F1**: Macro-averaged F1-Score

- **Entmt**: Entertainment (used to save space in the table)

These short forms allow for a more compact presentation of the results while maintaining readability.

4.2 One-vs-One (OvO) Results

4.2.1 One-vs-One (OvO) Results for Original Labels

Table 4.2 presents the performance metrics for One-vs-One (OvO) classification using **Original Labels** across all four models. Each row corresponds to a specific class combination, with metrics for both classes (Class1 and Class2) and the overall **Macro F1-Score**. The following are some observations drawn from it:-

- **Sbert** and **AllMini** generally achieve higher **Macro F1-Scores** compared to **SObert** and **Bert4RE**, indicating better overall performance generalized models in OvO classification in this model.
- Combinations involving **Other** class show lower performance across all models.
- **AllMini** consistently achieves the high **Macro F1-Scores** across most class combinations.

4.2.2 One-vs-One (OvO) Results for Expert-Curated Labels

Table 4.3 presents the performance metrics for OvO classification using **Expert-Curated Labels** across all four models. Some observations made from it are:-

- **Sbert** and **AllMini** again generally achieve higher **Macro F1-Scores** compared to **SObert** and **Bert4RE**, supporting our last claim.
- Those classes involving **Other** show lower performance across all models. This suggests that the **Other** class remains challenging to classify, even with expert-curated labels.
- **Sbert** consistently achieves the highest **Macro F1-Scores** across most class combinations, particularly for **Health vs Entertainment** (0.8385) and **Entertainment vs Energy** (0.8505).
- **SObert** shows the lowest performance, particularly in class combinations involving **Energy** and **Other**. For example, in **Entertainment vs Energy**, **SObert** achieves a **Macro F1-Score** of only 0.3633.

4.2.3 One-vs-One (OvO) Results for top 20 Word-Embedding based Labels

Table 4.4 shows OvO for Word-embedding-based labels (top 20 taken) for Word2Vec embedding, Table 4.5 shows results for GloVe Embeddings. Results align with prior observations: **Sbert** and **AllMini** generalize better, while **SObert** and **Bert4RE** face challenges, particularly with the **Other** class.

Model	Class1	Class2	P1	R1	P2	R2	Macro F1
Sbert	Health	Safety	0.6998	0.4755	0.7126	0.8643	0.6737
	Health	Entmt	0.8385	0.7791	0.7446	0.8110	0.7921
	Health	Energy	0.5501	0.8702	0.7260	0.3259	0.5620
	Health	Other	0.6315	0.9073	0.5600	0.1823	0.5098
	Entmt	**Safety**	**0.7595**	**0.7240**	**0.8578**	**0.8789**	**0.8048**
	Entmt	Energy	0.5899	0.8705	0.8483	0.5447	0.6833
	Entmt	Other	0.5852	0.8896	0.6259	0.2266	0.5193
	Safety	Energy	0.6542	0.9204	0.7300	0.3067	0.5984
	Safety	Other	0.6966	0.8957	0.2791	0.0938	0.4620
	Energy	Other	0.6797	0.8610	0.5991	0.3385	0.5962
AllMini	Health	Safety	0.8870	0.6088	0.7848	0.9484	0.7904
	Health	Entmt	0.8874	0.6509	0.6709	0.8960	0.7591
	Health	Energy	0.8797	0.4317	0.6369	0.9441	0.6699
	Health	Other	0.7284	0.7099	0.5689	0.5911	0.6494
	Entmt	**Safety**	**0.7510**	**0.8259**	**0.9030**	**0.8554**	**0.8326**
	Entmt	**Energy**	**0.9432**	**0.6348**	**0.7795**	**0.9712**	**0.8119**
	Entmt	Other	0.6409	0.8641	0.7091	0.4063	0.6263
	Safety	Energy	0.9094	0.6861	0.6686	0.9026	0.7751
	Safety	Other	0.7827	0.9002	0.6440	0.4193	0.6726
	Energy	Other	0.6608	0.9553	0.7333	0.2005	0.5481
SObert	Health	Safety	0.4007	0.9798	0.6571	0.0258	0.3092
	Health	Entmt	0.5575	0.7437	0.4432	0.2569	0.4813
	Health	Energy	0.8485	0.0472	0.5236	0.9920	0.3874
	Health	Other	0.6067	0.9831	0.3750	0.0156	0.3902
	Entmt	Safety	0.3471	0.9618	0.6897	0.0448	0.2972
	Entmt	Energy	0.5000	0.0021	0.5708	0.9984	0.3653
	Entmt	Other	0.5509	0.9533	0.4500	0.0469	0.3916
	Safety	Energy	1.0000	0.0067	0.4140	1.0000	0.2995
	Safety	Other	0.7188	0.3610	0.3116	0.6719	0.4532
	Energy	Other	0.6198	1.0000	0.0000	0.0000	0.3826
Bert4RE	Health	Safety	0.3789	0.4536	0.5819	0.5056	0.4770
	Health	Entmt	0.5763	0.1147	0.4450	0.8938	0.3927
	Health	Energy	0.4869	0.9106	0.5182	0.0911	0.3947
	Health	Other	0.6290	0.4317	0.4088	0.6068	0.5002
	Entmt	Safety	0.3375	0.7452	0.6285	0.2276	0.3994
	Entmt	Energy	0.4313	0.9936	0.7500	0.0144	0.3149
	Entmt	Other	0.5432	0.7749	0.4208	0.2005	0.4551
	Safety	Energy	0.6050	0.8464	0.4926	0.2125	0.5012
	Safety	Other	0.6565	0.4585	0.2603	0.4427	0.4339
	Energy	Other	0.7015	0.1502	0.3927	0.8958	0.3967

Table 4.2: One-vs-One (OvO) Results for Original Labels

4.2.4 One-vs-One (OvO) Results for top 50 Word-Embedding based Label

The results are shown in Table 4.6 and Table 4.7 reinforce previous claims: **Sbert** and **AllMini** consistently outperform **SObert** and **Bert4RE**, achieving higher

Model	Class1	Class2	P1	R1	P2	R2	Macro F1
Sbert	Health	Safety	0.8608	0.5632	0.7639	0.9395	0.7618
	Health	**Entmt**	**0.8368**	**0.8904**	**0.8499**	**0.7813**	**0.8385**
	Health	Energy	0.7523	0.6863	0.7257	0.7859	0.7362
	Health	Other	0.6352	0.9309	0.6204	0.1745	0.5137
	Entmt	Safety	0.9215	0.5732	0.8121	0.9742	0.7963
	Entmt	**Energy**	**0.8925**	**0.7580**	**0.8364**	**0.9313**	**0.8505**
	Entmt	Other	0.6598	0.8811	0.7522	0.4427	0.6560
	Safety	Energy	0.8534	0.8879	0.8305	0.7827	0.8381
	Safety	Other	0.7097	0.9978	0.9091	0.0521	0.4640
	Energy	Other	0.6638	0.9776	0.8409	0.1927	0.5521
AllMini	Health	Safety	0.9571	0.2631	0.6694	0.9922	0.6061
	Health	Entmt	0.8276	0.7774	0.7396	0.7962	0.7843
	Health	Energy	0.9150	0.3086	0.5976	0.9728	0.6010
	Health	Other	0.6811	0.4503	0.4427	0.6745	0.5384
	Entmt	Safety	0.9664	0.4883	0.7858	0.9910	0.7627
	Entmt	Energy	0.9515	0.5414	0.7394	0.9792	0.7664
	Entmt	Other	0.8139	0.6964	0.6836	0.8047	0.7449
	Safety	**Energy**	**0.8908**	**0.9148**	**0.8738**	**0.8403**	**0.8797**
	Safety	Other	0.7417	0.9720	0.7664	0.2135	0.5877
	Energy	Other	0.6983	0.9281	0.7472	0.3464	0.6351
SObert	Health	Safety	0.4258	0.1501	0.6050	0.8655	0.4671
	Health	Entmt	0.5703	0.2530	0.4469	0.7601	0.4567
	Health	Energy	0.5567	0.0911	0.5196	0.9313	0.4118
	Health	Other	0.5709	0.2648	0.3789	0.6927	0.4258
	Entmt	Safety	0.4205	0.0786	0.6596	0.9428	0.4543
	Entmt	Energy	0.0000	0.0000	0.5706	1.0000	0.3633
	Entmt	Other	0.5135	0.2824	0.4329	0.6719	0.4455
	Safety	Energy	0.6235	0.6368	0.4662	0.4521	0.5446
	Safety	Other	0.6872	0.5123	0.2881	0.4583	0.4704
	Energy	Other	0.6099	0.4744	0.3709	0.5052	0.4807
Bert4RE	Health	Safety	0.5000	0.0590	0.6057	0.9608	0.4243
	Health	Entmt	0.8082	0.0995	0.4612	0.9703	0.4012
	Health	Energy	0.7692	0.0169	0.5166	0.9952	0.3566
	Health	Other	0.7400	0.0624	0.4002	0.9661	0.3405
	Entmt	Safety	0.4051	0.0679	0.6581	0.9473	0.4465
	Entmt	Energy	0.5280	0.7601	0.7303	0.4888	0.6044
	Entmt	Other	0.6567	0.0934	0.4581	0.9401	0.3898
	Safety	Energy	0.5918	0.9070	0.4503	0.1086	0.4456
	Safety	Other	0.6157	0.1850	0.2788	0.7318	0.3441
	Energy	Other	0.7778	0.1006	0.3940	0.9531	0.3679

Table 4.3: One-vs-One (OvO) Results for Expert-Curated Labels

Macro F1-Scores, especially in combinations like **Health vs Entertainment** and **Entertainment vs Energy**. The **Other** class remains problematic.

Model	Class1	Class2	P1	R1	P2	R2	Macro F1
Sbert	Health	Safety	0.8367	0.5531	0.7575	0.9283	0.7501
	Health	Entmt	0.8229	0.8617	0.8149	0.7665	0.8159
	Health	Energy	0.8329	0.5632	0.6834	0.8930	0.7231
	Health	Other	0.7081	0.7201	0.5561	0.5417	0.6314
	Entmt	Safety	0.8833	0.5626	0.8062	0.9608	0.7821
	Entmt	**Energy**	**0.8872**	**0.7346**	**0.8232**	**0.9297**	**0.8385**
	Entmt	Other	0.8093	0.6667	0.6638	0.8073	0.7298
	Safety	**Energy**	**0.8673**	**0.8643**	**0.8076**	**0.8115**	**0.8377**
	Safety	Other	0.7622	0.9596	0.7647	0.3047	0.6427
	Energy	Other	0.7497	0.9185	0.7901	0.5000	0.7190
AllMini	Health	Safety	0.8030	0.5497	0.7525	0.9103	0.7383
	Health	Entmt	0.8197	0.7589	0.7223	0.7898	0.7713
	Health	Energy	0.8861	0.4722	0.6534	0.9425	0.6939
	Health	Other	0.6715	0.7099	0.5086	0.4635	0.5876
	Entmt	**Safety**	**0.8445**	**0.7495**	**0.8751**	**0.9271**	**0.8473**
	Entmt	Energy	0.9184	0.6688	0.7931	0.9553	**0.8203**
	Entmt	Other	0.6830	0.8599	0.7481	0.5104	0.6840
	Safety	**Energy**	**0.9439**	**0.7915**	**0.7584**	**0.9329**	**0.8488**
	Safety	Other	0.7782	0.9596	0.7955	0.3646	0.6797
	Energy	Other	0.6575	0.9569	0.7273	0.1875	0.5388
SObert	Health	Safety	0.4681	0.1113	0.6079	0.9159	0.4553
	Health	Entmt	0.5357	0.0253	0.4421	0.9724	0.3281
	Health	Energy	0.7122	0.1669	0.5426	0.9361	0.4787
	Health	Other	0.5974	0.6981	0.3697	0.2734	0.4791
	Entmt	Safety	0.3560	0.8238	0.6960	0.2130	0.4116
	Entmt	Energy	0.4697	0.9533	0.8440	0.1901	0.4698
	Entmt	Other	0.5504	0.9979	0.0000	0.0000	0.3547
	Safety	Energy	0.6487	0.7287	0.5310	0.4377	0.5831
	Safety	Other	0.6993	0.9854	0.3158	0.0156	0.4239
	Energy	Other	0.6268	0.9633	0.5208	0.0651	0.4376
Bert4RE	Health	Safety	0.4695	0.2597	0.6206	0.8049	0.5176
	Health	Entmt	1.0000	0.0118	0.4456	1.0000	0.3199
	Health	Energy	0.5094	0.0455	0.5146	0.9585	0.3766
	Health	Other	0.6192	0.5430	0.4070	0.4844	0.5105
	Entmt	Safety	0.3463	1.0000	1.0000	0.0034	0.2606
	Entmt	Energy	0.4314	0.9809	0.6538	0.0272	0.3257
	Entmt	Other	0.5509	1.0000	0.0000	0.0000	0.3552
	Safety	Energy	0.6378	0.5448	0.4630	0.5591	0.5471
	Safety	Other	0.6950	0.9350	0.2368	0.0469	0.4378
	Energy	Other	0.6324	0.8243	0.4330	0.2188	0.5032

Table 4.4: One-vs-One (OvO) Results for Word-Embedding-20 Labels (Word2Vec)

4.2.5 One-vs-One (OvO) Results for Combined Labels

Table 4.8 and Table 4.9 again reinforces previous claims.

Model	Class1	Class2	P1	R1	P2	R2	Macro F1
Sbert	Safety	Other	0.6997	0.9664	0.3182	0.0365	0.4385
	Health	Entmt	0.8415	0.7791	0.7456	0.8153	0.7940
	Entmt	**Energy**	**0.8581**	**0.8344**	**0.8779**	**0.8962**	**0.8665**
	Safety	Energy	0.6675	0.9182	0.7491	0.3482	0.6242
	Energy	Other	0.7113	0.9169	0.7438	0.3932	0.6578
	Entmt	Safety	0.6479	0.6093	0.8000	0.8251	0.7202
	Entmt	Other	0.5950	0.8577	0.6193	0.2839	0.5459
	Health	Other	0.6137	0.8735	0.4361	0.1510	0.4727
	Health	Energy	0.7454	0.6762	0.7181	0.7812	0.7287
	Health	Safety	0.5825	0.1012	0.6143	0.9518	0.4596
AllMini	Safety	Other	0.7446	0.7388	0.4041	0.4115	0.5747
	Health	Entmt	0.8982	0.4165	0.5615	0.9406	0.6361
	Entmt	Energy	0.7782	0.8493	0.8782	0.8179	0.8296
	Safety	Energy	0.7326	0.9798	0.9446	0.4904	0.7420
	Energy	Other	0.7276	0.6358	0.5076	0.6120	0.6168
	Entmt	Safety	0.8412	0.6412	0.8317	0.9361	0.8043
	Entmt	Other	0.6139	0.6178	0.5276	0.5234	0.5707
	Health	Other	0.6298	0.4132	0.4082	0.6250	0.4964
	Health	Energy	0.8711	0.3761	0.6158	0.9473	0.6359
	Health	Safety	0.8721	0.1265	0.6297	0.9877	0.4950
SObert	Safety	Other	0.6976	0.9854	0.1875	0.0078	0.4160
	Health	Entmt	0.6131	0.4570	0.4823	0.6369	0.5363
	Entmt	Energy	0.4223	0.6285	0.5581	0.3530	0.4688
	Safety	Energy	0.5938	0.8554	0.4464	0.1661	0.4716
	Energy	Other	0.5960	0.3818	0.3645	0.5781	0.4563
	Entmt	Safety	0.4532	0.1338	0.6667	0.9148	0.4889
	Entmt	Other	0.5294	0.2675	0.4408	0.7083	0.4494
	Health	Other	0.5608	0.3187	0.3688	0.6146	0.4337
	Health	Energy	0.5252	0.6678	0.5763	0.4281	0.5396
	Health	Safety	0.4043	0.0961	0.6012	0.9058	0.4390
Bert4RE	Safety	Other	0.7000	0.0628	0.3010	0.9375	0.2855
	Health	Entmt	0.7059	0.0202	0.4451	0.9894	0.3267
	Entmt	Energy	0.4307	0.9894	0.6667	0.0160	0.3157
	Safety	Energy	0.5928	0.9271	0.4715	0.0927	0.4390
	Energy	Other	0.7273	0.0383	0.3838	0.9766	0.3120
	Entmt	Safety	0.3199	0.4544	0.6297	0.4899	0.4633
	Entmt	Other	0.5000	0.0722	0.4447	0.9115	0.3620
	Health	Other	0.6806	0.0826	0.3989	0.9401	0.3537
	Health	Energy	0.4924	0.7605	0.5314	0.2572	0.4722
	Health	Safety	0.4222	0.1282	0.6038	0.8834	0.4570

Table 4.5: One-vs-One (OvO) Results for Word-Embedding-20 Labels (GloVe)

4.2.6 Varition across different label categories in OvO

Performance trends are stable across label categories, with **Sbert** and **AllMini** generalizing well, achieving higher **Macro F1-Scores** in key combinations like **Health vs**

Model	Class1	Class2	P1	R1	P2	R2	Macro F1
Sbert	Health	Safety	0.8492	0.4654	0.7267	0.9451	0.7115
	Health	Entmt	0.8491	0.7589	0.7322	0.8301	0.7898
	Health	Energy	0.8731	0.3946	0.6225	0.9457	0.6472
	Health	Other	0.7100	0.6813	0.5368	0.5703	0.6242
	Entmt	**Safety**	**0.8739**	**0.6327**	**0.8307**	**0.9518**	**0.8106**
	Entmt	**Energy**	**0.8538**	**0.7686**	**0.8380**	**0.9010**	**0.8386**
	Entmt	Other	0.7495	0.7941	0.7275	0.6745	0.7356
	Safety	**Energy**	**0.8663**	**0.8643**	**0.8073**	**0.8099**	**0.8370**
	Safety	Other	0.7390	0.9776	0.7917	0.1979	0.5792
	Energy	Other	0.7235	0.9281	0.7826	0.4219	0.6807
AllMini	Health	Safety	0.8300	0.3541	0.6891	0.9518	0.6479
	Health	Entmt	0.8628	0.4772	0.5788	0.9045	0.6602
	Health	Energy	0.8788	0.2934	0.5896	0.9617	0.5855
	Health	Other	0.7192	0.4621	0.4648	0.7214	0.5640
	Entmt	**Safety**	**0.8155**	**0.8068**	**0.8986**	**0.9036**	**0.8561**
	Entmt	**Energy**	**0.9263**	**0.6943**	**0.8065**	**0.9585**	**0.8348**
	Entmt	Other	0.7003	0.8981	0.8088	0.5286	0.7132
	Safety	Energy	0.9547	0.7085	0.6963	0.9521	0.8089
	Safety	Other	0.7989	0.9395	0.7621	0.4505	0.7149
	Energy	Other	0.6536	0.9553	0.7053	0.1745	0.5279
SObert	Health	Safety	0.0000	0.0000	0.6007	1.0000	0.3753
	Health	Entmt	0.0000	0.0000	0.4427	1.0000	0.3068
	Health	Energy	0.0000	0.0000	0.5135	1.0000	0.3393
	Health	Other	0.0000	0.0000	0.3930	1.0000	0.2821
	Entmt	Safety	0.3508	0.9660	0.7576	0.0561	0.3095
	Entmt	Energy	0.4297	1.0000	1.0000	0.0016	0.3022
	Entmt	Other	0.5504	0.9979	0.0000	0.0000	0.3547
	Safety	Energy	0.6116	0.9709	0.7451	0.1214	0.4796
	Safety	Other	0.6953	0.9159	0.2574	0.0677	0.4489
	Energy	Other	0.6549	0.6821	0.4441	0.4141	0.5484
Bert4RE	Health	Safety	0.0000	0.0000	0.6007	1.0000	0.3753
	Health	Entmt	0.5597	0.9882	0.5882	0.0212	0.3778
	Health	Energy	0.0000	0.0000	0.5135	1.0000	0.3393
	Health	Other	0.0000	0.0000	0.3930	1.0000	0.2821
	Entmt	Safety	0.0000	0.0000	0.6544	1.0000	0.3956
	Entmt	Energy	0.0000	0.0000	0.5706	1.0000	0.3633
	Entmt	Other	0.0000	0.0000	0.4491	1.0000	0.3099
	Safety	Energy	0.6568	0.4742	0.4634	0.6470	0.5454
	Safety	Other	0.7778	0.0078	0.3015	0.9948	0.2391
	Energy	Other	0.7895	0.0240	0.3835	0.9896	0.2996

Table 4.6: One-vs-One (OvO) Results for Word-Embedding-50 Labels (Word2Vec)

Entertainment and **Entertainment vs Energy**. In contrast, **SObert** and **Bert4RE** struggle with imbalanced precision and recall, particularly in challenging combinations involving the **Other** class.

Model	**Class1**	**Class2**	**P1**	**R1**	**P2**	**R2**	**Macro F1**
Sbert	Safety	Other	0.7036	0.9608	0.3966	0.0599	0.4582
	Health	Entmt	0.8398	0.8044	0.7661	0.8068	0.8038
	Entmt	**Energy**	**0.8206**	**0.8641**	**0.8935**	**0.8578**	**0.8585**
	Safety	Energy	0.6853	0.9204	0.7781	0.3978	0.6560
	Energy	Other	0.7150	0.9137	0.7429	0.4063	0.6637
	Entmt	Safety	0.6498	0.6815	0.8274	0.8061	0.7409
	Entmt	Other	0.5841	0.8917	0.6250	0.2214	0.5164
	Health	Other	0.6122	0.9157	0.4444	0.1042	0.4513
	Health	Energy	0.6819	0.7808	0.7593	0.6550	0.7156
	Health	Safety	0.7365	0.1838	0.6380	0.9563	0.5298
AllMini	Safety	Other	0.7216	0.9821	0.7419	0.1198	0.5191
	Health	Entmt	0.8528	0.5666	0.6164	0.8769	0.7024
	Entmt	Energy	0.7428	0.8705	0.8881	0.7732	0.8141
	Safety	Energy	0.7446	0.9709	0.9268	0.5256	0.7568
	Energy	Other	0.6948	0.8403	0.6047	0.3984	0.6205
	Entmt	Safety	0.7339	0.7261	0.8562	0.8610	0.7943
	Entmt	Other	0.6000	0.8726	0.6471	0.2865	0.5541
	Health	Other	0.6469	0.7538	0.4895	0.3646	0.5571
	Health	Energy	0.7700	0.6155	0.6940	0.8259	0.7192
	Health	Safety	0.8102	0.2951	0.6706	0.9540	0.6101
SObert	Safety	Other	0.6991	1.0000	0.0000	0.0000	0.4114
	Health	Entmt	0.7246	0.0843	0.4543	0.9597	0.3839
	Entmt	**Energy**	**0.5000**	**0.4756**	**0.6194**	**0.6422**	**0.5590**
	Safety	Energy	0.5876	0.9966	0.4000	0.0032	0.3728
	Energy	Other	0.5556	0.0559	0.3759	0.9271	0.3183
	Entmt	Safety	0.0000	0.0000	0.6537	0.9966	0.3948
	Entmt	Other	0.0000	0.0000	0.4485	0.9974	0.3094
	Health	Other	1.0000	0.0034	0.3938	1.0000	0.2859
	Health	Energy	0.6111	0.0742	0.5214	0.9553	0.4034
	Health	Safety	0.2500	0.0017	0.6003	0.9966	0.3763
Bert4RE	Safety	Other	0.6957	0.0717	0.3007	0.9271	0.2921
	Health	Entmt	0.7647	0.0219	0.4460	0.9915	0.3289
	Entmt	Energy	0.4294	1.0000	0.0000	0.0000	0.3004
	Safety	Energy	0.5876	1.0000	0.0000	0.0000	0.3701
	Energy	Other	0.2500	0.0016	0.3787	0.9922	0.2757
	Entmt	Safety	0.0000	0.0000	0.6537	0.9966	0.3948
	Entmt	Other	0.4375	0.0149	0.4470	0.9766	0.3210
	Health	Other	0.6800	0.0287	0.3950	0.9792	0.3089
	Health	Energy	0.4869	1.0000	1.0000	0.0016	0.3290
	Health	Safety	0.0000	0.0000	0.6007	1.0000	0.3753

Table 4.7: One-vs-One (OvO) Results for Word-Embedding-50 Labels (GloVe)

4.3 One-vs-Rest (OvR) Results

The One-vs-Rest (OvR) results (Tables 4.10, 4.11, 4.12, 4.14, and 4.16 4.13, 4.15, 4.17) across the eight label sets (Original, Expert-curated, Word-Embedding-20 for Word2Vec

Model	Class1	Class2	P1	R1	P2	R2	Macro F1
Sbert	Health	Entmt	0.8427	0.8314	0.7912	0.8047	**0.8175**
	Health	Safety	0.8638	0.5025	0.7412	0.9473	0.7335
	Health	Energy	0.8037	0.5801	0.6852	0.8658	0.7194
	Health	Other	0.6961	0.7302	0.5493	0.5078	0.6202
	Entmt	Safety	0.8875	0.6030	0.8207	0.9596	**0.8014**
	Entmt	Energy	0.8723	0.7834	0.8487	0.9137	**0.8528**
	Entmt	Other	0.7830	0.7431	0.7034	0.7474	0.7436
	Safety	Energy	0.8409	0.8890	0.8278	0.7604	**0.8285**
	Safety	Other	0.7355	0.9787	0.7865	0.1823	0.5679
	Energy	Other	0.7350	0.9217	0.7822	0.4583	0.6979
AllMini	Health	Entmt	0.8812	0.5126	0.5981	0.9130	0.6854
	Health	Safety	0.8619	0.3474	0.6894	0.9630	0.6494
	Health	Energy	0.8904	0.3423	0.6065	0.9601	0.6189
	Health	Other	0.7113	0.4570	0.4597	0.7135	0.5578
	Entmt	Safety	0.8589	0.7495	0.8761	0.9350	**0.8525**
	Entmt	Energy	0.9207	0.6900	0.8038	0.9553	**0.8309**
	Entmt	Other	0.7372	0.8280	0.7515	0.6380	0.7351
	Safety	Energy	0.9363	0.8240	0.7858	0.9201	**0.8621**
	Safety	Other	0.7962	0.9283	0.7288	0.4479	0.7060
	Energy	Other	0.6962	0.9297	0.7471	0.3385	0.6311
SObert	Health	Entmt	0.0000	0.0000	0.4427	1.0000	0.3068
	Health	Safety	0.0000	0.0000	0.6004	1.0000	0.3750
	Health	Energy	0.0000	0.0000	0.5131	1.0000	0.3389
	Health	Other	1.0000	0.0017	0.3934	1.0000	0.2840
	Entmt	Safety	0.0000	0.0000	0.6544	1.0000	0.3956
	Entmt	Energy	0.5551	0.5244	0.6564	0.6837	0.6045
	Entmt	Other	0.0000	0.0000	0.4491	1.0000	0.3099
	Safety	Energy	0.5889	0.9989	0.8000	0.0064	0.3768
	Safety	Other	1.0000	0.0090	0.3028	1.0000	0.2413
	Energy	Other	1.0000	0.0048	0.3813	1.0000	0.2808
Bert4RE	Health	Entmt	1.0000	0.0034	0.4435	1.0000	0.3106
	Health	Safety	0.0000	0.0000	0.6007	1.0000	0.3753
	Health	Energy	0.0000	0.0000	0.5127	1.0000	0.3386
	Health	Other	0.0000	0.0000	0.3930	1.0000	0.2821
	Entmt	Safety	0.3504	0.9278	0.7069	0.0919	0.3357
	Entmt	Energy	0.4355	0.9745	0.7209	0.0495	0.3473
	Entmt	Other	0.5865	0.8280	0.5737	0.2839	0.5332
	Safety	Energy	0.5934	0.9686	0.5484	0.0543	0.4174
	Safety	Other	0.6200	0.1043	0.2904	0.8516	0.3058
	Energy	Other	0.7273	0.0128	0.3814	0.9922	0.2880

Table 4.8: One-vs-One (OvO) Results for Combined Labels (Word2Vec)

and GloVe, Word-Embedding-50 for Word2Vec and GloVe, and Combined for Word2Vec and GloVe) are analyzed here about the performance of the four models: Sbert, AllMini, SObert, and Bert4RE. The following is a detailed analysis:

Model	Class1	Class2	P1	R1	P2	R2	Macro F1
Sbert	Safety	Other	0.7089	0.9910	0.7241	0.0547	0.4641
	Health	Entmt	0.8266	0.8685	0.8231	0.7707	0.8215
	Entmt	**Energy**	**0.9025**	**0.7665**	**0.8422**	**0.9377**	**0.8582**
	Safety	Energy	0.7222	0.9238	0.8196	0.4936	0.7134
	Energy	Other	0.7234	0.9233	0.7725	0.4245	0.6796
	Entmt	Safety	0.8092	0.4862	0.7759	0.9395	0.7287
	Entmt	Other	0.7023	0.7665	0.6774	0.6016	0.6851
	Health	Other	0.6636	0.8583	0.6000	0.3281	0.5864
	Health	Energy	0.7518	0.7049	0.7360	0.7796	0.7424
	Health	Safety	0.9301	0.2243	0.6572	0.9888	0.5755
AllMini	Safety	Other	0.7417	0.9563	0.6905	0.2266	0.5883
	Health	Entmt	0.8808	0.5481	0.6144	0.9066	0.7040
	Entmt	Energy	0.9082	0.6093	0.7644	0.9537	0.7890
	Safety	Energy	0.8643	0.9137	0.8661	0.7955	0.8588
	Energy	Other	0.6988	0.9042	0.7000	0.3646	0.6339
	Entmt	Safety	0.9075	0.5626	0.8077	0.9697	0.7880
	Entmt	Other	0.7238	0.6454	0.6161	0.6979	0.6684
	Health	Other	0.7663	0.3761	0.4606	0.8229	0.5476
	Health	Energy	0.9623	0.2580	0.5849	0.9904	0.5712
	Health	Safety	0.9545	0.1771	0.6451	0.9944	0.5406
SObert	Safety	Other	0.6627	0.0617	0.2984	0.9271	0.2822
	Health	Entmt	0.5333	0.0270	0.4420	0.9703	0.3293
	Entmt	Energy	0.0000	0.0000	0.5706	1.0000	0.3633
	Safety	Energy	0.5921	0.9619	0.5072	0.0559	0.4169
	Energy	Other	0.5811	0.0687	0.3771	0.9193	0.3289
	Entmt	Safety	0.0000	0.0000	0.6537	0.9966	0.3948
	Entmt	Other	0.2500	0.0021	0.4477	0.9922	0.3106
	Health	Other	0.6250	0.0084	0.3932	0.9922	0.2899
	Health	Energy	0.0000	0.0000	0.5127	0.9968	0.3386
	Health	Safety	0.4000	0.0067	0.6007	0.9933	0.3809
Bert4RE	Safety	Other	0.6111	0.0247	0.2984	0.9635	0.2515
	Health	Entmt	1.0000	0.0017	0.4431	1.0000	0.3087
	Entmt	Energy	0.4301	1.0000	1.0000	0.0032	0.3040
	Safety	Energy	0.5875	0.9978	0.3333	0.0016	0.3713
	Energy	Other	0.6429	0.0144	0.3805	0.9870	0.2887
	Entmt	Safety	0.6000	0.0064	0.6554	0.9978	0.4019
	Entmt	Other	0.0000	0.0000	0.4472	0.9922	0.3083
	Health	Other	0.2000	0.0017	0.3909	0.9896	0.2819
	Health	Energy	0.5798	0.1164	0.5236	0.9201	0.4306
	Health	Safety	0.0000	0.0000	0.6007	1.0000	0.3753

Table 4.9: One-vs-One (OvO) Results for Combined Labels (with GloVe)

- **Overall Trends:**
  - **Recall vs. Precision Trade-off:** Across all models and label sets, there

Model	Class	Precision	Recall	F1-Score	Macro F1
Sbert	Health	0.2084	0.8668	0.3359	0.3142
	Entertainment	0.1806	0.8684	0.2990	0.3494
	Energy	0.1914	0.5655	0.2859	0.3872
	Safety	0.2897	0.8240	0.4287	0.3230
	Other	0.1336	0.5990	0.2185	0.3943
AllMini	Health	0.2154	0.6324	0.3213	0.4405
	Entertainment	0.1381	0.7473	0.2331	0.2189
	Energy	0.2265	0.9073	0.3625	0.3242
	Safety	0.2827	0.7287	0.4074	0.3588
	Other	0.1349	0.9010	0.2347	0.2390
SObert	Health	0.0000	0.0000	0.0000	0.4445
	Entertainment	0.0000	0.0000	0.0000	0.4569
	Energy	0.0000	0.0000	0.0000	0.4410
	Safety	0.0000	0.0000	0.0000	0.4115
	Other	0.0000	0.0000	0.0000	0.4654
Bert4RE	Health	0.0000	0.0000	0.0000	0.4444
	Entertainment	0.1395	0.0382	0.0600	0.4771
	Energy	0.0000	0.0000	0.0000	0.4410
	Safety	0.0000	0.0000	0.0000	0.4112
	Other	0.1163	0.0260	0.0426	0.4801

Table 4.10: One-vs-Rest (OvR) Results for Original Labels

is a consistent trade-off between recall and precision. Models like Sbert and AllMini achieve high recall (often above 0.7 for classes like *Health* and *Safety*) but suffer from low precision (often below 0.3). This indicates that while these models are effective at capturing most of the true positives, they also include many false positives, leading to imbalanced performance.

- **Class-wise Performance:** The *Safety* class consistently achieves the highest F1-scores across all models and label sets (e.g., Sbert achieves an F1-score of 0.4547 in Word-Embedding-20 label config, Table 4.12). This suggests that the *Safety* class has more distinct textual features, making it easier for models to classify. In contrast, the *Other* class consistently underperforms, with low precision and F1-scores (e.g., 0.1810 for Sbert in Combined labels, Table 4.16), as expected because it lacks distinct features (In Other vs Not Other).
- **Impact of Label Configurations:** The choice of labels sets significantly impacts model performance. Expert-curated and word-embedding labels generally improve performance compared to the original labels. For example, Sbert's F1-score for the *Energy* class improves a bit from 0.2859 (Original labels, Table 4.10) to 0.3443 (Expert-curated labels, Table 4.11). This suggests that refined or domain-specific labels help models better capture class distinctions.

Model	Class	Precision	Recall	F1-Score	Macro F1
Sbert	Health	0.2071	0.8415	0.3324	0.3242
	Entmt	0.1905	0.8769	0.3130	0.3812
	Energy	0.2229	0.7556	0.3443	0.3891
	Safety	0.3067	0.8397	0.4493	0.3713
	Other	0.1316	0.4583	0.2045	0.4397
AllMini	Health	0.1681	0.1349	0.1497	0.4814
	Entmt	0.1088	0.0552	0.0732	0.4736
	Energy	0.2344	0.0479	0.0796	0.4728
	Safety	0.2578	0.1211	0.1648	0.4639
	Other	0.1367	0.7057	0.2290	0.3585
SObert	Health	0.2052	0.6223	0.3087	0.4209
	Entmt	0.1666	0.6157	0.2622	0.4117
	Energy	0.2119	0.6981	0.3251	0.3830
	Safety	0.3012	0.5706	0.3943	0.4637
	Other	0.1317	0.2370	0.1693	0.4927
Bert4RE	Health	0.2038	0.6374	0.3088	0.4116
	Entmt	0.1574	0.9384	0.2695	0.1830
	Energy	0.2466	0.4617	0.3215	0.5132
	Safety	0.3052	0.7410	0.4323	0.4141
	Other	0.1281	0.3047	0.1804	0.4755

Table 4.11: One-vs-Rest (OvR) Results for Expert-curated Labels

- **Model-specific:**
  - **Sbert:** Sbert consistently achieves high recall across all label sets, particularly for *Health* and *Safety* (e.g., 0.8668 recall for *Health* in Original labels, Table 4.10). However, its precision remains low, indicating a tendency to over-predict these classes. The model performs best on the *Safety* class, with F1-scores consistently above 0.4 across all label sets.
  - **AllMini:** AllMini shows a more balanced performance compared to Sbert, with higher precision and F1-scores for classes like *Energy* and *Safety* (e.g., 0.3625 F1-score for *Energy* in Original labels, Table 4.10). However, it struggles with the *Entertainment* class, achieving low recall and F1-scores (e.g., 0.2331 F1-score in Original labels, Table 4.10).
  - **SObert:** SObert performs very poorly on the original label set, with F1-scores of 0.0000 for all classes (Table 4.10). However, its performance improves significantly with expert-curated and word-embedding labels, particularly for the *Safety* class (e.g., 0.4624 F1-score in Word-Embedding-50 labels, Table 4.14). Notably, SObert exhibits extreme recall values (1.0000) in word-embedding label sets, suggesting overfitting.
  - **Bert4RE:** Bert4RE shows similar trends to SObert, with very poor performance on the original label set but improved results with refined labels.

Model	Class	Precision	Recall	F1-Score	Macro F1
Sbert	Health	0.2089	0.8465	0.3351	0.3283
	Entmt	0.1871	0.8408	0.3060	0.3845
	Energy	0.2114	0.7923	0.3338	0.3324
	Safety	0.3104	0.8498	0.4547	0.3775
	Other	0.1374	0.6771	0.2285	0.3741
AllMini	Health	0.1772	0.3322	0.2311	0.4605
	Entmt	0.2119	0.3482	0.2635	0.5339
	Energy	0.3034	0.2588	0.2793	0.5521
	Safety	0.3874	0.5975	0.4700	0.5710
	Other	0.1344	0.3542	0.1948	0.4735
SObert	Health	0.1997	0.9933	0.3325	0.1713
	Entmt	0.1594	1.0000	0.2750	0.1419
	Energy	0.2107	0.9936	0.3477	0.1781
	Safety	0.3008	0.9933	0.4618	0.2381
	Other	0.1299	1.0000	0.2300	0.1192
Bert4RE	Health	0.2202	0.4924	0.3043	0.4858
	Entmt	0.1565	0.8684	0.2652	0.2344
	Energy	0.2129	0.9409	0.3472	0.2374
	Safety	0.3000	0.9899	0.4605	0.2369
	Other	0.1124	0.2526	0.1556	0.4654

Table 4.12: One-vs-Rest (OvR) Results for Word-Embedding-20 Labels (Word2Vec)

For example, its F1-score for the *Energy* class improves from 0.0000 (Original labels, Table 4.10) to 0.3594 (Combined labels, Table 4.16). However, it still struggles with the *Other* class, achieving low F1-scores across all label sets. This could be due to newly initialized weights in the Bert4RE model which reduces its usefulness.

- **Observations:**
  - The *Safety* class is the most consistently well-performing class across all models and label sets, likely due to its distinct and well-defined features (it is also the class with most number of requirements).
  - The *Other* class remains the most challenging even in this case, with low precision and F1-scores.
  - Refined label sets (expert-curated and word-embedding) generally improve model performance, proving the importance of high-quality labeling in achieving better classification results.
  - AllMini and Sbert models show more balanced performance, while SObert and Bert4RE are prone to overfitting, particularly with word-embedding labels. This again leads us towards generalized models.

Model	Class	Precision	Recall	F1-Score	Macro F1
Sbert	Health	0.2111	0.7673	0.3311	0.3769
	Entmt	0.1950	0.8004	0.3136	0.4230
	Energy	0.2090	0.6741	0.3191	0.3856
	Safety	0.3075	0.7612	0.4381	0.4115
	Other	0.1414	0.5651	0.2262	0.4281
AllMini	Health	0.1859	0.1332	0.1552	0.4901
	Entmt	0.1404	0.2590	0.1821	0.4717
	Energy	**0.2806**	**0.1134**	**0.1615**	**0.5078**
	Safety	0.2221	0.2074	0.2145	0.4462
	Other	0.1365	0.6979	0.2283	0.3613
SObert	Health	0.2114	0.6239	0.3158	0.4346
	Entmt	0.1620	0.7771	0.2681	0.3221
	Energy	0.1884	0.4153	0.2592	0.4404
	Safety	0.2981	0.5975	0.3978	0.4507
	Other	0.1195	0.4557	0.1893	0.4111
Bert4RE	Health	0.2164	0.6459	0.3242	0.4383
	Entmt	0.1250	0.2760	0.1721	0.4446
	Energy	0.2100	0.6438	0.3167	0.4016
	Safety	0.2864	0.2668	0.2763	0.4900
	Other	0.1291	0.7917	0.2221	0.2776

Table 4.13: One-vs-Rest (OvR) Results for Word-Embedding-20 Labels (with GloVe)

4.4 Multi-class Results for 3-classes

The multiclass (3 classes) results (Tables 4.18, 4.19, 4.20, 4.22, 4.24, 4.21, 4.15, 4.25) across the eight label sets are analyzed here about the performance of the four models. The following is a detailed analysis:

- **Overall Trends:**
  - **Performance Across Models:** Sbert and AllMini consistently outperform SObert and Bert4RE across all label sets. Sbert achieves the highest F1-scores for most class combinations, particularly for *Safety* and *Energy* classes. For example, in the Expert-curated labels (Table 4.19), Sbert achieves a Macro F1-score of 0.6995 for the combination of *Health*, *Entmt*, and *Energy*, while SObert struggles with F1-scores as low as 0.0000 for some classes.
  - **Impact of Label Configuration:** Label Configurations (Expert-curated, Word-Embedding-20, and Word-Embedding-50) generally improve model performance compared to the original labels. For instance, Sbert's Macro F1-score for the *Health*, *Entmt*, and *Energy* combination improves from 0.5294 (Original labels, Table 4.18) to 0.6995 (Expert-curated labels, Table 4.19).
  - **Class-wise Performance:** The *Safety* and *Energy* classes here consistently achieve higher F1-scores across models and label sets, indicating that these

Model	Class	Precision	Recall	F1-Score	Macro F1
Sbert	Health	0.2019	0.7251	0.3158	0.3676
	Entmt	0.2037	0.7771	0.3228	0.4518
	Energy	0.1993	0.5863	0.2975	0.3987
	Safety	0.3157	0.7679	0.4474	0.4289
	Other	0.1360	0.5729	0.2198	0.4112
AllMini	Health	0.2077	0.6981	0.3202	0.3974
	Entmt	0.1881	0.5372	0.2786	0.4802
	Energy	0.2420	0.7220	0.3625	0.4500
	Safety	0.3512	0.8206	0.4919	0.4902
	Other	0.1409	0.5156	0.2214	0.4426
SObert	Health	0.1999	1.0000	0.3332	0.1666
	Entmt	0.1588	1.0000	0.2741	0.1370
	Energy	0.2111	1.0000	0.3486	0.1743
	Safety	0.3007	1.0000	0.4624	0.2312
	Other	0.1295	1.0000	0.2293	0.1146
Bert4RE	Health	0.2033	0.6206	0.3063	0.4170
	Entmt	0.1435	0.5775	0.2298	0.3592
	Energy	0.2284	0.5623	0.3249	0.4681
	Safety	0.3009	0.7948	0.4366	0.3774
	Other	0.0000	0.0000	0.0000	0.4654

Table 4.14: One-vs-Rest (OvR) Results for Word-Embedding-50 Labels (Word2Vec)

classes have more distinct and well-defined features. In contrast, the *Other* class consistently underperforms again, with F1-scores often below 0.3. This suggests that the class too broad and not well defined.

- **Model-specific:**
  - **Sbert:** Sbert demonstrates strong performance across all label sets, particularly for combinations involving *Safety* and *Energy*. For example, in the Word-Embedding-20 labels (Table 4.20), Sbert achieves an F1-score of 0.7609 for the *Safety* class in the *Entertainment*, *Safety*, and *Energy* combination. However, its performance on the *Other* class remains weak, with F1-scores often below 0.3.
  - **AllMini:** AllMini shows competitive performance, particularly for the *Safety* and *Energy* classes. For instance, in the Combined labels (Table 4.24), AllMini achieves a Macro F1-score of 0.7712 for the *Entertainment*, *Safety*, and *Energy* combination. However, like Sbert, it struggles with the *Other* class, indicating a common challenge across models.
  - **SObert:** SObert again performs very poorly on the original label set, with F1-scores of 0.0000 for many classes (Table 4.18). Its performance improves with different label configurations but remains inconsistent. Its performance in most of the cases is often subpar.

Model	Class	Precision	Recall	F1-Score	Macro F1
Sbert	Health	0.2074	0.7470	0.3247	0.3746
	Entmt	0.1863	0.8259	0.3040	0.3880
	Energy	0.1924	0.5927	0.2905	0.3771
	Safety	0.3009	0.8296	0.4417	0.3584
	Other	0.1449	0.5703	0.2311	0.4351
AllMini	Health	0.2099	0.3727	0.2685	0.4938
	Entmt	0.1496	0.5393	0.2342	0.3964
	Energy	**0.2363**	**0.4058**	**0.2986**	**0.5083**
	Safety	0.2976	0.4451	0.3567	0.4851
	Other	0.1323	0.5208	0.2110	0.4202
SObert	Health	0.1995	0.9933	0.3323	0.1703
	Entmt	0.1584	0.9915	0.2731	0.1417
	Energy	0.2054	0.8626	0.3318	0.2597
	Safety	0.3021	0.9787	0.4617	0.2573
	Other	0.1279	0.9688	0.2260	0.1304
Bert4RE	Health	0.2031	0.9848	0.3367	0.2012
	Entmt	0.1362	0.4947	0.2136	0.3780
	Energy	0.2200	0.5767	0.3185	0.4485
	Safety	0.2999	0.9226	0.4527	0.2930
	Other	0.1293	0.9922	0.2288	0.1205

Table 4.15: One-vs-Rest (OvR) Results for Word-Embedding-50 Labels (with GloVe)

- **Bert4RE:** Bert4RE shows the weakest performance among the models, particularly for the *Other* class, where F1-scores are consistently low. For example, in the Expert-curated labels (Table 4.19), Bert4RE achieves an F1-score of 0.0000 for the *Energy* class in the *Health*, *Entertainment*, and *Energy* combination. Its performance improves slightly with refined label configurations but remains inferior to Sbert and AllMini. This is again due to some of the weights newly initialized randomly.

- **Observations:**
  - The *Safety* and *Energy* classes are the most consistently well-performing classes across all models and label sets, likely due to their distinct textual features.
  - The *Other* class still remains the most challenging, with low F1-scores across all models, suggesting that it is too heterogeneous.
  - Label configurations (Expert-curated, Word-Embedding-20, and Word-Embedding-50) significantly improve model performance, highlighting the importance of high-quality labeling in multiclass classification.
  - Sbert and AllMini demonstrate the most balanced performance, while SObert and Bert4RE struggle with consistency and overfitting, particularly in the original label set.

Model	Class	Precision	Recall	F1-Score	Macro F1
Sbert	Health	0.2091	0.8196	0.3332	0.3441
	Entmt	0.1869	0.8662	0.3075	0.3734
	Energy	0.2072	0.7013	0.3199	0.3670
	Safety	0.3071	0.8632	0.4531	0.3596
	Other	0.1218	0.3516	0.1810	0.4529
AllMini	Health	0.1963	0.5363	0.2874	0.4317
	Entmt	0.1573	0.5159	0.2411	0.4252
	Energy	0.2769	0.3946	0.3254	0.5467
	Safety	0.3296	0.4910	0.3944	0.5160
	Other	0.1423	0.4661	0.2180	0.4594
SObert	Health	0.2001	1.0000	0.3334	0.1676
	Entmt	0.1598	1.0000	0.2755	0.1449
	Energy	0.2108	0.9984	0.3481	0.1740
	Safety	0.3006	0.9989	0.4621	0.2316
	Other	0.1296	1.0000	0.2295	0.1159
Bert4RE	Health	0.2261	0.3423	0.2723	0.5140
	Entmt	0.1577	0.9406	0.2701	0.1837
	Energy	0.2225	0.9345	0.3594	0.2903
	Safety	0.3003	0.9944	0.4613	0.2340
	Other	0.1262	0.8620	0.2202	0.2092

Table 4.16: One-vs-Rest (OvR) Results for Combined Labels (Word2Vec)

In conclusion, Sbert and AllMini are the most effective models for multiclass classification, particularly when using different configurations of labels. The *Safety* and *Energy* classes are the easiest to classify, while the *Other* class remains a significant challenge.

4.5 Multi-class Results for 4-classes

The multiclass (4 classes) results (Tables 4.26, 4.27, 4.28, 4.30, 4.32, 4.29, 4.31) across the eight label sets (Original, Expert-curated, Word-Embedding-20 for Word2Vec and GloVe labels, Word-Embedding-50 for Word2Vec and GloVe, and Combined) are analyzed here about the performance of the four models. The following is a detailed analysis:

- **Overall Trends:**
  - **Performance Across Models:** As claimed before, Sbert and AllMini consistently outperform SObert and Bert4RE across all label sets. For instance, in the Expert-curated labels (Table 4.27), Sbert achieves a Macro F1-score of 0.6254 for the combination of *Health*, *Entmt*, *Safety*, and *Energy*, while SObert struggles with F1-scores as low as 0.0000 for some classes. This trend is consistent across all label sets, reinforcing the superior performance of Sbert and AllMini.

Model	Class	Precision	Recall	F1-Score	Macro F1
Sbert	Health	0.2066	0.8246	0.3304	0.3318
	Entmt	0.1881	0.8195	0.3060	0.3962
	Energy	0.2034	0.6677	0.3118	0.3722
	Safety	0.3065	0.8206	0.4463	0.3808
	Other	0.1101	0.2839	0.1587	0.4525
AllMini	Health	0.2053	0.2496	0.2253	0.5024
	Entmt	0.1234	0.2633	0.1680	0.4462
	Energy	**0.2576**	**0.5575**	**0.3523**	**0.5138**
	Safety	0.3114	0.3610	0.3344	0.5072
	Other	0.1411	0.6146	0.2296	0.4105
SObert	Health	0.1997	0.9815	0.3319	0.1829
	Entmt	0.1591	0.9724	0.2734	0.1654
	Energy	0.2111	0.9808	0.3474	0.1925
	Safety	0.3040	0.9439	0.4599	0.2942
	Other	0.1282	0.9063	0.2246	0.1882
Bert4RE	Health	0.2038	0.7808	0.3232	0.3455
	Entmt	0.1574	0.9703	0.2709	0.1546
	Energy	0.2104	0.9505	0.3445	0.2154
	Safety	0.3008	0.9922	0.4617	0.2389
	Other	0.1320	0.9453	0.2316	0.1851

Table 4.17: One-vs-Rest (OvR) Results for Combined Labels (with GloVe)

  - **Impact of Label Configuration:** Label configurations continue to improve model performance, as seen in previous analyses. For example, Sbert's Macro F1-score for the *Health*, *Entmt*, *Safety*, and *Energy* combination improves from 0.4745 (Original labels, Table 4.26) to 0.6254 (Expert-curated labels, Table 4.27).
  - **Class-wise Performance:** The *Safety* and *Energy* classes consistently achieve higher F1-scores across models and label sets, as observed earlier. For example, in the Word-Embedding-20 labels (Table 4.28), Sbert achieves an F1-score of 0.7074 for the *Safety* class in the *Health*, *Entmt*, *Safety*, and *Other* combination. Conversely, the *Other* class continues to underperform, with F1-scores often below 0.3, indicating its persistent challenge across models.

- **Model-specific Insights:**
  - **Sbert:** Sbert maintains its strong performance across all label sets, particularly for combinations involving *Safety* and *Energy*. For example, in the Word-Embedding-20 labels (Table 4.28), Sbert achieves an F1-score of 0.7074 for the *Safety* class. However, its performance on the *Other* class remains weak, with F1-scores often below 0.3, consistent with previous observations.
  - **AllMini:** AllMini continues to show competitive performance, particularly for the *Safety* and *Energy* classes. For instance, in the Combined labels

Model	Class1	Class2	Class3	F1-Class1	F1-Class2	F1-Class3	Macro F1
Sbert	Health	Entmt	Energy	0.5947	0.6654	0.3282	0.5294
	Health	Entmt	Other	0.6712	0.6526	0.1184	0.4807
	Health	Safety	Energy	0.4344	0.6700	0.3088	0.4711
	Entmt	Safety	Other	0.6617	0.7198	0.0928	0.4914
	Health	Safety	Other	0.4901	0.6591	0.0783	0.4092
	Entmt	Energy	Other	0.5648	0.5622	0.2214	0.4495
	Safety	Energy	Other	0.6415	0.3948	0.0920	0.3761
	Entmt	Safety	Energy	0.6561	0.7005	0.3760	0.5775
	Health	Energy	Other	0.5603	0.3992	0.1778	0.3791
	Health	Entmt	Safety	0.5200	0.6638	0.6968	0.6269
AllMini	Health	Entmt	Energy	0.5300	0.6845	0.6934	0.6359
	Health	Entmt	Other	0.6381	0.6452	0.3484	0.5439
	Health	Safety	Energy	0.4893	0.7246	0.6375	0.6171
	Entmt	Safety	Other	0.6581	0.7874	0.2822	0.5759
	Health	Safety	Other	0.6029	0.7639	0.3561	0.5743
	Entmt	Energy	Other	0.6580	0.7259	0.2159	0.5333
	Safety	Energy	Other	0.7311	0.6541	0.1950	0.5267
	Entmt	Safety	Energy	0.6739	0.7368	0.7060	0.7056
	Health	Energy	Other	0.5262	0.6444	0.2177	0.4628
	Health	**Entmt**	**Safety**	**0.6198**	**0.6848**	**0.7906**	**0.6984**
SObert	Health	Entmt	Energy	0.0881	0.0042	0.5463	0.2128
	Health	Entmt	Other	0.5316	0.2840	0.0153	0.2770
	Health	Safety	Energy	0.0876	0.0134	0.4626	0.1879
	Entmt	Safety	Other	0.4190	0.0810	0.0733	0.1911
	Health	Safety	Other	0.4807	0.0473	0.0284	0.1854
	Entmt	Energy	Other	0.0042	0.5944	0.0000	0.1995
	Safety	Energy	Other	0.0134	0.4964	0.0000	0.1699
	Entmt	Safety	Energy	0.0042	0.0133	0.4800	0.1659
	Health	Energy	Other	0.0886	0.5671	0.0000	0.2186
	Health	Entmt	Safety	0.4380	0.2400	0.0285	0.2355
Bert4RE	Health	Entmt	Energy	0.1675	0.4284	0.0000	0.1986
	Health	Entmt	Other	0.0909	0.4485	0.2142	0.2512
	Health	Safety	Energy	0.3061	0.4549	0.1229	0.2946
	Entmt	Safety	Other	0.3906	0.2188	0.1140	0.2411
	Health	Safety	Other	0.3303	0.3714	0.1880	0.2965
	Entmt	Energy	Other	0.4431	0.0063	0.2213	0.2236
	Safety	Energy	Other	0.4578	0.2069	0.2197	0.2948
	Entmt	Safety	Energy	0.3468	0.3043	0.0126	0.2212
	Health	Energy	Other	0.3797	0.1026	0.3554	0.2793
	Health	Entmt	Safety	0.0926	0.3536	0.2837	0.2433

Table 4.18: Multiclass (3 Classes) Results for Original Labels

(Table 4.32), AllMini achieves a Macro F1-score of 0.6004 for the *Entmt*, *Safety*, *Energy*, and *Other* combination. However, like Sbert, it struggles with the *Other* class, indicating a common challenge across models.

- **SObert:** SObert's performance remains inconsistent, as noted earlier. While it shows slight improvement with refined label sets, its F1-scores are often subpar. For example, in the Word-Embedding-50 labels (Table 4.30), SObert achieves an F1-score of 0.2410 for the *Safety* class in the *Health*, *Entmt*, *Safety*, and *Other* combination, but its performance on other classes is often weak.
- **Bert4RE:** Bert4RE continues to show the weakest performance among the models, particularly for the *Other* class. For example, in the Expert-curated

Model	Class1	Class2	Class3	F1-Class1	F1-Class2	F1-Class3	Macro F1
Sbert	**Health**	**Entmt**	**Energy**	**0.6542**	**0.7469**	**0.6974**	**0.6995**
	Health	Entmt	Other	0.7026	0.7284	0.1684	0.5331
	Health	Safety	Energy	0.5416	0.7715	0.6629	0.6587
	Entmt	Safety	Other	0.6742	0.7571	0.0782	0.5032
	Health	Safety	Other	0.5841	0.7569	0.0542	0.4651
	Entmt	Energy	Other	0.7325	0.7435	0.2227	0.5662
	Safety	Energy	Other	0.7698	0.7240	0.0599	0.5179
	Entmt	**Safety**	**Energy**	**0.6624**	**0.7902**	**0.7562**	**0.7362**
	Health	Energy	Other	0.6100	0.6713	0.2087	0.4967
	Health	**Entmt**	**Safety**	**0.6424**	**0.6760**	**0.7725**	**0.6970**
AllMini	Health	Entmt	Energy	0.4305	0.6474	0.6623	0.5801
	Health	Entmt	Other	0.5071	0.7117	0.4590	0.5593
	Health	Safety	Energy	0.3124	0.7752	0.7305	0.6060
	Entmt	Safety	Other	0.6139	0.7674	0.3262	0.5692
	Health	Safety	Other	0.3381	0.7271	0.2901	0.4518
	Entmt	Energy	Other	0.6287	0.7150	0.4038	0.5825
	Safety	Energy	Other	0.8214	0.7555	0.2545	0.6105
	Entmt	**Safety**	**Energy**	**0.5552**	**0.8288**	**0.7713**	**0.7184**
	Health	Energy	Other	0.3560	0.6616	0.3626	0.4601
	Health	Entmt	Safety	0.3807	0.6017	0.7260	0.5695
SObert	Health	Entmt	Energy	0.1463	0.0000	0.5371	0.2278
	Health	Entmt	Other	0.3026	0.0846	0.3653	0.2508
	Health	Safety	Energy	0.1417	0.5012	0.4134	0.3521
	Entmt	Safety	Other	0.1276	0.4889	0.2807	0.2991
	Health	Safety	Other	0.2051	0.4501	0.2699	0.3083
	Entmt	Energy	Other	0.0000	0.4479	0.3409	0.2629
	Safety	Energy	Other	0.2613	0.3704	0.2628	0.2982
	Entmt	Safety	Energy	0.0000	0.5428	0.4149	0.3192
	Health	Energy	Other	0.1494	0.4282	0.3090	0.2956
	Health	Entmt	Safety	0.2032	0.0000	0.6019	0.2684
Bert4RE	Health	Entmt	Energy	0.0232	0.4756	0.4137	0.3041
	Health	Entmt	Other	0.0938	0.1604	0.4201	0.2247
	Health	Safety	Energy	0.0260	0.5807	0.1540	0.2536
	Entmt	Safety	Other	0.1153	0.2345	0.3296	0.2265
	Health	Safety	Other	0.0932	0.2617	0.2997	0.2182
	Entmt	Energy	Other	0.1350	0.1558	0.4130	0.2346
	Safety	Energy	Other	0.2403	0.1490	0.3013	0.2302
	Entmt	Safety	Energy	0.0949	0.5966	0.1628	0.2848
	Health	Energy	Other	0.0296	0.1634	0.3948	0.1959
	Health	Entmt	Safety	0.0941	0.1083	0.6180	0.2735

Table 4.19: Multiclass (3 Classes) Results for Expert-curated Labels

labels (Table 4.27), Bert4RE achieves an F1-score of 0.0000 for the *Energy* class in the *Health*, *Entmt*, *Safety*, and *Energy* combination. Its performance improves slightly with refined label sets but remains inferior to Sbert and AllMini.

- **Observations:**
  - The *Safety* and *Energy* classes remain the most consistently well-performing classes across all models and label sets, as previously observed. This is likely due to their distinct textual features.
  - The *Other* class continues to be the most challenging, with low F1-scores across

Model	Class1	Class2	Class3	F1-Class1	F1-Class2	F1-Class3	Macro F1
Sbert	Health	Entmt	Energy	0.6072	0.7424	0.7116	0.6871
	Health	Entmt	Other	0.6629	0.6738	0.4114	0.5827
	Health	Safety	Energy	0.4933	0.7679	0.6884	0.6498
	Entmt	Safety	Other	0.6398	0.7774	0.3748	0.5973
	Health	Safety	Other	0.5685	0.7589	0.2889	0.5388
	Entmt	Energy	Other	0.6659	0.7638	0.4701	0.6332
	Safety	Energy	Other	0.7801	0.7220	0.2851	0.5957
	Entmt	Safety	Energy	0.6424	0.7853	0.7609	0.7295
	Health	Energy	Other	0.5411	0.7044	0.4731	0.5729
	Health	Entmt	Safety	0.6203	0.6470	0.7579	0.6750
AllMini	Health	Entmt	Energy	0.5464	0.7066	0.7117	0.6549
	Health	Entmt	Other	0.6136	0.6944	0.5828	0.6303
	Health	Safety	Energy	0.4610	0.7746	0.6957	0.6437
	Entmt	Safety	Other	0.7119	0.8068	0.3643	0.6277
	Health	Safety	Other	0.5309	0.7637	0.3339	0.5428
	Entmt	Energy	Other	0.6958	0.7289	0.2321	0.5523
	Safety	Energy	Other	0.7946	0.7255	0.2194	0.5798
	Entmt	Safety	Energy	0.7055	0.8189	0.7663	0.7636
	Health	Energy	Other	0.5010	0.6765	0.2383	0.4719
	Health	Entmt	Safety	0.5720	0.6944	0.7727	0.6797
SObert	Health	Entmt	Energy	0.0357	0.4480	0.2908	0.2582
	Health	Entmt	Other	0.0481	0.4849	0.0000	0.1777
	Health	Safety	Energy	0.1234	0.5467	0.4302	0.3668
	Entmt	Safety	Other	0.4128	0.3089	0.0000	0.2406
	Health	Safety	Other	0.1676	0.6303	0.0652	0.2877
	Entmt	Energy	Other	0.4994	0.3040	0.0000	0.2678
	Safety	Energy	Other	0.5931	0.4425	0.0051	0.3469
	Entmt	Safety	Energy	0.3872	0.2871	0.2682	0.3142
	Health	Energy	Other	0.2347	0.5762	0.0767	0.2959
	Health	Entmt	Safety	0.0386	0.3811	0.2873	0.2357
Bert4RE	Health	Entmt	Energy	0.0067	0.4364	0.0422	0.1618
	Health	Entmt	Other	0.0232	0.4935	0.0000	0.1722
	Health	Safety	Energy	0.0920	0.5024	0.4134	0.3359
	Entmt	Safety	Other	0.4253	0.0067	0.0000	0.1440
	Health	Safety	Other	0.3094	0.5881	0.0652	0.3209
	Entmt	Energy	Other	0.4808	0.0429	0.0000	0.1746
	Safety	Energy	Other	0.5055	0.4527	0.0273	0.3285
	Entmt	Safety	Energy	0.3823	0.0067	0.0394	0.1428
	Health	Energy	Other	0.0587	0.5366	0.2390	0.2781
	Health	Entmt	Safety	0.0263	0.3904	0.0067	0.1411

Table 4.20: Multiclass (3 Classes) Results for Word-Embedding-20 Labels (Word2Vec)

all models, reinforcing the need for further refinement in its classification.

- Refined label configurations significantly improve model performance, as seen in earlier analyses, highlighting the importance of high-quality labeling in multiclass classification.
- Sbert and AllMini demonstrate the most balanced performance, while SObert and Bert4RE struggle with consistency and overfitting, particularly in the original label set.

In conclusion, Sbert and AllMini remain the most effective models for multiclass classification, particularly when using refined label sets. The *Safety* and *Energy* classes

Model	Class1	Class2	Class3	F1-Class1	F1-Class2	F1-Class3	Macro F1
Sbert	**Health**	**Entmt**	**Energy**	**0.5923**	**0.7404**	**0.7020**	**0.6783**
	Health	Entmt	Other	0.6495	0.6593	0.1165	0.4751
	Health	Safety	Energy	0.1574	0.6086	0.4554	0.4071
	Entmt	Safety	Other	0.5717	0.6905	0.0337	0.4319
	Health	Safety	Other	0.1646	0.6415	0.0562	0.2874
	Entmt	Energy	Other	0.6812	0.7801	0.2150	0.5588
	Safety	Energy	Other	0.6546	0.4489	0.0648	0.3894
	Entmt	Safety	Energy	0.6017	0.6379	0.4476	0.5624
	Health	Energy	Other	0.5712	0.6780	0.1785	0.4759
	Health	Entmt	Safety	0.1433	0.6045	0.6271	0.4583
AllMini	Health	Entmt	Energy	0.3855	0.6753	0.5899	0.5503
	Health	Entmt	Other	0.3594	0.5385	0.3083	0.4021
	Health	Safety	Energy	0.1940	0.6852	0.5899	0.4897
	Entmt	Safety	Other	0.5545	0.6891	0.3194	0.5210
	Health	Safety	Other	0.2591	0.6435	0.2977	0.4001
	Entmt	Energy	Other	0.5549	0.6557	0.4076	0.5394
	Safety	Energy	Other	0.6667	0.5341	0.3219	0.5076
	Entmt	Safety	Energy	0.6889	0.7638	0.5974	0.6834
	Health	Energy	Other	0.2591	0.6160	0.3847	0.4199
	Health	Entmt	Safety	0.1997	0.6837	0.7094	0.5309
SObert	Health	Entmt	Energy	0.3831	0.3558	0.2956	0.3448
	Health	Entmt	Other	0.3168	0.2694	0.0052	0.1971
	Health	Safety	Energy	0.1574	0.5661	0.1952	0.3062
	Entmt	Safety	Other	0.1675	0.6605	0.0052	0.2777
	Health	Safety	Other	0.1372	0.6265	0.0052	0.2563
	Entmt	Energy	Other	0.1694	0.2957	0.3286	0.2646
	Safety	Energy	Other	0.6065	0.4425	0.0000	0.3497
	Entmt	Safety	Energy	0.3872	0.2871	0.2682	0.3142
	Health	Energy	Other	0.2635	0.3330	0.2956	0.2974
	Health	Entmt	Safety	0.1274	0.1239	0.6039	0.2851
Bert4RE	Health	Entmt	Energy	0.0382	0.4361	0.0216	0.1653
	Health	Entmt	Other	0.0327	0.2923	0.0000	0.1083
	Health	Safety	Energy	0.1979	0.5771	0.0216	0.2655
	Entmt	Safety	Other	0.4128	0.0449	0.0000	0.1526
	Health	Safety	Other	0.0904	0.4635	0.3343	0.2961
	Entmt	Energy	Other	0.1694	0.0127	0.4117	0.1979
	Safety	Energy	Other	0.0721	0.0629	0.3343	0.1564
	Entmt	Safety	Energy	0.2884	0.0067	0.0126	0.1026
	Health	Energy	Other	0.1461	0.0630	0.3903	0.1998
	Health	Entmt	Safety	0.0512	0.2185	0.4603	0.2433

Table 4.21: Multiclass (3 Classes) Results for Word-Embedding-20 Labels (with GloVe)

are the easiest to classify, while the *Other* class continues to pose significant challenges. Future work should focus on improving the classification of the *Other* class and exploring additional refinements to labeling strategies.

4.6 Multi-class Results for 5-classes

The multiclass (5 classes) results across the eight label sets reinforce the trends observed earlier. Sbert and AllMini consistently outperform SObert and Bert4RE, with Sbert achieving the highest Macro F1-scores across all label sets. For example, in the

Model	Class1	Class2	Class3	F1-Class1	F1-Class2	F1-Class3	Macro F1
Sbert	Health	Entmt	Energy	0.4771	0.7326	0.6901	0.6332
	Health	Entmt	Other	0.6325	0.7018	0.4114	0.5819
	Health	Safety	Energy	0.3758	0.7625	0.6735	0.6039
	Entmt	Safety	Other	0.6398	0.7774	0.3748	0.5973
	Health	Safety	Other	0.5255	0.7393	0.2213	0.4954
	Entmt	Energy	Other	0.7137	0.7486	0.4701	0.6441
	Safety	Energy	Other	0.7760	0.7139	0.2851	0.5917
	Entmt	Safety	Energy	0.6757	0.7953	0.7567	0.7426
	Health	Energy	Other	0.4646	0.6723	0.4731	0.5367
	Health	Entmt	Safety	0.5657	0.6794	0.7523	0.6658
AllMini	Health	Entmt	Energy	0.3904	0.7237	0.6775	0.5972
	Health	Entmt	Other	0.4756	0.6735	0.5828	0.5773
	Health	Safety	Energy	0.3487	0.7559	0.6470	0.5839
	Entmt	Safety	Other	0.7224	0.8188	0.4483	0.6632
	Health	Safety	Other	0.4071	0.7471	0.3339	0.4960
	Entmt	Energy	Other	0.7303	0.7232	0.2321	0.5619
	Safety	Energy	Other	0.7595	0.6878	0.2194	0.5556
	Entmt	Safety	Energy	0.7247	0.7717	0.7410	0.7458
	Health	Energy	Other	0.3617	0.6311	0.2383	0.4104
	Health	Entmt	Safety	0.4354	0.6967	0.7681	0.6334
SObert	Health	Entmt	Energy	0.0000	0.4385	0.0407	0.1597
	Health	Entmt	Other	0.0000	0.4906	0.0000	0.1635
	Health	Safety	Energy	0.0000	0.6019	0.2111	0.2707
	Entmt	Safety	Other	0.4252	0.1031	0.0000	0.1761
	Health	Safety	Other	0.0000	0.6305	0.0633	0.2313
	Entmt	Energy	Other	0.4821	0.0032	0.0000	0.1618
	Safety	Energy	Other	0.6246	0.1948	0.0863	0.3019
	Entmt	Safety	Energy	0.3849	0.1014	0.0064	0.1642
	Health	Energy	Other	0.0000	0.5450	0.2796	0.2749
	Health	Entmt	Safety	0.0000	0.3949	0.2487	0.2145
Bert4RE	Health	Entmt	Energy	0.0000	0.0000	0.5406	0.1802
	Health	Entmt	Other	0.0000	0.0000	0.4192	0.1397
	Health	Safety	Energy	0.0000	0.5839	0.3342	0.3060
	Entmt	Safety	Other	0.0000	0.0800	0.3539	0.1446
	Health	Safety	Other	0.0000	0.0486	0.3356	0.1281
	Entmt	Energy	Other	0.0000	0.0251	0.4132	0.1461
	Safety	Energy	Other	0.0111	0.0308	0.3375	0.1264
	Entmt	Safety	Energy	0.0000	0.6091	0.2225	0.2772
	Health	Energy	Other	0.0000	0.0250	0.3882	0.1377
	Health	Entmt	Safety	0.0000	0.0000	0.6264	0.2088

Table 4.22: Multiclass (3 Classes) Results for Word-Embedding-50 Labels (Word2Vec)

Expert-curated labels (Table 4.35), Sbert achieves a Macro F1-score of 0.4775, while SObert and Bert4RE lag significantly behind. Refined label configuration, such as Expert-curated and Word-Embedding-20, continue to improve performance, as seen in Sbert's improvement from 0.3568 (Original labels) to 0.4775 (Expert-curated labels). The *Safety* and *Energy* classes consistently achieve higher F1-scores, while the *Other* class remains the most challenging, with F1-scores often below 0.3. This aligns with earlier findings, highlighting the distinctiveness of *Safety* and *Energy* and the heterogeneity of the *Other* class. In conclusion, Sbert and AllMini remain the most effective models, but further refinement is needed for the *Other* class to improve overall performance.

Model	Class1	Class2	Class3	F1-Class1	F1-Class2	F1-Class3	Macro F1
Sbert	Health	Entmt	Energy	0.6224	0.7380	0.6695	0.6766
	Health	Entmt	Other	0.6681	0.6798	0.1955	0.5145
	Health	Safety	Energy	0.2585	0.6369	0.4919	0.4624
	Entmt	Safety	Other	0.6058	0.6969	0.0578	0.4535
	Health	Safety	Other	0.2634	0.6584	0.0844	0.3354
	Entmt	Energy	Other	0.6816	0.7731	0.1955	0.5501
	Safety	Energy	Other	0.6675	0.4914	0.0925	0.4171
	Entmt	Safety	Energy	0.6291	0.6495	0.4895	0.5894
	Health	Energy	Other	0.5826	0.6494	0.1485	0.4602
	Health	Entmt	Safety	0.2497	0.6321	0.6417	0.5078
AllMini	Health	Entmt	Energy	0.5407	0.6552	0.7036	0.6331
	Health	Entmt	Other	0.5621	0.5939	0.2491	0.4684
	Health	Safety	Energy	0.3501	0.7030	0.5975	0.5502
	Entmt	Safety	Other	0.7119	0.7559	0.1643	0.5440
	Health	Safety	Other	0.3738	0.6994	0.1677	0.4136
	Entmt	Energy	Other	0.6412	0.7254	0.2491	0.5386
	Safety	Energy	Other	0.7441	0.6154	0.3219	0.5605
	Entmt	**Safety**	**Energy**	**0.6727**	**0.7463**	**0.6122**	**0.6771**
	Health	Energy	Other	0.5472	0.6615	0.2862	0.4983
	Health	Entmt	Safety	0.3219	0.6429	0.7054	0.5567
SObert	Health	Entmt	Energy	0.0099	0.3861	0.4994	0.2985
	Health	Entmt	Other	0.0034	0.0000	0.4190	0.1408
	Health	Safety	Energy	0.0033	0.5933	0.0000	0.1989
	Entmt	Safety	Other	0.0000	0.6449	0.0000	0.2150
	Health	Safety	Other	0.0034	0.6259	0.0000	0.2098
	Entmt	Energy	Other	0.0463	0.0155	0.3969	0.1529
	Safety	Energy	Other	0.6375	0.0000	0.0000	0.2125
	Entmt	Safety	Energy	0.0000	0.6184	0.0032	0.2072
	Health	Energy	Other	0.0515	0.0420	0.3835	0.1590
	Health	Entmt	Safety	0.0000	0.0000	0.6252	0.2084
Bert4RE	Health	Entmt	Other	0.0392	0.0124	0.4182	0.1566
	Safety	Energy	Other	0.0623	0.0000	0.3359	0.1327
	Entmt	Safety	Other	0.0000	0.0981	0.3583	0.1521
	Health	Entmt	Safety	0.0000	0.0000	0.6259	0.2086
	Health	Entmt	Energy	0.0067	0.4356	0.0000	0.1474
	Health	Energy	Other	0.0356	0.0032	0.3845	0.1411
	Entmt	Safety	Energy	0.0000	0.6184	0.0000	0.2061
	Health	Safety	Energy	0.0000	0.5941	0.0000	0.1980
	Health	Safety	Other	0.0000	0.0910	0.3441	0.1447
	Entmt	Energy	Other	0.0162	0.0000	0.4083	0.1415

Table 4.23: Multiclass (3 Classes) Results for Word-Embedding-50 Labels (with GloVe)

Model	Class1	Class2	Class3	F1-Class1	F1-Class2	F1-Class3	Macro F1
Sbert	Health	Entmt	Safety	0.5988	0.6789	0.7608	0.6795
	Health	Entmt	Energy	0.6068	0.7492	0.7076	0.6879
	Health	Entmt	Other	0.6629	0.7020	0.3767	0.5805
	Health	Safety	Energy	0.4842	0.7610	0.6713	0.6388
	Health	Safety	Other	0.5500	0.7492	0.2062	0.5018
	Health	Energy	Other	0.5514	0.6959	0.4350	0.5608
	Entmt	**Safety**	**Energy**	**0.6700**	**0.7899**	**0.7474**	**0.7358**
	Entmt	Safety	Other	0.6725	0.7711	0.2525	0.5653
	Entmt	Energy	Other	0.7116	0.7634	0.4317	0.6356
	Safety	Energy	Other	0.7720	0.7095	0.1907	0.5574
AllMini	Health	Entmt	Safety	0.4163	0.6970	0.7694	0.6276
	Health	Entmt	Energy	0.4209	0.7017	0.6886	0.6037
	Health	Entmt	Other	0.4962	0.6865	0.4380	0.5402
	Health	Safety	Energy	0.3686	0.7903	0.6887	0.6159
	Health	Safety	Other	0.3932	0.7424	0.3690	0.5015
	Health	Energy	Other	0.3820	0.6545	0.2917	0.4427
	Entmt	**Safety**	**Energy**	**0.7039**	**0.8334**	**0.7762**	**0.7712**
	Entmt	Safety	Other	0.7197	0.8091	0.4283	0.6524
	Entmt	Energy	Other	0.7071	0.7413	0.3721	0.6068
	Safety	Energy	Other	0.8072	0.7349	0.3031	0.6151
SObert	Health	Entmt	Safety	0.0000	0.0000	0.6259	0.2086
	Health	Entmt	Energy	0.0000	0.4008	0.5347	0.3118
	Health	Entmt	Other	0.0000	0.0000	0.4194	0.1398
	Health	Safety	Energy	0.0000	0.5946	0.0127	0.2024
	Health	Safety	Other	0.0000	0.0305	0.3413	0.1239
	Health	Energy	Other	0.0000	0.0064	0.3873	0.1312
	Entmt	Safety	Energy	0.0000	0.6196	0.0127	0.2108
	Entmt	Safety	Other	0.0000	0.0177	0.3621	0.1266
	Entmt	Energy	Other	0.0000	0.0095	0.4125	0.1407
	Safety	Energy	Other	0.0155	0.0032	0.3383	0.1190
Bert4RE	Health	Entmt	Safety	0.0000	0.3933	0.1236	0.1723
	Health	Entmt	Energy	0.0000	0.4364	0.0421	0.1595
	Health	Entmt	Other	0.0000	0.5000	0.1147	0.2049
	Health	Safety	Energy	0.0000	0.5914	0.0062	0.1992
	Health	Safety	Other	0.0000	0.3777	0.2741	0.2173
	Health	Energy	Other	0.0000	0.0063	0.3864	0.1309
	Entmt	Safety	Energy	0.3838	0.1201	0.0730	0.1923
	Entmt	Safety	Other	0.4425	0.0577	0.2073	0.2358
	Entmt	Energy	Other	0.5067	0.0250	0.2235	0.2517
	Safety	Energy	Other	0.1689	0.0249	0.3146	0.1695

Table 4.24: Multiclass (3 Classes) Results for Combined Labels (Word2Vec)

Model	Class1	Class2	Class3	F1-Class1	F1-Class2	F1-Class3	Macro F1
Sbert	Health	Energy	Other	0.6077	0.6883	0.3623	0.5528
	Entmt	Safety	Other	0.5711	0.7274	0.0802	0.4595
	Health	Entmt	Other	0.6852	0.6799	0.2535	0.5395
	Entmt	Safety	Energy	0.5767	0.7070	0.5820	0.6219
	Health	Entmt	Energy	0.6486	0.7366	0.7064	0.6972
	Health	Entmt	Safety	0.3392	0.5940	0.6922	0.5418
	Entmt	Energy	Other	0.7010	0.7682	0.3655	0.6115
	Safety	Energy	Other	0.7043	0.5779	0.1043	0.4621
	Health	Safety	Energy	0.3258	0.6775	0.5766	0.5266
	Health	Safety	Other	0.3368	0.6863	0.0981	0.3737
AllMini	Health	Energy	Other	0.3383	0.6486	0.3654	0.4508
	Entmt	Safety	Other	0.5907	0.7668	0.3186	0.5587
	Health	Entmt	Other	0.4656	0.6411	0.4514	0.5193
	Entmt	**Safety**	**Energy**	**0.6162**	**0.8174**	**0.7535**	**0.7291**
	Health	Entmt	Energy	0.3810	0.6626	0.6622	0.5686
	Health	Entmt	Safety	0.2760	0.6492	0.7168	0.5473
	Entmt	Energy	Other	0.6174	0.7234	0.3900	0.5769
	Safety	Energy	Other	0.7972	0.7346	0.2868	0.6062
	Health	Safety	Energy	0.2315	0.7551	0.7124	0.5663
	Health	Safety	Other	0.2493	0.7026	0.3105	0.4208
SObert	Health	Energy	Other	0.0034	0.0706	0.3870	0.1536
	Entmt	Safety	Other	0.0000	0.0881	0.3571	0.1484
	Health	Entmt	Other	0.0166	0.0000	0.4180	0.1449
	Entmt	Safety	Energy	0.0000	0.6162	0.0664	0.2275
	Health	Entmt	Energy	0.0000	0.0000	0.5393	0.1798
	Health	Entmt	Safety	0.0133	0.0000	0.6244	0.2126
	Entmt	Energy	Other	0.0000	0.0740	0.3993	0.1578
	Safety	Energy	Other	0.0833	0.0750	0.3347	0.1644
	Health	Safety	Energy	0.0000	0.5928	0.0692	0.2206
	Health	Safety	Other	0.0132	0.1099	0.3358	0.1530
Bert4RE	Health	Energy	Other	0.0000	0.0218	0.3850	0.1356
	Entmt	Safety	Other	0.0042	0.0345	0.3555	0.1314
	Health	Entmt	Other	0.0000	0.0042	0.4165	0.1402
	Entmt	Safety	Energy	0.0126	0.6181	0.0032	0.2113
	Health	Entmt	Energy	0.0000	0.4365	0.0064	0.1476
	Health	Entmt	Safety	0.0000	0.0084	0.6263	0.2116
	Entmt	Energy	Other	0.0279	0.0064	0.4074	0.1472
	Safety	Energy	Other	0.0807	0.0032	0.3365	0.1401
	Health	Safety	Energy	0.0000	0.5931	0.0032	0.1987
	Health	Safety	Other	0.0000	0.0618	0.3369	0.1329

Table 4.25: Multiclass (3 Classes) Results for Combined Labels (with GloVe)

Model	Class1	Class2	Class3	Class4	F1-C1	F1-C2	F1-C3	F1-C4	Macro F1
Sbert	Health	Entmt	Energy	Other	0.5180	0.5811	0.3050	0.1079	0.3780
	Health	Entmt	Safety	Energy	0.4111	0.6068	0.6193	0.2608	0.4745
	Health	Entmt	Safety	Other	0.4626	0.6090	0.6109	0.0515	0.4335
	Health	Safety	Energy	Other	0.3935	0.5832	0.2910	0.0639	0.3329
	Entmt	Safety	Energy	Other	0.6052	0.6025	0.3498	0.0703	0.4070
AllMini	Health	Entmt	Energy	Other	0.4994	0.6133	0.6096	0.1697	0.4730
	Health	**Entmt**	**Safety**	**Energy**	**0.4650**	**0.6345**	**0.6936**	**0.5992**	**0.5981**
	Health	Entmt	Safety	Other	0.5566	0.6005	0.7300	0.2091	0.5240
	Health	Safety	Energy	Other	0.4644	0.6872	0.5621	0.1429	0.4641
	Entmt	Safety	Energy	Other	0.6131	0.6950	0.6205	0.1440	0.5182
SObert	Health	Entmt	Energy	Other	0.0872	0.0042	0.4683	0.0000	0.1399
	Health	Entmt	Safety	Energy	0.0864	0.0042	0.0133	0.3948	0.1247
	Health	Entmt	Safety	Other	0.3869	0.2166	0.0263	0.0149	0.1612
	Health	Safety	Energy	Other	0.0868	0.0134	0.4055	0.0000	0.1264
	Entmt	Safety	Energy	Other	0.0042	0.0133	0.4186	0.0000	0.1090
Bert4RE	Health	Entmt	Energy	Other	0.0839	0.3429	0.0000	0.1818	0.1522
	Health	Entmt	Safety	Energy	0.0824	0.2821	0.2631	0.0000	0.1569
	Health	Entmt	Safety	Other	0.0736	0.3098	0.2011	0.0881	0.1681
	Health	Safety	Energy	Other	0.2589	0.3305	0.0951	0.1512	0.2089
	Entmt	Safety	Energy	Other	0.3057	0.2077	0.0063	0.0973	0.1543

Table 4.26: Multiclass (4 Classes) Results for Original Labels

Model	Class1	Class2	Class3	Class4	F1-C1	F1-C2	F1-C3	F1-C4	Macro F1
Sbert	Health	Entmt	Energy	Other	0.5708	0.6866	0.6362	0.1535	0.5118
	Health	**Entmt**	**Safety**	**Energy**	**0.5186**	**0.6396**	**0.7137**	**0.6296**	**0.6254**
	Health	Entmt	Safety	Other	0.5574	0.6559	0.7057	0.0489	0.4920
	Health	Safety	Energy	Other	0.4894	0.7031	0.6191	0.0493	0.4652
	Entmt	Safety	Energy	Other	0.6415	0.7109	0.6856	0.0543	0.5231
AllMini	Health	Entmt	Energy	Other	0.3502	0.6145	0.6069	0.3229	0.4736
	Health	Entmt	Safety	Energy	0.3019	0.5369	0.7217	0.6728	0.5583
	Health	Entmt	Safety	Other	0.3229	0.5787	0.6723	0.2621	0.4595
	Health	Safety	Energy	Other	0.2789	0.7263	0.6636	0.2171	0.4715
	Entmt	Safety	Energy	Other	0.5359	0.7649	0.6963	0.2296	0.5567
SObert	Health	Entmt	Energy	Other	0.1401	0.0000	0.4683	0.0000	0.1521
	Health	Entmt	Safety	Energy	0.1333	0.0000	0.4442	0.3713	0.2372
	Health	Entmt	Safety	Other	0.1890	0.0000	0.4065	0.2285	0.2060
	Health	Safety	Energy	Other	0.1358	0.2529	0.3257	0.2203	0.2337
	Entmt	Safety	Energy	Other	0.0000	0.2490	0.3355	0.2253	0.2025
Bert4RE	Health	Entmt	Energy	Other	0.0231	0.1406	0.1445	0.3157	0.1560
	Health	Entmt	Safety	Energy	0.0228	0.0923	0.5000	0.1411	0.1891
	Health	Entmt	Safety	Other	0.0846	0.1075	0.2170	0.2568	0.1665
	Health	Safety	Energy	Other	0.0259	0.2253	0.1353	0.2387	0.1563
	Entmt	Safety	Energy	Other	0.0936	0.2070	0.1375	0.2560	0.1735

Table 4.27: Multiclass (4 Classes) Results for Expert Curated Labels

Model	Class1	Class2	Class3	Class4	F1-C1	F1-C2	F1-C3	F1-C4	Macro F1
	Health	Entmt	Safety	Other	0.5406	0.6052	0.7074	0.2534	0.5267
	Health	Entmt	Energy	Other	0.5131	0.6364	0.6610	0.3712	0.5454
Sbert	**Health**	**Entmt**	**Safety**	**Energy**	**0.4714**	**0.6229**	**0.7065**	**0.6573**	**0.6145**
	Entmt	Safety	Energy	Other	0.5982	0.7256	0.6843	0.2491	0.5643
	Health	Safety	Energy	Other	0.4517	0.7075	0.6399	0.2437	0.5107
	Health	Entmt	Safety	Other	0.4883	0.6646	0.7305	0.2889	0.5431
	Entmt	Safety	Energy	Other	0.6621	0.7640	0.6847	0.1814	0.5731
AllMini	Health	Safety	Energy	Other	0.4136	0.7320	0.6282	0.1843	0.4895
	Health	**Entmt**	**Safety**	**Energy**	**0.4319**	**0.6738**	**0.7446**	**0.6531**	**0.6258**
	Health	Entmt	Energy	Other	0.4654	0.6712	0.6368	0.1974	0.4927
	Health	Entmt	Safety	Energy	0.0355	0.3126	0.2603	0.2619	0.2176
	Health	Entmt	Energy	Other	0.0355	0.3776	0.2854	0.0000	0.1746
SObert	Entmt	Safety	Energy	Other	0.3338	0.2735	0.2655	0.0000	0.2182
	Health	Safety	Energy	Other	0.1157	0.4871	0.3994	0.0051	0.2518
	Health	Entmt	Safety	Other	0.0385	0.3293	0.2736	0.0000	0.1603
	Health	Safety	Energy	Other	0.0895	0.4413	0.3791	0.0263	0.2340
	Health	Entmt	Energy	Other	0.0067	0.3701	0.0420	0.0000	0.1047
Bert4RE	Health	Entmt	Safety	Energy	0.0166	0.3087	0.0067	0.0330	0.0912
	Entmt	Safety	Energy	Other	0.3305	0.0067	0.0392	0.0000	0.0941
	Health	Entmt	Safety	Other	0.0261	0.3372	0.0067	0.0000	0.0925

Table 4.28: Multiclass (4 Classes) Results for Word-Embedding-Based-20 Labels (Word2Vec)

Model	Class1	Class2	Class3	Class4	F1-C1	F1-C2	F1-C3	F1-C4	Macro F1
	Health	Safety	Energy	Other	0.1538	0.5341	0.4372	0.0516	0.2942
	Health	Entmt	Energy	Other	0.5110	0.6405	0.6529	0.0917	0.4740
Sbert	Health	Entmt	Safety	Energy	0.1343	0.5854	0.5131	0.4275	0.4151
	Entmt	Safety	Energy	Other	0.5530	0.5619	0.4248	0.0337	0.3933
	Health	Entmt	Safety	Other	0.1370	0.5490	0.5517	0.0241	0.3155
	Health	Safety	Energy	Other	0.1321	0.5916	0.5106	0.2374	0.3679
	Health	Entmt	Energy	Other	0.2348	0.5436	0.5843	0.2957	0.4146
AllMini	Health	Entmt	Safety	Energy	0.1805	0.6381	0.6342	0.5711	0.5060
	Entmt	Safety	Energy	Other	0.5161	0.6289	0.5011	0.2723	0.4796
	Health	Entmt	Safety	Other	0.1288	0.5234	0.5899	0.2328	0.3687
	Health	Safety	Energy	Other	0.0706	0.5058	0.1672	0.0000	0.1859
	Health	Entmt	Energy	Other	0.1777	0.1498	0.2544	0.2514	0.2083
SObert	Health	Entmt	Safety	Energy	0.0557	0.0400	0.5021	0.1683	0.1915
	Entmt	Safety	Energy	Other	0.0551	0.5284	0.1643	0.0000	0.1870
	Health	Entmt	Safety	Other	0.1040	0.1080	0.5374	0.0103	0.1899
	Health	Safety	Energy	Other	0.1103	0.0562	0.0125	0.2693	0.1121
	Health	Entmt	Energy	Other	0.0409	0.1047	0.0063	0.3202	0.1180
Bert4RE	Health	Entmt	Safety	Energy	0.0464	0.2374	0.4045	0.0095	0.1744
	Entmt	Safety	Energy	Other	0.1107	0.0379	0.0158	0.2770	0.1104
	Health	Entmt	Safety	Other	0.0504	0.0842	0.0379	0.2831	0.1139

Table 4.29: Multiclass (4 Classes) Results for Word-Embedding-20 Labels (with GloVe)

Model	Class1	Class2	Class3	Class4	F1-C1	F1-C2	F1-C3	F1-C4	Macro F1
Sbert	Health	Safety	Energy	Other	0.3575	0.6970	0.6110	0.1562	0.4554
	Health	Entmt	Safety	Other	0.4995	0.6387	0.6899	0.1811	0.5023
	Entmt	Safety	Energy	Other	0.6396	0.7283	0.6727	0.1369	0.5444
	Health	Entmt	Energy	Other	0.4269	0.6667	0.6269	0.3178	0.5096
	Health	**Entmt**	**Safety**	**Energy**	**0.3588**	**0.6459**	**0.7092**	**0.6375**	**0.5878**
AllMini	Health	Entmt	Safety	Other	0.3778	0.6487	0.7313	0.3231	0.5202
	Entmt	Safety	Energy	Other	0.6869	0.7260	0.6486	0.1895	0.5627
	Health	Safety	Energy	Other	0.3149	0.7164	0.5805	0.1866	0.4496
	Health	Entmt	Safety	Energy	0.3256	0.6773	0.7268	0.6131	0.5857
	Health	Entmt	Energy	Other	0.3374	0.6823	0.5983	0.1944	0.4531
SObert	Health	Entmt	Energy	Other	0.0000	0.3717	0.0407	0.0000	0.1031
	Health	Entmt	Safety	Other	0.0000	0.3400	0.2410	0.0000	0.1453
	Entmt	Safety	Energy	Other	0.3326	0.1002	0.0064	0.0000	0.1098
	Health	Safety	Energy	Other	0.0000	0.5212	0.2084	0.0582	0.1969
	Health	Entmt	Safety	Energy	0.0000	0.3162	0.2326	0.0376	0.1466
Bert4RE	Health	Entmt	Safety	Energy	0.0000	0.0000	0.5122	0.2018	0.1785
	Health	Entmt	Safety	Other	0.0000	0.0000	0.0778	0.2767	0.0886
	Health	Entmt	Energy	Other	0.0000	0.0000	0.0250	0.3135	0.0846
	Health	Safety	Energy	Other	0.0000	0.0439	0.0094	0.2623	0.0789
	Entmt	Safety	Energy	Other	0.0000	0.0765	0.0032	0.2727	0.0881

Table 4.30: Multiclass (4 Classes) Results for Word-Embedding-Based-50 Labels (Word2Vec)

Model	Class1	Class2	Class3	Class4	F1-C1	F1-C2	F1-C3	F1-C4	Macro F1
Sbert	Health	Safety	Energy	Other	0.2331	0.5608	0.4630	0.0744	0.3328
	Health	Entmt	Energy	Other	0.5356	0.6499	0.6287	0.0744	0.4720
	Health	Entmt	Safety	Energy	0.2195	0.6014	0.5357	0.4639	0.4551
	Entmt	Safety	Energy	Other	0.5767	0.5759	0.4671	0.0523	0.4179
	Health	Entmt	Safety	Other	0.2283	0.5784	0.5718	0.0474	0.3565
AllMini	Health	Safety	Energy	Other	0.3133	0.6434	0.5645	0.1523	0.4184
	Health	Entmt	Energy	Other	0.4770	0.5644	0.6458	0.1700	0.4645
	Health	Entmt	Safety	Energy	0.2916	0.6033	0.6359	0.5735	0.5261
	Entmt	Safety	Energy	Other	0.5922	0.6787	0.5804	0.0797	0.4828
	Health	Entmt	Safety	Other	0.2964	0.5692	0.6421	0.0699	0.3944
SObert	Health	Safety	Energy	Other	0.0033	0.5259	0.0032	0.0000	0.1331
	Health	Entmt	Energy	Other	0.0000	0.0310	0.0186	0.3049	0.0886
	Health	Entmt	Safety	Energy	0.0000	0.0000	0.5128	0.0032	0.1290
	Entmt	Safety	Energy	Other	0.0000	0.5459	0.0032	0.0000	0.1373
	Health	Entmt	Safety	Other	0.0000	0.0000	0.5508	0.0000	0.1377
Bert4RE	Health	Safety	Energy	Other	0.0000	0.0601	0.0000	0.2677	0.0819
	Health	Entmt	Energy	Other	0.0067	0.0120	0.0000	0.3105	0.0823
	Health	Entmt	Safety	Energy	0.0000	0.0000	0.5132	0.0000	0.1283
	Entmt	Safety	Energy	Other	0.0000	0.0617	0.0000	0.2775	0.0848
	Health	Entmt	Safety	Other	0.0000	0.0000	0.0911	0.2842	0.0938

Table 4.31: Multiclass (4 Classes) Results for Word-Embedding-50 Labels (with GloVe)

Model	Class1	Class2	Class3	Class4	F1-C1	F1-C2	F1-C3	F1-C4	Macro F1
Sbert	Health	Entmt	Energy	Other	0.5196	0.6667	0.6560	0.3319	0.5435
	Health	Entmt	Safety	Energy	0.4654	0.6471	0.7066	0.6401	**0.6148**
	Health	Entmt	Safety	Other	0.5261	0.6399	0.6997	0.1772	0.5107
	Health	Safety	Energy	Other	0.4485	0.6962	0.6240	0.1706	0.4848
	Entmt	Safety	Energy	Other	0.6328	0.7209	0.6745	0.1628	0.5478
AllMini	Health	Entmt	Energy	Other	0.3607	0.6735	0.6171	0.2597	0.4777
	Health	Entmt	Safety	Energy	0.3355	0.6590	0.7613	0.6486	**0.6011**
	Health	Entmt	Safety	Other	0.3672	0.6720	0.7163	0.3128	0.5171
	Health	Safety	Energy	Other	0.3161	0.7404	0.6242	0.2063	0.4718
	Entmt	Safety	Energy	Other	0.6706	0.7749	0.6927	0.2635	**0.6004**
SObert	Health	Entmt	Energy	Other	0.0000	0.0000	0.0064	0.3129	0.0798
	Health	Entmt	Safety	Energy	0.0000	0.0000	0.5138	0.0127	0.1316
	Health	Entmt	Safety	Other	0.0000	0.0000	0.0303	0.2824	0.0782
	Health	Safety	Energy	Other	0.0000	0.0302	0.0000	0.2675	0.0744
	Entmt	Safety	Energy	Other	0.0000	0.0154	0.0032	0.2805	0.0748
Bert4RE	Health	Entmt	Energy	Other	0.0000	0.3813	0.0063	0.0957	0.1208
	Health	Entmt	Safety	Energy	0.0000	0.3133	0.1054	0.0062	0.1062
	Health	Entmt	Safety	Other	0.0000	0.3426	0.0776	0.0853	0.1264
	Health	Safety	Energy	Other	0.0000	0.3361	0.0032	0.2146	0.1385
	Entmt	Safety	Energy	Other	0.3568	0.0570	0.0248	0.1642	0.1507

Table 4.32: Multiclass (4 Classes) Results for Combined Labels (Word2Vec)

Model	Class1	Class2	Class3	Class4	F1-C1	F1-C2	F1-C3	F1-C4	Macro F1
Sbert	Health	Safety	Energy	Other	0.3075	0.6049	0.5425	0.0926	0.3869
	Health	Entmt	Energy	Other	0.5589	0.6485	0.6553	0.2385	0.5253
	Health	Entmt	Safety	Energy	0.3091	0.5677	0.6023	0.5493	0.5071
	Entmt	Safety	Energy	Other	0.5473	0.6281	0.5475	0.0749	0.4494
	Health	Entmt	Safety	Other	0.3207	0.5591	0.6144	0.0737	0.3920
AllMini	Health	Safety	Energy	Other	0.3265	0.6974	0.6551	0.2285	0.4769
	Health	Entmt	Energy	Other	0.3265	0.6065	0.6067	0.3032	0.4607
	Health	Entmt	Safety	Energy	0.2297	0.5908	0.7057	0.6616	0.5470
	Entmt	Safety	Energy	Other	0.5551	0.7456	0.6913	0.2455	0.5594
	Health	Entmt	Safety	Other	0.2460	0.5840	0.6553	0.2697	0.4388
SObert	Health	Safety	Energy	Other	0.0000	0.5128	0.1445	0.0000	0.1643
	Health	Entmt	Energy	Other	0.0000	0.0000	0.0532	0.3074	0.0902
	Health	Entmt	Safety	Energy	0.0000	0.0000	0.5128	0.0653	0.1445
	Entmt	Safety	Energy	Other	0.0000	0.1019	0.0576	0.2726	0.1080
	Health	Entmt	Safety	Other	0.0132	0.0000	0.2803	0.0000	0.0734
Bert4RE	Health	Safety	Energy	Other	0.0000	0.0727	0.0032	0.2680	0.0860
	Health	Entmt	Energy	Other	0.0000	0.0199	0.0064	0.3112	0.0844
	Health	Entmt	Safety	Energy	0.0000	0.0124	0.5130	0.0032	0.1322
	Entmt	Safety	Energy	Other	0.0083	0.0706	0.0032	0.2784	0.0901
	Health	Entmt	Safety	Other	0.0000	0.0042	0.0553	0.2780	0.0844

Table 4.33: Multiclass (4 Classes) Results for Combined Labels (with GloVe)

Model	F-1 (Health)	F-1 (Entmt)	F-1 (Safety)	F-1 (Energy)	F-1 (Other)	Macro F1
Sbert	0.3756	0.5650	0.5509	0.2488	0.0435	0.3568
AllMini	0.4480	0.5873	0.6591	0.5393	0.1141	0.4696
SObert	0.0856	0.0042	0.0133	0.3524	0.0000	0.0911
Bert4RE	0.0680	0.2540	0.1920	0.0000	0.0761	0.1180

Table 4.34: Multiclass (5 Classes) Results for Original Labels

Model	F-1 (Health)	F-1 (Entmt)	F-1 (Safety)	F-1 (Energy)	F-1 (Other)	Macro F1
Sbert	0.4712	0.6236	0.6587	0.5900	0.0441	0.4775
AllMini	0.2754	0.5210	0.6826	0.6199	0.1997	0.4597
SObert	0.1281	0.0000	0.2391	0.2971	0.1927	0.1714
Bert4RE	0.0227	0.0912	0.1956	0.1252	0.2102	0.1290

Table 4.35: Multiclass (5 Classes) Results for Expert Curated Labels

Model	F-1 (Health)	F-1 (Entmt)	F-1 (Safety)	F-1 (Energy)	F-1 (Other)	Macro F1
Sbert	0.4357	0.5796	0.6638	0.6115	0.2154	0.5012
AllMini	0.3937	0.6473	0.7083	0.5969	0.1605	0.5013
SObert	0.0354	0.2768	0.2491	0.2592	0.0000	0.1641
Bert4RE	0.0166	0.2740	0.0067	0.0328	0.0000	0.0660

Table 4.36: Multiclass (5 Classes) Results for Word-Embedding-Based-20 Labels (Word2Vec)

Model	F-1 (Health)	F-1 (Entmt)	F-1 (Safety)	F-1 (Energy)	F-1 (Other)	Macro F1
Sbert	0.1254	0.5373	0.4603	0.4069	0.0240	0.3108
AllMini	0.1209	0.5284	0.5558	0.4859	0.2220	0.3826
SObert	0.0495	0.0360	0.4508	0.1558	0.0000	0.1384
Bert4RE	0.0310	0.0754	0.0333	0.0095	0.2294	0.0757

Table 4.37: Multiclass (5 Classes) Results for Word-Embedding-20 Labels (with GloVe)

Model	F-1 (Health)	F-1 (Entmt)	F-1 (Safety)	F-1 (Energy)	F-1 (Other)	Macro F1
Sbert	0.3424	0.6107	0.6590	0.5809	0.1226	0.4631
AllMini	0.2947	0.6460	0.6935	0.5571	0.1654	0.4713
SObert	0.0000	0.2800	0.2258	0.0376	0.0000	0.1087
Bert4RE	0.0000	0.0000	0.0746	0.0032	0.2244	0.0604

Table 4.38: Multiclass (5 Classes) Results for Word-Embedding-Based-50 Labels (Word2Vec)

Model	F-1 (Health)	F-1 (Entmt)	F-1 (Safety)	F-1 (Energy)	F-1 (Other)	Macro F1
Sbert	0.2026	0.5545	0.4888	0.4447	0.0425	0.3466
AllMini	0.2719	0.5428	0.5912	0.5517	0.0776	0.4070
SObert	0.0000	0.0000	0.4619	0.0032	0.0000	0.0930
Bert4RE	0.0000	0.0000	0.0595	0.0000	0.2297	0.0578

Table 4.39: Multiclass (5 Classes) Results for Word-Embedding-50 Labels (with GloVe)

Model	F-1 (Health)	F-1 (Entmt)	F-1 (Safety)	F-1 (Energy)	F-1 (Other)	Macro F1
Sbert	0.4329	0.6117	0.6574	0.5976	0.1463	0.4892
AllMini	0.3039	0.6473	0.7170	0.5948	0.1916	0.4909
SObert	0.0000	0.0000	0.0299	0.0000	0.2300	0.0520
Bert4RE	0.0000	0.2826	0.0753	0.0032	0.0721	0.0866

Table 4.40: Multiclass (5 Classes) Results for Combined Labels (Word2Vec)

Model	F-1 (Health)	F-1 (Entmt)	F-1 (Safety)	F-1 (Energy)	F-1 (Other)	Macro F1
Sbert	0.2946	0.5392	0.5466	0.5204	0.0690	0.3939
AllMini	0.2199	0.5574	0.6580	0.6243	0.2081	0.4536
SObert	0.0000	0.0000	0.0814	0.0567	0.2271	0.0731
Bert4RE	0.0000	0.0041	0.0646	0.0032	0.2291	0.0602

Table 4.41: Multiclass (5 Classes) Results for Combined Labels (with GloVe)

Chapter 5

Analysis of Results

5.1 Broader Analysis

5.1.1 Identification of Best-Performing Models and Best Label Configurations

For each experimental setup (One-vs-One, One-vs-Rest, 3-Class, 4-Class, and 5-Class classification), the following steps were undertaken:

- **Comparison of F1-Scores**: The primary metric used to evaluate performance was the **Macro F1-Score**, which provides a balanced measure of precision and recall across all classes. For each combination of model and label configuration, the Macro F1-Score was computed and compared.
- **Selection of Best Performers**: Within each experimental setup, the best model keeping label-configuration constant and label configuration keeping model constant was chosen for the one with highest Macro F1-Score were identified. For example:
  - In One-vs-One classification, the **AllMini** model was the winner with **Expert-Curated Labels** by achieving the highest Macro F1-Score of 0.8797 for the *Safety vs Energy* class combination. (Table 5.1)
  - In 5-Class classification, the **AllMini** model with **Word-20 (Word2Vec)** labels achieved the highest Macro F1-Score of 0.5013 (Table 5.9).
- **Dominant Combinations**: Specific combinations of models and label configurations that consistently performed well across multiple setups were highlighted. For instance, **AllMini** emerged as the dominant model in most cases, while **Word-20 (Word2Vec)** labels often outperformed other label configurations.

5.1.2 Computation of Win Ratios

To quantify the dominance of specific models and label configurations across all experimental setups, **Win Ratios** were computed. The process involved:

- **Counting Wins**: For each model or label configuration, the number of cases where it achieved the highest Macro F1-Score was counted.

- **Calculating Ratios**: The Win Ratio was calculated as the proportion of cases won relative to the total number of cases. Mathematically:

$$\text{Win Ratio (Model)} = \frac{\text{Number of Cases where Model Wins}}{\text{Total Number of Cases}}$$

  Similarly, the Win Ratio for label configurations was computed by counting the number of times a specific label configuration outperformed others.

- **Interpreting Results**: The Win Ratios provided a clear indication of overall performance. For example:
  - **AllMini** achieved the highest Win Ratio of 0.725 among models, indicating its dominance in 72.5% of the cases (Table 5.11).
  - **Word-20 (Word2Vec)** achieved the highest Win Ratio of 0.35 among label configurations, suggesting its effectiveness in 35% of the cases (Table 5.12).

5.1.3 Conclusion of Analysis

The combination of F1-Score comparisons and Win Ratio computations allowed us to quantitatively assess the strengths of different models and label configurations. This approach not only highlighted the best-performing combinations for specific tasks but also provided insights into their generalizability across diverse experimental setups.

5.1.4 Analysis Tables

Label Config.	Best Model	Class Config.	F1
Original	AllMini	Entmt vs Safety	0.8326
Expert-Curated	AllMini	Safety vs Energy	0.8797
Word-20 (Word2Vec)	AllMini	Safety vs Energy	0.8488
Word-50 (Word2Vec)	AllMini	Entmt vs Safety	0.8561
Combined (Word2Vec)	AllMini	Safey vs Energy	0.8621
Word-20 (GloVe)	Sbert	Entmt vs Energy	0.8665
Word-50 (GloVe)	Sbert	Entmt vs Energy	0.8585
Combined (GloVe)	Sbert	Entmt vs Energy	0.8582

Table 5.1: Comparison of Models for One-vs-One (OvO) for all label configurations

Model	Best Label Config.	Class Config.	F1
Sbert	Word-20 (GloVe)	Entmt vs Energy	0.8665
AllMini	Expert-Curated	Safety vs Energy	0.8797
SObert	Word-50 (GloVe)	Entmt vs Energy	0.5590
Bert4RE	Expert-Curated	Entmt vs Energy	0.6044

Table 5.2: Comparison of Label Config. for One-vs-One (OvO) for all Models

Label Config.	Best Model	Class Config.	F1
Original	Bert4RE	Other	0.4801
Expert-Curated	SObert	Other	0.4927
Word-20 (Word2Vec)	AllMini	Safety	0.5710
Word-50 (Word2Vec)	AllMini	Safety	0.4902
Combined (Word2Vec)	AllMini	Energy	0.5467
Word-20 (GloVe)	AllMini	Energy	0.5078
Word-50 (GloVe)	AllMini	Energy	0.5083
Combined (GloVe)	AllMini	Energy	0.5138

Table 5.3: Comparison of Models for One-vs-Rest (OvR) for all label configurations

Model	Best Label Config.	Class Config.	F1
Sbert	Combined (Word2Vec)	Other	0.8665
AllMini	Word-20 (Word2Vec)	Safety	0.5710
SObert	Expert-Curated	Other	0.5590
Bert4RE	Combined (Word2Vec)	Health	0.5140

Table 5.4: Comparison of Label Config. for One-vs-Rest (OvR) for all Models

Label Config.	Best Model	Class Config.	F1
Original	AllMini	Health vs Entmt vs Safety	0.6984
Expert-Curated	Sbert	Entmt vs Safety vs Energy	0.7362
Word-20 (Word2Vec)	AllMini	Entmt vs Safety vs Energy	0.7636
Word-50 (Word2Vec)	AllMini	Entmt vs Safety vs Energy	0.7458
Combined (Word2Vec)	AllMini	Entmt vs Safety vs Energy	0.7712
Word-20 (GloVe)	Sbert	Health vs Entmt vs Energy	0.6783
Word-50 (GloVe)	AllMini	Entmt vs Safety vs Energy	0.6771
Combined (GloVe)	AllMini	Entmt vs Safety vs Energy	0.7291

Table 5.5: Comparison of Models for 3-Class Classification (Multiclass) for all label configurations

Model	Best Label Config.	Class Config.	F1
Sbert	Word-50 (Word2Vec)	Entmt vs Safety vs Energy	0.7426
AllMini	Combined (Word2Vec)	Entmt vs Safety vs Energy	0.7712
SObert	Word-20 (Word2Vec)	Health vs Safety vs Energy	0.3668
Bert4RE	Word-20 (Word2Vec)	Health vs Safety vs Energy	0.3359

Table 5.6: Comparison of Label Config. for 3-Class Classification (Multiclass) for all Models

Label Config.	Best Model	Class Not Included	F1
Original	AllMini	Other	0.5981
Expert-Curated	Sbert	Other	0.6254
Word-20 (Word2Vec)	AllMini	Other	0.6258
Word-50 (Word2Vec)	Sbert	Other	0.5878
Combined (Word2Vec)	Sbert	Other	0.6148
Word-20 (GloVe)	AllMini	Other	0.5060
Word-50 (GloVe)	AllMini	Other	0.5261
Combined (GloVe)	AllMini	Health	0.5594

Table 5.7: Comparison of Models for 4-Class Classification (Multiclass) for all label configurations

Label Config.	Best Model	Class Not Included	F1
Sbert	Expert Curated	Other	0.6254
AllMini	Word-20 (Word2Vec)	Other	0.6258
SObert	Word-20 (Word2Vec)	Entmt	0.2518
Bert4RE	Original	Entmt	0.2089

Table 5.8: Comparison of Label Config. for 4-Class Classification (Multiclass) for all Models

Label Config.	Best Model	F1
Original	AllMini	0.4696
Expert-Curated	Sbert	0.4775
Word-20 (Word2Vec)	AllMini	0.5013
Word-50 (Word2Vec)	AllMini	0.4713
Combined (Word2Vec)	AllMini	0.4909
Word-20 (GloVe)	AllMini	0.3826
Word-50 (GloVe)	AllMini	0.4070
Combined (GloVe)	AllMini	0.4536

Table 5.9: Comparison of Models for 5-Class Classification (Multiclass) for all label configurations

Model	Best Label Config.	F1
Sbert	Word-20 (Word2Vec)	0.5012
AllMini	Word-20 (Word2Vec)	0.5013
SObert	Expert Curated	0.1714
Bert4RE	Expert Curated	0.1290

Table 5.10: Comparison of Label Config. for 5-Class Classification (Multiclass) for all Models

Model	Win Ratio
Sbert	0.225
AllMini	0.725
SObert	0.025
Bert4RE	0.025

Table 5.11: Overall Model Comparison: **According to this analysis, the best performing combination is *AllMini*** with a win ratio of 0.75.

Label Config.	Win Ratio
Original	0.05
Expert-Curated	0.30
Word-20 (Word2Vec)	0.35
Word-50 (Word2Vec)	0.05
Combined (Word2Vec)	0.15
Word-20 (GloVe)	0.05
Word-50 (GloVe)	0.05
Combined (GloVe)	0

Table 5.12: Overall Label Comparison: **According to this analysis, the best performing model is *AllMini* and the best label configuration is *Word-embedding based top 20 (Word2Vec)***

5.2 Micro Analysis

This analysis considers 31 cases (5 OvR, 10 OvO, 10 three-class, 5 four-class, and 1 five-class) across 32 environments (4 models × 8 label configurations). In each case, the best of all these 32 environments is considered a winner. Finally, a win-ratio is calculated out of the 31 possible cases.

Case	Classes	Winner (Model + Label Config)
One-vs-Rest	Health	Bert4RE (Combined W2V)
	Entmt	AllMini (Word-20 W2V)
	Energy	AllMini (Word-20 W2V)
	Safety	AllMini (Word-20 W2V)
	Other	SObert (Expert)
One-vs-One	Health vs Entmt	Sbert (Expert-Curated)
	Health vs Energy	Sbert (Combined GloVe)
	Health vs Safety	AllMini (Original)
	Health vs Other	AllMini (Original)
	Entmt vs Energy	Sbert (Word-20 GloVe)
	Entmt vs Safety	AllMini (Word-50 W2V)
	Entmt vs Other	AllMini (Expert)
	Energy vs Safety	AllMini (Expert)
	Energy vs Other	Sbert (Word-20 W2V)
	Safety vs Other	AllMini (Word-50 W2V)
3-Class	Health vs Entmt vs Energy	Sbert (Expert)
	Health vs Entmt vs Safety	AllMini (Original)
	Health vs Entmt vs Other	AllMini (Word-20 W2V)
	Health vs Energy vs Safety	Sbert (Word-20 W2V)
	Health vs Energy vs Other	Sbert (Word-20 W2V)
	Health vs Safety vs Other	AllMini (Original)
	Entmt vs Energy vs Safety	AllMini (Combined W2V
	Entmt vs Energy vs Other	Sbert (Word-50 W2V)
	Entmt vs Safety vs Other	AllMini (Word-50 W2V)
	Energy vs Safety vs Other	AllMini (Combined W2V)
4-Class	Health vs Entmt vs Energy vs Safety	AllMini (Word-20 W2V)
	Health vs Entmt vs Energy vs Other	AllMini (Combined W2V)
	Health vs Entmt vs Safety vs Other	AllMini (Word-20 W2V)
	Health vs Energy vs Safety vs Other	AllMini (Word-50 W2V)
	Entmt vs Energy vs Safety vs Other	AllMini (Combined W2V)
5-Class	All	AllMini (Word-20 W2V)

Table 5.13: Winner for each case (in terms of both model and label configuration)

Model + Label Config	**Win Ratio**
AllMini (Word-20 W2V)	0.23
AllMini (Original)	0.10
AllMini (Word-50 W2V)	0.13
AllMini (Combined W2V)	0.13
AllMini (Expert)	0.06
Sbert (Expert)	0.06
Sbert (Expert-Curated)	0.03
Sbert (Combined GloVe)	0.03
Sbert (Word-20 GloVe)	0.03
Sbert (Word-20 W2V)	0.10
Sbert (Word-50 W2V)	0.03
Bert4RE (Combined W2V)	0.03
SObert (Expert)	0.03

Table 5.14: Overall Model + Label Configuration Comparison: **According to this analysis, the best performing combination is *AllMini (Word-20 W2V)*** with a win ratio of 0.23.

5.3 Conclusion

By leveraging Macro F1-Scores and Win Ratios, we identified the best-performing combinations for each setup and quantified their dominance. The **AllMini** model emerged as the most consistent performer, achieving the highest Win Ratio of 0.725 among all models. Similarly, the **Word-20 (Word2Vec)** label configuration demonstrated superior effectiveness with a Win Ratio of 0.35. These findings underscore the importance of selecting appropriate models and label configurations to optimize performance in classification tasks. Overall, the combination of **AllMini** and **Word-20 (Word2Vec)** proved to be the most robust and generalizable choice across diverse experimental setups.

Supplementary Information

The code and supplementary materials for this project are available on:

- GitHub Repository: `https://github.com/rohmeh/zsl4re`

Bibliography

[1] Waad Alhoshan, Alessio Ferrari, and Liping Zhao. Zero-shot learning for requirements classification: An exploratory study, 2023.

[2] Tobias Hey, Jan Keim, Anne Koziolek, and Walter Tichy. Norbert: Transfer learning for requirements classification. 09 2020.

[3] Pradeep K. Murukannaiah, Nirav Ajmeri, and Munindar P. Singh. Acquiring creative requirements from the crowd: Understanding the influences of individual personality and creative potential in crowd RE. In *Proceedings of the 20th IEEE International Requirements Engineering Conference (RE)*, pages 176–185, Beijing, September 2016. IEEE Computer Society.

[4] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[5] Aarón Rodríguez García. Bert4re: Domain-specific bert model for requirements engineering. `https://huggingface.co/thearod5/bert4re`, 2024. Accessed: 2025-03-06.

[6] Jeniya Tabassum. Bertoverflow: A bert model trained on stack overflow questions. `https://huggingface.co/jeniya/BERTOverflow`, 2024. Accessed: 2025-03-06.

[7] Sentence Transformers Team. all-minilm-l12-v2. `https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2`, 2024. Accessed: 2025-03-06.

[8] Liping Zhao, Waad Alhoshan, Alessio Ferrari, Keletso J Letsholo, Muideen A Ajagbe, Erol-Valeriu Chioasca, and Riza T Batista-Navarro. Natural language processing for requirements engineering: A systematic mapping study. *ACM Computing Surveys (CSUR)*, 54(3):1–41, 2021.