Public Forgetfulness? Reputation Visualization in the Wake of #MeToo

Michael Cai - mcai88@stanford.edu | Alex Goodman - alexgood@stanford.edu

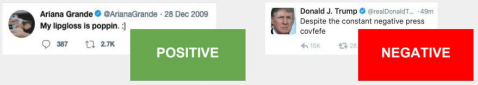
Content Warning: Sexual Assault

Introduction

During the ongoing #MeToo movement that reached a fever pitch last summer, many celebrities rightfully had their reputations drastically damaged by accusations of sexual assault. In this project, we built a deep convolutional neural network sentiment analyzer to classify tweets as either positive or negative in order to visualize how the conversations about accused celebrities on Twitter have changed as a result of those accusations. We were curious to see if public opinions about these people have bounced back since the accusations surfaced.

Problem

Tweet Classification: Given a tweet, classify it as either positive (1.0) or negative (0.0). For example...



Data Extraction and Visualization: Run classification on tweets and visualize distribution of positive vs negative tweets

Dataset

Sentiment140 Dataset: Dataset containing ~1.6 million tweets extracted by Stanford PhD students in 2009 labeled as positive or negative. Train set: ~1.5 million tweets, dev set: 100,000 tweets, test set: 500 tweets.
Link: http://www.sentiment140.com/

GloVe Twitter 27B Vectors: Pretrained word embeddings extracted from over 2 billion tweets containing 27 billion tokens. Vocabulary includes 1.2 million words. For this project, we used the 100d word vectors in to train our model.
Link: https://nlp.stanford.edu/projects/glove/

Model

Deep Convolutional Neural Network: We fine tuned the deep CNN architecture Michael designed in his CS224N final project, which was inspired by the work of Facebook AI's Conneau et. al.

Evaluation Metrics: Our model was evaluated using via the F1 score. Our top model achieved a test set F1 score of **0.853**

Hyperparameters: We primarily tuned the weight decay for L2 regularization and early stopping

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

$$L2 = L - \lambda||(w)||^2$$

Model	Top Dev Set F1 Score
Weight Decay = 1.0	0.730
WD = 0.1	0.668
WD = 0.0001	0.807
WD = 0.00001	0.805
Early Stopping, no WD	**0.817**

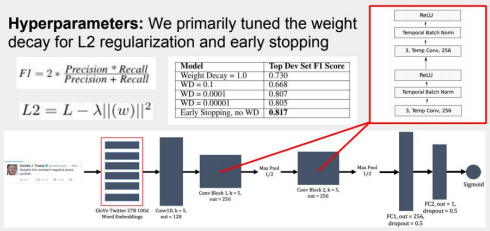


Figure 1: A visualization of the neural network we used. The architecture of the "Conv Block" layers is shown in the diagram to the top right

Analysis

Overfitting Model: Our model tends to overfit the train set - weight decay helps solve the problem of overfitting but doesn't improve performance. We solved overfitting problem with early stopping

Reputations: Tweets about celebrities tended to be overall more positive than negative regardless of timing relative to the accusations. Some celebrities' reputations recovered better than others.

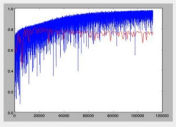


Figure 2: Model train (blue) and dev (blue) F1 scores plotted vs iterations showing overfitting

Results + Discussion

Overall, we found mixed results without a clear trend in how public opinion changed from immediately after accusations to today. For some perpetrators, like James Franco, public opinion has bounced back, while for others like Aziz Ansari, it has gotten worse. We were expecting a more obvious downward trend for all accused celebrities, but that doesn't seem to be the case. We also expected more negative tweets overall.

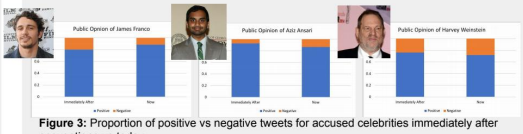


Figure 3: Proportion of positive vs negative tweets for accused celebrities immediately after accusations vs today

Future Directions

In the future, we would like an opportunity to explore more powerful, deeper neural network architectures such as Google's BERT architecture for this problem. In addition, we would do more to filter extracted tweets for retweets or nonsense tweets so that our numbers are more reflective of reality.

References

[1] Y. Kim, "Convolutional neural networks for sentence classification," Sep 2014. 4
[2] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, "Very deep convolutional networks for text classification," Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Jan 2017.
[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," oct 2018.
[4] S. M. Mohammad, S. Kiritchenko, and X. Zhu, "Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets," 2013.
[5] J. Barnes, R. Klinger, and S. S. I. Walde, "Assessing state-of-the-art sentiment models on stateof-the-art sentiment datasets," Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2017.
[6] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," 2009.
[7] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.
[8] "Pytorch."