Movie Review Dataset

|  |  |
| --- | --- |
| **Size** | 63 MB |
| **Dataset Characteristics:** | Text |
| **Attribute Characteristics:** | N/A |
| **Associated Tasks:** | Classification |
| **Number of Instances:** | 50k |
| **Number of Attributes:** | N/A |
| **Missing Values?** | N/A |
| **Number of Classes (for classification)** | 2 |
| **Area:** | Movie |
| **Additional Details:** | N/A |

# Dataset Information:

This dataset contains movie reviews along with their associated sentiment polarity labels. It is intended to serve as a benchmark for sentiment classification. For each review the naming convention is: **[[id]\_[rating].txt]** where [id] is a unique id and [rating] is the rating for that review on a 1-10 scale (larger denotes positive). For example, the file [data/pos/100\_7.txt] is the text for a positive-labeled test set example with a unique id 100 and a rating of 7/10.

# Dataset Organization:

data/pos/

1\_9.txt

2\_8.txt

3\_10.txt …

data/neg/

1\_3.txt,

2\_4.txt.

3\_4.txt …

Class Information:

|  |  |
| --- | --- |
| **Name** | **Number of instances** |
| Pos | 25,000 |
| Neg | 25,000 |
| Total | 50,000 |

Expected Task Description:

When performing the classification, consider three cases – (1) Only the text is given, don’t use any rating, (2) Only the rating is given and (3) Both text and rating are given. It will be great if you can manage to get good validation scores using configuration (1). Make sure to look at some papers that talk about document classification, because this task is highly relevant to such problems. Make sure to analyze the validation samples where the model makes mistakes in prediction and tell us about what kind of challenging properties these mistaken documents have.