Yelp Review Sentiment Classification Machine Learning Report

Vadim Kroytor

Credentials: York University Bsc. Hons. Computer Science Degree

January 17, 2022

Introduction

Natural language processing **(**NLP) is a field of computer science which focuses on providing computers the capability to better understand the meaning behind words closer to a level that humans can. NLP has a large amount of use cases when dealing with text or unstructured text data. For instance, say that you wanted to determine which topics random articles online discussed, such as law, sports, entertainment, etc. By applying Natural Language Processing, the text within specific articles can be dissected to determine what the article topic is.

With this project, NLP fits perfectly as the central goal is to determine whether a review - i.e. words that combine to form one textual sentence(s) - is positive, negative, or neutral. In the following sections, an intricate explanation of how this is done through data mining classification will be provided. More specifically, this project paper will discuss how pre-processing, feature-selection, learning, and classifying algorithms all play a role in classification.

Preprocessing

Pre-processing data can seem to be overwhelming as it consumes the majority of the time when making classifier models. However, with the help of the programming language Python and its vast libraries, the pre-processing step in classification can be a much smoother process. In this project, stemming, lemmatization, and stop word removal are all viable options for pre-processing the training data in order to remove **noise** from the training data set. In order to understand the importance of pre-processing, all of these options will be now covered.

Stemming refers to the process of removing the suffix or prefix of words. For example: fishing, fished, and fisher would all be reduced to simply fish through a stemming algorithm. It is self-evident that stemming plays a crucial in removing redundancy in data sets, hence why it has

been used in this python program for this project. While stemming is useful towards removing redundancy, it does not aid towards removing meaningless words that do not contribute towards determining the sentiment of a sentence(s). This is where lemmatization has its advantages in pre-processing. Lemmatization actually guarantees that a meaningful word is returned in proper form (Yse, 2019).

Speaking of the meaningfulness of words, the elimination of stop words is a great approach towards removing words that carry very little useful information (Yse, 2019). In my python program, stop words were removed by using the nltk (Natural Language Toolkit) library and downloading the library's stop words. With the use case of this project being to determine the sentiment of a review, it is clear that the meaningfulness of words is extremely important towards classification analysis. For instance, strong positive words such as "love", "adore", "fantastic", "amazing", would likely indicate positive reviews and vice versa for negative reviews. Neutral feeling words could indicate neutral reviews, but classifying neutral reviews is a more difficult task. Now that the pre-processing step has been discussed for this project, the next step will be to discuss feature selection.

Feature/Attribute Selection

The traditional bag of words method was used to represent the data sets. The traditional bag of words method ultimately keeps track of the total occurrences of most frequently used words in the data. In my python program, the top 10,000 most frequently seen words were selected from the training data set and converted into a numeric format. This conversion to a numeric format is known as vectorization. Vectorization is useful in order for a machine learning algorithm to determine the weight/value and presence of a word in a dataset more efficiently (Pantola, 2018).

Learning Algorithm Used

In this project, Neural Networks have been utilized through the use of the Multi-Layer Perceptron classifier in my python program. Supervised learning Neural Networks work well with this project because of how well neural net models can memorize data given an adequate capacity. All of the semantics, rules and meaning behind sentences cannot all be decomposed into code due to the vast complexity of the human language. Neural Networks can bridge and reduce the gap behind this complexity by memorizing the data through an analysis of interaction effects in a nonlinear manner. Neural Network classification yields competitively high classification accuracy when being fed large quantities of data and computational power (Bhatia, 2018). In this case, because of the large amount of data available (56,000 separate reviews), Neural Network classification is a clear choice in order to get a high classification accuracy when predicting the test set class label.

Neural Networks were chosen to classify the sentiment of yelp reviews based on their classification accuracy in comparison to other machine learning models. These other models include the K-NN Classifier (with k set to 100), and the Random Forests Classifier with classification accuracies of 70% and 81% respectively. As Classification Accuracy with the Neural Networks Classifier yields a classification accuracy of 82%, this model was designated as the most suitable for this classification problem (based on the designated criteria).

Conclusion

In conclusion, this project outlines how sentiment predictions can be made through python programs and a provided training data set. With this model fully implemented, the

sentiment of new reviews can be classified with an 82% accuracy. Pre-processing, feature selection, and choosing which classification algorithm to use are all crucial steppingstones towards building this strong classification model. All in all, NLP is the field of computer science which has many use cases. The field of data science is definitely a force to be reckoned with given its capabilities, as demonstrated by how something as simple as yelp reviews can be used to make predictive models.

Appendix Regarding Python Program

In order to run the yelp_review_sentiment_classifier.py, you must have test_yelp_data.csv, train_yelp_data.csv, and stopwords.csv in the same directory as the python program. Please note that Python version 3.9 was used when compiling this program. On the following page, the output of the classification reports for the three mentioned models are displayed for this classification problem.

```
Multi-Layer Perceptron (Neural Networks) Classifier Classification Report:
              precision    recall  f1-score   support

    negative       0.77      0.78      0.78      3480
     neutral       0.35      0.31      0.33      1685
    positive       0.90      0.91      0.91     11635

    accuracy                           0.82     16800
   macro avg       0.67      0.67      0.67     16800
weighted avg       0.82      0.82      0.82     16800


Process finished with exit code 0
```

```
Random Forests Classifier Classification Report:
              precision    recall  f1-score   support

    negative       0.86      0.62      0.72      3480
     neutral       0.73      0.01      0.03      1685
    positive       0.81      0.99      0.89     11635

    accuracy                           0.81     16800
   macro avg       0.80      0.54      0.55     16800
weighted avg       0.81      0.81      0.77     16800
```

```
KNN Classifier Classification Report:
              precision    recall  f1-score   support

    negative       0.96      0.02      0.04      3480
     neutral       0.00      0.00      0.00      1685
    positive       0.70      1.00      0.82     11635

    accuracy                           0.70     16800
   macro avg       0.55      0.34      0.29     16800
weighted avg       0.68      0.70      0.58     16800
```

References

Bhatia, R. (2018). *When not to use neural networks*. Medium. Retrieved from https://medium.datadriveninvestor.com/when-not-to-use-neural-networks-89fb50622429

Pantola, P. (2018, June 14). *Natural language processing: Text data vectorization*. Medium. Retrieved from https://medium.com/@paritosh_30025/natural-language-processing-text-data-vectorization-af2520529cf7

Yse, D. L. (2019). *Your guide to natural language processing (NLP)*. Medium. Retrieved from https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1