CS 450
Semester Project
Natural Language Processing

Table of Contents

Contact information

Cannon Bray, Jashon Newlun, Saffra Parks, Avery Robbins

Kaggle competition:
Real or Not? NLP with Disaster Tweets - Predict which Tweets are about real disasters and which ones are not.

https://www.kaggle.com/c/nlp-getting-started/data

Introduction

The scope of this project is to determine which Tweets contain information about real disasters and which ones don't. The data we used is from a current Kaggle competition. This is an interesting problem because it allows us to learn and explore Natural Language Processing methods while working with fairly simple data. We used R Studio as our platform for this project.

Our hope is that we will be able to apply sentiment analysis, confusion matrices, and traditional Machine Learning algorithms to correctly classify texts with at least 70% accuracy.

At the completion of the project we were able to achieve 74.8% accuracy, exceeding our goal.

Pre-processing

The data provided by Kaggle was pre-split into test and train sets. The test set contained columns titled "id", "keyword", "location", and "text". The train set also included a target column. There are over 7,000 observations.

The "text" column contains the actual text of each Tweet and is the column we are mostly concerned with.

Example of what the data looks like:

```
> head(train)
# A tibble: 6 x 5
     id keyword location text                                                target
  <int> <chr>   <chr>    <chr>                                                <int>
1     1 NA      NA       Our Deeds are the Reason of this #earthquake May ALL~      1
2     4 NA      NA       Forest fire near La Ronge Sask. Canada                     1
3     5 NA      NA       All residents asked to 'shelter in place' are being ~      1
4     6 NA      NA       13,000 people receive #wildfires evacuation orders i~      1
5     7 NA      NA       Just got sent this photo from Ruby #Alaska as smoke ~      1
6     8 NA      NA       #RockyFire Update => California Hwy. 20 closed in bo~      1
```

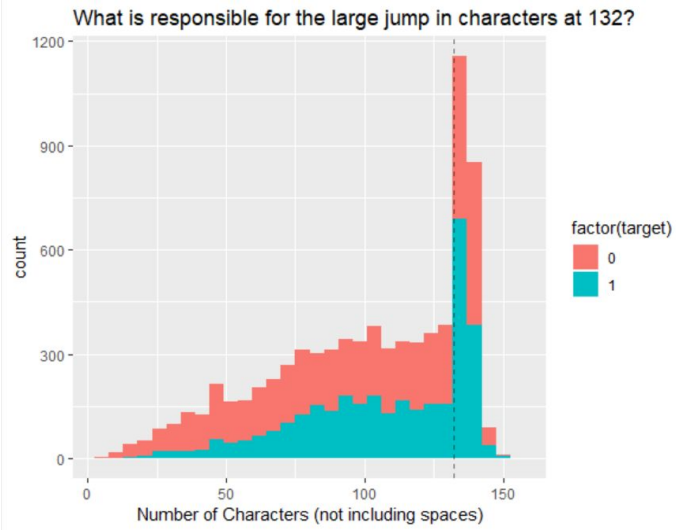
Algorithms and Computation

Initial exploration of the data consisted of several sections 1-8, as explained below.

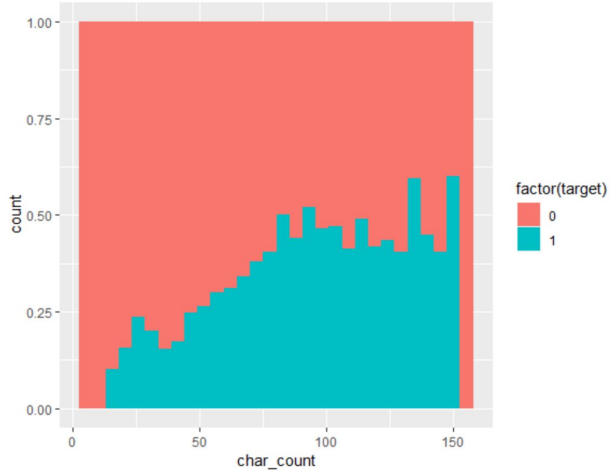
1. Ratio of Target Class: 57% of the provided Tweets do not reference a disaster, and the remaining 43% do reference a disaster. The ratio is fairly even, which is a good thing.

2. Additional Features: 61 Tweets have both a location and keyword associated with them. We can determine that these extra features will likely not be very helpful in our analysis.

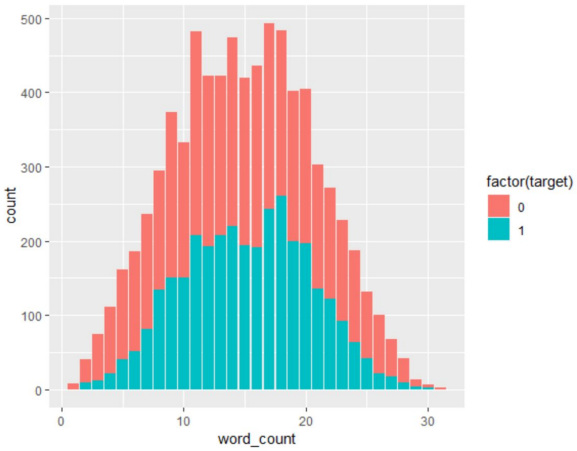
3. Character Length: The count of characters seems to have a fairly consistent positive trend, until the 132 character mark. Additional exploration would be necessary to determine what causes the large jump.



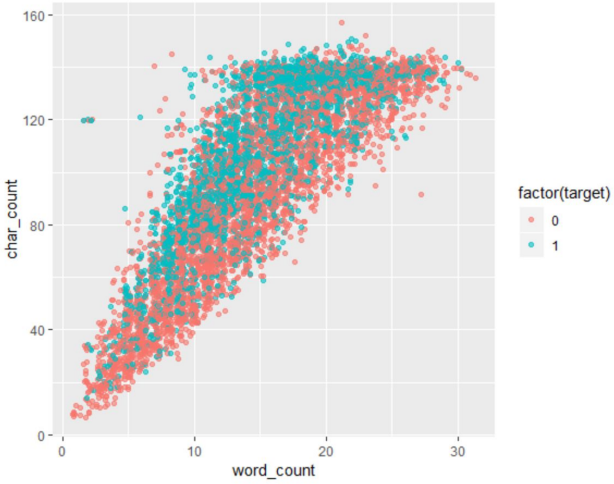
The number of characters expressed through proportional stacked bar charts is another way to consider this factor. If the number of characters in a Tweet is non-predictive, we should see a perfect 50/50 split among all Tweet lengths. After creating the graph, it looks like longer Tweets are more likely to be an actual disaster Tweet. At this point we can note that the length of a Tweet will likely be helpful for a future model.



4. Word Count: The distribution of word count is the same for Tweets related to an actual disaster and those that aren't, in short, word count alone is not a predictor of Tweet classification.



5. Combination of Word Count and Character Count: Next, we wanted to determine if the combination of word count and character count may have more predictive power than either of them alone. It does appear that there could be an interesting relationship between the two variables. We will make a note that perhaps both variables should be used in a future model.



6. Hashtags, Numbers, and Capital Letters: Similar to word count, none of these seem to be very powerful predictors of Tweet classification.

7. Link: As it turns out, links appear to be decent indicators of Tweet classification, with Tweets related to real disasters tending to contain a link.

	contains_link	target	n
	<dbl>	<int>	<int>
1	0	0	2543
2	0	1	1099
3	1	0	1799
4	1	1	2172

8. Finally, we looked at how many Tweets contain the word “breaking”. There are such a small count of Tweets (14) that contain the word that it’s likely to not contribute much to the predictive power.

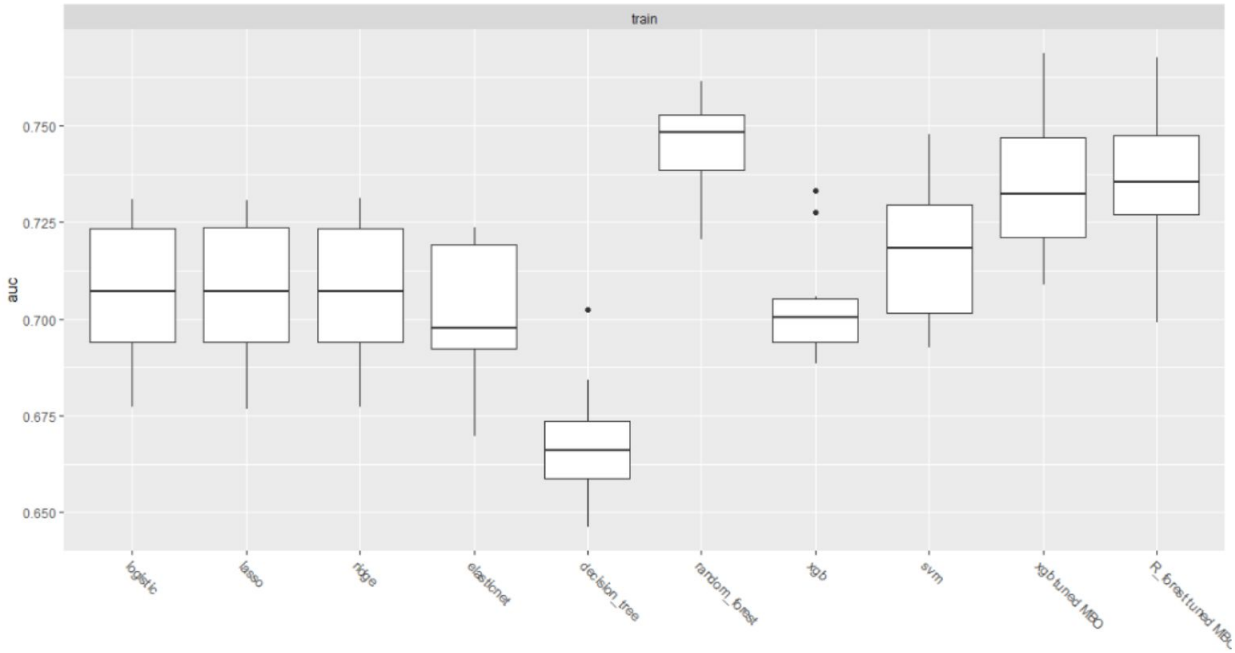
After exploring each of these variables, we combined them into a single dataset, normalized the columns, and split into train and test data.

This is an example of what the normalized data now looks like:

```
> head(data_training)
# A tibble: 6 x 6
  breaking contains_link capital_count number_count hashtag_count target
     <int>         <dbl>         <dbl>        <dbl>         <dbl>  <int>
1        0             0        -0.750       -0.665        -0.404      1
2        0             0        -0.846        0.967         0.497      1
3        0             0        -0.654       -0.665         1.40       1
4        0             0        -0.0778      -0.0126        2.30       1
5        0             0        -0.750       -0.665        -0.404      1
6        0             0        -0.846       -0.665        -0.404      1
```

An additional technique that we implemented is sentiment analysis which includes correlation for each of our new columns.

After the exploration, we used the Area Under the Curve (AUC) metric to determine which algorithm appears to have the best predictive power. Random Forest has the highest average AUC score so we picked that as the model to tune.



Conclusion

Once the data was created, processed, and normalized, we applied a Random Forest to it and began tuning the parameters. The accuracy score is currently 74.8%.
The ability to predict relative disaster tweets has application in several areas. One area that is particularly worth noting is how it can allow disaster relief organizations and news agencies to more programmatically monitor Tweets during times of actual emergencies.

This project has been interesting because it has given all of us exposure to NLP techniques, as well as deepened our knowledge of data preprocessing. It holds true to the idea that the majority of work a data scientist does is in preparing the data to be worked with.

No particular ethical issues seem to be apparent with this work.