UN Cybersecurity Project Midterm Report

Group Members:

Shimeng Lin, Bennong Liu, Shaoying Tang, Chuanze Cai

Contents

[Summary: 3](#_Toc480796106)

[Introduction: 3](#_Toc480796107)

[Transition from Phase I to Phase II: 3](#_Toc480796108)

[Keywords Expansion: 3](#_Toc480796109)

[Phase II Progress: 3](#_Toc480796110)

[Phase II Achievements: 3](#_Toc480796111)

[Transition from Phase I to Phase II 3](#_Toc480796112)

[Phase II Progress: Keywords Expansion 4](#_Toc480796113)

[Methodologies: Supervised Model 4](#_Toc480796114)

[Keywords in Phase I: 4](#_Toc480796115)

[Keywords in Phase II: 6](#_Toc480796116)

[Summary 7](#_Toc480796117)

[Results Comparison: Phase II vs. I: 7](#_Toc480796118)

[Original Model: 7](#_Toc480796119)

[Updated Model: 7](#_Toc480796120)

[Accuracy of the Updated Model: 8](#_Toc480796121)

[Phase II Achievements and Future: 8](#_Toc480796122)

# Summary:

The project is to text analysis the cybersecurity strategy reports from different countries and show the tagged sentences to the end-users. In this phase, we expand the important topics of the table. And do the same visualization work.

# Introduction:

## Transition from Phase I to Phase II:

Tagging Applied to a Specific Country

## Keywords Expansion:

Methodologies

Category & Subcategory Expansions

## Phase II Progress:

    Updated Tagging Procedure & Tagging Results

    Comparison and Summary Charts

## Phase II Achievements:

Improvements and Calculation, and Future Plan

## Transition from Phase I to Phase II

* **Tagging to UK Documents**

We applied the methodologies and codes from Phase I to a specific country – United Kingdom to

check the tagging performance.

 Transition from Phase I to Phase II:

* **Tagging Results**

Out of 523 sentences contained in the UK text, there were 18 sentences tagged by the category and subcategory selected. The tagging rate was only 3.44%, which was considered relatively low.

However, this only applied to UK case. We could also check this rate using other countries' text data.

We figured there were only a few key words captured. With the supervised model, we proceeded to expand the category/ subcategory selections.

# Phase II Progress: Keywords Expansion

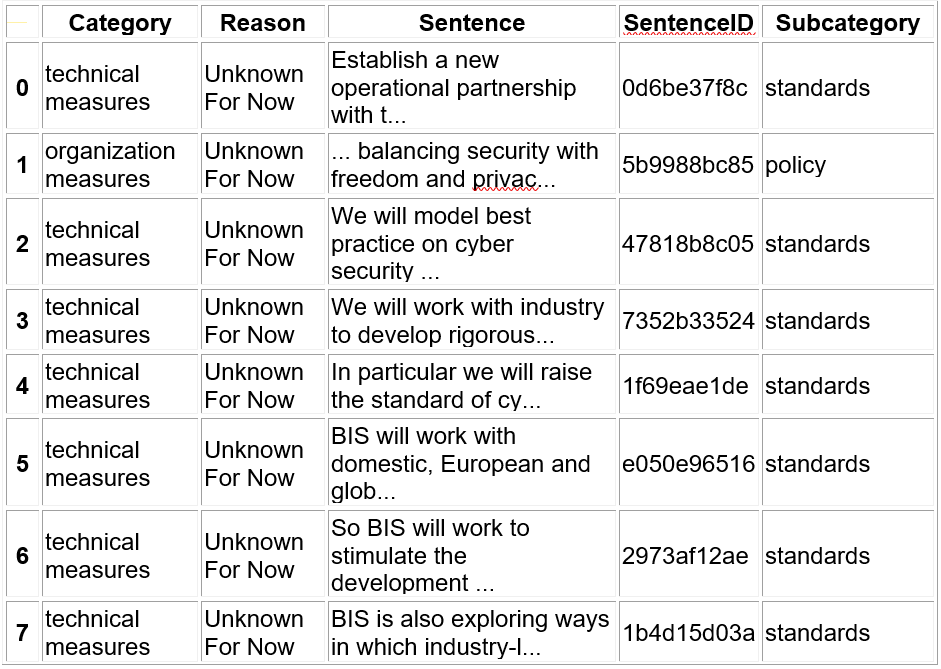
## Methodologies: Supervised Model

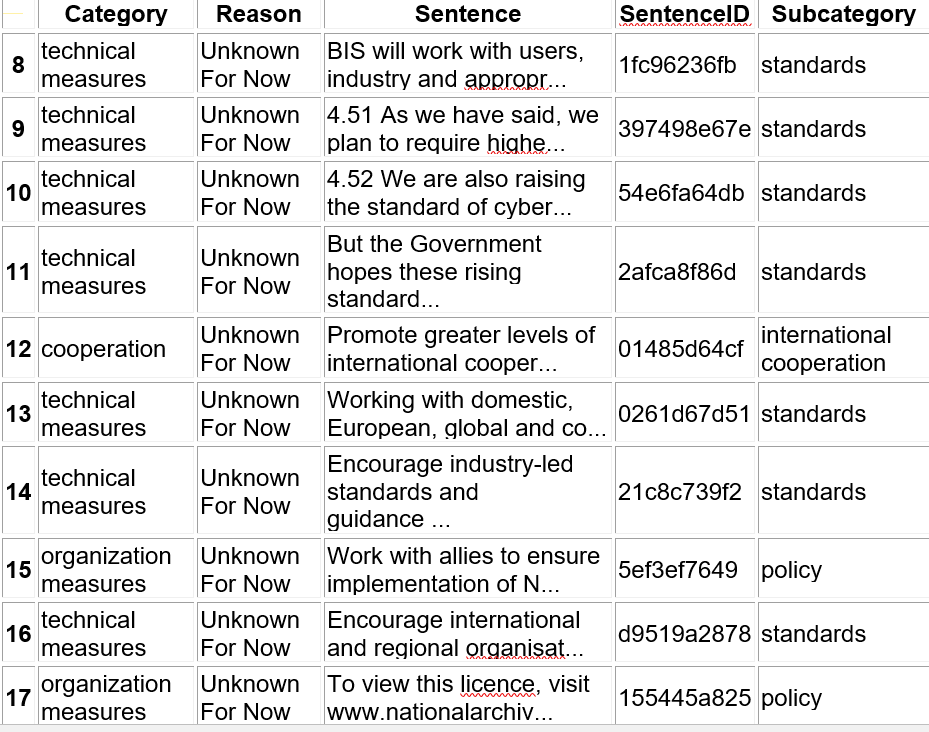
After reading the UK files, we expanded the keywords in category/subcategory to improve the accuracy of our model.

## Keywords in Phase I:

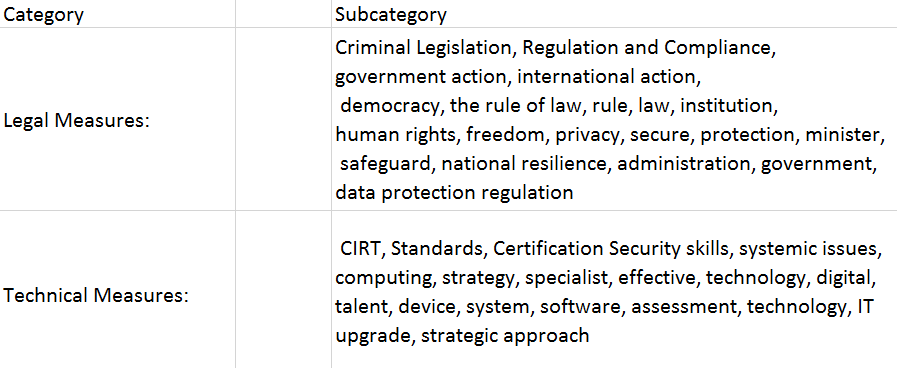


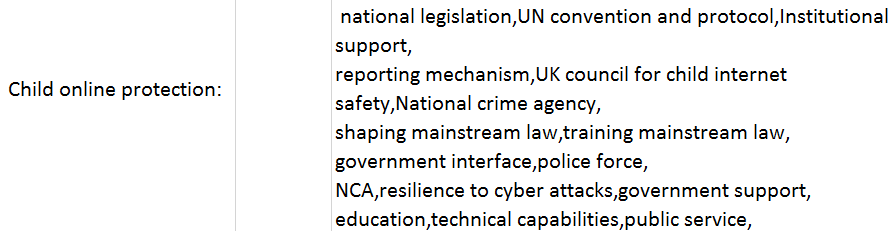
Only 18 sentences in the UK file were tagged using the Phase I keywords.





## Keywords in Phase II:



## Summary

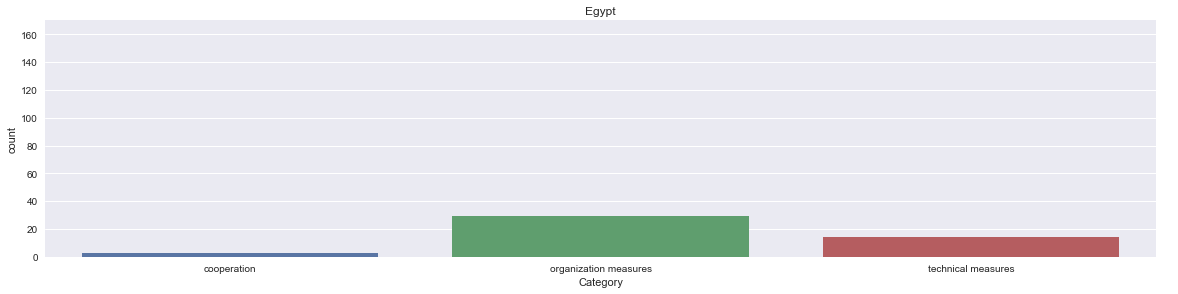
We added new keywords and the number of sentences tagged was increased resulting from the updated model. The tagging rate was approximately 50%, which was acceptable.

However, certain keywords may not be applicable. In the next stage, we will further improve the words selection in the categories.

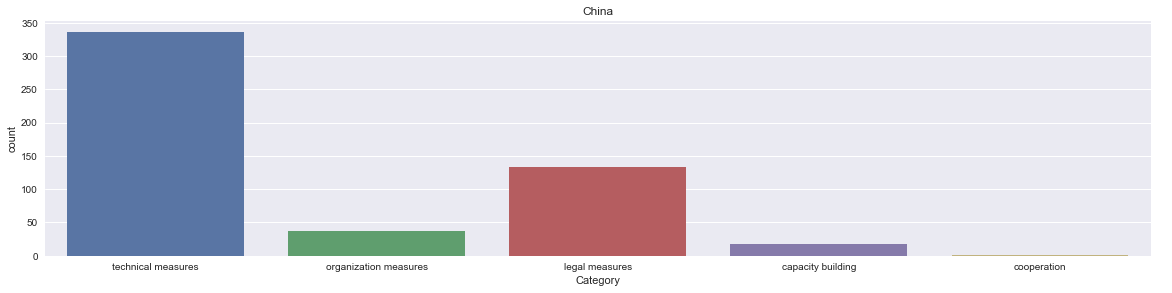
# Results Comparison: Phase II vs. I:

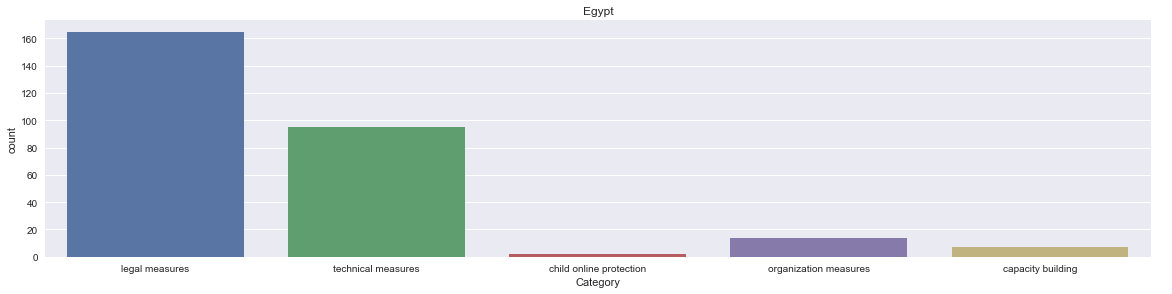
## Original Model:





## Updated Model:





## Accuracy of the Updated Model:

Problems:

Sentences with keywords of mixed categories. e.g. Establishing a regulatory framework a) MCIT will work together with the Ministry of Justice (MoJ) to develop legal framework and conduct for addressing cybersecurity issues…...    Organization Measure or Technical Measure or Legal Measure?

# Phase II Achievements and Future:

From the above conclusions, we could observe that the results of distribution were only for informative purpose although we improved the results from Phase I.

According to the tagging results, the categories/keywords may not be accurate enough to provide solid evidence to support the next research stage.

Our future plan includes updating the model with more restrictive categorizing/clustering criteria to optimize the results. We will also calculate the tentative accuracy rates in terms of measurement.