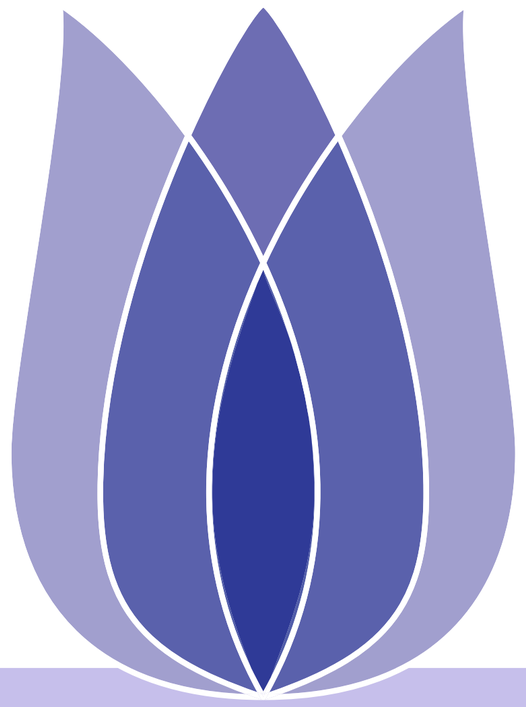
FLIP(01) Mid-term Presentation

Rongxin Xu
Hunan University

19 January 2019



Outline

Introduction

Data Description

Exploratory Data Analysis

Data Preprocessing

Modeling and Model Evaluation

Conclusion and Future research

Introduction

Problem Description

- The ubiquitousness of smartphones enables people to announce an emergency they're observing in real-time.
  - Predict whether a real disaster has occurred based on keywords, location, and Twitter text.

Data Description

Attribute Information

- Attributes Information

1. There are 3 data sets.

   train.csv the training set.
   test.csv the test set.
   sample_submission.csv a sample submission file in the correct format.

2. There are 3 data sets with a total of 5 attributes.

Table 1: Attribute Information

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

Exploratory Data Analysis

Scatter plot of keywords

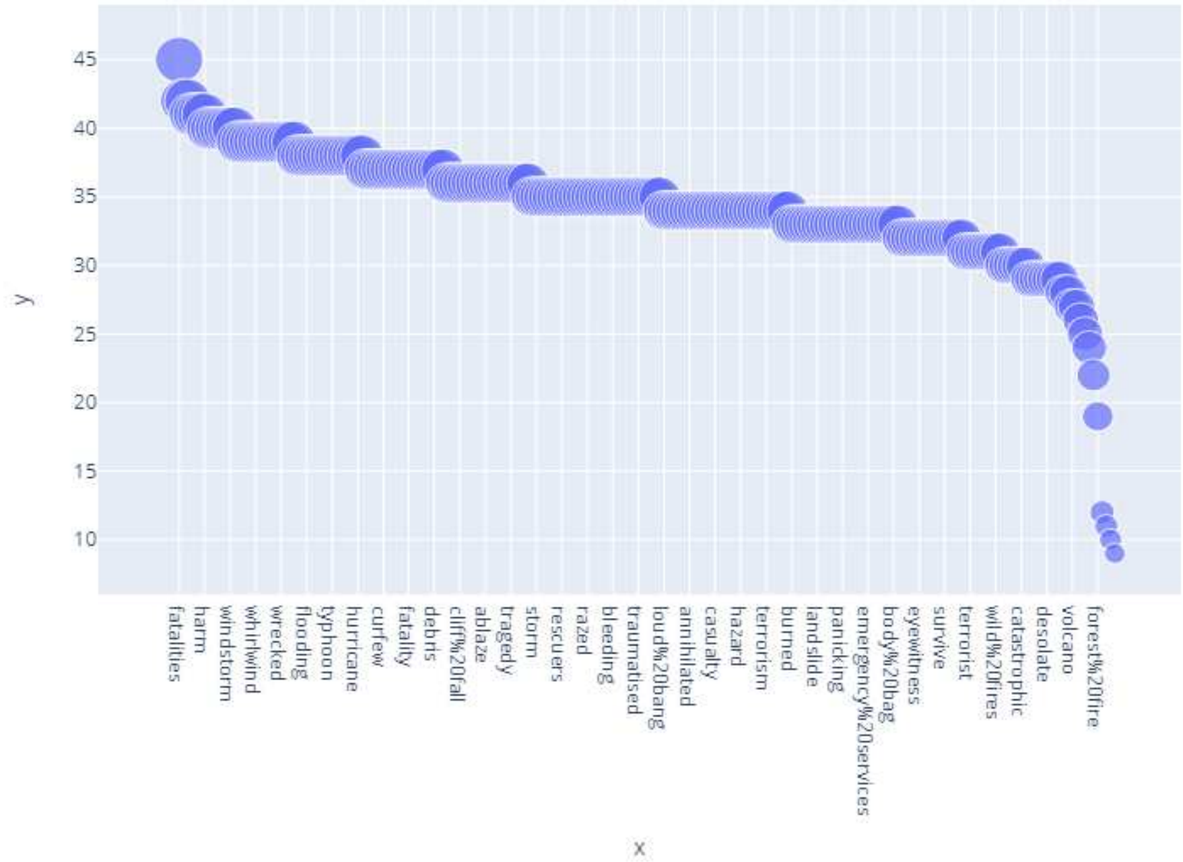


Figure 1: Scatter plot of keywords

Scatter plot of location

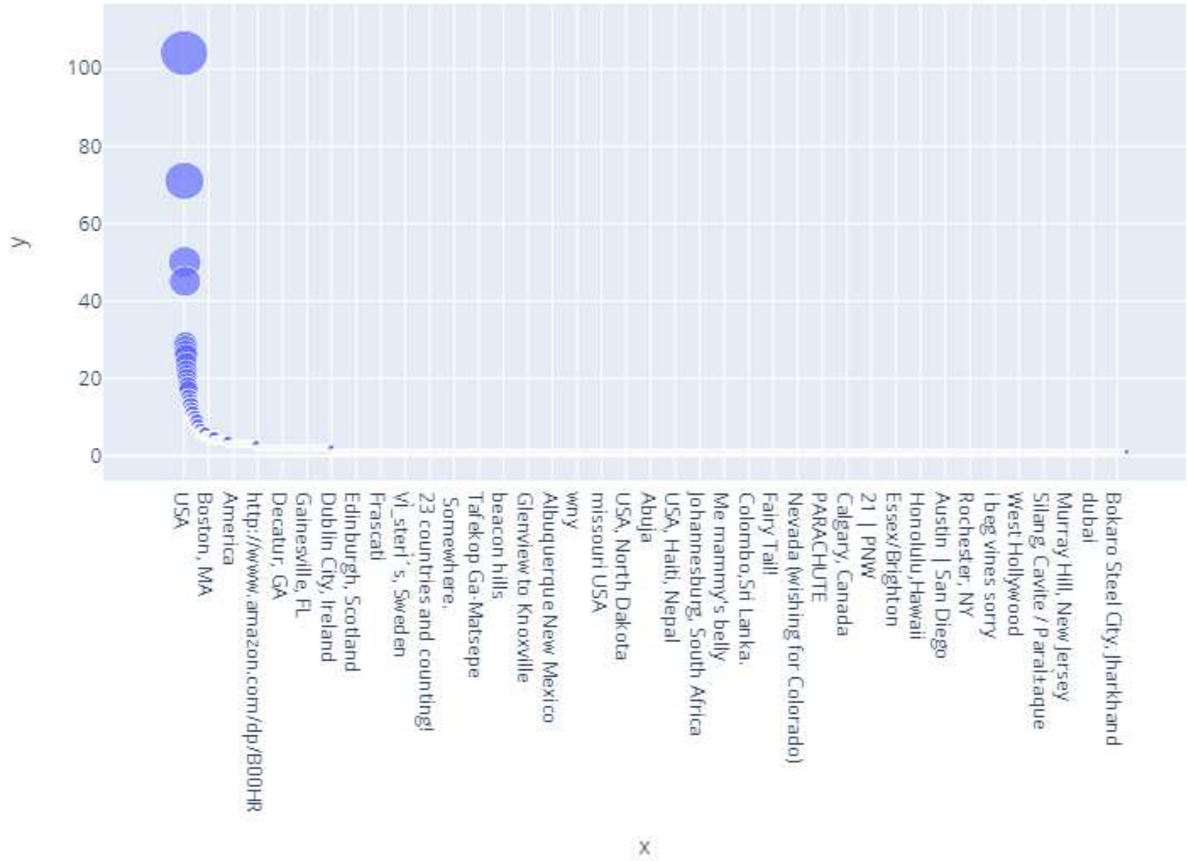


Figure 2: Scatter plot of location

Missing Values

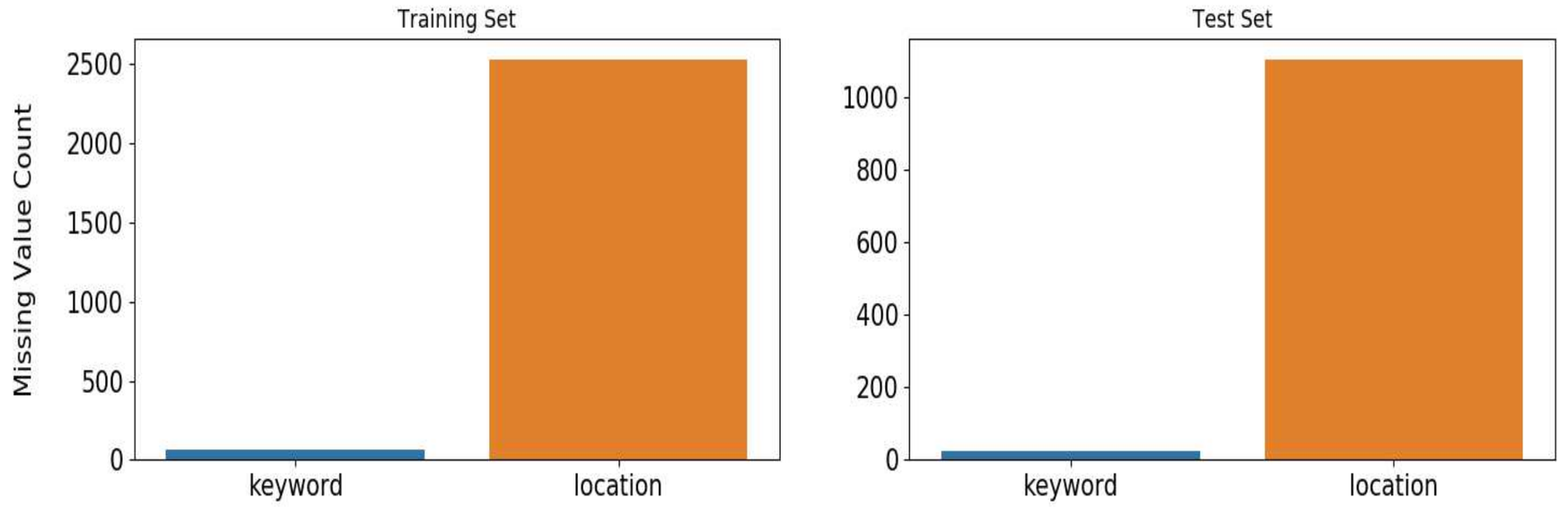


Figure 3: Missing Values of location and keywords

Summary

- Both training and test set have same ratio of missing values in keyword and location.

1. 0.8% of keyword is missing in both training and test set.
2. 33% of location is missing in both training and test set.

- From Data description and above analysis we can conclude :

1. Locations are not automatically generated, they are user inputs and that's why data is not clean and there are too many incoorect and missing values.We can skip the 'location' column from our feature list.
2. We can consider the 'keyword' column as a feature because there are a lot of unique keywords and missing values are very insignificant (< 1 percentage).

Data Preprocessing

Basic NLP Techniques

Create a Corpus from 'Text' coloumn, Corpus is a simplified version of our text data that contain clean data. To create Corpus should have to perform the following actions

Remove unwanted words Removal of unwanted words such as special characters and numbers to get only pure text.
Transform words to lowercase Transform words to lowercase because upper and lower case have diffirent ASCII codes.
Remove stopwords Stop words are usually the most common words in a language and they will be irrelevant in determining the nature.
Stemming words Stemming is the process of reducing words to their word stem, base or root form.

Modeling and Model Evaluation

Evaluation

Use F1 score.

- F1 is calculated as follows:

$$F_1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

- where:

$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{recall} = \frac{TP}{TP + FN}$$

Modeling

Since this NLP problem is also a binary classification problem, we might as well try some traditional classification models, such as:

Gaussian Naive Bayes
Gradient Boosting
K - Nearest Neighbor
Decision Tree
Logistic Regression
XGBOOST
Voting Classifier

Model Evaluation

Use F1 score and accuracy to evaluate model performance.

Table 2: Model Evaluation

model	Accuracy Score		F1 Score
	Train Data Set	Test Data Set	
Gaussian Naive Bayes	0.783314021	0.75843812	0.653371
Gradient Boosting	0.853666611	0.75446724	0.647673
K - Nearest Neighbors	0.975004138	0.726009265	0.56051
Decision Tree	0.975004138	0.729318332	0.663374
Logistic Regression	0.840258235	0.80344143	0.752706
XGBOOST	0.840258235	0.80344143	0.752706
Voting Classifier	0.898857805	0.782925215	0.752706

Conclusion and Future research

Conclusion

- Locations are not automatically generated, they are user inputs and that's why data is not clean and there are too many incoorect and missing values.We can skip the 'location' column from our feature list.
- In traditional classification models, the logistic regression model, XGBOOST, and Voting Classifier have higher F1 scores, which means that the performance is better than other models.

Future research

- Embeddings & Text Cleaning.
- Design meta features and extract features using N-gram.
- Use word2vec, glove, BERT, etc. to construct word vectors and compare their performance.

Thank you & Question