“Am I the Asshole?”: A Deep Learning Approach for Evaluating Moral Scenarios

Ivy Wang Department of Computer Science
Stanford University
wangivy@stanford.edu

1 Introduction

“Am I the Asshole” (AITA) is a reddit forum where users describe a moralistic scenario that they were involved in. Other users then make a moralistic judgement about whether the original poster is an “asshole” based on the text provided.

The project will investigate whether deep neural methods can be used to predict moralistic judgements. The task is to predict the moralistic judgement made by the majority of other users based on the original post. It will receive the post title and post content as input, and output whether the final judgement of the other users are “Not the asshole” (NTA), “No assholes here” (NAH), “Everyone sucks here” (ESH), or “You’re the asshole” (YTA).

2 Related work

This project builds upon pre-existing work that uses deep learning methods to perform attribute prediction tasks. For example, Tigunova, et. al.[1] combines NLP techniques such as word embeddings[2], attention mechanisms[3] and CNNs[4] to obtain a “Hidden Attribute Model” (HAM) that is able to make predictions of a user’s attributes (such as profesion, gender, age, etc.) based solely on user-generated text on social media.

The benefit of the this approach is that it addresses the lack of research for attribute prediction based solely on user-generated texts. The other approaches which use engineered features require prior knowledge and assumptions about the predictive ability of the features. Furthermore, neural network approaches to these problems have been rare.

3 Dataset and Features

The text used for this project has been obtained from https://www.reddit.com/r/amitheasshole. I have written a scraper using PushshiftAPI[5] to obtain the first 100 posts of each day from 2016 to 2019. After removing posts that have had their content removed or have not been tagged with a verdict, there were 19,540 posts with each entry containing the text of the post’s title, story, and verdict. The dataset was then split into training/validation/test sets using a 60/20/20 split.

The breakdown of the classes in the training set are as follows:

Verdict	**NTA**	**YTA**	**NAH**	**ESH**
Samples	7396	2545	1070	692
Percentage	63.2%	21.7%	9.1%	5.9%

CS230: Deep Learning, Spring 2020, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

3.0.1 Implications

The dataset is biased towards 'NTA' verdicts, which presents a challenge for the project because a reflex baseline model that constantly predicts 'NTA' would achieve an accuracy of approximately 58%. This bias is seen in the result of the baseline model as well, which tends to severely over-predict 'NTA'. This is addressed by undersampling the training set so that all classes are more equally represented.

Another problem is the insufficient amount of data due the the limitations of the Reddit API. In particular, the smallest class only contains 692 posts, which is insufficient for training more compelex neural networks due to the curse of dimensionality. To increase the number of samples in each category, the four categories are consolidated into two:

$$\{\text{NTA}, \text{NAH}\} \to \text{NTA}$$
$$\{\text{YTA}, \text{ESH}\} \to \text{YTA}$$

This is reasonable since the classes now reflect the user's culpability in the scenario. This increases the number of training samples in each class to be:

Verdict	**NTA**	**YTA**
Samples	8466	3237
Percentage	72.3%	27.7%

which will be combined with the undersampling techniques mentioned above. (Unfortunately, it will later be shown that this is still insufficient.)

4 Methods

The architecture used is modified from the research presented in 'An analysis of the user occupational class through Twitter content'[2], which combines Word2Vec word embeddings with an logistic classifier.

4.1 Post2Vec (P2V)

In order to avoid the curse of dimensionality with bag-of-words approaches, I first construct an embedding for each post:

1. Train a skip-gram model with negative sampling (Word2Vec) on the corpus of all training titles and stories to obtain embedding e
    - An optimal embedding size of 30 was found when using the Gensim[6] implementation of Word2Vec
2. A title embedding is the average embedding of all the words in a title

$$v_{\text{title}} := \frac{1}{|\text{title}|} \sum_{w \in \text{title}} e_w$$

3. A story embedding is the average embedding of all the words in a story

$$v_{\text{story}} := \frac{1}{|\text{story}|} \sum_{w \in \text{story}} e_w$$

4. Concatenate the vectors to obtain a post vector

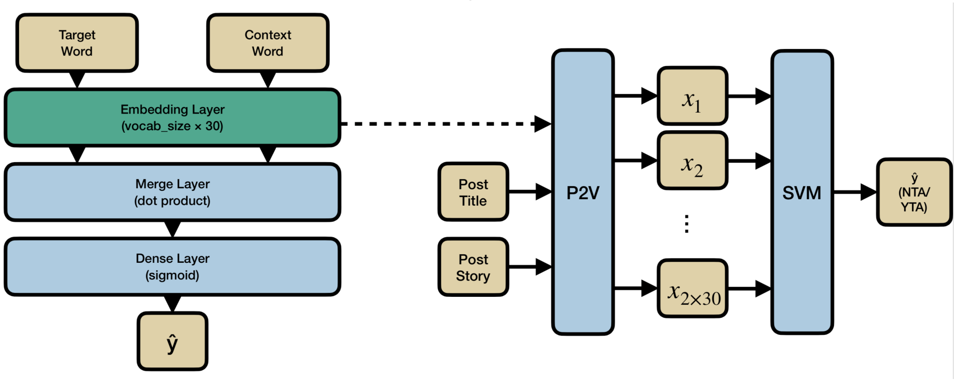
$$v_{\text{post}} := [v_{\text{title}} \quad v_{\text{story}}]$$

- Since optimal embedding size was found to be 30, a post embedding has size 60.

4.2 Post2Vec with SVM (P2V-SVM)

The P2V embeddings are used as an input to Scikit-learn library's implementation of a support vector machine (SVM) with a radial basis function (RBF) kernel. This model is motivated by the fact that Word2Vec minimizes the cosine similarity between words that appear in similar contexts. RBF would be a more appropriate choice of capturing this relationship since both cosine similarity and RBF are based on Euclidean norms.

Architecture Diagram of P2V-SVM:

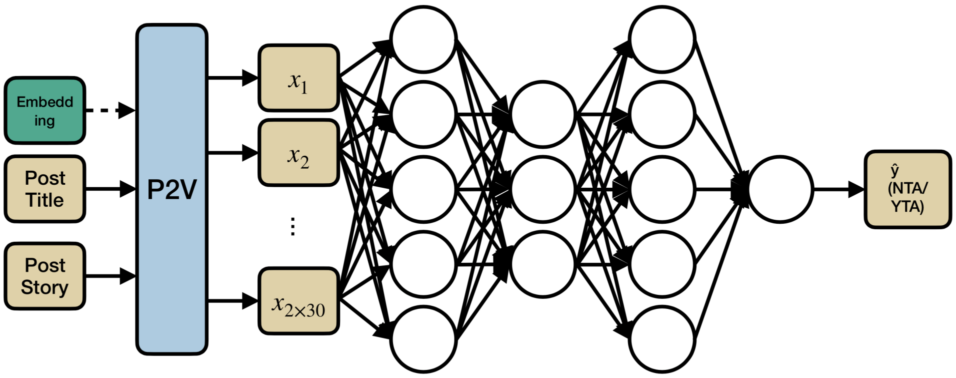


The model is then tuned based on the experimental results of the model applied to the validation set.

4.3 Post2Vec with Deep Neural Networks (P2V-DNN)

Similar to P2V-SVM, the P2V embeddings are used as an input to a fully connected deep neural network with 3 hidden layers of size 50,30, and 50, respectively. Each hidden layer uses a $\tanh()$ activation function, and the output layer uses a sigmoid $\sigma()$ activation function.

Architecture Diagram of P2V-DNN:



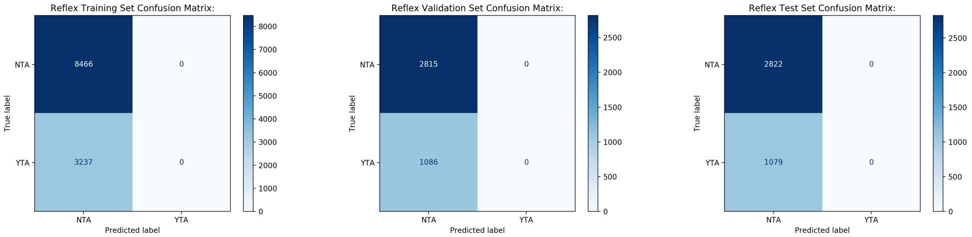
While many model architectures and hyperparameters were tried, the differences they made were caused mostly by the insufficient training data. We will discuss this below.

5 Experiments/Results/Discussion

We can compare the performance of our model to the baseline reflex model by examining the confusion matrix of the model performance on the training, validation, and test sets.

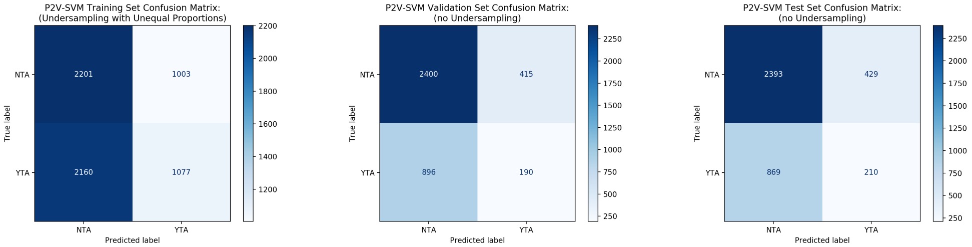
Note that the vast difference between the training and validation sets for the two P2V models is due to the undersampling done on the training set.

Reflex Model Performance:



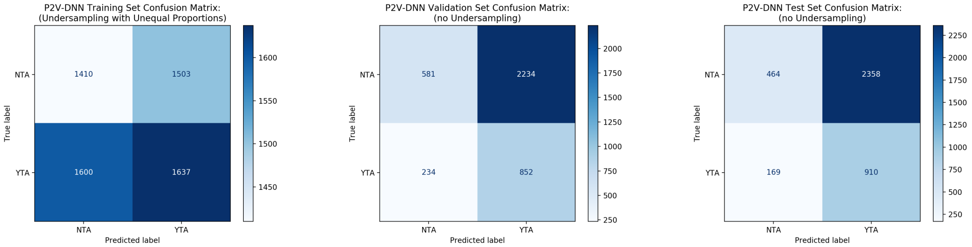
As expected, the reflex model gives high recall but low precision on 'NTA' samples, and low recall on 'YTA' samples.

P2P-SVM Model Performance:



While the recall for 'YTA' has slightly increased, the precision remains very low.

P2P-DNN Model Performance:



The performance of P2V-DNN is comparable to P2V-SVM. However, the most significant result is how much the model changes when the training data changes caused by the randomness of undersampling. This strongly indicates that the model has extremely high variance and that the performance improvements made by changing the training data can partly be attributed to just randomness.

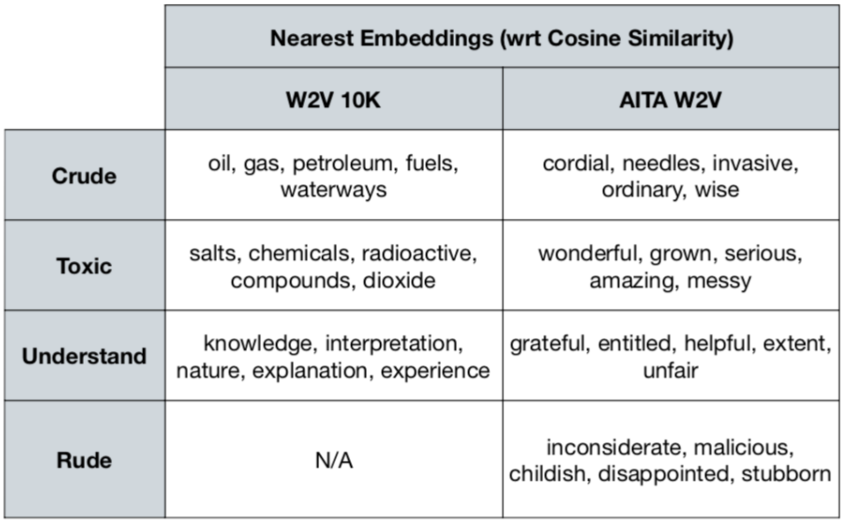
6 Discussion

The results of our model give the following two key takeaways from this project.

6.1 Significance of Training Word2Vec

Discussion on AITA is very specific in scope, and usually related to inter-personal scenarios. Therefore the words used may often have different meaning than in typical usage, and their embeddings from being trained on the different corpuses would reflect this difference. We can examine the most similar (with respect to cosine similarity) embeddings to specific words to see this. For example, we compare a pre-trained Word2Vec embeddings of the 10,000 most common words[7] to our Word2Vec embeddings:

Word Embedding Comparisons:



	Nearest Embeddings (wrt Cosine Similarity)	
	W2V 10K	AITA W2V
Crude	oil, gas, petroleum, fuels, waterways	cordial, needles, invasive, ordinary, wise
Toxic	salts, chemicals, radioactive, compounds, dioxide	wonderful, grown, serious, amazing, messy
Understand	knowledge, interpretation, nature, explanation, experience	grateful, entitled, helpful, extent, unfair
Rude	N/A	inconsiderate, malicious, childish, disappointed, stubborn

It is clear that the word embeddings in our model use these words in very different contexts. Furthermore, important words like ‘Rude’ may be missing in pre-trained models.

6.2 Importance of Data

The potential of P2V-DNN remains unrealized because its variance is too high. This highlights the need for ample amounts of quality data when it comes to deep learning tasks: More complex models require more training data. The unfortunate corollary of this is that complex models are quite useless without sufficient data.

7 Conclusion/Future Work

It seems like the main failure mode right now is caused by the high variance of the model. Small changes in the training data gives massive changes in performance of both training and validation sets. This suggests that there isn’t enough training data, which also makes intuitive sense due to the small number of samples of each class. However, more training data is hard to come by due to the limitations of the Reddit API. This is particularly unfortunate since more complex models such as ones that consider the position of words in the text would require more training data as well.

8 Contributions and Code

All work was done on my own unless otherwise cited.

Code for this project can be found at: https://github.com/IvyCodes/CS230-Project

References

[1] A. Tigunova, A. Yates, P. Mirza, and G. Weikum, “Listening between the lines: Learning personal attributes from conversations,” in *The World Wide Web Conference*, 2019, pp. 1818–1828. [Online]. Available: https://arxiv.org/pdf/1904.10887.pdf

[2] D. Preoţiuc-Pietro, V. Lampos, and N. Aletras, “An analysis of the user occupational class through twitter content,” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 1754–1764.

[3] P. Vijayaraghavan, S. Vosoughi, and D. Roy, “Twitter demographic classification using deep multi-modal multi-task learning,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2017, pp. 478–483.

[4] R. K. Bayot and T. Gonçalves, "Age and gender classification of tweets using convolutional neural networks," in *International Workshop on Machine Learning, Optimization, and Big Data*. Springer, 2017, pp. 337–348.

[5] "Pushshiftapi." [Online]. Available: https://github.com/pushshift/api

[6] [Online]. Available: http://radimrehurek.com/gensim/models/word2vec.html

[7] D. Smilkov, N. Thorat, C. Nicholson, E. Reif, F. B. Viégas, and M. Wattenberg, "Embedding projector: Interactive visualization and interpretation of embeddings," *arXiv preprint arXiv:1611.05469*, 2016.