Austin Beaulieu: abeau96
Ryan Rowe: rfrowe
Glenn Hanawalt: grh95

FraudBGone Dataset

User-generated product reviews have the potential to make online shopping more democratic: trustworthy reviews benefit consumers by letting them make choices more confidently and reward sellers who provide quality goods and services. Unfortunately, unethical sellers can exploit this system by generating fraudulent reviews. Our dataset consists of Amazon UK product reviews with review rating distributions for the product and reviewer, labeled to indicate whether or not the review is fake (non-genuine, paid for). The objective is binary classification: given a user review, is it real or fake?

This data is perfect for machine learning applications. Patterns in text as well as review history for the product and the reviewer carry valuable information about the validity of a review. From a natural language perspective, review text is rich for multiple vectors of analysis including NLP processing and n-grams. A learner trained on this data could be particularly useful in filtering out fake reviews automatically.

To ensure that our data is of high quality, we hand-crawled AmazonUK hand-crawled for products and reviewers, choosing reviews with ratings across the spectrum. Fake reviewers were identified by their product purchase history, interval between purchases (e.g. over a dozen headphone purchases in one month), review distribution (consistent 5-star reviews), and similar or identical purchase and review history of other users (indicative of fake reviews being purchased for particular product). A scraping tool was used to extract the following info for the selected users.

Each data point pertains to a single user review for a product and conforms to this schema:

```
[rating,                      This star rating of this user review, an integer from 1-5
 reviewText,                  The full text of this user review
 productRatingsDistribution,  [#1, #2, #3, #4, #5-star reviews] by ALL users for THIS product
 reviewerRatingsDistribution] [#1, #2, #3, #4, #5-star reviews] by THIS user on ALL products
```

For example, a typical data point might be:

```
[2,
 “Terrible, I never ordered this. Broke when I picked it up.”,
 [126, 83, 23, 26, 1],
 [0, 1, 0, 0, 0]]
```

Each file contains a single data point as shown above, and can be imported as a list of these features:

```
import ast
x = ast.literal_eval(d)
```

Labels are binary, with 1 signifying a fake review and 0 signifying a real review.

We chose a conventional 90/10 split for the training/testing data: the training set consists of ~90% of the fake reviews and ~90% of the real reviews, and the testing set consists of the remaining ~10% fake and ~10% real. Users were randomly shuffled and split between sets so that no user would have comments in both the test and training set. This prevents classification by simply associating specific users' review distributions (which are likely unique) with their fake/real label, which could result in high test accuracy without any actual generalization.

From an ethical standpoint, we don't feel that this dataset violates users' privacy: all the user-generated information contained in our dataset is publicly available, and the data does not contain identifying information. If this dataset contains an unfair bias, it comes from the natural-language review text: it is conceivable that some natural language factor that correlates with the real/fake split might also have some correlation with the writing style of some specific group(s) of people. However, we have no reason to believe that such a correlation exists, and we suspect that even if it does it will be heavily outweighed by more fundamental differences between fake and real reviews.