Boston University

Graduate School of Arts and Sciences

Master's Project

Singular Value Decomposition-Based Detection Methods for Anomalous Reviews

By

Sophia Celeste Antoinette van Valkenburg

Submitted in partial fulfillment of the
requirements for the degree of
Master in Computer Science

Approved By

Advisor ____________________________
Evimaria Terzi, PhD.
Assistant Professor of Computer Science

1 Introduction

Online reviews for products, services, and places have become increasingly important resources for consumers to make informed decisions about how to spend their money. Thus, given a set of reviews, it may be useful to detect which of them are outstanding or anomalous in some way. For instance, fake reviews are harmful for the integrity of online reviewing systems, so they should be detected and removed.

In this paper, we describe an anomalous review detection method. We use a singular value decomposition (SVD)-based method from [3] and compare three different feature spaces: TF-IDF on whole the review text, TF-IDF on the given product attributes (referred to as 'attribute TF-IDF'), and opinion scores of the given product attributes, using an opinion extraction method from [2]. We tested the method using two types of anomalous data: reviews from a single product ('single-product anomalous reviews'), or reviews from a few different products ('mixed-product anomalous reviews').

We consider an 'anomalous' review in a review dataset to be a review for a product that is different from the majority of product(s) in the dataset. For example, a review for diapers embedded in a dataset that is mostly reviews for mp3 players. However, it is possible that this method could be extended to detect fake reviews.

We find that SVD-based anomaly detection of reviews works best when there is a small percentage (10%) of anomalous reviews in the dataset. In addition, using attribute TF-IDF achieves the best results on the mixed-product anomalous data, while using opinion scores works well for detecting single-product anomalous data.

Based on these results, we conclude that the SVD-based method is a possible approach for anomaly detection in review datasets. However, more work needs to be done to fully automate our approach and expand the method to detect other kinds of anomalous data, such as fake reviews.

The following section discusses the anomaly detection algorithm. Section 3 describes how the review data is processed, section 4 discusses properties of the singular values, and section 5 shows our results. Finally, we discuss the future direction of this research, and conclude the paper.

2 Anomaly Detection Using Singular Value Decomposition

The anomaly detection method using singular value decomposition is based on a similar method from "Diagnosing Network-Wide Traffic Anomalies" [3] for detecting anomalous traffic in computer networks. In this section I will discuss how this method works, and what conditions the data must meet in order for it to work.

Singular Value Decomposition (SVD) is a factorization of a matrix $M = U\Sigma V^T$, where U contains the left singular vectors of M, V^T contains the right

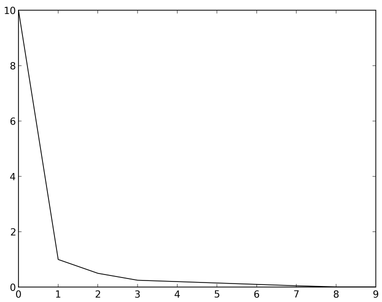


Figure 1: An example of a ‘knee’ or ‘L’ shape plot.

singular vectors of M, and Σ contains the singular values of M in decreasing order. Singular values capture the magnitude of variance along a particular dimension of the data.

We can use SVD to determine which parts of the data contribute to noise. If most of the variance in the data is from a small number of dimensions, then we can conclude that the rest of the variance is due to noise. In other words, this would mean the SVD has a small number of very high singular values and a long tail of very low singular values. A plot of these values would look like the ‘knee’ or ‘L’ shape in Figure 1.

We set all singular values after the ‘knee’ to zero and multiply the matrix factors to get a new data matrix M' that contains the non-noisy, or ‘normal’ part of the data. Conversely, if we set all singular values before the ‘knee’ to zero, we can multiply the factors to get a new data matrix M_R that contains the anomalous, or ‘residual’ part of the data. Once we have the residual part, we can compute a residual score for each data point using some metric; for this method, we use the L2-norm. We then compare the score to a threshold to determine which data points are anomalous.

In sum, the SVD-based algorithm for anomaly detection consists of the following steps:

1. Compute the SVD of the data matrix $M = U\Sigma V^T$.

2. Determine the ‘knee’ k of the singular values. In order for the method to work, the plot of singular values must have a ‘knee’ or ‘L’ shape; in other words, a small number of high singular values and a large number of low singular values.

3. Set the first k singular values to zero and multiply the matrix factors together to get residual matrix M_R.

4. Compute the score of each data point in M_R using the L2-norm of the data point.

5. Compare the score of each data point to a threshold to determine if it is anomalous or not. For the review data, we found that the method

Table 1: The six different methods of anomaly injection.

Percentage	Type
10%	mixed
20%	mixed
40%	mixed
10%	single
20%	single
40%	single

had more true positives and fewer false positives if we used a 'less-than' threshold. For each data point with a score less than the threshold, mark it as anomalous.

3 Data Processing Methods

3.1 Anomaly Injection

To test the anomaly detection method, we use a simple approach. Given a review dataset for one product or one category of products ('majority' reviews), we inject the dataset with reviews from different products or categories ('anomalous' reviews) and test how well the method detects the reviews from different products.

There are two injection methods: inject the majority with reviews from a single product ('single-product anomalous reviews'), or reviews from a few different products ('mixed-product anomalous reviews'). We expect single-product anomalous reviews to be harder to detect since the variance caused by one product will be less noisy than the variance caused by multiple products.

Finally, we varied the number of injections to compare detection results where 10%, 20%, or 40% of reviews are anomalous. We expect that the dataset with 10% anomalous reviews will be easiest to detect since it has the largest SNR (signal-to-noise ratio).

3.2 Feature Spaces

We also compared three different feature spaces to represent the review data: TF-IDF on the whole review text; TF-IDF on specific (pre-defined) product attributes, such as battery life or picture quality (which I will call 'attribute' TF-IDF); and opinion scores of product attributes, using the opinion extraction algorithm discussed in [2].

3.2.1 TF-IDF

TF-IDF, or *term frequency - inverse document frequency*, measures how important a particular word is to the given review. If R is the set of reviews, and

$f(w, r)$ is the frequency of word w in review r, then

$$TFIDF(w, r) = f(w, r) \times \log \frac{|R|}{|\{r' \in R | w \in r'\}|} .$$

So, for example, if the word 'diaper' appears in an anomalous diaper review, but never in the majority mp3 reviews, its TF-IDF score for that review would be very high. Anomalous reviews may have higher TF-IDF scores per word because their content is different from the majority reviews' content.

For each product review in each dataset, we first stem the text using a Lancaster Stemmer from NLTK and remove any punctuation and stop words. Then we compute the TF-IDF of each stemmed word in the text and store it in a dictionary for that review. Finally, all TF-IDF scores for all words in all reviews are aggregated into a $n \times m$ data matrix M, where n is the number of reviews in the dataset, and m is the number of unique words in the dataset. So, for example, the TF-IDF score for the jth word of the dataset in review i will be stored at $M_{i,j}$.

3.2.2 Attribute TF-IDF

We compute attribute TF-IDF the same way as normal TF-IDF, except that the dictionary is filtered to only include words in a pre-defined product attribute list. For example, the attribute TF-IDF data matrix for camera reviews would only include the TF-IDF scores for words like battery, picture, lens, etc. Anomaly detection using this feature space may be easier because we limit the data to terms that are more unique to the product review category.

3.2.3 Opinion Scores of Attributes

An opinion score of a given product attribute f in review r is a quantification of the positive, negative, or neutral opinion that the reviewer has of f given its context in r. If the opinion is positive, the opinion score will be positive; likewise, if the opinion is negative, the opinion score will be negative. The greater the magnitude of the opinion score, the stronger the opinion. A score that is zero or very close to zero indicates a neutral opinion.

Our opinion extraction algorithm is loosely based on the algorithm from [2], which assigns an opinion score to each attribute in the review based on the polarity of surrounding words as well as some other lexical information[1]. We are given a set of reviews R, a list of product attributes F_r for each review r, and a list of positive and negative opinion words (POS and NEG, respectively. These are words that are intrinsically positive or negative, such as 'love' or 'hate'). The opinion scores of attributes in each review are computed independently of other reviews. The steps of the opinion extraction algorithm are below.

1. Initially, for each (sentence, attribute) pair, we compute the orientation of that attribute in the sentence. Suppose a sentence s has a set of opinion

[1] [2] also includes method to determine the polarity of 'context dependent' opinion words, such as 'long', but this method is not included in our method; thus, context dependent opinion words are ignored and do not contribute to the opinion score of the attribute in question.

words $OP_s = \{p, n \in s | p \in POS | n \in NEG\}$.

$$orientation(f, s) = \sum_{w \in OP_s} \frac{orientation(w)}{distance(f, w)}$$

Where $orientation(w)$ is -1 or 1, and distance is the number of words between f and w in the sentence. If a negation word like ‘not’ precedes the opinion word, then its orientation is negated.

2. If f is in a BUT clause of the sentence (as in, ‘the camera is ugly but the picture quality is excellent’), then only the opinion words in the BUT clause will factor into the orientation of f. If there are no opinion words in the BUT clause, then we use the negated opinion words of the previous clause.

3. If, after the initial pass, f still has an opinion score of zero (so there are no opinion words in the sentence), compute its orientation using the opinion words of the previous and next sentences:
$$orientation(f, s_i) = orientation(f, s_{i-1}) + orientation(f, s_{i+1})$$

4. Sum the opinion scores of each attribute per sentence to get the aggregate opinion score of that attribute for the whole review.

5. Store the opinion scores in an $n \times 2m$ data matrix M, where n is the number of reviews and m is the total number of product attributes. Each attribute f is represented using 2 columns: f_{pos} and f_{neg}. For a given review, if attribute f has an aggregate positive opinion score, we store it in f_{pos}; if it has an aggregate negative opinion score, we store it in f_{neg}.

3.3 Data

The data for this experiment comes from product reviews on Amazon.com. It includes majority reviews, single-product anomalous reviews, and mixed-product anomalous reviews.

The majority products consist of 10 electronics (3 cameras, 2 mp3 players, 2 phones, 2 routers, and 1 dvd player), for a total of 515 reviews [2]. The majority reviews were grouped into six test groups: 1 dvd player, 1 mp3 player, 3 cameras, 2 phones, 2 routers, and 2 mp3 players, where we inject anomalous data into each of the six groups.

The single-product anomalous review group consists of up to 25 reviews from a diaper product. The mixed-product anomalous review group consists of up to 25 reviews from a combination of products (CDs, books, vacuums, kitchen/dining products, games, GPS, furniture, toys, guitars, thermometer, shoes, radio, and diapers), with 1 to 3 reviews per product. The number of reviews in these groups per dataset varies according to the percentage of anomalous reviews we want to inject.

For opinion analysis, each review in all three groups specifies a list of the product attributes (such as battery life, picture quality, etc.) it discusses.

[2] obtained from http://www.cs.uic.edu/l̃iub/FBS/sentiment-analysis.html

Dataset	#Products	#Reviews
DVD	1	95
1-mp3	1	91
2-mp3	2	141
camera	3	120
phone	2	86
router	2	73

Table 2: Breakdown of majority product reviews. 1-mp3 and 2-mp3 share a product.

4 Taking A Look At Singular Values

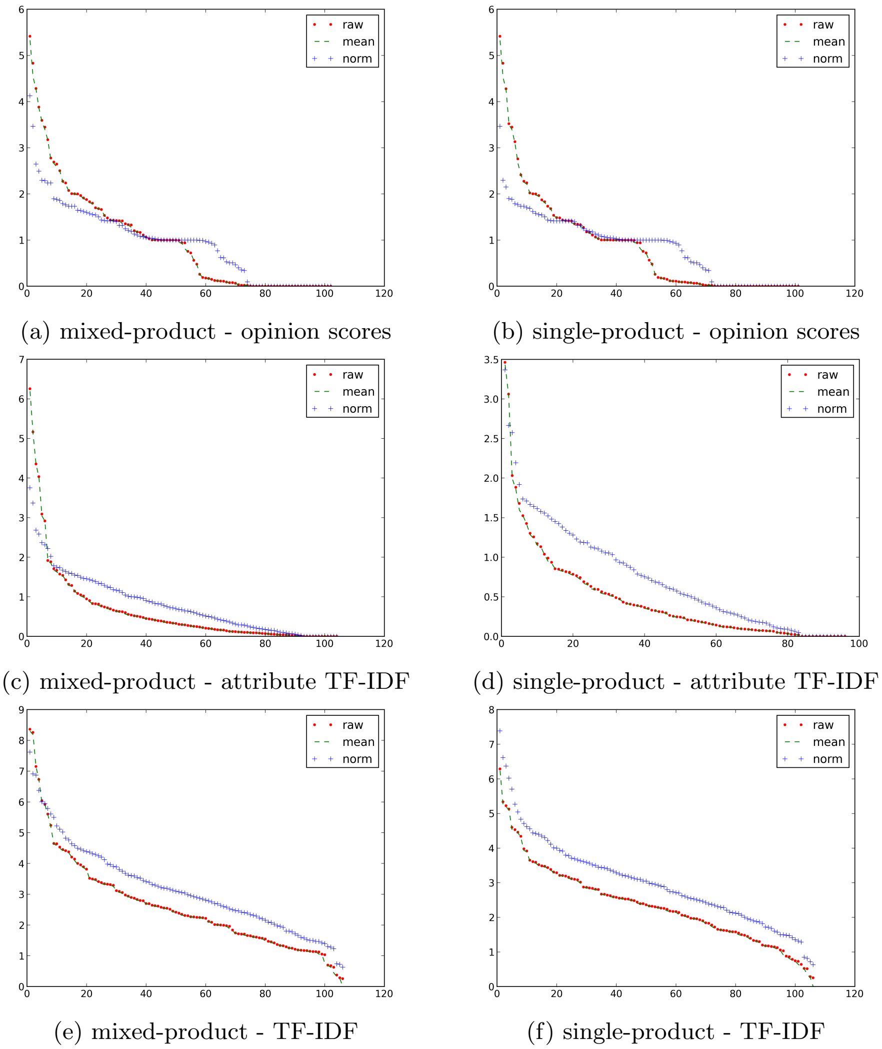
In order to determine whether SVD-based anomaly detection makes sense for the TF-IDF and opinion scored review data, we can look at the singular values computed using steps 1 and 2 of our anomaly detection algorithm. If the data has a strong 'knee', or a handful of high singular values and a long tail of low singular values, then using the SVD-based method makes sense. In addition, by comparing the singular value plots of different injection methods and feature spaces, we can predict how well the detection method will work.

Figure 2 shows six singular value plots from the DVD dataset injected with 10% single-product and mixed-product anomalous data. The plots compare the effect of using TF-IDF, attribute TF-IDF, and opinion scores to process the data. In addition, each plot shows the difference between raw scores, mean-centered scores, or normalized scores.

While all of the plots display a 'knee' shape, some combination of injection methods and feature spaces have a stronger knee than others. Mixed-product anomalous data has somewhat higher singular values, resulting in a sharper knee for the attribute TF-IDF feature space. In both the single-product and mixed-product cases, attribute TF-IDF has the sharpest knee, while regular TF-IDF has the weakest. This may indicate that the anomaly detection test will perform best using attribute TF-IDF.

In addition, normalized scores do not result in a very strong knee, while raw and mean-centered scores result in nearly identical singular values. Thus, we just display the raw scores in the next experiments.

Figure 3 shows six singular value plots from the camera dataset injected with 10%, 20%, 40% anomalous data, using the attribute TF-IDF feature space. Each plot compares the singular values of the data before and after anomaly injection. As the injection percentages increase from 10% to 40%, the singular values of the datasets after injection also increase. This would indicate that the magnitude of variance is greater as we add anomalous data. In addition, mixed-product anomalous data has higher singular values than single-product anomalous data. This may be because single-product anomalous data, coming from a single product, would be less noisy than mixed-product anomalous data, so there would be less variance. From this we can predict that the mixed-product



(a) mixed-product - opinion scores

(b) single-product - opinion scores

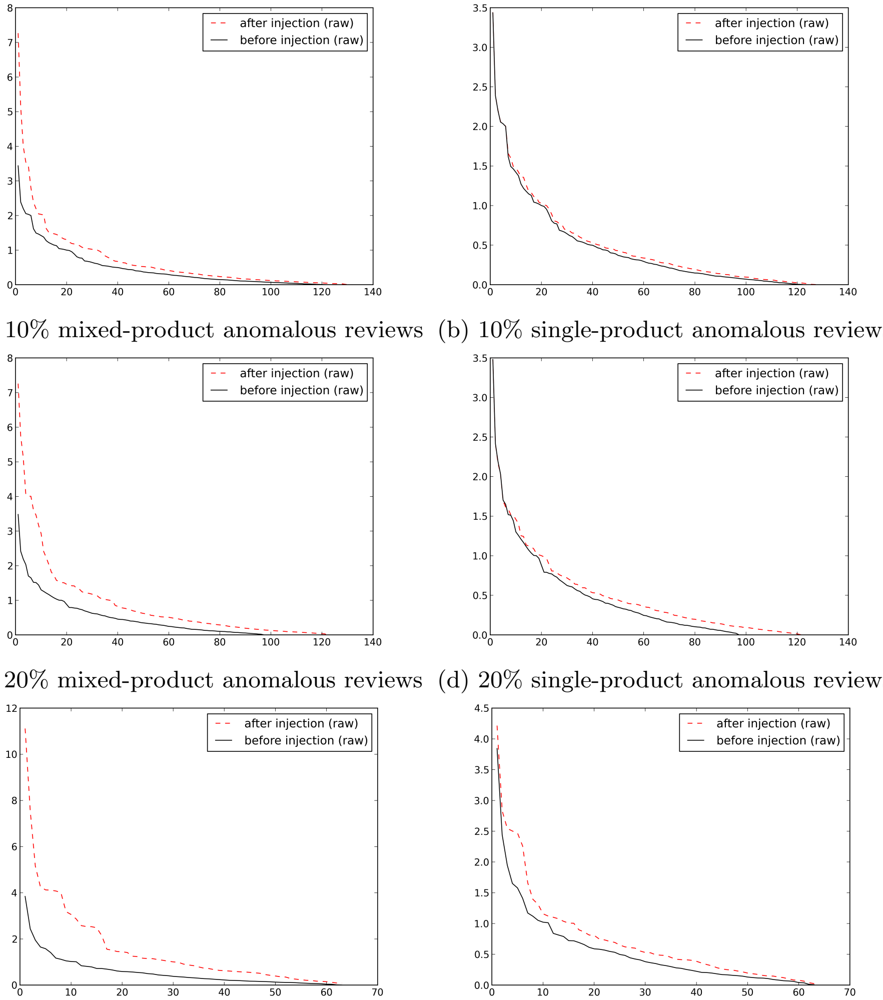
(c) mixed-product - attribute TF-IDF

(d) single-product - attribute TF-IDF

(e) mixed-product - TF-IDF

(f) single-product - TF-IDF

Figure 2: Singular values of the DVD dataset injected with 10% anomalous data, using raw, mean-centered and normalized data. The x-axis is the number of singular values and the y-axis is the magnitude of the singular values.



(a) 10% mixed-product anomalous reviews

(b) 10% single-product anomalous reviews

(c) 20% mixed-product anomalous reviews

(d) 20% single-product anomalous reviews

(e) 40% mixed-product anomalous reviews

(f) 40% single-product anomalous reviews

Figure 3: Singular values of the camera dataset, before and after anomaly injection. The x-axis is the number of singular values, and y-axis is magnitude of singular values.

anomalous data may be easier to detect than single-product anomalous data.

5 Anomaly Detection Results

To determine whether SVD-based anomaly detection is effective on review data, we ran the test on a number of thresholds for a fixed k (knee) value, all chosen manually. Recalling section 2, if the L2-norm of a data point in the residual matrix M_R is less than the threshold, that point is considered anomalous. Then, we plotted the resulting TPR (true positive rate) and FPR (false positive rate) on a ROC curve. As the curve approaches the upper lefthand corner of the plot, the FPR decreases and TPR increases, so we consider a "good" result to be one that is close to the upper lefthand corner. Specifically, the TPR must be greater than 60% and the FPR must be less than 40%.

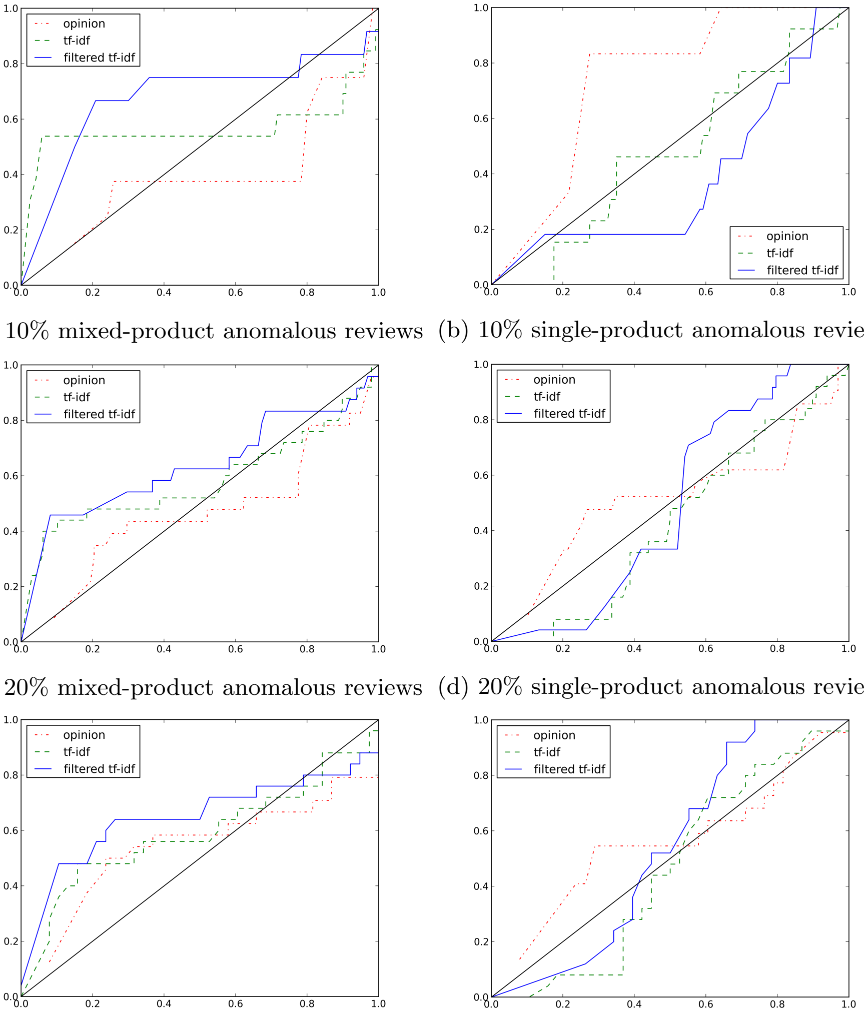
Our results indicate that it is possible to use SVD-based anomaly detection for review data. We found that anomaly detection works best when the percentage of anomalous reviews is small (10%). Attribute TF-IDF tends to work well for mixed-product anomalous reviews, while opinion scores work well for single-product anomalous reviews. Regular TF-IDF often has second-best results in both cases, indicating that it could be an alternative if an attribute list is not available; however using attribute TF-IDF or opinion scores is preferable.

Table 3 shows the best TPR/FPR pair in each of the datasets. The scores were obtained by manually selecting them from the ROC curves. First, we can see that the best results are mostly from the datasets injected with 10% anomalous data (the only exception is the router dataset with single-product anomalous reviews). In addition, most use raw or mean-centered data. Single-product anomalous datasets unanimously achieve best results using opinion scores, while mixed-product anomalous datasets do not have a consistent preference, but a majority achieve best results using attribute TF-IDF. The three exceptions are from the 1-mp3, 2-mp3, and router datasets. Opinion scores work best for 1-mp3 and router, while no method had sufficient performance for 2-mp3. In addition, single-product anomalous datasets tend to have higher TPR (0.8+) than mixed-product (0.6+).

We will examine the ROC curves and singular values plot of the camera dataset in more detail.

Figure 4 shows the ROC curves for TF-IDF, attribute TF-IDF, and opinion scores of the camera dataset injected with both types of anomalous data. For both single-product and mixed-product injected datasets, injecting a small percentage of anomalous reviews (10%) is easiest to detect, with a TPR/FPR of 0.67/0.20 for mixed-product, and 0.83/0.24 for single-product. In addition, we can see that for mixed-product anomalous reviews, attribute TF-IDF works best, while for single-product anomalous reviews, opinion scores work best.

Figure 3 shows that the singular values for the 20% and 40% anomalous datasets are higher than those of the 10% anomalous datasets, indicating higher variance, or greater noise. 10% anomalous datasets are easier to detect than those with higher percentages of anomalous reviews because the dataset as a



(a) 10% mixed-product anomalous reviews (b) 10% single-product anomalous reviews

(c) 20% mixed-product anomalous reviews (d) 20% single-product anomalous reviews

(e) 40% mixed-product anomalous reviews (f) 40% single-product anomalous reviews

Figure 4: ROC curves for TF-IDF, attribute TF-IDF, and opinion scores of the camera dataset injected with both types of anomalous data. The x-axis is FPR and y-axis is TPR.

Dataset	% Anomalous	Processing	TPR	FPR
DVD-single	10	opinion scores, raw	0.83	0.26
1-mp3-single	10	opinion scores, raw	0.80	0.10
2-mp3-single	10	opinion scores, mean	0.70	0.15
camera-single	10	opinion scores, raw	0.83	0.24
phone-single	10	opinion scores, raw	0.80	0.12
router-single	20	opinion scores, mean	0.70	0.38
DVD-mixed	10	attribute TF-IDF, raw	0.78	0.19
1-mp3-mixed	10	opinion scores, norm	0.67	0.15
2-mp3-mixed	-	-	-	-
camera-mixed	10	attribute TF-IDF, raw	0.67	0.20
phone-mixed	10	attribute TF-IDF, raw	0.67	0.23
router-mixed	10	opinion scores, mean	1.00	0.08

Table 3: Best results in all datasets

whole is less noisy, so noisy points will stand out.

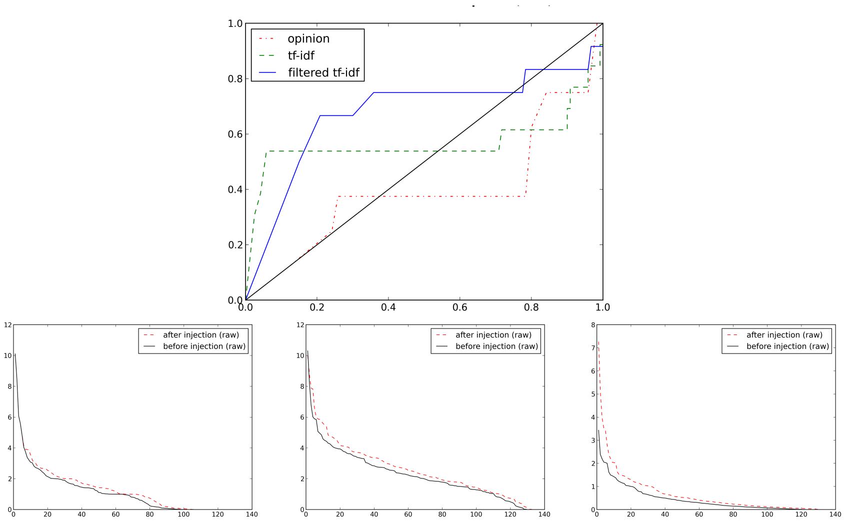
Why does attribute TF-IDF work best for the mixed-product anomalous camera dataset? Figure 5 shows the ROC curves and singular value plots for the 10% dataset. The singular values of the dataset after injection using attribute TF-IDF are the highest relative to the dataset before injection, indicating highest variance. It is possible that opinion scores and regular TF-IDF of anomalous data do not differ enough from non-anomalous data to result in accurate anomaly detection. However, high variance does not explain why 10% anomalous datasets are easier to detect than 20% and 40%, because 10% anomalous datasets have lower variance.

In addition, as we can see in Figure 6, which shows the ROC curves and singular value plots for the 10% *single-product* anomalous camera dataset, high variance does not explain why using opinion scores achieves the best result. In this case, attribute TF-IDF has the highest singular values, but does not result in the best performance. So, it is unclear why opinion scores work best in this case.

Further experiments are needed to determine why certain data processing methods achieve better results than others. Nonetheless, it is clear from these results that the SVD-based method is a viable approach to anomaly detection in review datasets.

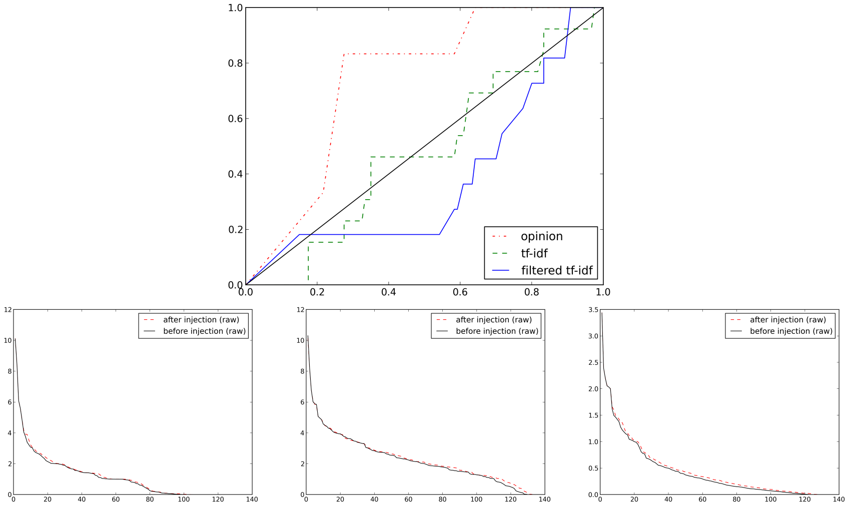
6 Future Work

While our results show that the SVD method can work, we need to use a method that will automatically pick a value for the knee and threshold, instead of choosing them manually. In addition, it would be useful to determine why attribute TF-IDF and opinion scores work, perhaps by looking at other properties of the feature spaces besides singular values. We can also try more nuanced feature



(a) Opinion Scores (b) TF-IDF (c) Attribute TF-IDF

Figure 5: ROC curves and singular values for 10% mixed-product anomalous camera dataset



(a) Opinion Scores (b) TF-IDF (c) Attribute TF-IDF

Figure 6: ROC curves and singular values for 10% single-product anomalous camera dataset

spaces, other than just attribute TF-IDF and opinion scores, to improve the TPR/FPR ratio. Finally, we need to expand the method to support different kinds of anomalous data, such as fake reviews.

7 Conclusion

In this paper, we described an approach for anomalous review detection using an SVD-based method. We compared three different feature spaces for the method: TF-IDF, attribute TF-IDF, and opinion scores, and used two different anomaly injection types: single-product anomalous reviews, and mixed-product anomalous reviews. We found that our method works best when there is a small percentage (10%) of anomalous reviews in the dataset. Using attribute TF-IDF achieves good results on the mixed-product anomalous data, while using opinion scores works well for detecting single-product anomalous data.

The SVD-based method is a viable approach for anomaly detection in review datasets. However, more work needs to be done to automate the entire process and improve the TPR and FPR. In addition, it would be useful to expand the method to detect other kinds of anomalous data, such as fake reviews.

References

[1] Bird, Steven, Edward Loper and Ewan Klein. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

[2] Xiaowen Ding, Bing Liu, and Philip S. Yu. 2008. "A Holistic Lexicon-Based Approach to Opinion Mining". In Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM '08). ACM, New York, NY, USA, 231-240.

[3] Anukool Lakhina, Mark Crovella, and Christophe Diot. 2004. "Diagnosing network-wide traffic anomalies". SIGCOMM Comput. Commun. Rev. 34, 4 (August 2004), 219-230.