Knowledge • 0 teams

Review Sentiment Classification

Tue 6 Oct 2015 Wed 11 Nov 2015 (28 days to go)

Dashboard

- Home
  - Data
  - Make a submission
- Information
  - Description
  - Evaluation
  - Rules
- Forum
- Leaderboard
- My Team
- My Submissions

Forum (0 topics)

Competition Details » Get the Data » Make a submission

This competition is private-entry. You've been invited to participate.

The goal of the competition is to predict the ratings (1.0-5.0) for a given text review of a product from Amazon.com's website

Users have opinions and they express it. Be it books, movies or even professors, end user can express opinion about them. Sometimes the reviews are accompanied by a rating.

In this competition, given a text review we would like to automatically predict it's 5-star rating. (5-highest and 1-lowest).

You are provided the following data files that you should use.

A training file train.dat is available that includes 5195 reviews in a sparse term-frequency format. The class label is given by the last column and is either 1.0, 2.0, 4.0 or 5.0.

A test file test.dat is provided to you that includes 2000 reviews in the same format as the training file. There are no labels in the last column. In fact, your goal is to predict the same.

An additional unlabeled.dat file is provided to you with 2000 reviews but there are no ratings. You may want to use these reviews to help your classification models. Using unlabeled examples within your classification models is called semi-supervised classification. This may or may not help your overall classification results.

Goal: For each row (example) in the test file predict one of the ratings (one per row). Ratings can be 1, 2, 4 or 5. There should be 2000 predicted ratings. Note, it should be integer ratings.

Format of test file submission:

id,category

....

...

See file all1s.csv for a sample format. It assumes all the predicted classes are 1s.

Caveats.

- The words (features) that exist across the training, testing and unlabeled text files may not necessarily be the same i.e., you may have to figure out a way to represent the mappings.
- You are free to use WEKA but this is not in WEKA-supported format. If you want to use Weka, you need to write a program to convert the data.
- Other options include using packages like MALLET, ORANGE, SVM LIGHT or write your own algorithm.
- Not all features may be important. Think of feature selection, reduction, engineering.
- Also remember the leaderboard reflects ranking based on a 50% sample of the test set.
- Use Data Mining wisely and correctly.
- Grading will be based on your effort i.e., how you tried to achieve the best predictions in this challenge along with your final leaderboard ranking.

Started: 6:41 pm, Tuesday 6 October 2015 UTC
Ends: 11:59 pm, Wednesday 11 November 2015 UTC (36 total days)
Points: this competition does not award ranking points
Tiers: this competition does not count towards tiers

Terms & Conditions • Privacy Policy