**Machine Learning Engineer Nanodegree**

**Capstone Project**

Nabanita Roy  
August 2021

**I. Definition**

**Project Overview**

Natural Language Processing with Disaster Tweets is a Kaggle Challenge where tweets are collected with labels indicating whether the tweets are about a disaster that occurred or not. Since tweets are social media language, therefore, it is a challenge to automatically identify them. Besides, ambiguity in texts makes it more difficult to achieve automatic identification of tweets containing information on real disaster. The objective of this project is to predict using machine learning if a tweet contains information on occurrence of a real disaster or not.

**Problem Statement**

Given a tweet, this task is designed to identify if it contains information on occurrence of a real disaster or not.

**Metrics**

Chart, bar chart

Description automatically generatedThis challenge is a classification problem therefore accuracy, precision, recall are relevant measures to assess the predictions. Since the training data is unbalanced, the accuracy will not be considered as a useful metric in this case and only F1-score will be considered to avoid accuracy paradox. F1-score is helpful to monitor both precision and recall as metrics since it represents their harmonic mean. An F1-score of more than 80% on the test set of this dataset is desirable.

**II. Analysis**

**Data Exploration**

This dataset has 7613 data points where each data point contains the tweet text, an unique id for the tweet, a keyword from that tweet, the location from where the tweet was posted and also if the tweet is about real disaster or not. Here is a column name and summary of the information available:

|  |  |
| --- | --- |
| Column Name | Description |
| id | A unique identifier for each tweet |
| keyword | A keyword from that tweet |
| location | The location from where the tweet was posted |
| text | Tweet text |
| target | Indicates whether a tweet is about real disaster or not. Present only in training set. |

This dataset contains missing values in columns ‘location’ and ‘keyword’. Following is a summary of missing values:

|  |  |
| --- | --- |
| Column Name | Missing Values (Percentage) |
| id | 0.0 |
| keyword | 0.8 |
| location | 33.3 |
| text | 0.0 |
| Target | 0.0 |

The ‘location’ column is the one with the most missing values. Since Named Entity Recognition (NER) systems can detect locations, these missing values could be filled.

Following is a sample of the dataset:

A picture containing graphical user interface

Description automatically generated

In the column ‘text’, typical social media texts are observed which includes tweet mentions, hashtags, urls and inconsistent casing. Besides, in the ‘location’ column, the location information is also in consistent. It could contain a city name, a country name, name of a state or province, or a combination of these. It could also contain other non-spatial information like ‘Est. September 2021’ as in the second row of the data table. The target column contains 0 or 1, where 1 indicates that the tweet is about a real disaster and 0 indicates that the tweet is not about a real disaster.

Below is another example of the data instances. Clearly, the ‘location’ column contains hashtags as well and state abbreviations. The ‘text’ column contains contractions and incorrect punctuations.

A picture containing graphical user interface

Description automatically generatedTable

Description automatically generated with medium confidence

Finally, 57% of the dataset contains information on tweets about real disasters and 42% contains information on tweets not about real disasters. Therefore, the dataset is almost balanced and no further balancing the target datapoints is required.

**Exploratory Visualization**

Chart, bar chart

Description automatically generatedIn this dataset, ‘keywords’ is an important information about the tweets. Following is a bar chart about the top 10 keywords in the tweets about real disasters and those not about real disasters.

Chart, bar chart

Description automatically generated

It is observed that derailment, wreckage and outbreak are the top three keywords in the tweets which are about real disasters. In the tweets not about real disasters, body bags, harm, armageddon are the top three keywords. The keywords – ‘wreckage’ and ‘wrecked’, both originating from the work ‘wreck’, occur in both target categories. Hence, one of the steps in pre-processing would be using stemming or lemmatizing in this use-case.

Following are the probabilities of the keywords occurring in real disaster tweets, where ‘derailment’, ‘debris’ and ‘wreckage’ have a 100% probability. This means that in the training set, every tweet having these three keywords were about real disasters. Similarly, the tweets having the keyword ‘outbreak’, 97% of them were about real disasters. Therefore, it could be assumed that if these keywords, with over 90% probability occur in a tweet, the probability of them being about a real disaster is much higher than any other keyword. However, this conclusion is based on non-normalised keywords data because of which keywords like ‘suicide bombing’ and ‘suicide bomber’ are occurring separately in this list.

|  |  |
| --- | --- |
| Keyword | Probability the Keyword is in a Tweet about Real Disaster |
| derailment | 1.000000 |
| debris | 1.000000 |
| wreckage | 1.000000 |
| outbreak | 0.974359 |
| typhoon | 0.972973 |
| oil spill | 0.972973 |
| suicide bombing | 0.968750 |
| suicide bomber | 0.966667 |
| bombing | 0.925926 |
| rescuers | 0.906250 |

The next column is ‘location’, for which a bar chart has been used to plot the top 20 most frequently occurring locations.

Chart, histogram

Description automatically generated As in this bar chart, USA appeared most in tweet locations. Again, the inconsistency in location texts is evident because ‘USA’, states within USA and ‘United States’ are all appearing in the top 20 list. Therefore, several locations from the USA has appeared most number of times here in general, followed by Asian countries, UK, Nigeria, Canada and Australia, if aggregated.

Finally, the ‘target’ column distribution is of interest to check if the data is balanced. Following is the bar plot of the counts of data instances for tweets about and not about real disasters.

Chart, bar chart

Description automatically generated

Apart from these techniques, other textual features were explored in the tweets, such as:

1. Number of punctuations: Punctuations could be overused in tweets which could be a supportive indicator of discerning the target classes. Therefore, a punctuation count analysis was conducted, as shown below:

Chart, box and whisker chart

Description automatically generated

It is evident in this box plot that the IQR of the boxplots are comparable but for tweets not about real disasters have way more outliers of higher orders than those about real disasters.

1. Appearance of numbers in Tweets: Numbers of casualties are often mentioned in texts about disasters.

Chart, bar chart

Description automatically generated

In the bar chart on the left, it is evident that the probability of finding numbers in real disaster tweets is higher.

1. Sentiment Analysis of Tweets: Tweets about real disasters are more likely to have less words about positive emotions than negative ones.

Chart, bar chart

Description automatically generatedThe bar chart on the left clearly shows that even though negative emotions are dominant in the entire corpus, the percentage of positive emotions in tweets about real disasters is about 50% lower than the tweets not about real disasters.

1. Chart, histogram

   Description automatically generatedChart, histogram

   Description automatically generatedTweet Length Analysis: The following analysis is conducted to verify if there is any distinguishable element in the lengths of the texts. Hence, I have compared the lengths of the texts, in terms of words and characters per tweet, between the two types of tweets. The distribution of the tweets about real disasters are more like a symmetric distribution while the one not about real disaster are somewhat left skewed.
2. Text

   Description automatically generatedTerm Frequency Analysis using Wordcloud: Wordcloud is the most used visualization technique for natural language texts to search for frequently occurring words. On the left is a wordcloud of most frequently occurring words in tweets about real disasters. Words about disasters are primarily present here. From an unbiased view, the words like fire, death, wildfire, storm, burning, flood, damage, casualty, police, bomb, evacuate, drought, earthquake – the collection of these words could be attributed to the topic of natural disasters. On the contrary, in the below wordcloud about tweets not about real disasters, the words cannot be easily assigned a particular topic. They are more random. Hence, it could be safely assumed that machine learning techniques based on term frequencies would be a great fit to use on this corpus.

Text

Description automatically generated

**Algorithms and Techniques**

To predict if a tweet is about real disasters or not, using Bag of Words approach of using term-frequencies seems promising.

**Benchmark**

In this section, you will need to provide a clearly defined benchmark result or threshold for comparing across performances obtained by your solution. The reasoning behind the benchmark (in the case where it is not an established result) should be discussed. Questions to ask yourself when writing this section:

* *Has some result or value been provided that acts as a benchmark for measuring performance?*
* *Is it clear how this result or value was obtained (whether by data or by hypothesis)?*

**III. Methodology**

*(approx. 3-5 pages)*

**Data Preprocessing**

In this section, all of your preprocessing steps will need to be clearly documented, if any were necessary. From the previous section, any of the abnormalities or characteristics that you identified about the dataset will be addressed and corrected here. Questions to ask yourself when writing this section:

* *If the algorithms chosen require preprocessing steps like feature selection or feature transformations, have they been properly documented?*
* *Based on the****Data Exploration****section, if there were abnormalities or characteristics that needed to be addressed, have they been properly corrected?*
* *If no preprocessing is needed, has it been made clear why?*

**Implementation**

In this section, the process for which metrics, algorithms, and techniques that you implemented for the given data will need to be clearly documented. It should be abundantly clear how the implementation was carried out, and discussion should be made regarding any complications that occurred during this process. Questions to ask yourself when writing this section:

* *Is it made clear how the algorithms and techniques were implemented with the given datasets or input data?*
* *Were there any complications with the original metrics or techniques that required changing prior to acquiring a solution?*
* *Was there any part of the coding process (e.g., writing complicated functions) that should be documented?*

**Refinement**

In this section, you will need to discuss the process of improvement you made upon the algorithms and techniques you used in your implementation. For example, adjusting parameters for certain models to acquire improved solutions would fall under the refinement category. Your initial and final solutions should be reported, as well as any significant intermediate results as necessary. Questions to ask yourself when writing this section:

* *Has an initial solution been found and clearly reported?*
* *Is the process of improvement clearly documented, such as what techniques were used?*
* *Are intermediate and final solutions clearly reported as the process is improved?*

**IV. Results**

*(approx. 2-3 pages)*

**Model Evaluation and Validation**

In this section, the final model and any supporting qualities should be evaluated in detail. It should be clear how the final model was derived and why this model was chosen. In addition, some type of analysis should be used to validate the robustness of this model and its solution, such as manipulating the input data or environment to see how the model’s solution is affected (this is called sensitivity analysis). Questions to ask yourself when writing this section:

* *Is the final model reasonable and aligning with solution expectations? Are the final parameters of the model appropriate?*
* *Has the final model been tested with various inputs to evaluate whether the model generalizes well to unseen data?*
* *Is the model robust enough for the problem? Do small perturbations (changes) in training data or the input space greatly affect the results?*
* *Can results found from the model be trusted?*

**Justification**

In this section, your model’s final solution and its results should be compared to the benchmark you established earlier in the project using some type of statistical analysis. You should also justify whether these results and the solution are significant enough to have solved the problem posed in the project. Questions to ask yourself when writing this section:

* *Are the final results found stronger than the benchmark result reported earlier?*
* *Have you thoroughly analyzed and discussed the final solution?*
* *Is the final solution significant enough to have solved the problem?*

**V. Conclusion**

*(approx. 1-2 pages)*

**Free-Form Visualization**

In this section, you will need to provide some form of visualization that emphasizes an important quality about the project. It is much more free-form, but should reasonably support a significant result or characteristic about the problem that you want to discuss. Questions to ask yourself when writing this section:

* *Have you visualized a relevant or important quality about the problem, dataset, input data, or results?*
* *Is the visualization thoroughly analyzed and discussed?*
* *If a plot is provided, are the axes, title, and datum clearly defined?*

**Reflection**

In this section, you will summarize the entire end-to-end problem solution and discuss one or two particular aspects of the project you found interesting or difficult. You are expected to reflect on the project as a whole to show that you have a firm understanding of the entire process employed in your work. Questions to ask yourself when writing this section:

* *Have you thoroughly summarized the entire process you used for this project?*
* *Were there any interesting aspects of the project?*
* *Were there any difficult aspects of the project?*
* *Does the final model and solution fit your expectations for the problem, and should it be used in a general setting to solve these types of problems?*

**Improvement**

In this section, you will need to provide discussion as to how one aspect of the implementation you designed could be improved. As an example, consider ways your implementation can be made more general, and what would need to be modified. You do not need to make this improvement, but the potential solutions resulting from these changes are considered and compared/contrasted to your current solution. Questions to ask yourself when writing this section:

* *Are there further improvements that could be made on the algorithms or techniques you used in this project?*
* *Were there algorithms or techniques you researched that you did not know how to implement, but would consider using if you knew how?*
* *If you used your final solution as the new benchmark, do you think an even better solution exists?*

**Before submitting, ask yourself. . .**

* Does the project report you’ve written follow a well-organized structure similar to that of the project template?
* Is each section (particularly **Analysis** and **Methodology**) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?
* Would the intended audience of your project be able to understand your analysis, methods, and results?
* Have you properly proof-read your project report to assure there are minimal grammatical and spelling mistakes?
* Are all the resources used for this project correctly cited and referenced?
* Is the code that implements your solution easily readable and properly commented?
* Does the code execute without error and produce results similar to those reported?