TROLLS ON TWITTER

Gene Tanaka
gtanaka@stanford.edu

Tyler Consigny
tconsig@stanford.edu

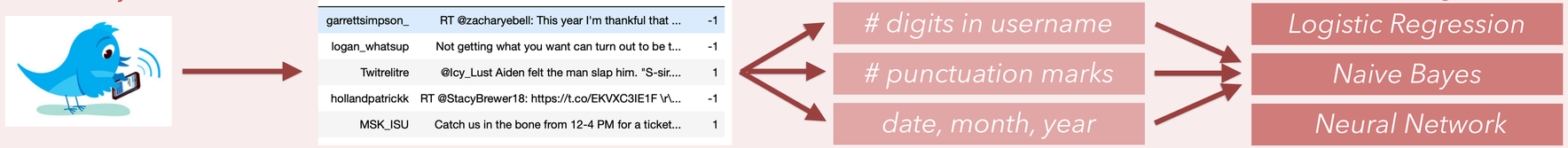
Motivation

- Divisive power of 'troll' tweets increasing at a rapid pace
- With targeted tweets that are intended to further divide opinions, integrity of our democracy and free elections are threatened.
- Essential to be able to differentiate real vs. 'troll' tweets
- We developed classifier to predict whether a tweet is real or not.
- After iterative process, neural network produced best results

Challenges

- *Productive Feature Extraction*: Difficult to find patterns found in the majority of troll tweets as they target different groups
- *Evaluation*: Percentage accuracy was not enough, needed to take into account false negatives and positives as well [f1 + confusion matrix]
- *Datasets:* No datasets of both troll and real tweets, so it seemed there were some implicit similarities that couldn't be avoided in the two separate datasets we had to use

Pipeline



name	text	truth
garrettsimpson_	RT @zacharyebell: This year I'm thankful that ...	-1
logan_whatsup	Not getting what you want can turn out to be t...	-1
Twitrelitre	@Icy_Lust Aiden felt the man slap him. "S-sir....	1
hollandpatrickk	RT @StacyBrewer18: https://t.co/EKVXC3IE1F \r\...	-1
MSK_ISU	Catch us in the bone from 12-4 PM for a ticket...	1

Approaches

Logistic Regression:

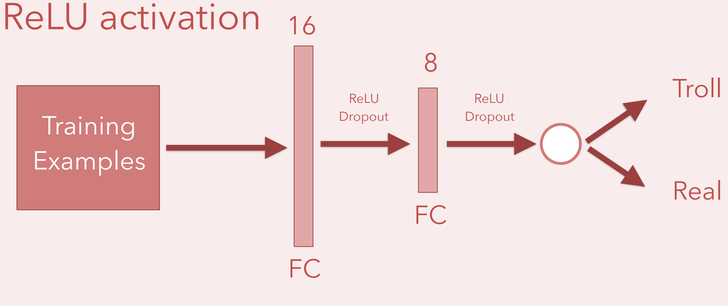
- Used features such as word counts and n-grams as baseline

Naive Bayes:

- Added numerical features such as length of username, # punctuation marks, # special characters, date, month, year

Neural Network (MLP):

- Feedforward, 2 hidden layers
- Dropout, L2 Regularization to reduce overfitting
- ReLU activation



Results & Analysis

	Train Error (n = 50000)	Test Error (n = 2000)
Logistic Regression	0.05912	0.275
Naive Bayes	0.03218	0.177
Neural Net (First)	0.00174	0.1905
Neural Net w/ Dropout, L2	0.0092	0.0795

Fig. A: Percent Error for Each Model

Fig. B: Loss/Accuracy vs. Epoch

n = 2000	Predicted Real	Predicted Troll
Actual Real	963	37
Actual Troll	122	878

Fig. C: Confusion Matrix for Neural Net w/ Dropout, L2

Remarks:

- NN without Dropout and L2 severely overfit
- Confusion matrix shows that improved NN correctly classified 96.3% of non-troll tweets, but 87.8% of troll tweets

References

[1] Ollie. (2018, July 31). Why We're Sharing 3 Million Russian Troll Tweets. Retrieved from https://fivethirtyeight.com/features/why-were-sharing-3-million-russian-troll-tweets/.
[2] Vikas. (2018, February 15). Russian Troll Tweets. Retrieved from https://www.kaggle.com/vikasg/russian-troll-tweets.
[3] Kudugunta, S., & Ferrara, E. (2018). Deep neural networks for bot detection. Information Sciences, 467, 312-322. doi:10.1016/j.ins.2018.08.019
[4] Zheng, A. (2015). Evaluating Machine Learning Models. Retrieved from https: //www.oreilly.com/ideas/evaluating-machine-learning-models/page/5/ hyperparameter-tuning