Checkpoint 1 - analiza danych

Warsztaty z technik uczenia maszynowego

Jakub Rymuza Karol Nowiński

23 marca 2022

1 Użyte dane

W projekcie wykorzystano dwa zbiory danych z [1]:

- zbiór treningowy *train.csv*, który zostanie wykorzystany do wytrenowania modelu do rozwiązywania problemu. Zawiera on 7613 rekordów. Rekordy zawierają następujące pola:
  - *id* - identyfikator rekordu,
  - *keyword* - słowo kluczowe uprzednio wyciągnięte z wiadomości - wszystkie słowa kluczowe dotyczą katastrof, są tą słowa takie jak na przykład "crash", "earthquake" i tym podobne. To pole jest opcjonalne, tzn. nie każdy rekord je zawiera,
  - *location* - lokalizacją z której wiadomość została wysłana. Podobnie jak *keyword*, jest to pole opcjonalne,
  - *text* - najważniejsza pole - zawiera treść wiadomości,
  - *target* - wartość logiczna stwierdzająca czy dana wiadomość mówi o katastrofie czy nie.
- zbiór testowy *test.csv* - zbiór na którym model będzie testowany. Zawiera on 3263 rekordów. Rekordy wyglądają tak samo, jak w przypadku zbioru treningowe, za wyjątkiem oczywiście braku pola *target*, którego obliczenie jest celem projektu.

2 Analiza danych

2.1 Podstawowe informacje

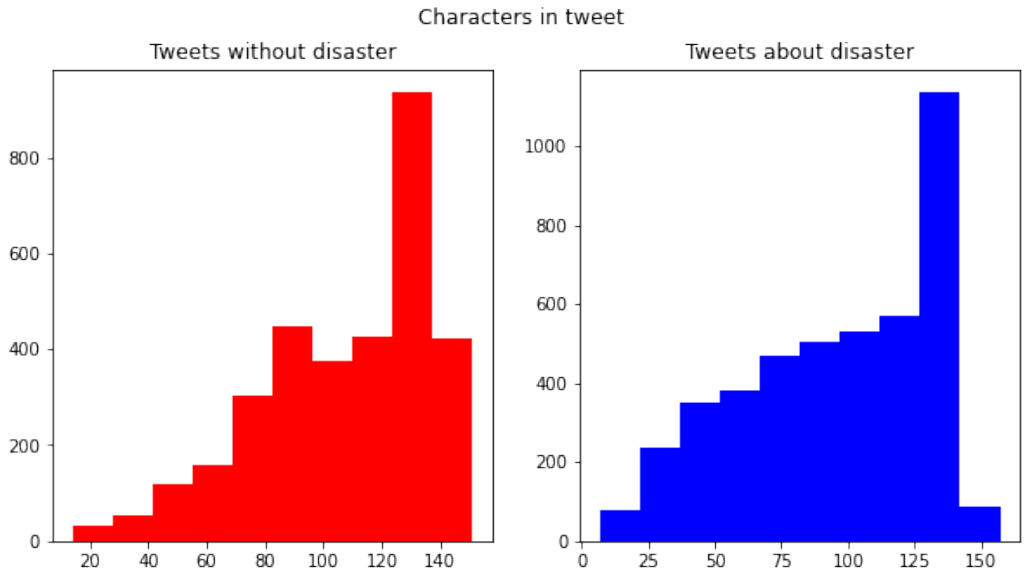
W zbiorze *train* składającym się z 7613 rekordów, tylko 5080 posiada podaną lokację, a 7552 posiada podane słowo kluczowe (*keyword*).

Natomiast w zbiorze *test* na 3263 rekordów, tylko 2158 posiada podaną lokację, a 3237 posiada podany *keyword*.

Wynika stąd, że w obu zbiorach prawie wszystkie rekordy mają podane słowo kluczowe, natomiast tylko około dwie trzecie z nich mają podaną lokację.

2.2 Długość tweetów

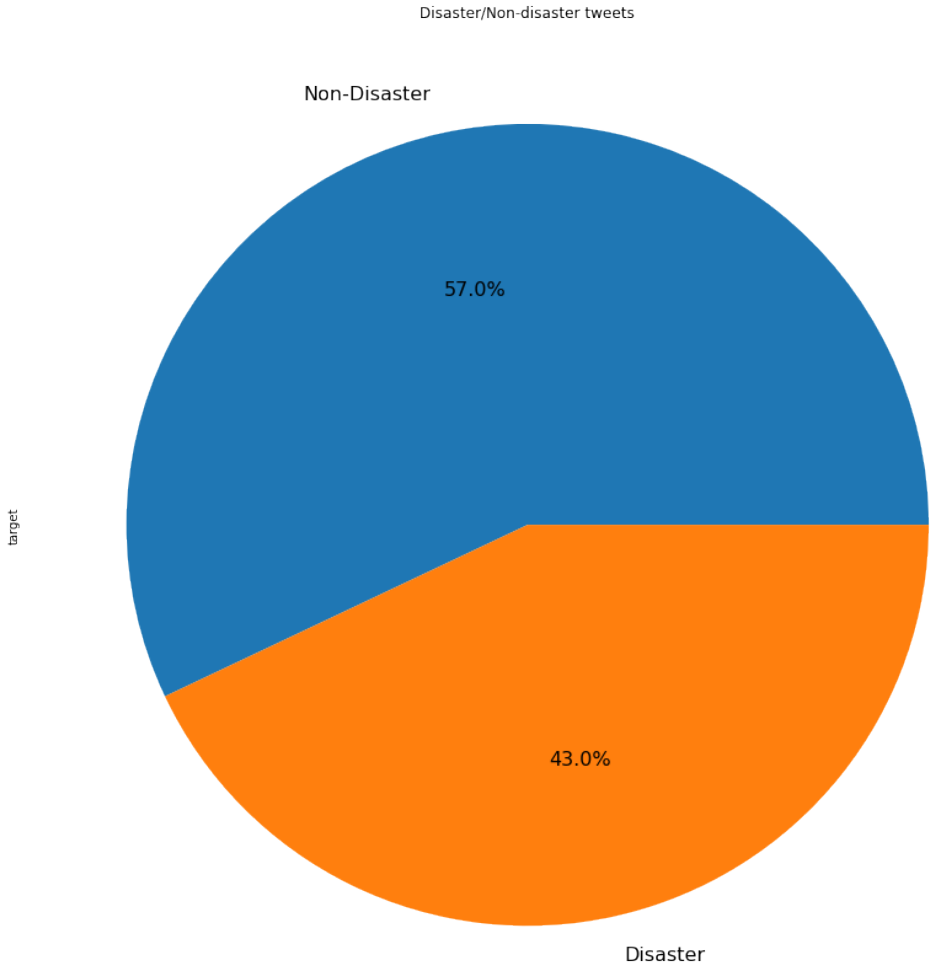
Na rysnku 1 umieszczono histogramy liczby tweetów w danym przedziale długości w zależności od tego czy mówią o katastrofie czy nie dla zbioru *train*. Można zauważyć, tweety o katastrofie częściej są krótkie lub średniej długości, natomiast rzadko są bardzo długie. Natomiast w obu przypadkach pik występuje dla długości około 120 znaków.



Rysunek 1: Znaki w tweet'cie

2.3 Skład danych treningowych

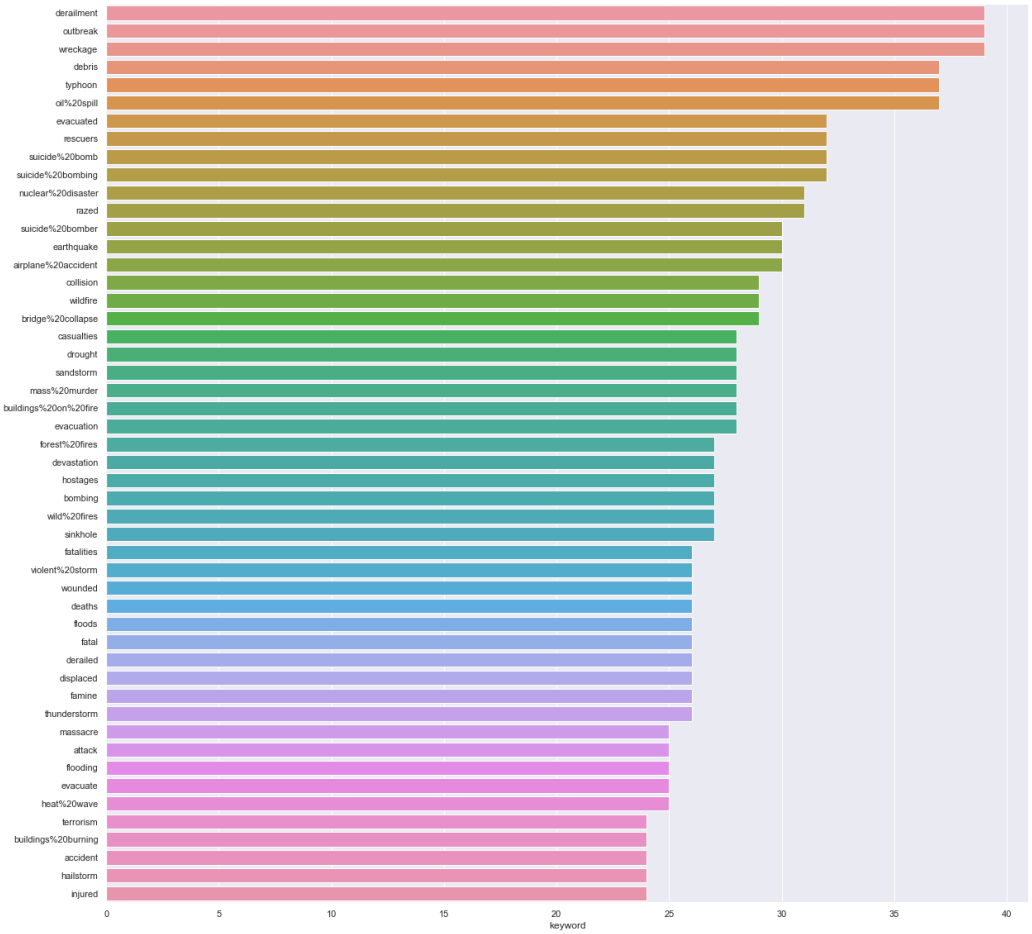
Na rysnku 2 zaznaczono skład zbioru *train* ze względu na ilość tweetów o katastrofach. Jak widać dane są rozmieszczone dość równomiernie, z niewielką przewagą tweetów niemówiących o katastrofach. Taki skład danych treningowych umożliwia dobre wytrenowanie modelu.



Rysunek 2: Stosunek tweetów o katastrofach i nie o katastrofach

2.4 Słowa kluczowe

Słowa kluczowe mogą pełnić ważna rolę przy klasyfikacji tweetu. Na rysunku 3 przedstawiono popularność poszczególnych słów kluczowych w zbiorze *train*. Jak widać najpopularnijsze słowa kluczowe to *derailment*, *outbreak* i *wreckage*.



Rysunek 3: Popularność słów kluczowych

3 Bibliografia

[1] https://www.kaggle.com/c/nlp-getting-started