Predicting Political Affiliation from Tweets

Jorge Cordero, Eddie Sun, Zoey Zhou
{icoen, eddiesun, cuizy} @stanford.edu

Motivation

Knowledge of public opinion is a powerful tool for governments and companies to have, but is difficult to measure on a large scale. Currently, opinion polls are used, but these only sample a small portion of the population. However, deep learning can potentially be used on social media, such as Twitter, to gauge public opinion on a large scale.

For this project, we develop a classification algorithm to predict political party affiliation from tweets.

Data & Features

We used Kyle Pastor's "Democrat vs. Republican Tweets" dataset from Kaggle [1]. The dataset contains 86,460 labeled tweets from the 200 most recently elected congressmen (49% Democrat, 51% Republican). The dataset was split 80/10/10 between train/dev/test.

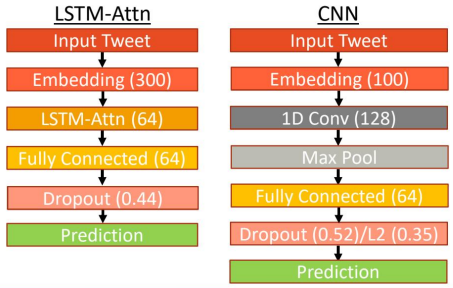
Sample Republican Tweet

Sample Democrat Tweet

To preprocess the raw tweets, we removed hyperlinks, symbols, punctuation, and twitter handles. We then padded each tweet to the maximum tweet length and embedded each tweet using the Keras embedding layer.

Model

We created and optimized two models: a LSTM-Attention and a CNN model. We also experimented with other types of recurrent network types. The models were coded in Keras with Tensorflow backend.



Results

Model	Train Acc (%)	Valid Acc (%)	Test Acc (%)
CNN*	96.9	84.5	82.4
LSTM-Attn*	91.2	85.2	83.9
LSTM	91.3	84.6	83.3
GRU	88.6	84.8	83.6
LSTM (Bi-Dir)	90.0	84.6	83.3
GRU (Bi-Dir)	90.6	84.4	83.6

Confusion Matrix

LSTM-Attn Model	Predicted Dem.	Predicted Rep.
Labeled Dem.	3442 (39.9%)	787 (9.1%)
Labeled Rep.	608 (7.0%)	3808 (44.0%)

*performed random hyperparameter search

Discussion

Model Performance

- LSTM-Attn. performs better than CNN due to its ability to learn long-term dependencies
- CNN still performs remarkably well for having such a simple architecture

Model Confidence (LSTM-Attn)

Predicted Democrat with high confidence ($p = 0.99$):

Predicted Republican with high confidence ($p = 0.04$):

Unable to discern ($p = 0.5$):

Error Analysis (LSTM-Attn)

- Tweets with no political content, short tweets (3-6 words), and tweets with multiple people mentioned were commonly misclassified

Future Work

- **Character-level Embeddings:** Twitter language is often quite different from standard English, so character-level embeddings have been suggested to work better for tweets [3].
- **Different NN Architectures:** RCNN's [4], Bi-directional LSTM's with max-pooling [5], and CNN-Attention [6] show promise in sentence classification tasks.
- Further hyperparameter tuning, text preprocessing, use existing embeddings (i.e. GloVe) for unseen words.

References

[1] Pastor, Kyle. "Democrat Vs. Republican Tweets." Kaggle: 27 May 2018.
[2] Kim, Yoon. "Convolutional neural networks for sentence classification." arXiv preprint arXiv:1408.5882 (2014).
[3] Vosoughi, Soroush, Prashanth Vijayaraghavan, and Deb Roy. "Tweet2vec: Learning tweet embeddings using character-level cnn-lstm encoder-decoder." ACM, 2016.
[4] LAI, S.; XU, L.; LIU, K.; ZHAO, J.. Recurrent Convolutional Neural Networks for Text Classification. AAAI Conference on Artificial Intelligence, North America, Feb. 2015
[5] Zhou, Peng, et al. "Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling." arXiv preprint. (2016).
[6] Elbayad, Maha, Laurent Besacier, and Jakob Verbeek. "Pervasive Attention: 2D Convolutional Neural Networks for Sequence-to-Sequence Prediction." arXiv preprint (2018).