**Dataset Datasheet: IMDB reviews**

**Summary:**

The IMDB dataset also known as the Internet Movie Database dataset is a dataset consisting of 50,000 samples of movie reviews with the associated sentiment regarding the movie. There are 25,000 training samples and 25,000 test samples. The data was crawled from the IMDB webpage, reviews section.

The main use of the dataset is for Natural Language Processing (NLP) tasks. The dataset consists of samples of raw text, namely strings and their corresponding sentiment which is a positive or negative tag expressed as the string ‘positive’ and ‘negative’ accordingly.

**Uses, Distribution and Maintenance:**

This dataset is suitable for sentiment analysis and for detecting whether a piece of text is positive or negative in nature. The dataset is distributed online and can be accessed from the following link: http://ai.stanford.edu/~amaas/data/sentiment/. There are no issues with the distribution as the data was crawled directly from the IMDB website, which is a public website. The maintenance is conducted by the university of Stanford.

**Advantages & Disadvantages:**

Since the dataset consists of mostly positive or negative reviews, the form of text lacks neutral representation which can affect the word embeddings to be more polarised than for normal text. Since there is little neutral representation one might consider to use normal text to bridge this gap as the required model will be able to process any text.

Negative comments using offensive language may be present, however it is expected that the IMDB site takes care of offensive material and would not host it.

**Bibliography**

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). [Learning Word Vectors for Sentiment Analysis.](http://ai.stanford.edu/~amaas/papers/wvSent_acl2011.pdf) *The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).*