**TECHNICAL TASK**

**Candidate :** Zeki Gültekin

**Start Date :** 02.12.2022

**Deadline :** 12.12.2022

**Description:**

Smart Navigator is an intelligent decision system that transforms big data into actionable business intelligence and bespoke intelligent threat & risk analysis. We provide a flagship product, Smart Navigator, which does not replace your internal teams but enriches their capabilities through AI-powered insights, predictions and recommendations so they can more effectively tackle tough challenges. Our solution safeguards symbiosis of humans and machines by leveraging trusted digital transformation, ensures digital sovereignty and helps you to safeguard your business in an era of cyber security threats and attacks.

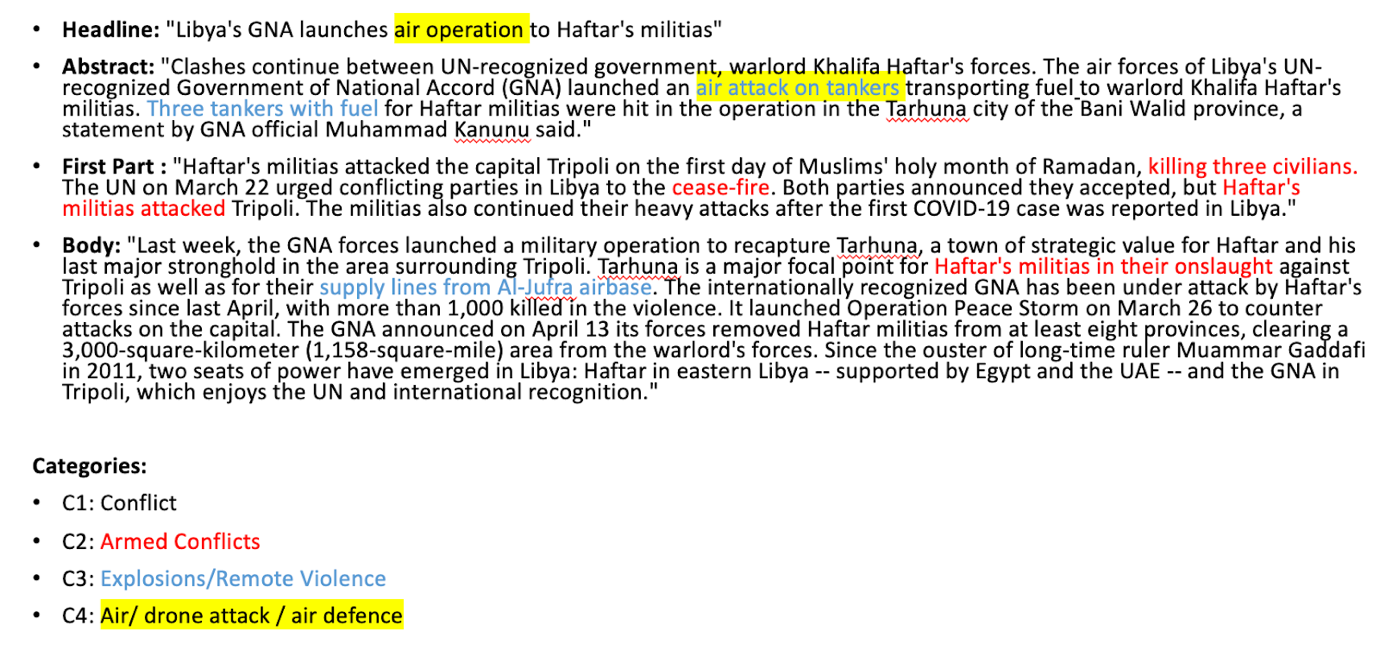
**Data:**

The given data includes the following:

1. **DocumentID**
2. **EventInstanceId**
3. **Source**
4. **Country**
5. **Headline**
6. **Abstract**
7. **First Part**
8. **Body**
9. **Event Category**
10. **1st Level Sub Category**
11. **2nd Level Sub Category**
12. **3rd Level Sub Category**

Each row or observation is one document and has unique Document ID. Each document has essential features like Headline, Abstract, First Part and Body to develop ML/DL model based on them to be able to categorize it.

For instance, the figure below shows one document with above-mentioned features and through this document, we will explain the idea of classification. The "black" words of the document focus more on Conflict in C1 (Event Category), and the words on the "red" underline further describe Armed Conflicts in C2 (1st Level Sub Category). Therefore, when understanding the semantic of each document, it is necessary to capture these associations between texts and the hierarchical structure. As shown in the figure, the document concentrates on the armed conflict in C2 since its parent category conflict in C1 is focused. And Explosions / Remote Violence in C3 (2nd Level Sub Category) with “blue” words is the subclass of armed conflict and Air / drone attack / air defence in C4 (3th Level Sub Category) which are highlighted with “yellow” is the subclass of explosions/remote violence, which should also be highly heeded.



In the given dataset, there are different type of event categories, such as: Agreements (25 observation), Conflicts (133 observations), Diplomatic (84 observations), Disasters (55 observations), Economic Events (62 observations), Political Events (40 observations), Societal Challenges (42 observations), Terror (39 observations) and Uprisings (10 observations). Not each event but most of them have also their sub level categories. For example, in conflict event category, there are 48 observations for Non-Violent Conflicts (6 retention, 30 military preparations, 12 economic conflicts) and 85 observations armed conflicts (6 battles, 42 explosions/remote violence, 6 violence against civilians, 31 battles).

**Challenge:**

1. Select one domain as event category, develop your text classification model and clearly present your model scores through evaluation metrics and ideal hyperparameters. Please keep in mind that, if you select the event category which has sub-level categories, your model should predict the relevant sub level categories. For example, event category: Conflict, 1st Level Sub Category: Armed Conflicts, 2nd Level Sub Category: Explosions/Remote Violence, 3rd Level Sub Category: Air/drone attack / air defence.
2. You will work with a very small dataset. Please report the challenges while you are working with such kind of small datasets during developing your ML/DL model and explain your approaches to overcome this problem.
3. Which text classification model have you used and why?
4. The shared data is whole dataset (train + test dataset). Please clearly state the size of your train and test dataset.