Capstone Project

Rakshit Pratap Singh
Machine Learning Engineer Nanodegree

Definition

Domain background

Sentiment Analysis is a fundamental task in Natural Language Processing (NLP). Its uses are many: from analysing political sentiment on social media , gathering insight from user-generated product reviews or even for financial purposes, such as developing trading strategies based on market sentiment . The goal of most sentiment classification tasks is to identify the overall sentiment polarity of the documents in question, i.e. is the sentiment of the document positive or negative.

In this project we are doing the sentiment analysis of amazon fine food review dataset. Our task is to label the reviews as either positive or negative.

Related academic research and academic work:-

https://gist.github.com/abhigrover101/dff3ebd06a0c30c7155f
http://aics2017.dit.ie/papers/AICS2017_paper_21.pdf

Problem Statement

Our goal is to correctly label the amazon fine food reviews as either positive or negative. This is a supervised learning task so we will be employing various supervised learning techniques. The aim is to maximize the accuracy of the classification.

Dataset and inputs

For this project, the dataset is obtained from Kaggle (https://www.kaggle.com/snap/amazon-fine-food-reviews/data).

The Amazon Fine Food Reviews dataset is approx. ~300 MB large dataset which consists of around 568k reviews about amazon food products written by reviewers between 1999 and 2012. Each review has the following 10 features:

- Id
- ProductId - unique identifier for the product
- UserId - unqiue identifier for the user
- ProfileName
- HelpfulnessNumerator - number of users who found the review helpful
- HelpfulnessDenominator - number of users who indicated whether they found the review helpful
- Score - rating between 1 and 5
- Time - timestamp for the review
- Summary - brief summary of the review
- Text - text of the review

Here, for each review we have score rating between 1 and 5.for simplification we will label the review as negative if it has a score of 1-3 and positive for a score of 4 or 5. There are many features that are provided in the dataset. However we will use 'Text' as an input and 'Score' as an output.

Solution Statement

Our aim is to build a model that will predict the reviews as either positive or negative. In order to do so, first we have to extract the features from the dataset and we have to convert it in a form suitable to feed in to the algorithm.for this project our input feature is the 'Text' and the output feature(label) is the 'Score'. Score features has rating between 1 and 5 for each review here we will label the review as negative(0) if it has a score of 1-3 and positive(1) for a score of 4 or 5.

Our input feature is textual, so we have to apply various methods to make it suitable for the algorithm.

1. First we will apply the tokenization method for the purpose of feature extraction and also to clean the text.
2. Stemming the text
3. Now the more cleaned data will be fed to the vectorizers which will convert the data in to numerical form which is suitable for the algorithm to understand.
4. We can use various methods to vectorize our data like Bag of Words, countvectorizer or tf-idf vectorizer.
5. Now the data(Features + label) will be fed to the supervised learning algorithms like Naïve Bayes, Logistic Regression, SVM etc.
6. The model will be tested on accuracy or f1 score metric and the model with highest metric score will be selected.

Benchmark

Benchmarking of our model can be done by comparing our results by other kagglers. Kaggler ‘’amandeep is able to get an accuracy of 89% for naïve bayes and 92% for logistic regression. A link for his notebook is:- https://www.kaggle.com/amanai/amazon-fine-food-review-sentiment-analysis

Methodology

Implementation

Data Visualization: Visualizing the data to generate the useful insights.

Data Preprocessing: Removal of unnecessary things like punctuation, special characters etc are done.

Encoding the label: The label is a score between 1 and 5. We want to predict if a review is positive or not. We will consider a review positive if it is greater than or equal to 4, and negative if it is between 1 and 3.

Feature Extraction: Text data requires special preparation before we can start using it for predictive modeling. words need to be encoded as integers or floating point values for use as input to a machine learning algorithm. Hence tokenization and vectorization is done to fulfill this purpose.

Model Selection: Here we will experiment with various algorithms to find out the best algorithm for this use case.

Model Tuning: tuning of model to increase the accuracy

Testing: Test the model on testing dataset.

Results

Model Evaluation and Validation

Evaluation is done by Confusion Metrix which is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

It is often convenient to combine precision and recall into a single metric called the F1 score, in particular if you need a simple way to compare two classifiers. The F1 score isthe harmonic mean of precision and recall. Whereas the regular meantreats all values equally, the harmonic mean gives much more weight to low values.As a

result, the classifier will only get a high F1 score if both recall and precision are high.

References

https://en.wikipedia.org

https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/

https://www.kaggle.com/snap/amazon-fine-food-reviews/kernels

https://www.kaggle.com/snap/amazon-fine-food-reviews/data

https://cseweb.ucsd.edu/classes/wi17/cse258-a/reports/a022.pdf