Sentiment Analysis of Amazon Reviews

Machine Learning

Mohd Naki (2018052), Vaibhav Gupta (2019341), Tushar Singh (2019394)

1 MOTIVATION

The objective of our project is to classify the positive and negative reviews of consumers over different types of products and build a supervised learning model to polarize huge amounts of reviews. The motivation behind implementing this technique is that organisations always want to find consumer opinions and emotions about their services and products. Prospects also want to know the opinions and emotions of existing consumers before they purchase the product or service.

2 DATA ACQUISITION

Public Dataset: Datafiniti's Consumer Reviews of Amazon Products

3 PREPROCESSING TECHNIQUES

- Handling empty values in the dataset.
- Lowercasing the text data.
- A new column for sentiment.
- Punctuation mark removal.
- Eliminate stop words. These words occur frequently but do not add anything of value like "the", "he", "him", "his", "her" etc.
- Stemming: removing suffixes and prefixes to reduce variations down to root word.
- Lemmatization. (Alternative to stemming).
- Making a dictionary of all words used in the dataset and then converting them to vectors.
- GloVe dictionary (alternative to above step).

4 LEARNING TECHNIQUES

4.1 Naive Bayes (Baseline)

We will be using Naive Bayes as the Baseline Method for comparing the results.

4.2 K-nearest Neighbor

We will be using the K-nearest neighbor method as an advanced technique for the classification.

4.3 Linear Support Vector Machine

LSVM or Linear Support Vector Machine will be another advanced technique used.

4.4 Long Short Term Memory

We will be using LSTM or Long Short Term Memory, which is a unit of Recurrent Neural Network (RNN) as another advanced learning technique.

5 MODEL SELECTION STRATEGY

We will select the best performing model by using cross validation. We will consider all the classification algorithms and perform the model selection process. We will be using Bayesian optimization for tuning the hyperparameters.

6 TRAINING APPROACHES

We will be using following techniques for training and optimization:

- Stochastic Gradient Descent
- RMSprop
- Adam

7 EVALUATION METRICS AND ERROR ANALYSIS

We will be using following evaluation metrics for analyzing the performance of our models:

- Confusion Matrix
- F1 Score
- Area Under the ROC curve (AUC – ROC)
- Log loss

8 DELIVERABLES

8.1 Mohd Naki

- Pre-Processing
- Implementing LSVM
- Implementing LSTM
- Implementing RMSprop
- Implementing Adam
- Confusion Matrix
- Area under the ROC Curve
- Log Loss

8.2 Vaibhav Gupta

- Pre-Processing
- Implementing Naive Bayes
- Implementing K-nearest neighbor
- Stochastic Gradient
- Confusion Matrix
- Log Loss

8.3 Tushar Singh

- Pre-Processing
- Implementing Naive Bayes
- Implementing K-nearest neighbor
- Implementing RMSprop
- F1 Score