Cooper Denning, Zane Anderson

Dr. William Hooper

November 15, 2020

CSC 3400

Predicting Song Ratings with AI

Data processing has become a huge part of our world, whether it’s business related, entertainment, or economics; understanding, writing, and figuring out how to predict data is a key aspect to growth in our technological world. At the level of technological advancements we have today, Charles Nyce PhD explains in his essay, *Predictive Analytics White Paper*, “some [databases] require preforming thousands or millions of calculations. Advances in computer hardware and software design have yielded software packages that quickly perform such calculations” (Nyce, 2). Predictive analytics is a simple way of describing the process of taking a set of information and creating an output of predicted sets. This is where the use of AI and machine learning come into play.

According to Aarushi Singh of MarTech, “Predictive analytics is the process of using data mining, statistics, and modeling to make predictions.” We can use software to look through urls, databases, surveys, and many different types of datasets to train someone or something to understand the hidden patterns that are described in the datasets. Using these patterns, it becomes easy to make predictions of what might come next.

With patterns collected, predictions are ready to be made. One way of using AI to predict an output is with reviews. The reviews can be music, movies, food, restaurants, or any creative work that is up for criticism. Using a database of review scores from a website and the written reviews that go along with them, machine learning can help predict review scores to near perfection.

But how does an AI read a review and know whether the reviewer feels positively or negatively towards what they’re reviewing? A first thought might be adding an entire dictionary’s worth of adjectives to your program and lean the review score based on the given adjectives. When in reality, we can train the AI to learn each word using datasets provided. Using the reviews, an AI can learn the vocabulary used on its own. Andrew Trask in his book, *Deep Learning*, describes the process of, “creating an input matrix that represents the vocabulary of a review” (Trask, 192) to learn the correlation between vocabulary and the given rating. This process is called Natural Language Processing or NLP. According to Trask, “NLP tasks seem to do one of three things: label a region of text…; link two or more regions of text…; or try to fill in missing information based on context” (189). In this case, the AI will be trying to fill in the missing information, or the rating, based on the written review.

NLPs are becoming extremely popular due to the fact they play a huge role in advancing artificial intelligence to thinking like humans. Since “AI seeks to create machines that can think and engage with the word as humans do,” being able to decipher human written reviews is a huge step forward. Using supervised learning, we can make this step of receiving an input and outputting what we want to know.

To decipher the vocabulary of the review, a good practice is creating vectors for each unique word used. Each row in the matrix can be the individual words used, while the columns are a unique set of numbers distinguishing each word and how many times they are used in a review. The columns are set to binary ones and zeros to translate the original input into a numerical input that can be easily read by the machine. The correlation of matrices to the review scores given by the training examples is what accomplishes the goal of creating an AI that can predict future review scores. Through each iteration of reviews, the dataset is getting closer and closer to having a near zero percent error rate. Interestingly, this neural network is training not to learn what the words of a review means, but to reduce its error to zero.

At the beginning of our project we wanted to go in depth to produce an algorithm that would take the user’s input on a certain album and output another album that the user may like. This is quite difficult unless you have a lot of information on the contents of the music such as “bass heavy”, “melodic”, or “fast rhythm”. Without a lot of information to judge the music off of it would be almost impossible to train a neural network that would output what we wanted. The most logical step forward for us was to find someway that we could judge the music. That is how we discovered NLPs. Starting with a rough idea of what we wanted we looked through some code tutorials from TensorFlow’s website which led us to the Text Classification using IMDB reviews. It was difficult to understand how we could go about changing the judgement of the system from either positive or negative to a 10-scale. After talking with you Dr. Hooper (sorry for being so meta), we decided to follow your advice and just for now replace all the negative reviews in the IMDB file with reviews from our Pitchfork dataset whose score was less than 5.0 and vice versa for the positive. This led to great success after some work to make sure the formatting was correct. We currently don’t have any new data to test on and are hoping to create some by the time presentation comes around to showcase our weights will work on different data rather than the ones trained on. We currently are able to get to a 95.3% accuracy which we are proud of. We also have a partial example of a vectorized review that will showcase the concept. One problem thought that we weren’t able to fix was the char ‘\’ showing up in our stripped-down versions of our reviews right before they were tokenized and vectorized. We believe that the 1s in the vectorized review array correspond to these ‘\’. When trying to remove using the method “tf.strings.regex\_replace()” we found it impossible to replace a single character and therefore impossible to take them out. We feel our project has come a long way in a short amount of time but it does lack some originality in a sense that could have been made up for if we were able to judge our output on a 10 point scale. That being said TensorFlow is a really deep platform that is difficult to understand in the short amount of time we gave ourselves. This project has led to a stronger understanding of a concept in AI that neither of us knew before coming into this. We feel good about making a presentation that revolves around NLP and are excited to share what we found with our peers.

**Works Cited**

Charles Nyce PhD. “Predictive Analytics,” *White Paper, 2006.* <https://www.the-digital-insurer.com/wp-content/uploads/2013/12/78-Predictive-Modeling-White-Paper.pdf>

Conaway, Nolan. “18,393 Pitchfork Reviews,” Version 1, 2017. https://www.kaggle.com/nolanbconaway/pitchfork-data/version/1

Singh, Aarushi. “How AI Is Taking Predictive Analytics to the Next Level.” *MarTech Vibe*, 21 Aug. 2019, www.martechvibe.com/insights/staff-articles/how-ai-is-improving-predictive-analytics/.

Trask, Andrew W. *Grokking Deep Learning*. Manning, 2019.

“Basic Text Classification  :   TensorFlow Core.” *TensorFlow*, www.tensorflow.org/tutorials/keras/text\_classification.