Nick Videtti

nvidetti@syr.edu

Assignment 1

Sentiment Analysis Tools

NLTK Sentiment Analyzer / VADER Sentiment Intensity Analyzer

4/10/2023

| NLTK vs VADER: Which Sentiment Analysis Tool is Better? | |
| --- | --- |
| **Introduction** | Sentiment Analysis is a great analysis method for text data that are created by humans, such as reviews, emails, and social media posts. What Sentiment Analysis does is it looks at data and essentially tries to figure out the context behind what was written, oftentimes in the form of a binary (positive or negative) classification of the “mood” of the data.  There are many sentiment analysis tools available for use. However, much like any product, service, tool, or other entity in which a consumer has multiple options to choose from, finding out which sentiment analysis tool is best is a big topic of interest in Data Science. The good news is that there are many ways to compare tools, whether it be the accuracy (which itself can be interpreted using different measures), reliability, processing times, and seemingly endless other ways. The bad news is that even when a comparison method is picked, results can vary. Most, if not all, Sentiment Analysis tools have many parameters, which can vastly alter accuracies if they are not tuned optimally. Also, the data used and use case in general also have a big role in how the tool will perform.  With all that said, for both feasibility and possibility of any meaningful comparison between tools, only two specific tools will be compared in specific ways using specific data. That is, comparisons will be done between the tools, but in no way give insight towards which is truly better, while knowing that “better” is a subjective term itself.  The tools that will be compared are NLTK’s Sentiment Analyzer and VADER’s Sentiment Intensity Analyzer. These are both available in Python. The data set, also available in Python, will be examples of tweets with their assigned sentiment of either Positive or Negative. In total there are 5,000 positive tweets and 5,000 negative tweets. The tools will each be used 10 times and their accuracy will be compared each time and then will be averaged across the 10 trials at the end. Here, accuracy is defined as the tool finding the correct sentiment of the tweets. |
| **Analysis** | The overall process of this accuracy analysis is to find the accuracy of each tool over 10 iterations of using the tool. During each iteration, the accuracies of each tool will be found and compared to the other tool. These accuracies will be stored in a list and those lists will be averaged after the 10 iterations and then those average accuracies will also be compared. All of these results will be briefly summarized at the end of the analysis.  During each of the 10 iterations, all 10,000 tweets and their corresponding sentiment will be stored in a pandas DataFrame. From there, the data will be randomly sorted and the indices will be reset. Then, the NLTK and VADER analyses will be done and their accuracies will be compared.  For the NLTK analysis, the NLTK Sentiment Analyzer (SentimentAnalyzer() from nltk.sentiment) will be used with a NLTK Naive Bayes Classifier (nltk.classify.NaiveBayesClassifier). The training and testing data need to be in the form of feature sets in order to do this, so features were created and stored in the aforementioned pandas DataFrame. The training data will be the first 80% of the features and the testing data will be the remaining 20%. Ensuring these are not always the same two sets of data is the main reason for randomly sorting the data before this step during each of the 10 iterations. The accuracy is reported in a few different ways (accuracy, recall, precision, F-measure, etc.), but traditional accuracy will be used.  For the VADER analysis, the VADER Sentiment Intensity Analyzer (SentimentIntensityAnalyzer from vaderSentiment.vaderSentiment) will be used. The compound sentiment score that is generated for each tweet will be used to determine the predicted sentiment. These scores range from -1 to +1. Anything above 0.5 is considered positive, anything below -0.5 is considered negative, and anything between -0.5 and 0.5 will be considered neutral. Both the compound score and the sentiment based on the score will be stored in the pandas DataFrame. |
| **results** | The results were as follows:  DataFrame during first iteration      Accuracy Summary - First Iteration    Accuracy Summary - Second Iteration    Final Summary    These results are actually quite disappointing and a sign that parameters and train/test processes were not optimized. While the results of not every iteration are shown in the screenshots above, it is clear to see what ended up happening. First, the NLTK tool seemed to only ever pick one sentiment value and then predict every tweet to have that sentiment (see Precision and Recall scores). While it is not certain, perhaps the model did not find any significant relationships between tweets and sentiment and simply picked whichever of the two sentiment values was more prevalent during that iteration. Even more disappointing, the VADER accuracy was worse every time. The VADER accuracy was actually the same for every iteration since the random sorting and train/test splitting did not affect the VADER tool. One major flaw could be the fact that a large portion of tweets were automatically “incorrectly” labeled by VADER when they were labeled as “neutral”, since the “true” sentiments were either positive or negative. It would be interesting to see how these performed when an even amount of positive and negative tweets are used in the NLTK tool, and sentiments are only positive or negative in the VADER tool. Cross-validating for the NLTK tool, or at least more consistent train/test processes between the two tools may have also helped to find more meaningful results. |
| **conclusions** | Technically, in this very specific scenario, the NLTK tool was better than the VADER tool, but results indicated that the tools were not used to their best ability.  For the 10 iterations, the NLTK tool was more accurate all 10 times with an average accuracy of 49.52%. The VADER tool had the same accuracy of 47.53% for all 10 iterations. Even more surprising, the NLTK tool picked the same sentiment for every single tweet during all 10 iterations and was still more accurate each time. This is likely due in part to the fact that a large portion of the tweets were labeled “neutral” by the VADER tool, which automatically fails since there are no true “neutral” sentiment tweets, only positive or negative.  Finding better ways to use each tool with these data, ideally ways that include an optimal combination of efficiency, accuracy, and similarity, would result in much more interesting findings in a future comparison of the two tools. |