Natural Language Processing to Triage Maritime Distress Signals - Department of Mathematics - Spring 2024

Background

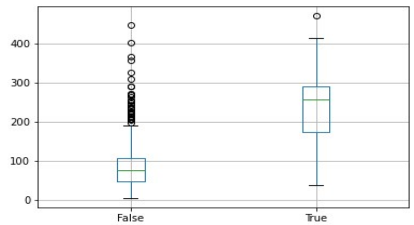
Search and Rescue Satellite Aided Tracking (SARSAT) alerts are distress signals received by United States Coast Guard (USCG) command centers. They are activated by Emergency Position Indicating Radio Beacons (EPIRBs), which may be activated manually or from water intrusion into the device casing due to partial or full submersion of a vessel. The image below outlines the SARSAT response process. The CG Command Centers will often receive false alerts due to accidental activations. Responding to false alerts regularly waste valuable time, money, and resources. When responding to a SARSAT alert, a watch stander records details about the case such as the timeline of events, location, vessel details, and USCG response actions.

Problem Statement

USCG watch-standers have built a personal understanding of common causes and indicators of false alerts through years of experience. However, the Coast Guard would like to validate, or redefine, this intuition with data-driven findings. Using prior MISLE case narratives, Natural Language Processing can be utilized to find key situational indicators that can help determine whether an alert has a high likelihood of being a false alarm. These indicators may allow watch standers to prioritize response to more urgent alerts and increase operational confidence through data-backed decisions.

Analysis

Labeling the Data



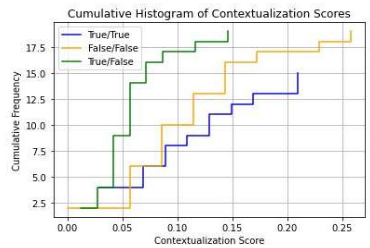
The figure above displays a boxplot of the length of case narratives.

Criteria Scores	Lives at Risk	Length of Nar.	"False" Mentioned?
True Distress	> 0 : +10	> 143 Words : +4	No : +0
False Distress	0 : -5	< 143 Words : -4	Yes : -13

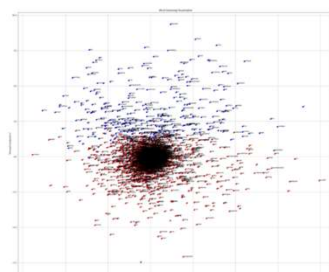
The figure above displays the labeling criteria used for the weighted label.

The figure above displays a word cloud of the words that were always related to false cases.

Natural Language Processing



The figure above displays a cumulative histogram of the contextualization scores.



The figure above displays the true and false word cluster.

```
word_vectors.most_similar(positive=["false"])
[('flase', 0.7344209551811218),
 ('accidental', 0.7011280655860901),
 ('fale', 0.56069016456604),
 ('unitl', 0.5586419105529785),
 ('inadvertant', 0.5584424138069153),
 ('non', 0.5549355149269104),
 ('malfunction', 0.5459740161895752),
 ('deactivate', 0.5320846438407898),
 ('repeat', 0.5278744697570801),
 ('accidently', 0.5274808406829834),
 ('set', 0.5188876986503601),
 ('safe', 0.5164341330528259),
```

The figure above displays the Word2Vec results with the word "false", the most similar words to false are shown here.

Group Members:
1/c Lillia Farlow
1/c Lleyton Herb
1/c Dana O'Brien

Advisors:
Dr. Zachary Kudlak
LCDR Justin Maio
LT Justin Sherman

Sponsors:
CDR Greg Higgins
LTJG Rachel Samotis
Mr. Douglas Samp

Methodology

Labeling the Data

To create a labeled dataset, we needed to identify what features were associated with a false alert. We created a weighted label based on narrative length, number of lives at risk, and whether the word "false" was mentioned in the narrative. Through the labeling process, we identified a list of words and phrases that are never found in true cases and always appeared in confirmed false cases. Ultimately, we were able to use the "Lives at Risk" column from the MISLE dataset to determine whether a case was true or false, labeling cases with 0 lives at risk as false.

Natural Language Processing

To analyze the data, we used NLP techniques including contextualization scoring, word clustering, logistic regression, and word2vec. For contextualization scoring we performed pairwise comparisons with true vs true, false vs false and true vs false cases. We randomly selected each case, with no repetition, and compared them to each other to receive a similarity score out of 100. For word clustering, we applied k-means clustering with two clusters (true vs false), and then reduced the dimensionality of the word vectors to 2 using Principal Component Analysis. The word cluster is shown below, with true words labeled in blue and false words labeled in red. Next, we used logistic regression to see what words were of high weight and importance when determining if a case was true or false. Additionally, we used Word2Vec to find words that are contextually similar and were commonly found near each other.

Results

We were able to identify a list of key indicators that never occur in true cases and were always associated with false cases. This list will help watch standers to make data-driven decisions and ultimately protect the U.S. Coast Guard from future liability. The list is as follows:

no registration information, registration expired, unreliable beacon, unreliable hex (ID), stolen, old EPIRB, out of service, was thrown out, vessel was sold, the vessel sank, landfill

Policy Recommendations

Based on our results, we have the following recommendations:

- Have harbormasters ensure EPIRB registration is up to date for the vessels in their marinas.
- Have mariners register their EPIRB before leaving the store after purchasing one.
- Increase USCG public affairs activity surrounding EPIRBs: registration, placement of EPIRBs, expiration, etc.
- Provide a whitepaper for watch standers which includes the key indicators list to allow confidence in data driven decisions.

To increase integrity of our conclusions, we recommend that this analysis is expanded to include data from all U.S. Coast Guard districts.