# DRy Model

*Team Lilac*: Hai Le, Enan Srivastava

# Overview

- We wanted to analyze whether or not certain text was Democrat or Republican.
- We acquired 18 Congressional Session Records and ran a LSTM Classifier.
- While we were not able to hit our targeted accuracy, upon hyperparameter tuning and stopword elimination performance will improve greatly.

# Dataset

We obtained the dataset from the **Stanford's Social Science Data Collection**: https://data.stanford.edu/congress_text

The dataset contains speeches given on the floor of each chamber, the House and the Senate, from *the 97$^{th}$ to the 114$^{th}$ Congress*.

Train set

D 287,292

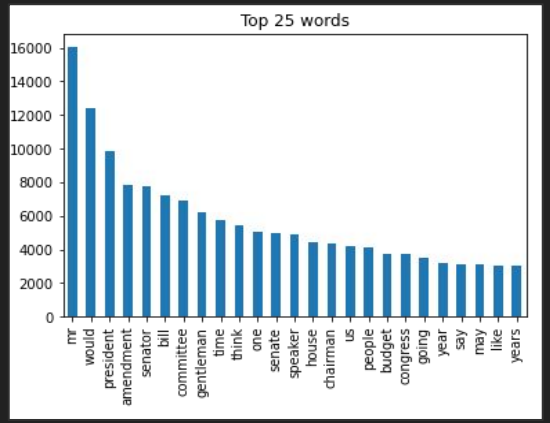
R 257,822

# Preprocessing

- Data was read in and concat from 18 csvs
- Only text >1000 *characters* was kept
- Only words were kept and then *tokenized* to list
- Token Map was generated, single occurrences of tokens were deleted
- Data was encoded via *token map* and split 70/25/5 to train/test/val
- When put into the data loader, the max size is 2000 words

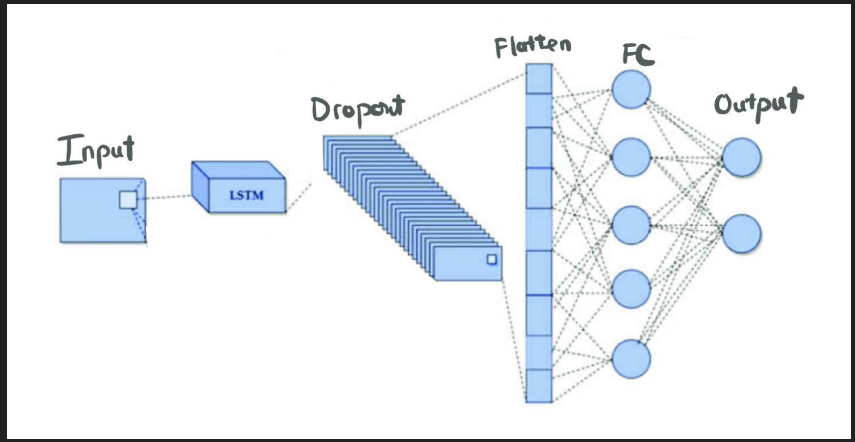
# Token Map

- A unigram token map was generated via only the training data
- <<PAD>> and <<UNK>>  tokens are first 2
- Single instances of tokens were misspellings/corrupt data and were omitted



# Model

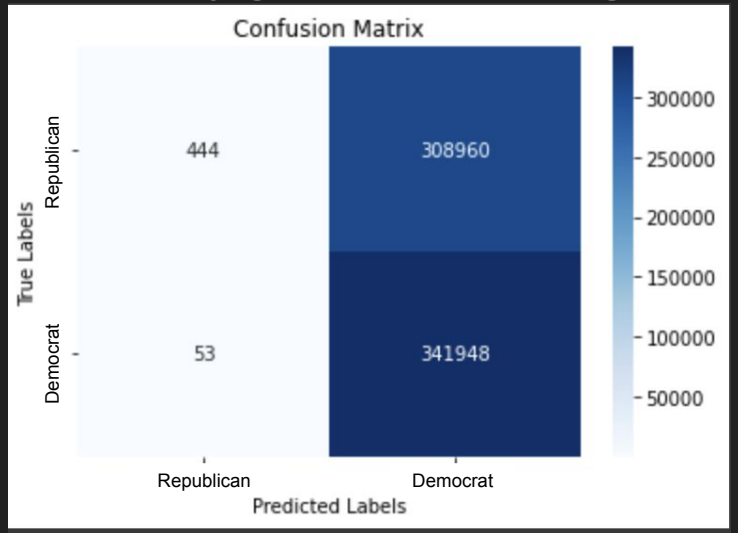
- We Chose an LSTM model due to the forgetful nature of Congress.
- The Data is input by a Batch Size of 512, Text Size of 2000, and a Target Size of 1
- We have a Fully Connected layer and Dropout Layer (dropout rate=0.5, input_size=300).
- Data is not shuffled in order to add time dimension in future

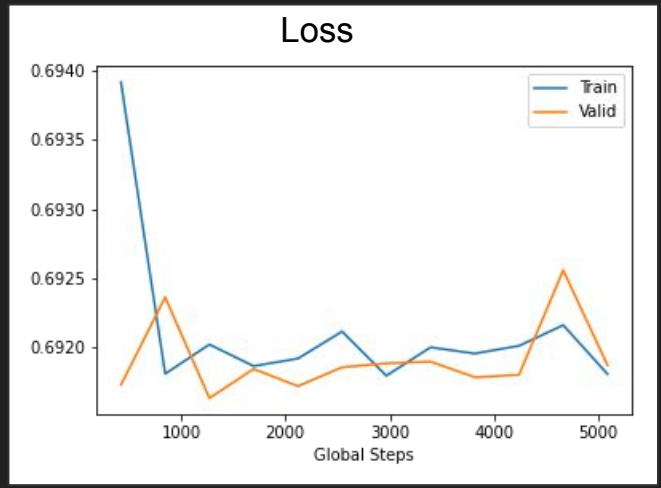


# Performance

Learning rate=.007, epochs=5, batch_size=512

Extremely good at predicting Democrat speeches



# Future Improvements

- Implement file based loading to train on longer than 2000 word speeches
- Running hyperparameter tuning
- Truncate token map and classify stopword tokens to <<STP>>
- Implement an RNN layer or TF-IDF behind the LSTM to give general language processing
- Add in a year dimension to the model during training and use