Ben Prescott
Fourth Research/Programming Assignment
MSDS 453 2021SP
6/04/2021

Reddit Corpus Unsupervised Sentiment Analysis Using Word Embeddings and NLTK VADER

Abstract

The focus of this research is performing sentiment analysis on unlabeled data originating from a reduced corpus consisting of Reddit utterances. The goal of this research is to develop an approach that assists in identifying 'high risk' subreddits, providing parents and guardians with information to protect children and young adults from potential cyberbullying or harmful information.

The corpus was further narrowed to a single subreddit named 'Gaming' to provide focus into a single topic and group of individuals. Very limited pre-processing was done to preserve the overall semantics of each utterance, playing a critical role in accurate sentiment analysis. NLTK's VADER model was leveraged to perform initial sentiment analysis on the data, followed using a basic autoencoder to generate word embeddings. The model was then trained using a supervised method with word embeddings and VADER scores, then evaluated to determine the feasibility of leveraging word embeddings to determine an utterance's sentiment.

Introduction

With the increased dependence on technology and more youth turning to the Internet at a young age, the risk of emotional harm increases for those left ungoverned. Due to the sheer size of the Internet, it is impossible for parents to know every corner their children may explore or be contributing to. Reddit is a popular source of information and discussion for many, but also comes with its share of harmful and adult-rated content buried within different subreddits.

Due to the growth in unstructured data, much of publicly available data is unlabeled. In the case of this research, the unstructured and unlabeled data exists within utterances (comments) from Reddit's subreddit named 'Gaming'. Manual labeling efforts are both costly and subjective in nature, relying on the opinion of the individuals labeling the data. Many comments are not black and white, making labeling efforts even more difficult. This paper explores methods to leverage both pre-trained models and learned word embeddings to determine the sentiment of a Reddit comment.

Literature Review

My research shows that the concept of leveraging both learned and pre-trained word embeddings for sentiment analysis is not a new concept. There are many guides describing methods that use word embeddings to predict class labels for sentiment analysis, including Jason Brownlee's book *Deep Learning for Natural Language Processing*, used as part of this course.

There are also a few research articles that focus on generating word embeddings for unlabeled data. One paper titled "Learning Word Vectors for Sentiment Analysis" by Maas et al., touches on very similar methods, using labeled movie review data to improve the word embeddings. This is a very similar approach to what is used in this article.

However, specific use cases of generating word embeddings for comparison with pre-trained VADER scores against a Reddit corpus seems to be a new application. I have not been able to find other examples of applying this approach to Reddit data, but the process is well known.

Methods

My approach started with loading the reduced Reddit corpus that consisted of 297,132 utterances across 100 subreddits. I reduced the scope of the research to focus on a single subreddit, "Gaming", which consists of 3,542 utterances. I performed minimal pre-processing and cleaning to retain as much of the text structure as possible. The only observations removed were empty utterances, those marked as deleted or removed, and those posted by the Reddit bot “AutoModerator”. The utterances were then processed using NLTK’s VADER model to receive sentiment scores.

VADER returns various polarity scores, consisting of Positive, Neutral, Negative, and Compound. The Compound score represents the overall sentiment within a range of -1 to +1, with +1 being extremely positive and -1 being extremely negative. To simplify this approach for my research, I decided on a threshold of a Compound score of 0.05 to change the VADER scores into binary representations. Any score greater than 0.05 is considered Positive (1) and below 0.05 is Negative (0).

Each utterance was then tokenized and zero-padded to ensure consistency in sequence lengths. Filtering out words and special characters, as well as not converting the case to lowercase, were both excluded from tokenization to ensure parity with VADER’s understanding of the data. These padded sequences were then split into train, validation, and test datasets for training and evaluation purposes. An autoencoder consisting of a two-layer encoder, single bottleneck, and two-layer decoder was then used to learn word embeddings for the training data.

This trained model was then used to predict the embeddings for the train, validation, and test utterances. The predicted embeddings were then used to train a simplified classification neural network, along with the binary VADER scores, to determine prediction accuracy using word embeddings. This model was later used to predict the test data sentiment scores and measured using a ROC curve and confusion matrix.

Results

Using an unsupervised method to generate word embeddings, followed by using word embeddings and VADER scores in a binary classification neural network, proved to provide only slightly better results than guessing, with a ROC AUC of 0.56. The confusion matrix reinforced the AUC score, with 200 being correctly classified and 155 incorrectly classified, using the test dataset.

However, the size of the reduced dataset only consists of 3,542 observations, which may not be enough data for the model to learn. The methods mentioned in this paper were also tested against four included subreddits, rather than one, to compare performance using a larger overall dataset of 12,080 utterances. The performance then shifted from a ROC AUC score of 0.56 to 0.60, with 378 correctly classified and 296 incorrectly classified.

Conclusions

This research shows positive indication that word embeddings can be used to determine sentiment. However, performance was much lower than expected, which may show greater improvement with a much larger dataset, or mixture of subreddits with similar content. Other approaches to potentially improve performance may be to convert VADER scores to categorical values rather than binary, create a more complex neural network using LSTM cells and an embedding layer rather than autoencoder, further refine the number of embedding dimensions and vocabulary length, and explore more pre-processing efforts prior to labeling with VADER and learning word embeddings.

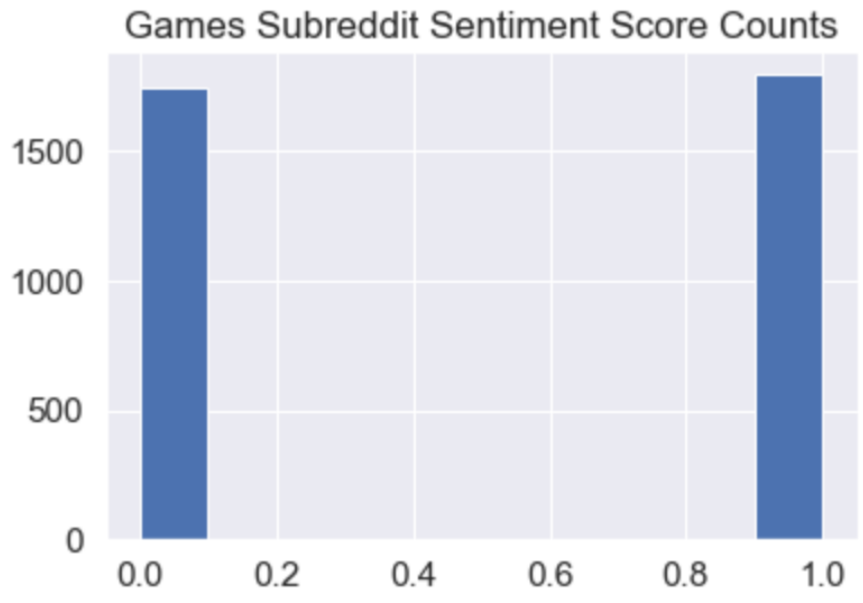
References

Brownlee, J. 2019. Deep Learning for Natural Language Processing: Develop Deep Learning Models for Natural Language in Python.

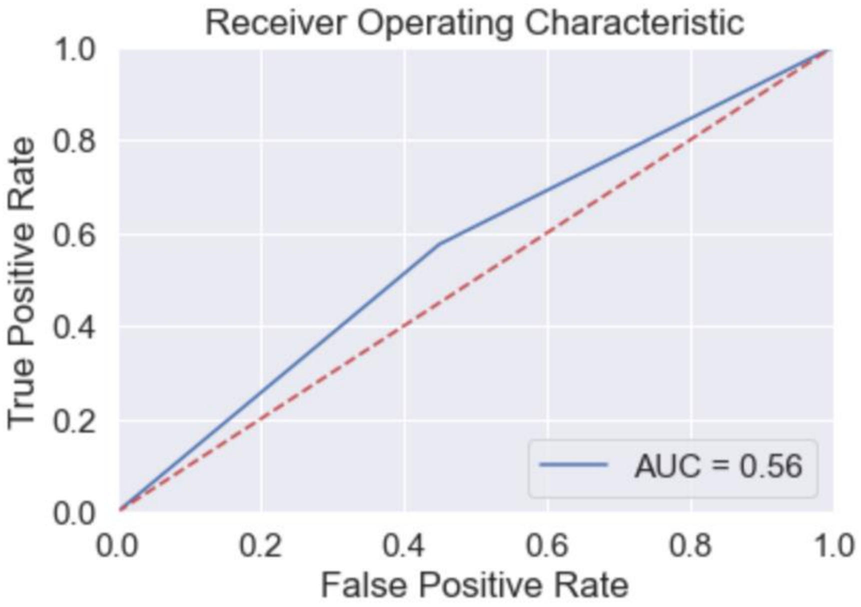
Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A., & Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. *ACL*. https://ai.stanford.edu/~ang/papers/acl11-WordVectorsSentimentAnalysis.pdf

Appendix

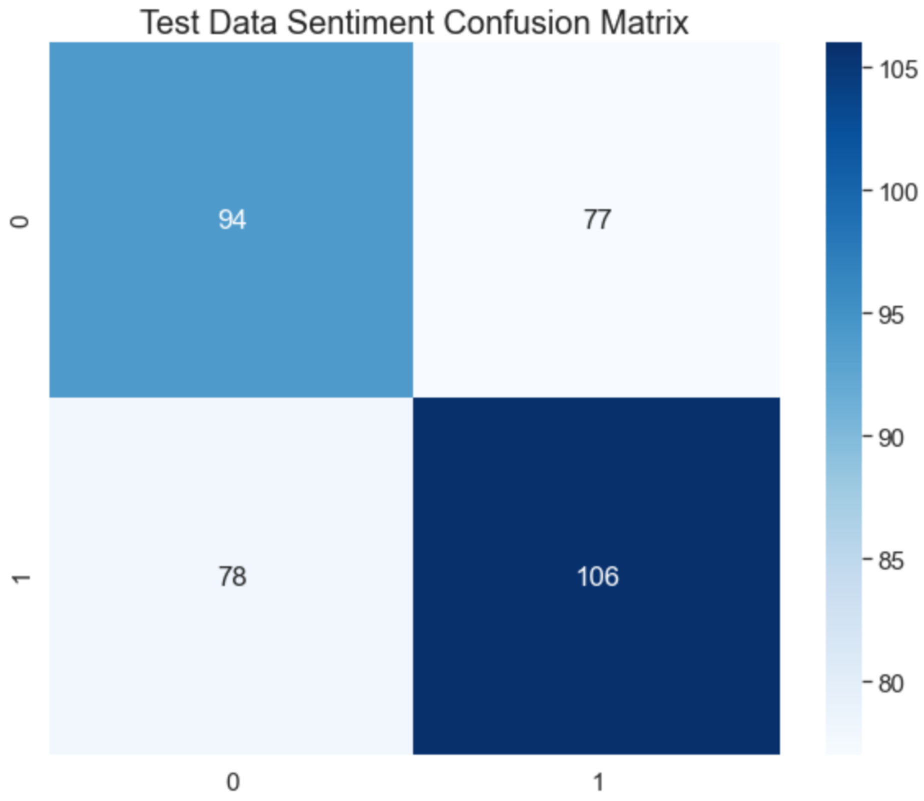
‘Games’ Subreddit Sentiment Score Couns



Test Set Predictions ROC Curve



Test Set Predictions Confusion Matrix


Test Data Sentiment Confusion Matrix