

## Motivation

**For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to enable research on predicting sentiment polarity—i.e., given a piece of English text, predict whether it has a positive or negative affect—or stance—toward its topic. The dataset was created intentionally with that task in mind, focusing on movie reviews as a place where affect/sentiment is frequently expressed.<sup>1</sup>

**Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

The dataset was created by Bo Pang and Lillian Lee at Cornell University.

**Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.

Funding was provided from five distinct sources: the National Science Foundation, the Department of the Interior, the National Business Center, Cornell University, and the Sloan Foundation.

**Any other comments?**

None.

## Composition

**What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are movie reviews extracted from newsgroup postings, together with a sentiment polarity rating for whether the text corresponds to a review with a rating that is either strongly positive (high number of stars) or strongly negative (low number of stars). The sentiment polarity rating is binary {positive, negative}. An example instance is shown in figure 1.

**How many instances are there in total (of each type, if appropriate)?**

There are 1,400 instances in total in the original (v1.x versions) and 2,000 instances in total in v2.0 (from 2014).

**Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset is a sample of instances. It is intended to be a random sample of movie reviews from newsgroup postings, with the

these are words that could be used to describe the emotions of john sayles' characters in his latest , limbo . but no , i use them to describe myself after sitting through his latest little exercise in indie egomania . i can forgive many things . but using some hackneyed , whacked-out , screwed-up \* non \* - ending on a movie is unforgivable . i walked a half-mile in the rain and sat through two hours of typical , plodding sayles melodrama to get cheated by a complete and total copout finale . does sayles think he's roger corman ?

Figure 1. An example “negative polarity” instance, taken from the file neg/cv452\_tok-18656.txt.

exception that no more than 40 posts by a single author were included (see “Collection Process” below). No tests were run to determine representativeness.

**What data does each instance consist of?** “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance consists of the text associated with the review, with obvious ratings information removed from that text (some errors were found and later fixed). The text was down-cased and HTML tags were removed. Boilerplate newsgroup header/footer text was removed. Some additional unspecified automatic filtering was done. Each instance also has an associated target value: a positive (+1) or negative (-1) sentiment polarity rating based on the number of stars that that review gave (details on the mapping from number of stars to polarity is given below in “Data Preprocessing”).

**Is there a label or target associated with each instance?** If so, please provide a description.

The label is the positive/negative sentiment polarity rating derived from the star rating, as described above.

**Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Everything is included. No data is missing.

**Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** If so, please describe how these relationships are made explicit.

None explicitly, though the original newsgroup postings include poster name and email address, so some information (such as threads, replies, or posts by the same author) could be extracted if needed.

**Are there recommended data splits (e.g., training, development/validation, testing)?** If so, please provide a description of these splits, explaining the rationale behind them.

The instances come with a “cross-validation tag” to enable replication of cross-validation experiments; results are measured in classification accuracy.

**Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.

See preprocessing below.

**Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** If it links

<sup>1</sup>All information in this datasheet is taken from one of the following five sources; any errors that were introduced are the fault of the authors of the datasheet: <http://www.cs.cornell.edu/people/pabo/movie-review-data/>; <http://xxx.lanl.gov/pdf/cs/0409058v1>; <http://www.cs.cornell.edu/people/pabo/movie-review-data/rt-polaritydata/README.1.0.txt>; <http://www.cs.cornell.edu/people/pabo/movie-review-data/poldata/README.2.0.txt>.

to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is entirely self-contained.

**Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals’ non-public communications)?** If so, please provide a description.

Unknown to the authors of the datasheet.

**Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** If so, please describe why.

Some movie reviews might contain moderately inappropriate or offensive language, but we do not expect this to be the norm.

**Does the dataset identify any subpopulations (e.g., by age, gender)?**

If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

**Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** If so, please describe how.

Some personal information is retained from the newsgroup posting in the “raw form” of the dataset (as opposed to the “preprocessed” version, in which these are automatically removed), including the name and email address the author posted under (note that these are already public on the internet newsgroup archive).

**Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** If so, please provide a description.

Aside from the aforementioned name/email addresses, no.

**Any other comments?**

None.

## Collection Process

**How was the data associated with each instance acquired?** Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data was mostly observable as raw text, except that the labels were extracted by the process described below. The data was collected by downloading reviews from the IMDb archive of the `rec.arts.movies.reviews` newsgroup, at <http://reviews.imdb.com/Reviews>.

**What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software**

**programs, software APIs)?** How were these mechanisms or procedures validated?

Unknown to the authors of the datasheet.

**If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

The sample of instances collected is English movie reviews from the `rec.arts.movies.reviews` newsgroup, from which a “number of stars” rating could be extracted. The sample is limited to forty reviews per unique author in order to achieve broader coverage by authorship. Beyond that, the sample is arbitrary.

**Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

Unknown to the authors of the datasheet.

**Over what timeframe was the data collected?** Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

Unknown to the authors of the datasheet.

**Were any ethical review processes conducted (e.g., by an institutional review board)?** If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

Unknown to the authors of the datasheet.

**Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

As described above, the data was collected from newsgroups.

**Were the individuals in question notified about the data collection?** If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

No. The data was crawled from public web sources, and the authors of the posts presumably knew that their posts would be public, but the authors were not explicitly informed that their posts were to be used in this way.

**Did the individuals in question consent to the collection and use of their data?** If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

No (see previous question).

**If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A.

**Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?** If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A.

**Any other comments?**

None.

### Preprocessing/cleaning/labeling

**Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** If so, please provide a description. If not, you may skip the remaining questions in this section.

Instances for which an explicit rating could not be found were discarded. Also only instances with strongly-positive or strongly-negative ratings were retained. Star ratings were extracted by automatically looking for text like “\*\*\*\* out of \*\*\*\*\*” in the review, using that as a label, and then removing the corresponding text. When the star rating was out of five stars, anything at least four was considered positive and anything at most two negative; when out of four, three and up is considered positive, and one or less is considered negative. Occasionally half stars are missed which affects the labeling of negative examples. Everything in the middle was discarded. In order to ensure that sufficiently many authors are represented, at most 20 reviews (per positive/negative label) per author are included.

In a later version of the dataset (v1.1), non-English reviews were also removed.

Some preprocessing errors were caught in later versions. The following fixes were made: (1) Some reviews had rating information in several places that was missed by the initial filters; these are removed. (2) Some reviews had unexpected/unparsed ranges and these were fixed. (3) Sometimes the boilerplate removal removed too much of the text.

**Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** If so, please provide a link or other access point to the “raw” data.

Yes. The dataset itself contains all the raw data.

**Is the software that was used to preprocess/clean/label the data available?** If so, please provide a link or other access point.

No.

**Any other comments?**

None.

### Uses

**Has the dataset been used for any tasks already?** If so, please provide a description.

At the time of publication, only the original paper (<http://xxx.lanl.gov/pdf/cs/0409058v1>). Between then and 2012, a collection of papers that used this dataset was maintained at <http://www.cs.cornell.edu/people/pabo/movie%2Dreview%2Ddata/otherexperiments.html>.

**Is there a repository that links to any or all papers or systems that use the dataset?** If so, please provide a link or other access point.

There is a repository, maintained by Pang/Lee through April 2012, at <http://www.cs.cornell.edu/people/pabo/movie%2Dreview%2Ddata/otherexperiments.html>.

**What (other) tasks could the dataset be used for?**

The dataset could be used for anything related to modeling or understanding movie reviews. For instance, one may induce a lexicon of words/phrases that are highly indicative of sentiment polarity, or learn to automatically generate movie reviews.

**Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

There is minimal risk for harm: the data was already public, and in the preprocessed version, names and email addresses were removed.

**Are there tasks for which the dataset should not be used?** If so, please provide a description.

This data is collected solely in the movie review domain, so systems trained on it may or may not generalize to other sentiment prediction tasks. Consequently, such systems should not—without additional verification—be used to make consequential decisions about people.

**Any other comments?**

None.

### Distribution

**Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** If so, please provide a description.

Yes, the dataset is publicly available on the internet.

**How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** Does the dataset have a digital object identifier (DOI)?

The dataset is distributed on Bo Pang’s webpage at Cornell: <http://www.cs.cornell.edu/people/pabo/movie-review-data>. The dataset does not have a DOI and there is no redundant archive.

**When will the dataset be distributed?**

The dataset was first released in 2002.

**Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The crawled data copyright belongs to the authors of the reviews unless otherwise stated. There is no license, but there is a request to cite the corresponding paper if the dataset is used: *Thumbs up? Sentiment classification using machine learning techniques*. Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Proceedings of EMNLP, 2002.

**Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

**Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

Unknown to authors of the datasheet.

**Any other comments?**

None.

tion. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

Others may do so and should contact the original authors about incorporating fixes/extensions.

**Any other comments?**

None.

## Maintenance

**Who will be supporting/hosting/maintaining the dataset?**

Bo Pang is supporting/maintaining the dataset.

**How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

The curators of the dataset, Bo Pang and Lillian Lee, can be contacted at <https://sites.google.com/site/bopang42/> and <http://www.cs.cornell.edu/home/lllee>, respectively.

**Is there an erratum?** If so, please provide a link or other access point.

Since its initial release (v0.9) there have been three later releases (v1.0, v1.1, and v2.0). There is not an explicit erratum, but updates and known errors are specified in higher version README and diff files. There are several versions of these: v1.0: <http://www.cs.cornell.edu/people/pabo/movie-review-data/README>; v1.1: <http://www.cs.cornell.edu/people/pabo/movie%2Dreview%2Ddata/README.1.1> and <http://www.cs.cornell.edu/people/pabo/movie-review-data/diff.txt>; v2.0: <http://www.cs.cornell.edu/people/pabo/movie%2Dreview%2Ddata/poldata.README.2.0.txt>. Updates are listed on the dataset web page. (This datasheet largely summarizes these sources.)

**Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?

This will be posted on the dataset webpage.

**If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?** If so, please describe these limits and explain how they will be enforced.

N/A.

**Will older versions of the dataset continue to be supported/hosted/maintained?** If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

The dataset has already been updated; older versions are kept around for consistency.

**If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** If so, please provide a descrip-