Sentimental analysis for yelp dataset

Bhaskar Jupudi NJUPUDI@UCSC.EDU
Trivikram Bollempalli TBOLLEMP@UCSC.EDU
Chandrahas CJAGADIS@UCSC.EDU
Karthikeyan KARTHIK@UCSC.EDU

Abstract

In this project, we aim to perform sentiment analysis i.e., classifying whether the review is postive or negative using the yelp dataset based on reviews and ratings. The classification problem can be solved by a set of algorithms. Every algorithm has its own advantages and disadvantages in terms of accuracy and model complexity.

For example, Naive Bayes classifier is faster to compute than Logistic Regression classifier for huge datasets. But the disadvantage with the former classifier is that it assumes that features are independent where as the latter has no such assumptions which can lead to better precdiction. Our work mainly concentrates on implementing these two classifiers and techniques to make them perform much better. We have adopted multi-processing for feature extraction to make it way faster and also implemented two different approaches of Logistic regression for both binary and multi-class classfication. We have also implemented Naive Bayes classifier. Finally, we contrast these two algorithms based on time taken for execution and performance metrics like accuracy, precision and recall.

1. Problem statement

asasd

2. Feature Extraction

asdasdas

3. Model Formulation

4. Evaluations

We performed all our experiments on a server that has 24 physical cores (with hyperthreading 2) and 128GB of DRAM.

5. Results

5.1. Effect of parallelism

Table 1. Execution time for extraction of features in Logistic regression classifier.

Parallelism	Size	Features	Time
No	100k	9049	65m36.271s
No	50k	5323	18m32.441s
Yes	100k	9049	7m8.291s
Yes	50k	5323	2m36.947s

Table 2. Parallelism vs countvectorizer()

Method	Features	Time
Parallel	17083	42m27.394s
countvectorizer	10k	Run-Time error

Table 3. Performance analysis of LR and NB for binary classification

Classifier	Features	Time	Accuracy	Precision	Recall
LR with gradient	9049	102m39.110s	84.81	FILL	FILL
LR with SGD	13084	13m30.578s	86.87	[89.41, 80.76]	[91.75, 76.11]
Naive Bayes	199118	1m34.385s	82.98	[86.80, 73.96]	[88.73, 70.34]

Table 4. Performance analysis of LR and NB for multiclass classification

Classifier	Features	Time	Accuracy	Precision	Recall
LR with gradient	–	102m39.110s	84.81	FILL	FILL
LR with SGD	–	13m30.578s	86.87	[89.41, 80.76]	[91.75, 76.11]
Naive Bayes	219383	3m29.505s	60.60	[58.5, 33.9, 34.2, 44.6, 75.6]	[77.8, 10.1, 17.2, 55.3, 74.3]

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).

References

Langley, P. Crafting papers on machine learning. In Langley, Pat (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.