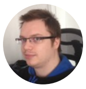
Steam Review Dataset - new, large scale sentiment dataset



Antoni Sobkowicz
Ośrodek Przetwarzania Informacji

Steam Review Dataset - new, large scale sentiment dataset

Antoni Sobkowicz[*], **Wojciech Stokowiec**[*]
[*]Ośrodek Przetwarzania Informacji - Państwowy Instytut Badawczy
al. Niepodległości 188b, 00-608 Warszawa, Poland
{antoni.sobkowicz, wojciech.stokowiec}@opi.org.pl

Abstract

In this paper we present new binary sentiment classification dataset containing over 3,640,386 reviews from Steam User Reviews, with detailed analysis of dataset properties and initial results of sentiment analysis on collected data.

1. Introduction

This paper introduces binary sentiment classification dataset containing over 3,640,386 reviews in English. Contrary to other popular sentiment corpora (like Amazon reviews dataset (McAuley et al., 2015) or IMBD reviews dataset (Maas et al., 2011)) Steam Review Dataset(Antoni Sobkowicz, 2016)[1] is also annotated by Steam community members providing insightful information about what other users consider helpful or funny. Additionally, for each game we have gathered all available screen-shots which could be used for learning inter-modal correspondences between textual and visual data. We believe that our dataset opens new directions of research for the NLP community.

Steam User Reviews, online review part of Steam gaming platform, developed by Valve Corporation, are one of more prominent ways of interaction between Steam Platfrom users, allowing them to share their views and experiences with games sold on platform. This allows users to drive sales of a game up or slow them down to the point of product being removed from sale, as online user reviews are known to influence purchasing decisions, both by their content: (Ye et al., 2009) and volume: (Duan et al., 2008).

Each review is manually tagged by author as either positive or negative before posting. It also contains authors user name (Steam Display Name), number of hours user played the game, number of games owned by the user and number of reviews written by user.

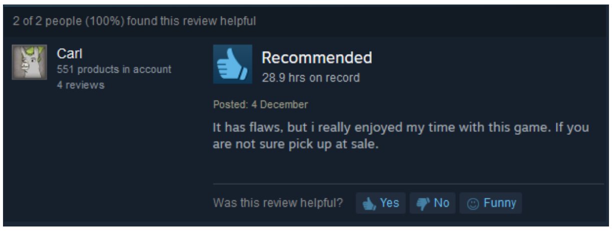
After the review is online, other Steam users can tag review as Useful/Not Useful (which add to Total score) or Funny. Useful/Not Useful score is used to generate Usefulness score (percentage of Useful score to Total). Funny score is different – it does not count into total, and allows user to tag review as Funny only.

In the rest of paper we describe dataset in detail and provide basic analysis, both based on review scores and texts. We also provide baselines for sentiment analysis an topic modelling on dataset. We encourage everyone to explore dataset, especially:

- relations between games, genres and reviews
- dataset network properties – connection between users, groups of people
- inter-modal correspondences between reviews and game screen-shots

[1] Availability information is described in section 6.

2. Detailed dataset description and analysis



2 of 2 people (100%) found this review helpful

Carl
551 products in account
4 reviews

Recommended
28.9 hrs on record

Posted: 4 December

It has flaws, but i really enjoyed my time with this game. If you are not sure pick up at sale.

Was this review helpful?

Figure 1: Typical Steam game review.

We gathered over 3,640,386 reviews in English for 6158 games spanning multiple genres, which, to the best of our knowledge, consist of over 80% of all games in steam store. We have also gathered screen-shots and basic metadata for each game that we have processed. For each review we extracted Review Text, Review Sentiment, and three scores - Usefulness and Total scores and Funny score. Detailed description of each of the scores is as follows:

- **Usefulness Score** - the number of users who marked a given review as useful
- **Total Score** - the number of users rating usefulness of a given review
- **Funny Score** - the number of users who marked a given review as funny
- **Funny Ratio** - the fraction of Funny Score to Total Score

We stored all extracted data, along with raw downloaded HTML review (for extracting more information in future) in database. Here by score, we understand the number of users who marked given review as

2.1. Review sentiment/score

We calculated basic statistics for gathered data: from collected 3,640,386 reviews written by 1,692,556 unique users. Global positive to negative review ratio was 0.81 to 0.19. Average review Total Score was 6.39 and maximum was 22,649. Average Useful Score/Total Score ratio for reviews with Total Score >1 was 0.29, with maximum of 1.0 and minimum of 0.0. Average Funny Score was 0.95 (with 329,278 reviews with Funny Score at least 1), and maximum was 20,875.

Sentiment	Usefulness average	σ
Positive	0.624	0.369
Negative	0.394	0.307

Table 1: Usefulness average comparison for positive and negative reviews.

Analysis of Usefulness (Useful Score to Total Score ratio) for positive and negative reviews showed that average Usefulness for positive reviews is statistically higher than for negative reviews (according to unpaired t-Test, with P-value > 0.0001). Averages and standard deviations are shown in table 1.

Distribution of Usefulness and Funny Score to Length of review are shown in figure 5. Additionally, as shown in figure 4, we binned Usefulness into 100 logarithmic beans. The utility of the review is roughly (except some outliers) exponential function of the length of the comment, for both positive and negative reviews - fitted log function has $R^2 = 0.954$ for positive reviews, $R^2 = 0.979$ for negative reviews. Funny Score seems to be unrelated to Length of the review.

After analysis, both qualitative and quantitative, we have decided to mark reviews as popular when they they are in the 20% of reviews with largest Total Score (per game). Reviews were marked as funny if they are popular and have Funny Ratio (Funny Score to Total Score ratio) score greater or equal to 20% (after excluding reviews with zero Funny Score). The distribution of Funny Ratio is shown in figure 2.

2.2. Review content

Average review length was 371 characters/78 words long, with longest review being 8623 characters long. The distribution of review length measured in characters is log-normal with $\mu = 4.88$ and $\sigma = 1.17$, with $R^2 = 0.990$, which is consistent with findings by (Sobkowicz et al., 2013). Histogram of review length with fitted distribution is shown in figure 6. Long tail of distribution (reviews over 1500 characters long) consists of 4,7% of all reviews. However, there is a large number of reviews with lengths above 8000 characters that do not fit this distribution. A closer inspection showed that these texts are the result of a “copy/paste” of the Martin Luther Kings ‘I Have a Dream’ speech, posted 16 times by one unique user (who, beside that posted only one relevant review). Rest of the these very long reviews are not informative, like one word repeated many times, or other, long non-review stories. These outliers in the length distribution pointed out (without reliance on contextual analysis) the existence of *trolling* behavior, even in a community of supposedly dedicated users sharing common interests.

Average length (in characters and words) for positive and negative reviews are aggregated in table 2. Performed t-Test on data converted to log scale showed that length difference is statistically significant (P-value > 0.0001), with negative reviews being longer.

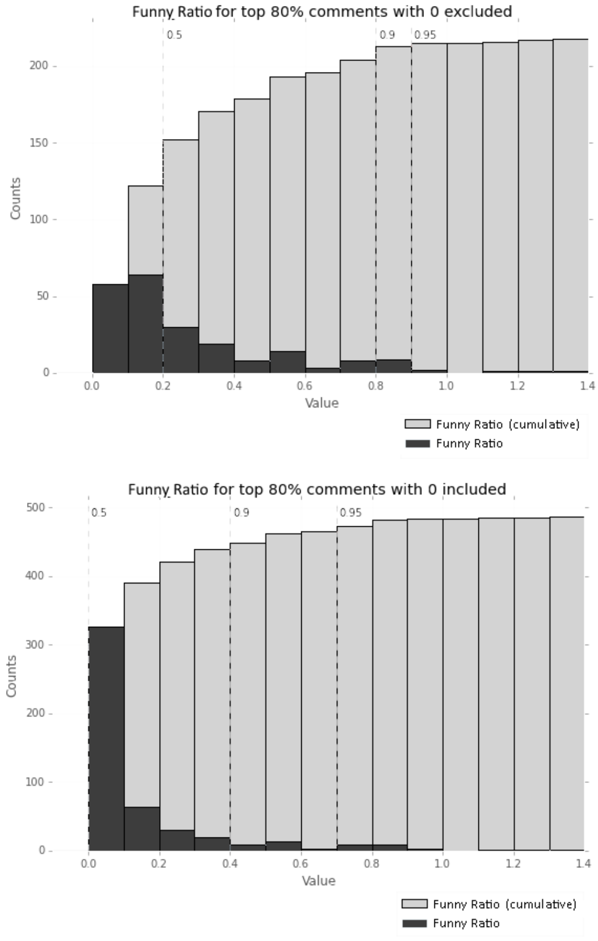


Figure 2: Distribution of funny ratio

Sentiment	Avg. words	σ	Avg. chars	σ
Positive	73.5	134.2	348.7	633.3
Negative	98.9	162.0	464.9	763.4

Table 2: Average length in words and characters comparison for positive and negative reviews.

2.3. Users

There were 1,692,556 unique users, with 35369 users writing more than 10 reviews, average 2.15 review per user. We also identified group of 94 users, who each had their own one or two prepared reviews and posted them repeatedly – reviews in this group ranged from short informative ones to ”copy/paste” – like the aforementioned Martin Luther King speech or recipes for pancakes.

There were 6252 users who wrote more than ten reviews, all of them being positive, and only 47 users who wrote more than 10 reviews, all of them being negative.

3. Sentiment Analysis

We performed basic sentiment analysis on collected dataset to establish baseline for future works and comparisons.

3.1. Experiment description

We used full dataset with 30/70 split - 1,120,325 out of 3,640,386 reviews used as test data, and rest as training data. Each review was represented as TF-IDF vector from

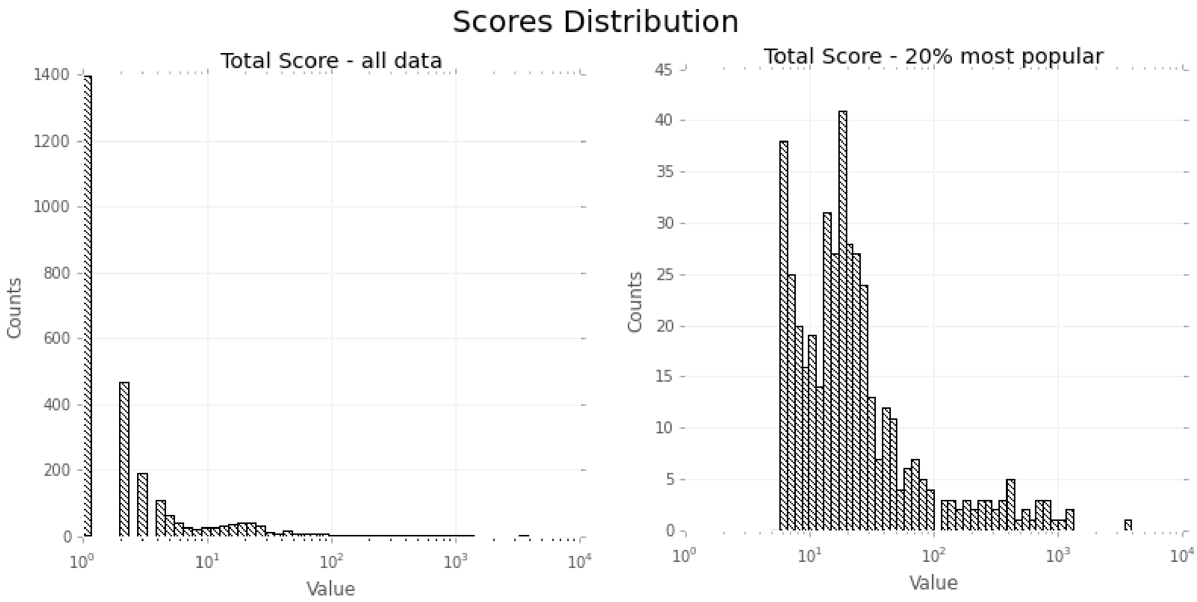


Figure 3: Distribution of extracted Total scores

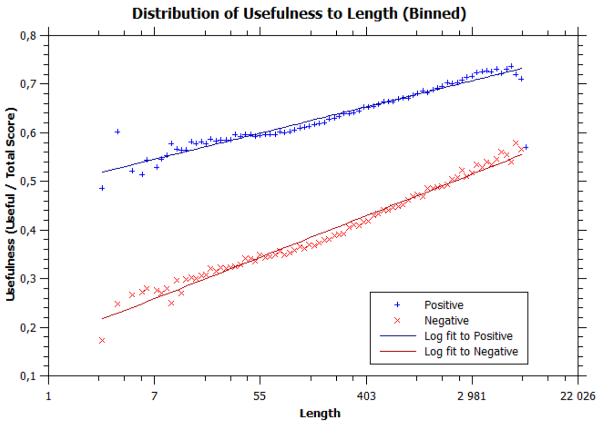


Figure 4: Usefulness of review to length, binned by length, with fitted log function for positive and negative.

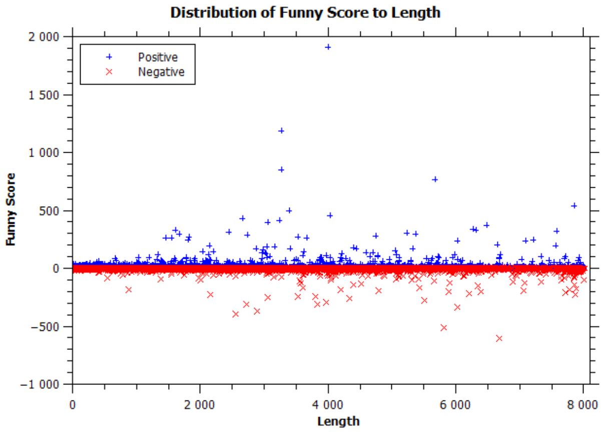


Figure 5: Funny Score of review to length. Funny Score for negative reviews is shown on negative to provide better readability.

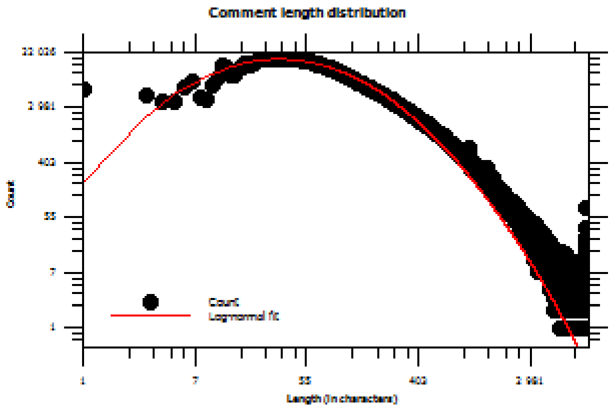


Figure 6: Review text length histogram with fitted log-normal distribution.

space of all available reviews. Using obtained vectors, we trained two models - one based on Maximum Entropy classifier (descibed in (Menard, 2002)) and other on Multinomial Naive Bayes classifier (described in (McCallum et al., 1998))

Model evaluation details are shown in tables 3 and 4.

4. Toolset

Steam Review Dataset (SRD) was gathered using custom toolset written in Python and Selenium. We also created basic analytical tools using Python with Gensim (Řehůřek and Sojka, 2010) and Scikit-learn (Pedregosa et al., 2011) packages.

4.1. Data gatherer

Data dathering package was creating using Python with Selenium. Package reads game id list from CSV file, and for each found id it scrapes game front page and two review pages - for positive and negative reviews. Package handles large number of reviews for each game (restricted by RAM

Emotion	precision	recall	f1-score	support
-1	0.8	0.64	0.71	212704
1	0.92	0.96	0.94	907621
Avg / Total	0.9	0.9	0.9	1120325

Table 3: Results for Maximum Entropy model

Emotion	precision	recall	f1-score	support
-1	0.9	0.05	0.09	212704
1	0.82	1	0.9	907621
Avg / Total	0.83	0.82	0.75	1120325

Table 4: Results for Multinomial Naive Bayes model

of machine it runs on), age verification pages, cache cleaning and, with additional tools, gathering of screenshots for each game. For each scraped game, it created two json files - one for front page information and one with all review data. Json files can then be parsed using provided scripts and saved into database (currently SQLite, but few changes are needed to use other SQL based DB engines).

4.2. Analytical and auxiliary tools

For performing basic analysis, we created several python scripts.

Classification script which was used for sentiment analysis part of this work, allows for easy text classification using one of several algorithms provided by scikit-learn package. Tool allows for simple algorithm evaluation (with training and test set) as well as 10-fold cross validation.

Word2vec and doc2vec scripts which can be used to perform word2vec and doc2vec(Mikolov et al., 2013) analysis on gathered review and game description data, implemented using gensim package. Tools are interactive and allow for easy comparison of terms/reviews.

CSV export tool used for exporting CSV from dataset database. Can be used to export any columns with additional SQL modifier, and split resulting file in two (with 70/30 ratio) for easy use in model training and validation.

5. Results and discussion

From two tested models, Maximum Entropy model works better (with f1-score of 0.9). This seems to be because of unbalanced training set (as dataset is split 0.81/0.19 between positive and negative classes) - Naive Bayes models tend to train poorly on unbalanced sets.

6. Availability and future work

Sentiment part of described dataset is available online in form of CSV file. Full dataset (in form of sqlite/mysql database), with all accompanying tools, will be provided at a later date.

In the near future we are going to add more user related data to the dataset – this should allow this dataset to be more useful in network-related research.

References

Duan, W., Gu, B., and Whinston, A. B. (2008). Do online reviews matter?—an empirical investigation of panel data. *Decision support systems*, 45(4):1007–1016.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June. Association for Computational Linguistics.

McAuley, J., Pandey, R., and Leskovec, J. (2015). Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.

McCallum, A., Nigam, K., et al. (1998). A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer.

Menard, S. (2002). *Applied logistic regression analysis*, volume 106. Sage.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. `http://is.muni.cz/publication/884893/en`.

Sobkowicz, P., Thelwall, M., Buckley, K., Paltoglou, G., and Sobkowicz, A. (2013). Lognormal distributions of user post lengths in internet discussions-a consequence of the weber-fechner law? *EPJ Data Science*, 2(1):1–20.

Ye, Q., Law, R., and Gu, B. (2009). The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, 28(1):180–182.

Language Resources

Antoni Sobkowicz. (2016). *Steam Review Dataset - game related sentiment dataset*. Ośrodek Przetwarzania Informacji, 1.0, ISLRN 884-864-189-264-2.