<http://www.cs.jhu.edu/~mdredze/datasets/sentiment/index2.html>

This sentiment dataset was used in our paper:

John Blitzer, Mark Dredze, Fernando Pereira. **Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification.** *Association of Computational Linguistics (ACL)*, 2007. [[PDF](http://www.cs.jhu.edu/~mdredze/publications/sentiment_acl07.pdf)]

If you use this data for your research or a publication, please cite the above paper as the reference for the data. Also, please drop me a line so I know that you found the data useful.

The Multi-Domain Sentiment Dataset contains product reviews taken from Amazon.com from 4 product types (domains): Kitchen, Books, DVDs, and Electronics. Each domain has several thousand reviews, but the exact number varies by domain. Reviews contain star ratings (1 to 5 stars) that can be converted into binary labels if needed. This page contains some descriptions about the data. If you have questions, please email me directly ([email found here](http://www.cs.jhu.edu/~mdredze/)).

A few notes regarding the data.

1) There are 4 directories corresponding to each of the four domains. Each directory contains 3 files called positive.review, negative.review and unlabeled.review. (The books directory doesn't contain the unlabeled but the link is below.) While the positive and negative files contain positive and negative reviews, these aren't necessarily the splits we used in the experiments. We randomly drew from the three files ignoring the file names.

2) Each file contains a pseudo XML scheme for encoding the reviews. Most of the fields are self explanatory. The reviews have a unique ID field that isn't very unique. If it has two unique id fields, ignore the one containing only a number.

There are always small details and I am sure that I omitted many of them. If you have a question after reading the paper and this page, please let me know.

**Link to download the data:**

Multi-Domain Sentiment Dataset (30 MB) [[domain\_sentiment\_data.tar.gz](http://www.cs.jhu.edu/~mdredze/datasets/sentiment/domain_sentiment_data.tar.gz)]

Books unlabeled data (2 MB) [[book.unlabeled.gz](http://www.cs.jhu.edu/~mdredze/datasets/sentiment/book.unlabeled.gz)]