Detection of Cyberbullying and Threats Using LSTM Networks and Multilayer Perceptrons

Alex Donovan, Ben Auslin, Trishiet Ray

Stanford Computer Science

Motivation

- **47%** of adolescents say they have experienced cyberbullying.
- Cyberbullying is one of the main contributors to **teen suicides**.
- Detecting textual threats on social media will recognize cyberbullying early and alleviate the negative consequences.

Problem Definition

- The EmoBank data set contains **10,000 cross-domain sentences, each annotated with averaged values for 3 human ratings**
- Sentences were annotated according to the **Valence-Arousal-Dominance** scheme, with a score from 1 to 5 for each of the three categories: V, A, and D.

Goals

1. Predict Valence, Arousal, and Dominance for any given sentence input.
2. Classify **sentences** as threats or non-threats with high accuracy and low average error.

Challenges and Solutions

Text is sequential, so the order of the words is important to the sentiment	Use RNN / LSTM and MLP that analyze sequential data well
The data was averaged from 1 to 5 and had mostly neutral examples	Normalized data to better train RNN, penalize extreme errors
There are features specific to the platform that cannot be accounted for in any single model (replies, emoticons, photo tagging, etc.)	Robust processing of sentences, punctuation

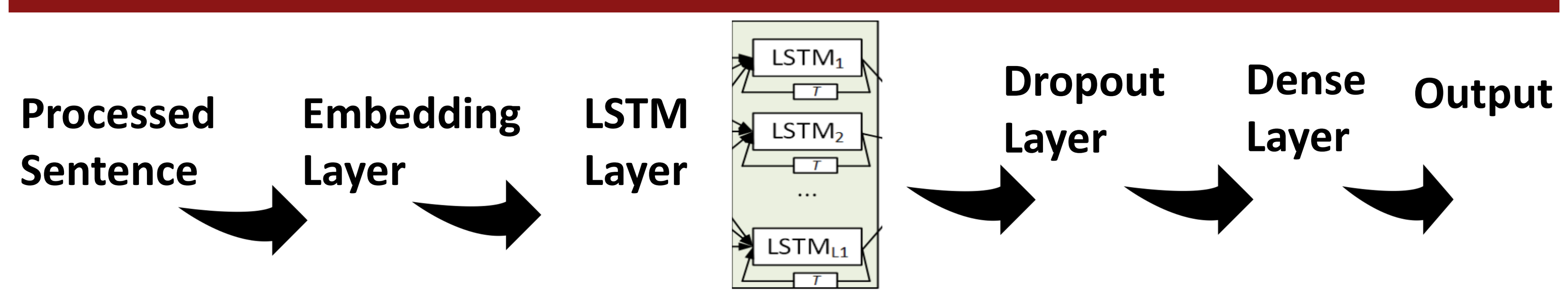
Approach and Model Implementation

We designed 3 models to predict V, A, D: **Multiple Linear Regression, LSTM,** and **MLP.** We preprocessed the sentences using a bag of words model, then architected and tuned each of our models with grid search and compared results.

Multiple Linear Regression

Our baseline was a linear regression along each emotional dimension which allowed us to play with different feature extractors (word counts, indicators, varying treatments of punctuation and numeric text, word2vec). This model was the simplest and also the least accurate.

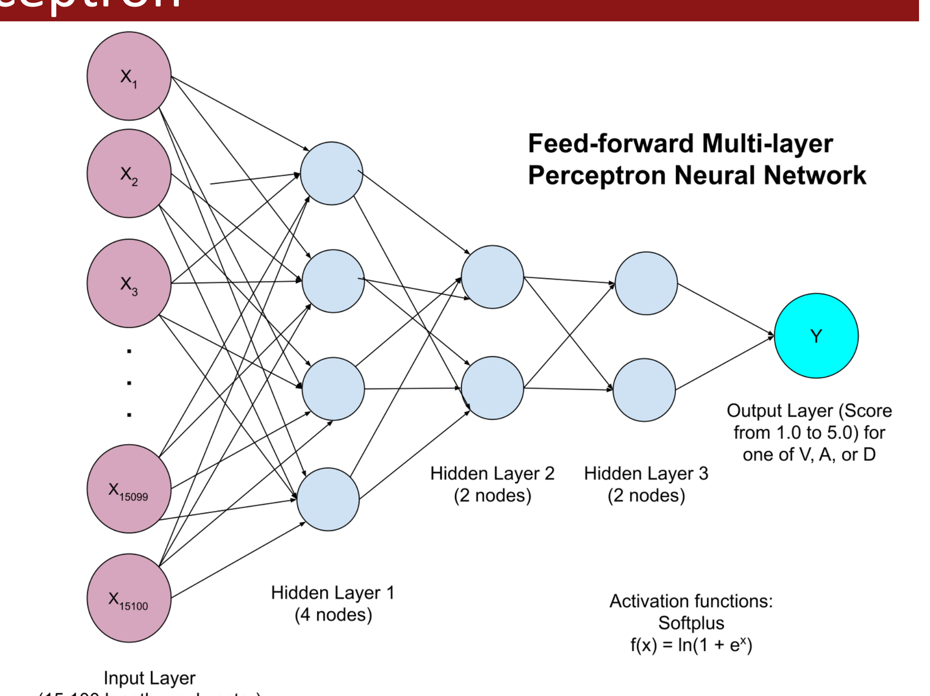
Long Short-Term Memory Recurrent Neural Network



1. Tune LSTM / RNN on validation set 2. Train using MSE loss, Adam optimizer
3. Output prediction for valence, arousal, dominance

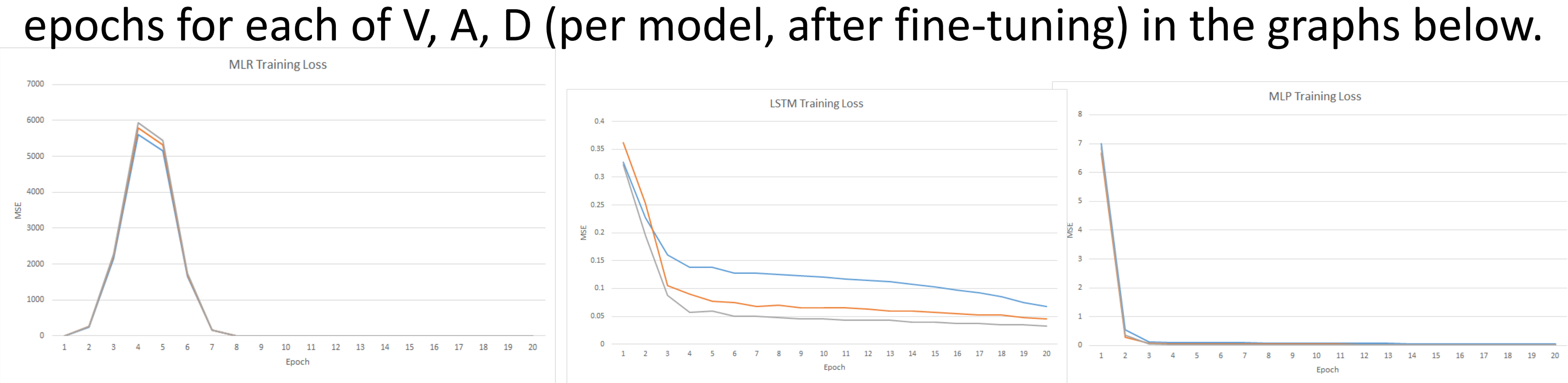
Multilayer Perceptron

Feed-forward MLP is well suited for regression problems with given inputs The final network had 3 hidden layers of 4, 2, and 2 nodes, with softmax activations. We then defined a secondary single layer network that takes in the V, A, D predictions from the MDP and the text matrices, and predicts whether a comment is a threat.



Results

We had an 80-20 train-test split. We present the training losses per number of epochs for each of V, A, D (per model, after fine-tuning) in the graphs below.



Analysis

Model	Valence	Arousal	Dominance
Linear Regression	MAE: 0.276 Accuracy: 66.48%	MAE: 0.216 Accuracy: 74.00%	MAE: 0.187 Accuracy: 81.92%
LSTM	MAE: 0.253 Accuracy: 69.23%	MAE: 0.202 Accuracy: 78.54%	MAE: 0.166 Accuracy: 84.0%
MLP	MAE: 0.229 Accuracy: 72.48%	MAE: 0.194 Accuracy: 79.38%	MAE: 0.164 Accuracy: 85.64%

True Positives: 30	False Positives: 286
False Negatives: 18	True Negatives: 1679

Accuracy: 84.90%
Precision: 9.49%
Recall: 62.50%
F-1: 0.165

We report the best mean absolute error as well as accuracy for V, A, and D per model. A prediction for V, A, or D is considered correct if it is within ± 0.3 of the actual averaged human rating out of 5. Finally, for threat detection, we defined a comment as a potential threat if the actual dominance >= 3.2 and valence <= 2.5 and classify using MLP results.

Analysis

- The **MLP results were best**, followed by the LSTM, and the Linear Regression.
- **MLP and LSTM outperform Linear Regression** since they use more sophisticated features and are better suited for sequential data. MLP outperformed LSTM because it was easier to optimize. The LSTM network had a dropout layer but still can be optimized more. We expect LSTM to outperform MLP when the model is improved, and there may be overfitting.
- All models produced relatively similar results; improvements in metrics are gradual but universal between models. **Dominance was the easiest to classify (0.229 MAE), followed by arousal (0.194 MAE), followed by valence (0.164 MAE).**
- **Arousal had little effect** on threat detections but can predict sexual language
- Our models were very accurate at predicting the valence, arousal, and dominance within 0.3. The predicted VAD wasn't enough to classify; a secondary network was needed. There were only 48 potential threats out of 10,000 texts (in the threshold range).
- True Positive: **"If he refuses, threaten him."**
  - Predicted VAD: (3.01, 3.08, 3.20); Actual VAD: (2.44, 3.33, 3.44)
- False Negative: **"It was not an attractive face..."**
  - Predicted VAD: (3.34, 3.10, 3.06); Actual VAD: (2.33, 3.00, 3.22)
    - Resulting from over-predicting V and under-predicting D (lack of training data for these extremities), does not deal with negations like "not" well
- True Negative: **"Noriega is close to Castro and may once have been his agent."**
  - Predicted VAD: (2.89, 2.87, 3.08); Actual VAD: (3.00, 2.89, 3.11)
    - Models good at predicting true negatives with non-extreme VAD around 3

Future Work

- Train on data sets with more variation in VAD values, especially on themed/domain-specific cyberbullying and platform specific (Twitter, Youtube) datasets
- Identify features other than VAD that correlate with cyber harm
- Automated sentiment-based cyberbully detection, warning systems

Stanford University