

The background features a light gray field with various geometric elements. On the left, a complex network graph is visible, consisting of numerous black dots (nodes) connected by thin gray lines. Scattered across the entire background are several triangles of different sizes and orientations, some outlined in gray and others in black. In the upper right corner, there is a sparse collection of small, faint gray dots.

NLP DISASTER TWEETS

A binary classification problem
By Cristina Sahoo

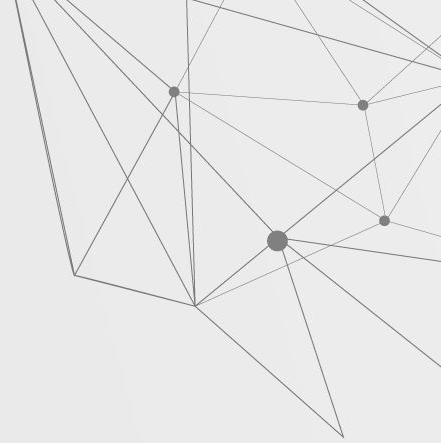
PROBLEM STATEMENT

Kaggle: In this competition, you're challenged to *build a machine learning model that predicts which Tweets are about real disasters and which one's aren't*. You'll have access to a dataset of 10,000 tweets that were hand classified.



WHY IS THIS IMPORTANT?

Twitter has become an important communication channel in times of emergency. The ubiquitousness of smartphones enables people to announce an emergency they're observing in real-time. Because of this, more agencies are interested in programatically monitoring Twitter (i.e. disaster relief organizations and news agencies).





DATA SOURCE

(1) Kaggle/Data for Everyone

Files:
train.csv with 7,613 records.
test.csv with 3,263 records.

Train data has 7,613 rows.

Test data has 3,263 rows.

Text data was run through several cleanup functions to remove punctuation, numbers, misspelled words, stop words, etc. The cleaned data was then split into features and target, vectorized, and transformed into tensors/multidimensional arrays to be fed into the neural network model.



DATA DICTIONARY

Feature	Type	Description
id	int64	a unique identifier for each tweet
text	object	the text of the tweet
location	object	the location the tweet was sent from (may be blank)
keyword	object	a particular keyword from the tweet (may be blank)
target	int64	in train.csv only, this denotes whether a tweet is about a real disaster (1) or not (0)

WORD CLOUDS

A word cloud shaped like the word 'Deeds'. The word 'Deeds' is the largest and most prominent, written in a dark blue font. Other words are arranged around it in various sizes and colors (green, yellow, blue, and purple). The words include: shelter, wildfires, Ronge, place, got, earthquake, residents, Reason, evacuation, La, Two, Canada, cranes, fire, receive, asked, M, near, people, Ruby, Forest, Alaska, sent, photo, and Sask.

A word cloud shaped like the word 'love'. The word 'love' is the largest and most prominent, written in a dark blue font. Other words are arranged around it in various sizes and colors (green, yellow, blue, and purple). The words include: fast, fruits, lovely, Cramer, wrecked, Li, Disney, ready, British, Great, boxes, Summer, atmosphere, Iger, explode, words, car, and goooooooooooooooooal.



MODELS

Models used:

- Dense Neural Network with three layers, 16 units each, activation function 'relu', and output function 'sigmoid'; optimizer 'rmsprop', loss function 'binary crossentropy', and metric 'accuracy'. Accuracy 79%.
- Multinomial Naive Bayes. Accuracy 80%.



CONCLUSIONS AND RECOMMENDATIONS

- (1) Disaster tweets contain words like forest, evacuation, residents, shelter, wildfires, earthquake.
- (2) Not Disaster tweets contain words like love, lovely, man, car ,summer, fruits.
- (3) The Naive Bayes models performed best, with 80% accuracy.
- (4) Neural Networks requires very large amounts of data, so they may not be the best option for working with this dataset.



