Real or Not? NLP with Disaster Tweets
PMLDL D1.1 progress report

Ivan Golov, Ilnaz Magizov, Alexey Shulmin

September 21, 2024

1 General Problem

Fake news and misinformation have become rampant in today's digital age. Social media platforms, like Twitter, often become breeding grounds for false information due to the speed at which content is shared and the lack of rigorous fact-checking. Misinformation can lead to public confusion, panic, and even dangerous situations, especially during disasters or emergencies. The challenge is distinguishing real, impactful events from posts that simply use emotionally charged language without any basis in actual events.

2 Business Problem / Social Value

The main social value of our project is to reduce panic and confusion by helping users, emergency responders, and organizations differentiate between real disaster news and emotional or sensationalized content. Businesses, particularly those in emergency services, news verification, or social media monitoring, can use this tool to quickly filter real news from noise. This could also aid in automating disaster response and resource allocation by ensuring the information acted upon is reliable.

3 Machine Learning Problem

The ML problem here is a classification task. The goal is to classify tweets into two categories:

- True disaster news
- Emotionally charged, non-news content

This is a text classification problem where our model will learn from patterns in the text to determine whether a tweet refers to an actual disaster or just emotionally laden content.

4 Data Requirements

For this problem, we found a dataset on Kaggle that includes:

- Tweets with information related to disasters (such as natural disasters, accidents, or other emergencies).
- Each tweet is labeled as either:
  - Real disaster news
  - Non-news emotional tweets

Additionally, this dataset contains metadata such as keyword from the tweet and location, which might be helpful to identify trends and patterns related to real events.

5 Exploratory Data Analysis (EDA)

The dataset comprises information about tweets with features `['location', 'keyword']` and a target variable indicating the presence of a real disaster (0 for no disaster and 1 for a real disaster). The training set consists of 7,613 rows, providing a suitable size for text analysis and disaster prediction tasks.

There are missing values for the keyword (61 missing) and location (2,533 missing) features. The `location` column contains 3,342 unique values, while the `keyword` column has 222 unique values. The text features contain numerous special characters, URLs, and tags, signaling the need for various NLP techniques for data cleaning and preparation for training.

The target variable distribution in the training set is relatively balanced. Analysis of features using word clouds reveals patterns in the `keyword` feature that help distinguish between disaster and non-disaster tweets. However, the distribution of the `location` feature is problematic due to a significant amount of unidentified regions, leading us to plan on dropping this feature because of extensive missing values and its limited predictive significance.

Please check out full EDA report in our GitHub repository.

6 Future Project Workflows

- **Stage 1:** Business problem and Data understanding | Infrastructure plan.
- **Stage 2:** Data preparation | ETL and data storage.
- **Stage 3:** Model development and training | Pipeline of transformations and data preparations.
- **Stage 4:** Model evaluation and optimization | Experiment tracking system
- **Step 5:** Deployment and monitoring.

7 Formal information

Our project: **Real or Not? NLP with Disaster Tweets**
Our team:

- Ivan Golov - i.golov@innopolis.university
- Ilnaz Magizov - i.magizov@innopolis.university
- Alexey Shulmin - a.shulmin@innopolis.university

Our GitHub repository: https://github.com/IVproger/PMDL-DisasterTweets
Dataset we are going to use: https://www.kaggle.com/competitions/nlp-getting-started/overview
Don't forget to check out our EDA:
https://github.com/IVproger/PMDL-DisasterTweets/blob/main/notebooks/eda.ipynb