Model Cards: Ahmad Soni, Nasiem Ayob, Nicholas Muir, Simon Rosen

# Model Card - Logistic Regression

## Model Details

- Developed and popularized primarily by Joseph Berkson
- Sklearn's logistic regression model(i.e. LogisticRegression). Pretrained to predict whether a tweet relates to a disaster or not.
- Created by: Ahmad Soni, Nasiem Ayob, Nicholas Muir, Simon Rosen

## Intended Use

- To predict when a disaster has occured so that the relevant authorities can be alerted.
- The users of this model should be the relevant authorities that handle natural disasters in the country where the model is tracking.
- This model should not be used instead of existing procedures to report disasters but rather as an aid to them.

## Factors

- Identify terms or phrases that correlate to disasters.

## Metrics

- Classification report.
- Confusion matrix.

## Training Data

- A subset of the original dataset was used as training data.

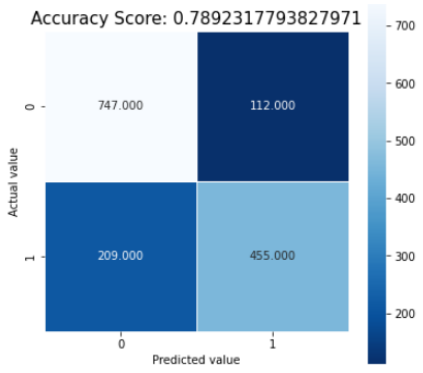
## Evaluation Data

- The data used was a collection of tweets that corresponded to disasters or not disasters. The tweets were manually labeled.
- A subset of the original dataset that was not used in training was used for evaluation.

## Caveats and Recommendations

- Use grid search on a range of parameters to find more suitable parameters for the model.

## Quantitative Analysis



Model Cards: Ahmad Soni, Nasiem Ayob, Nicholas Muir, Simon Rosen

# Model Card - Support Vector Model

## Model Details

- Developed by Vladimir Vapnik at AT&T Bell Laboratories with colleagues.
- Sklearn's support vector model(i.e. SGDClassifier).
- Pretrained to predict whether a tweet relates to a disaster or not.
- Created by: Ahmad Soni, Nasiem Ayob, Nicholas Muir, Simon Rosen

## Intended Use

- To predict when a disaster has occured so that the relevant authorities can be alerted.
- The users of this model should be the relevant authorities that handle natural disasters in the country where the model is tracking.
- This model should not be used instead of existing procedures to report disasters but rather as an aid to them.

## Intended Use

- To predict when a disaster has occured so that the relevant authorities can be alerted.
- The users of this model should be the relevant authorities that handle natural disasters in the country where the model is tracking.
- This model should not be used instead of existing procedures to report disasters but rather as an aid to them.

## Factors

- Identify terms or phrases that correlate to disasters.

## Metrics

- Classification report.
- Confusion matrix.

## Training Data

- A subset of the original dataset was used as training data.

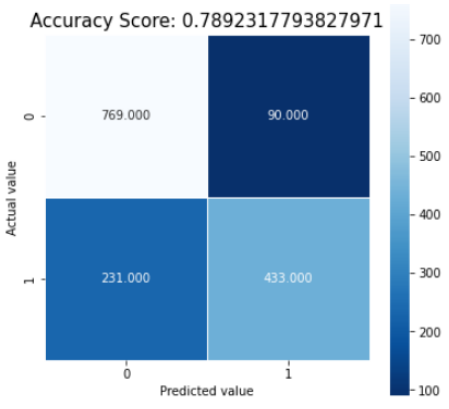
## Evaluation Data

- The data used was a collection of tweets that corresponded to disasters or not disasters. The tweets were manually labeled.
- A subset of the original dataset that was not used in training was used for evaluation.

## Caveats and Recommendations

- Use grid search on a range of parameters to find more suitable parameters for the model.

## Quantitative Analysis



Model Cards: Ahmad Soni, Nasiem Ayob, Nicholas Muir, Simon Rosen

# Model Card - Multinomial Naive Bayes Classifier

## Model Details

- Developed by Vladimir Vapnik at AT&T Bell Laboratories with colleagues.
- Sklearn's Multinomial Naive Bayes classifier(i.e. MultinomialNB ).
- Pretrained to predict whether a tweet relates to a disaster or not.
- Created by: Ahmad Soni, Nasiem Ayob, Nicholas Muir, Simon Rosen

## Intended Use

- To predict when a disaster has occured so that the relevant authorities can be alerted.
- The users of this model should be the relevant authorities that handle natural disasters in the country where the model is tracking.
- This model should not be used instead of existing procedures to report disasters but rather as an aid to them.

## Factors

- Identify terms or phrases that correlate to disasters.

## Metrics

- Classification report.
- Confusion matrix.

## Training Data

- A subset of the original dataset was used as training data.

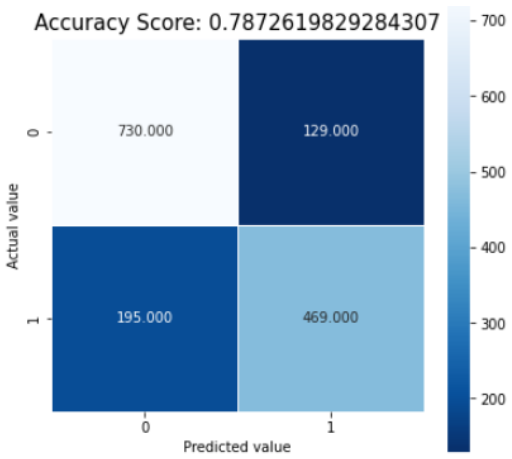
## Evaluation Data

- The data used was a collection of tweets that corresponded to disasters or not disasters. The tweets were manually labeled.
- A subset of the original dataset that was not used in training was used for evaluation.

## Caveats and Recommendations

- Use grid search on a range of parameters to find more suitable parameters for the model.

## Quantitative Analysis



Model Cards: Ahmad Soni, Nasiem Ayob, Nicholas Muir, Simon Rosen

# Model Card - Latent Dirichlet Allocation

## Model Details

- Developed by David Blei, Andrew NG, and Michael I.
- Gensim's multicore lda model.
- Pretrained to extract topics from a corpus created by the tweets.
- Created by: Ahmad Soni, Nasiem Ayob, Nicholas Muir, Simon Rosen

## Intended Use

- To create a corpus using the given tweets.
- To extract the topics from the corpus.
- The next step is for us to classify which topic corresponds to a disaster or not.
- We then classify the tweets based on the probability it belongs to the disaster category or not.

## Factors

- Identify terms or phrases that correlate to disasters.

## Metrics

- Topic coherence score.
- Confusion matrix.

## Training Data

- A subset of the original dataset was used as training data.

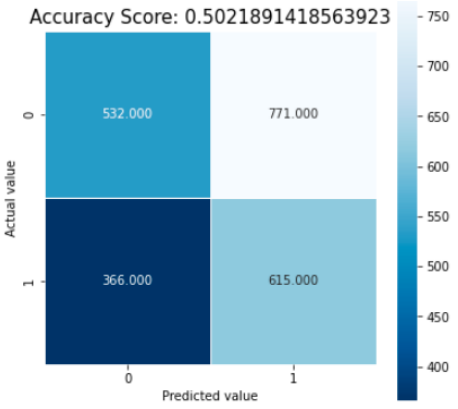
## Evaluation Data

- The data used was a collection of tweets that corresponded to disasters or not disasters. The tweets were manually labeled.
- A subset of the original dataset that was not used in training was used for evaluation.

## Caveats and Recommendations

- Tune the hyperparameters using accuracy as a metric.

## Quantitative Analysis



# Model Card - LSTM

## Model Details

- This classifier was created and trained to predict if a tweet corresponds to a disaster or not.
- LSTM.
- A pre-trained word embedding model was used. This word embedding the GloVe pre-trained word vectors that were trained off a Twitter dataset.
- Tensorflow and SpaCy were used.
- Created by: Ahmad Soni, Nasiem Ayob, Nicholas Muir, Simon Rosen

Intended Use

- To predict when a disaster has occured so that the relevant authorities can be alerted.
- The users of this model should be the relevant authorities that handle natural disasters in the country where the model is tracking.
- This model should not be used instead of existing procedures to report disasters but rather as an aid to them.

Factors

- Identify terms or phrases that correlate to disasters.

Metrics

- Confusion matrix.
- Accuracy score

Training Data

- The data used was a collection of tweets that corresponded to disasters or not disasters. The tweets were manually labeled.
- A subset of the original dataset was used as training data.

Evaluation Data

- The data used was a collection of tweets that corresponded to disasters or not disasters. The tweets were manually labeled.
- A subset of the original dataset that was not used in training was used for evaluation.

Caveats and Recommendations

- Some of the data used for training was of low quality and contained nonsensical strings of characters that was most likely caused by the data being corrupted. Better training could be achieved by better data cleaning.

Quantitative Analysis