# IMDB Review Sentiment Prediction

### Using natural language processing, determine the sentiment of a corpus of reviews

## Introduction

The aim of this competition is to predict the sentiment of IMDB movie reviews.

## Problem Statement

Given a training dataset of reviews and corresponding sentiment, try to predict the sentiment of the reviews in the test dataset. The predicted sentiment that is closest to the actual sentiment, as calculated by the Binary Cross Entropy, scores highest on the leaderboard.

## Data

The training data for the competition consist of the following columns:

1. id : Record identifier
2. text : IMDB movie review
3. labels : Whether or not the sentiment is positive

The test data is missing the labels column and the submitted dataset should only include id and labels columns

There are three files provided for this competition:

1. train.csv : Training data
2. test.csv : Test data
3. sample\_script.py : Sample script for loading data and predicting sentiment
4. sample\_submission.csv : Format for submission of the predictions on the test dataset

## Submission

The submission file for the competition should consist of two columns:

1. id : Record identifier for the test observation
2. labels : Prediction score for the record belonging to class ‘1’. Type: Float (between 0 to 1)

Submissions not meeting the requirement stated above will be disqualified. The training, test data and sample submission files can be downloaded here.

## Evaluation

The evaluation metric for the competition is the Binary Cross Entropy score.

*p*: Predicted probability between 0 and 1

*y:* True value of either 0 or 1

Binary Cross Entropy: -(y\*log(p) + (1-y)\*log(1-p))

During the competition, a public leaderboard will be updated based on the submissions and will be visible to all participants. The public leaderboard will be calculated based on 50% of the test data. The final results will be based on the other 50% of the data, so the final standings may be different.