Cyberbullying Tweets

Will Sharpe

Data

- 47,692 tweets stored as comma-separated-values (.csv)
- Table contains just two columns: tweet_text and cyberbullying_type
- There are 6 classes that each tweet may be classified as:
  - Religion
  - Age
  - Gender
  - Ethnicity
  - Not cyberbullying
  - Other cyberbullying
- This is a multiclass and not multilabel classification

Approach and Techniques

- To cast a wide net in techniques tried, thereby trying many but not going too deep into any one
- All models received data that was vectorized by sklearn's CountVectorizer
  - stop_words = "english"
  - 47,692 rows by 59,970 (except for n-grams)
- All models received data that was split by sklearn's train_test_split the same way:
  - 70% training 30% testing
  - Same random state number: 132

Overview - 5 Main Sections

1. Initial phase of testing with no preprocessing, only looking over data for validity
    a. Check for missing values
    b. Check for types of data
    c. Check for a near equal amount of representation for each of the 6 classes
2. Cleaning and modifying the data before running some of the same tests from before
3. Unigrams - Bigrams - Trigrams
4. Building a Neural Network
5. Cross Validation of top performer

Initial Phase

The first batch of testing 6 models used:

1. Logistic Regression: 0.8364
2. Random Forest: 0.8265
3. Support Vector Machine: 0.8254
4. Decision Tree: 0.8079
5. Multinomial Naïve Bayes: 0.7807
6. K-nearest Neighbor: 0.618

Just the accuracy is provided here. Each model's accuracy is within a small percentage of all other fields.

Preprocessing

- In this second approach 3 of the 6 previous models were selected
  - The top performer, a middle performer and the worst performer
- The approach this time was to manipulate the data more than we have so far
- All capitalizations were removed for lowercase to make more uniform and simplified
- Punctuations were removed to be replaced by a space
- These two steps were done in an effort simplify data and improve models.

Preprocessing

Logistic Regression: 0.8299 previous of: 0.8364

Decision Tree: 0.8014 previous of: 0.8079

K-Nearest Neighbor: 0.6665 previous of: 0.6818

Issues for Sections 3 and 5

- The volume of data created issues when attempting to use the whole dataset in N-grams and Cross Validation
- The kernel would die due to RAM being overloaded
- Attempted but failed to resolve:
  - Use of different IDE's, different computers and modifying allotted RAM
  - Google Colab and Google Colab Pro
- Resolution:
  - Seperate a random sample of 10,000 tweets from the data
  - Random_state of 132 used consistently
  - Random sample was checked for balance of classes

Unigrams - Bigrams - Trigrams

- N-grams: made 1, 2 and 3 word tokens
  - 10,000 rows by 217,279 columns
- Bigrams: only made of 2 word tokens
  - 10,000 rows by 93,392 columns
- 3 of 6 models from initial section used for testing:
  - K-nearest Neighbors
  - Decision Tree
  - Logistic Regression

Unigrams - Bigrams - Trigrams Results

N-grams

- K-Nearest Neighbors: 0.3987
- Decision Tree: 0.7990
- Logistic Regression: 0.8223

Bigrams

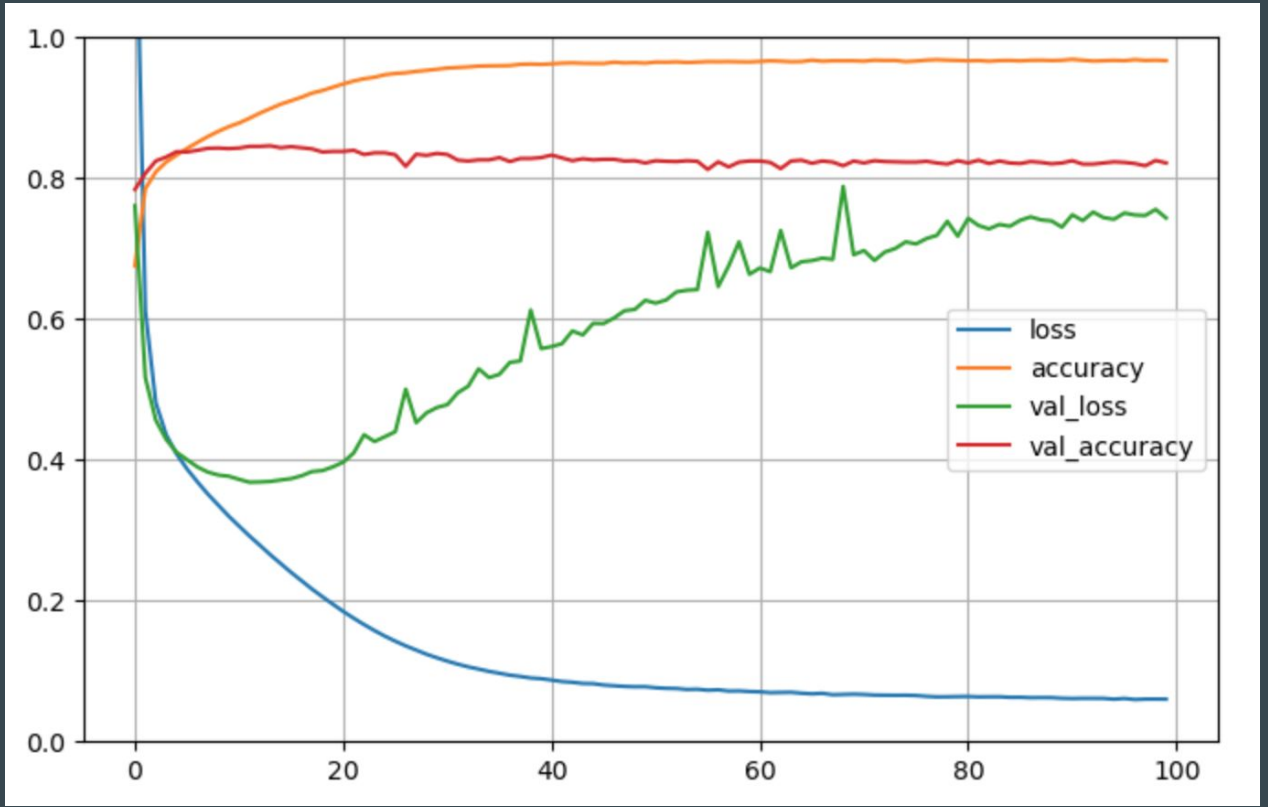
- K-Nearest Neighbors: 0.2717
- Decision Tree: 0.6100
- Logistic Regression: 0.5907

Neural Network

- 70% training 30% testing
  - Out of the training group 2,500 was put into a validation group
- Set 1st hidden layer to 300 nodes
- Set 2nd hidden layer to 100 nodes
- Initial training: 100 epochs
- Second attempt, new model, training: 16 epochs

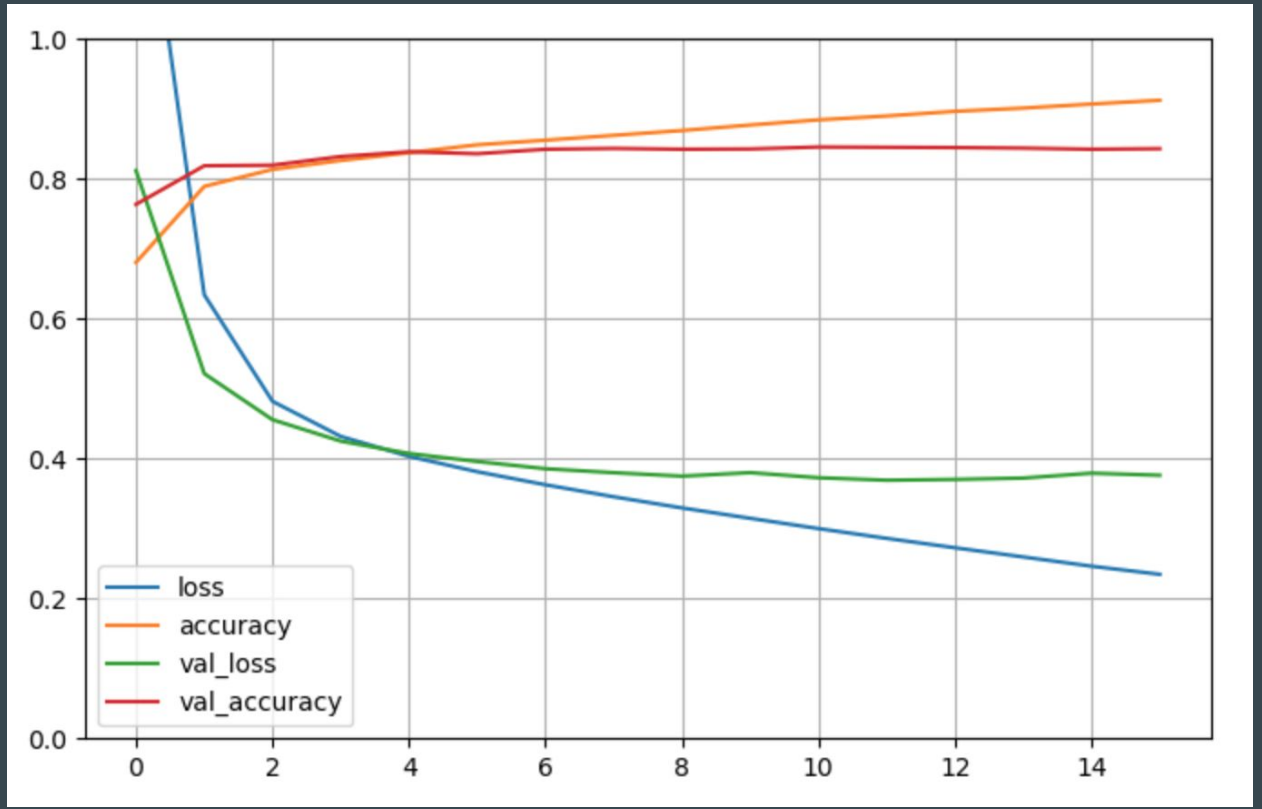
100 epochs

Training Accuracy: 0.9758
Testing Accuracy: 0.8228
Validation Accuracy: 0.8216



16 epochs

Training Accuracy: 0.9209
Testing Accuracy: 0.8402
Validation Accuracy: 0.8432



Cross Validation

- Only Logistic Regression was explored for cross validation
- 10,000 taken out of entire dataset
- Cross validation set to 10
- Test Score Accuracies for each of the 10
  - 0.809; 0.837; 0.840; 0.832; 0.815; 0.833; 0.823; 0.825; 0.827; 0.833;
  - Mean of scores: 0.83

Logistic Regression Cross Validation Results

	Precision	Recall	F1-Score	Support
Age	0.98	0.98	0.98	1,668
Ethnicity	0.98	0.97	0.98	1,713
Gender	0.91	0.83	0.87	1,634
Not Cyberbullying	0.57	0.56	0.56	1,639
Other	0.59	0.68	0.63	1,642
Religion	0.97	0.94	0.95	1,704
Accuracy			0.83	10,000

Source of Data used for this Project

- J. Wang, K. Fu, C.T. Lu, “SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection,” Proceedings of the 2020 IEEE International Conference on Big Data (IEEE BigData 2020), December 10-13, 2020.