Eric Zhou 4685-2070

CIS4930 Individual Coding Assignment

Spring 2023

1. Problem Statement

Problem: How can we determine the sentiment of online text messages, classified as either positive or negative? We want to be able to determine how a person feels through their text message. I solved this problem by taking a large data set of text messages labeled 0 or 1 for their sentiment. Then I processed and extracted their features to be fed to classification models.

2. Data Preparation

*Data was generously provided in the assignment so collection was already completed. To prepare the data, I followed the general steps for cleaning text, including converting to lowercase and removing special characters and numbers. I kept stop words to preserve useful words that could indicate sentiment. I then lemmatized and tokenized the data for use by classification models. Features that I implemented were bag of words, tf*idf, and word2vector sets. Note: I only use 10% of the total training data set as there are too many entries for my computer to run.*

3. Model Development

- Model Training
    - *I selected Logistic Regression, SVM, Multinomial Naive Bayes, and Random forest classification models for training. For bag of words and tf*idf I was able to take the provided train.csv and test.csv and feed them to vectorizers from the sklearn module. For word2vec I had to merge train.csv and test.csv to leverage the train_test_split() function to create my sets. Overall, my training/test sets were either already separated or regenerated by module functions.*
- Model Evaluation
    - *Looking at the different feature extraction methods, word2vec provides the highest accuracy among the three when testing with logistic regression as the classification model (w2v: 80.8%, bow: 72.1%, tf*idf: 72.1%). Precision and F1 scores also appear to be significantly higher in word2vec while bag of words and tf*idf are about the same in all stats.*

    - *When comparing across 4 different classification models using word2vec, however, we find that there is slightly less variation in the results. The accuracy scores from highest to lowest are as follows: SVM: 80.87%, Logistic Regression: 80.85%, Random Forest: 79.99%, Naive Bayes: 76.56%. Additionally, SVM scored highest F1 score of 0.89 while LR scored highest precision of 0.82.*

  - *Based on the results of this experiment, the variable with the greatest effect on accuracy and F1 scores is the feature extraction method used for training sets, in this case word2vec. Finally, out of the 4 classification models used, SVM performed the best when trained with word2vec.*

4. Discussion

- *With the highest achieved accuracy score of 81%, we can say that the model correctly predicts the sentiment of text messages about 4 out of 5 times. In some contexts, this may be enough to consider the problem solved while other times, this may not be enough at all. For the scope of this assignment, incorrect predictions pose no real risks, therefore, I would consider the problem solved. If the predictions of my model were to have a direct impact on some critical event, then perhaps 80% accuracy is much too low.*

- *During data preparation and model development, I had to make decisions on how to process the data, such as choosing whether or not to remove stop words or choosing stemming vs. lemmatization. Since the words in text messages play such a crucial role in determining sentiment, I realized that preserving meaning would be my best bet for the assignment. I kept stop words and chose lemmatization to keep slight nuances that some variations of words may have.*

- *I've learned a lot about the machine learning process from this assignment. The steps from data exploration, processing, feature extraction, classification model training, and evaluation are deeply ingrained in my mind now. I feel more prepared to take on machine learning tasks in the future.*

5. Appendix

https://scikit-learn.org/stable/