Anomaly Detection Challenges - Challenge II

Hamza Tahir (03670002) and Muhammad Hamza Usmani (03669506)

Technical University of Munich

1 Introduction

This brief report serves as a purpose to present and explain the methodologies applied to tackle the second challenge in the Practical: Anomaly Detection Challenges. Section 2 discusses the challenge task and the data set for the machine learning/anomaly detection task. Section 3 explains the approaches adopted for the task. Finally, Section 4 summarizes the results.

2 The Challenge

Machine Learning Task The machine learning task for this challenge is to determine if hotel review from Yelp dataset is 'fake' or not 'fake'

Table 1. Decision Classes

Class	Representation
Fake Review	Y
Genuine Review	N

The Data set The dataset consists of hotel reviews from Yelp data set. The reviews are hotel reviews from Illinois Chicago area. The training data set of the challenge consists of 2969 reviews. The training data set has uneven class distribution, there are 2319 genuine reviews while there are 377 fake samples. The test data has 2950 reviews.

Hotel data about all reviews is also available, while almost all data about reviewers is also part of the data set, the data of 4 reviewers in the training set and 9 reviewers in test set is however missing.

3 Methodology

This section explains the data analysis and the machine learning process to build the model for classifying the given samples as fake or not fake reviews.

3.1 Features - Linguistic or Behavioral?

For classification of reviews, there are generally two types of features that are used [1]: Linguistic or Behavioral. The former focuses on the content of the reviews themselves, generally using a n-gram feature approach, while the latter is focused on the 'meta-data' of the review, such as information about the reviewer etc.

Even though linguistic features have been shown to get an accuracy of up to 90% [3], linguistic features are at times not useful in finding if a review is fake or not, [1]. Behavioral features however are generally more robust and helpful with classification of reviews as fake or not. With this in mind, we opted to go for a behavioral spam analysis approach rather than a linguistic one.

Following is a list of features we used through our experiments. Some of them are taken straight from suggestions by [1]. Table 2 also summarizes the features used for our analysis.

Maximum Number of Reviews (MNR): Maximum number of reviews (MNR), or maximum reviews per reviewer per day are number of reviews of a reviewer in a day. According to [1] , spammers have more reviews per day, in comparison to non-spammers.

Percentage of Positive Reviews (PR): Percentage Positive Reviews (PR) is the percentage of reviews of a reviewer that are positive. A review is considered to be positive by [1] when it is rated 4+. Majority of spammers have most of their reviews as positive.

Review Length (RL): Review Length (RL) is another behavioral feature considered by [1] to distinguish between spammers and non-spammers. Majority of spammers have higher average review word length (¿200), according to [1].

Reviewer Deviation (RD): The spammers are likely to deviate from the general opinion [1]. The reviewer deviation is the difference between a rating and the average rating of a hotel. According to [1]; majority of non-spammers are bounded by an absolute deviation of 0.6.

Review 'Metadata' (RWM): For each review, we had its 'rating', 'usefulCount', 'coolCount' and 'funnyCount'. According to our analysis, spam reviews generally had a low value for each of these features.

Reviewer 'Metadata' (RRM): For each reviewer, we had his or her 'friendCount', 'firstCount', 'usefulCount', 'coolCount', 'funnyCount', 'complimentCount', 'tipCount' and 'fanCount'. Again, our analysis showed that the values for users who wrote 'fake' reviews were generally low across these features.

Table 2. Behavioral Features Comparison

Feature	Claim	In Given Dataset
Max Number of Reviews (MNR)	Spammers have multiple reviews per day.	Most reviewers post only once per day
Review Length (RL)	Fake reviews have greater lengths.	Average length of fake reviews is 122 while that of genuine one is 157 words.
Percent Positive Reviews (PR)	Spammers generally rate more, in most of their reviews.	Average of Percentage Positive Reviews of Spammers is: 0.52, while that of Non-spammers is 0.62.
Reviewer Deviation (RD)	Spammers deviate more than general opinion.	Spammers on average deviate by 1.17 stars, while genuine reviewers deviate on average by 0.85.
Review Metadata (RWM)	Review spams have less 'counts; than normal.	Our analysis shows that reviews that were marked fake generally had less 'usefulCounts', 'funnyCounts' etc.
Reviewer Metadata (RRM)	Spammers have less 'counts' than normal.	Our analysis shows that reviewers that made fake reviews generally had less 'usefulCounts', 'funnyCounts' etc.

3.2 Importing data to a relational database

As the data we had was scattered across four different files, which had relationships amongst one another, we decided to pre-process the data and put it directly in a relational database. This was to make further analysis and feature extraction easier. We decided on a SQLite database, with four tables, namely: 'reviews_test', 'reviews_train', 'reviewer' and 'hotel'.

3.3 Feature Scaling

The given features and the extended behavioral features have different scales, this required to normalize all features on one scale, so that the different ranges and scales of features do not contribute to relative weights of those features. To normalize, the following method was used:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

3.4 Re-sampling

The training data set is higher number of non-fake to fake reviews, thus the data was re-sampled to build a robust classifier. The re-sampled data set had

377 genuine reviews that were randomly chosen from the data set, and 377 fake reviews from the training data set.

3.5 Machine Learning Techniques

Following machine learning techniques were used to classify given samples as "Genuine Review" aka. "Non-fake review" (represented as 0 or "N") or "Fake Review"(represented as 1 or "Y"):

1. Naive Bayes
2. Support Vector Classifier (SVC)
3. Random Forest

3.6 Generalizing - Developing a Robust Classifier

To build a robust classifier, that is relatively general and is not restricted only to the given training set following techniques were used:

Three-fold Cross Validation: Three-fold cross validation was used to overcome the problem of over-fitting, and to build a model to that will generalize to an independent dataset, Hawkins et. al. (see [2]).

4 Results

Results of the challenge are summarized in this section. It must be noted that the Random Forest classier was trained without normalization (as it used entropy as the spiting measure). The results shown here are the best accuracies we could get for each classifier.
Our analysis shows that the **Random Forest** approach works best. Our analysis also shows that training only on the metadata features (**RRM** and **RWM**) yeild the best results. Therefore, we ignored the rest of the features (suggested by [1]). This could be because the data we had was particularly limited and had only a few thousand reviews. Therefore the extended features in [1] were limiting as there was not enough data for them to be discriminatory.

Table 3. Training and Testing Results

Technique	Training Accuracy	Test Accuracy	Average
Naive Bayes	87.402	86.768	**87.04**
SVC	85.279	87.103	**86.191**
Random Forest	88.589	89.539	**89.064**
Average	**87.09**	**87.773**	

References

1. Mukherjee A., Venkatarman V., Liu Bing and Glance N.: What Yelp Fake Review Filter Might Be Doing? Seventh International AAAI Conference Weblogs and Social Media
2. Hawkins D. , Basak S. , and Denise M. Assessing Model Fit by Cross-Validation J. Chem. Inf. Comput. Sci., 2003, 43 (2), pp 579586 (2003)
3. Ott, Myle and Choi, Yejin and Cardie: Finding deceptive opinion spam by any stretch of the imagination; 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, 2011, 309–319