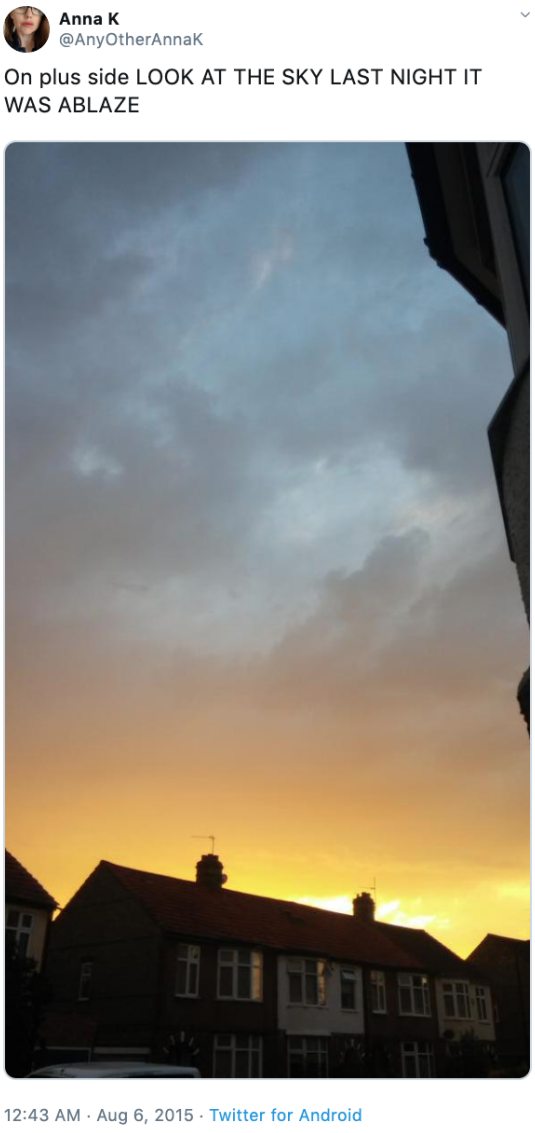
From the Kaggle competition description:

“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).”

However, there is a lot of ambiguity in language, so it can be hard to tell when a person is actually discussing a disaster. Words like “ablaze”,”bomb”,”terror”, and “quake” could reference multiple things. Here is another example from Kaggle:



Problem statement:

Twitter is a place with real time updates and could be a source to alerts for disasters. In order to do this successfully, I need to create a model that identifies genuine disaster indicators and distinguishes them from variant word usage. The final model will be applied to tweets 24 hours old or less to create disaster alerts.

The model will be trained on tweets, and the success is if the model can determine if a tweet is seriously alerting a real disaster and NOT alert for false positives.

It will be difficult because some sentences will be syntactically identical and so it’s likely there will be false positives that are impossible to prevent. This is also a machine learning model and there can’t be any manual tweaking, unlike some more traditional NLP methods that are less black box. It’s possible there will be an accuracy threshold that cannot be passed.

The primary data sources used are from Kaggle, the annotated training data set and the test data set. They came pre-split and the training data was pre-annotated.

This project is created for the Springboard Data Science Bootcamp as one of the capstone projects, and my mentor Varun will be consulted and have approval on it. The final project will be submitted to him and Springboard.