Sentiment Analysis for Amazon Musical Instruments Reviews

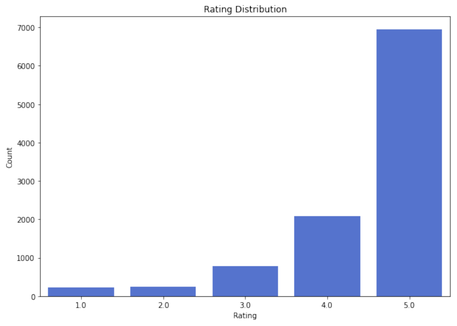
Eduard Mihranyan

Introduction

- Sentiment analysis of product reviews, an application problem, has recently become very popular in text mining and computational linguistics research.
- Here, we want to study the correlation between the Amazon musical instruments reviews and the rating given by the customers
- The objective of this paper is to classify the positive and negative reviews of the customers over different products and build a supervised learning model to polarize large amounts of reviews.

Dataset

- Our dataset comes from Kaggle. It is Amazon Musical Instruments Reviews. There are 10,261 rows in total. Each row consists of a review followed by a rate, which is an integer from 1 to 5. The distributions of the rates are shown in the figure below.



As we can see, the distribution of the dataset is super imbalanced, which will be discussed later. There are rows without rate, which we just treat as missing data

Features

- The features we extracted include two types.
- We first removed punctuation,then we tokentized the sentences in the text, andeventually lemmatized each word to its lemma. We haveused keras Tokenizer and converted texts to sequences. We have also done padding to equalize the lengths of allinput reviews.
- Alternatively, we have used standard count vectorization and tf-idf vectorization techniques. We have built also bigrams and trigrams to also capture word combinations.

Models

- Naive Bayes: This algorithm assumes that x_i's are conditionally independent given y.

$$p(x_1, ..., x_k|y) = \prod_{i=1}^{k} p(x_i|y)$$

- Logistic regression: This algorithm tries to maximimize the following likelihood function:

$$l(\theta) = \sum_{i=1}^{m} y^{(i)} \log h(x^{(i)}) + (1 - y^{(i)}) \log(1 - h(x^{(i)}))$$

- SVM: Geometrically given two types of points, circles and xi, in a space, it tries to maximize the minimum distance from one of the points to the other.

$$\underset{\gamma,\omega,b}{\operatorname{argmax}} \frac{1}{2} \|\omega\|^2$$

$$s.t. y^i(w^T x + b) >= 1, i = 1, 2, ..., m$$

- XGBoost:

$$\mathcal{L}(\phi) = \Sigma_i l(\hat{y}_i, y_i) + \Sigma_k \Omega(f_k)$$

$$where : \Omega(f) = \gamma T + \frac{1}{2}\lambda ||\omega||^2$$

- LSTM:

A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate.

Analysis

- The dataset is unbalanced. Based on the result, the model may not have a good generalization of these data. That's why some of the models have poor accuracy when weight classes or synthetically generate data.
- One of the methods used for balancing the training data was SMOTE. SMOTE synthetically generates new data points of minority class considering k-nearest neighbors of each point in the class. However, results were not as effective as it was expected.
- As the result have shown all models performed best by transforming text to sequences rather than count vectorizing or tf-idf vectorization.

Results

- The entire dataset of 10,261 reviews was divided into a training set of size 8208 (80%)and a test set of size 2053 (20%).
- With 2067-d input features representing review text, we implemented Multinomial Naive Bayes, SVM with Linear Kernel, Logistic Regression, XGBoost and LSTM.
- As we can see from the table of performance LSTM gives significantly higher F1 score and ROC and it's also good in terms of accuracy.

Models	Accuracy	F1 score	ROC
Multinomial NB	80.4%	0.52	0.52
Logistic Regression	51.7%	0.43	0.51
Linear SVM	79.0%	0.52	0.52
XGBoost	88.4%	0.50	0.52
XGBoost+SMOTE	82.0%	0.53	0.53
LSTM	86.8%	0.69	0.72
LSTM+SMOTE	87.3%	0.72	0.74

Table1. Performance of different models

Future Work

If we have more time, we want to change to another dataset which has a relatively more balanced dataset. The training at the moment is not that satisfactory. We also want to go deeper in the LSTM neural network in which case we might get better accuracy.

References

[1] Dataset Link: `https://www.kaggle.com/eswarchandt/amazon-music-reviews`

[2] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528. ACM, 2003..

[3] B. Liu and L. Zhang.*A Survey of Opinion Mining and Sentiment Analysis*, pages 415–463. Springer US, Boston, MA, 2012.

[4] C. Rain. Sentiment analysis in amazon reviews using probabilistic machine learning. *Swarthmore College*, 2013.