Amazon - Sentiment Analysis and Time Series Forecast

Naren Thanikesh - nthanik
Sruthi Venkatesh Bapu - svenka15
Vishal Ramaswamy Chittoor Venkatasubramanian - vchitto

1 Abstract

To predict the score of a new review and classify them as good, bad or moderate while detecting fake reviews based on the data set. To analyze and forecast review score on a monthly basis to improve marketing strategies.

2 Data set

We obtained the amazon fine food review dataset from SNAP that contains about 600,000 tuples and 8 columns with the distribution as shown in the Figure 1.The dataset consisted of data from 1999 to 2012.

- Helpfulness Numerator, Helpfulness Denominator - To determine how helpful the review was.
- Score - The star given by the user
- Summary Text - A summary of the review
- Review - Review comments
- Product id, Used id, Profile Name - Information about product and user

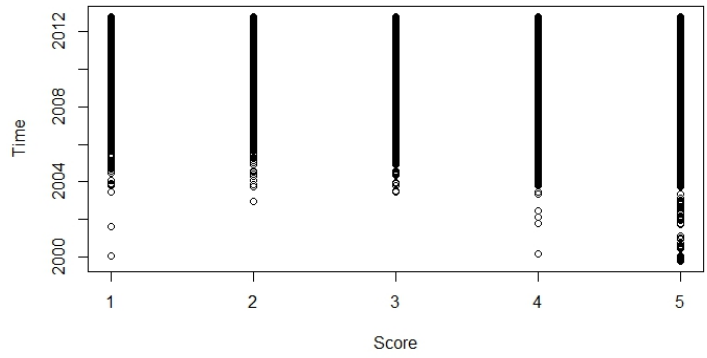

Figure 1: Dataset

3 Business Value

Reviews and customer feedback are considered to be an important factor for the sale of a product. As Good reviews can improve the future sales, a producer must aim to satisfy the customer needs that prompts them to give a good review by improving the quality of the product. Thus a review helps the producer concentrate on the required products (products with low reviews) and learn the aspect of the product that needs to be improved.

4 Business Question

Sentiment analysis is a vast field and a lot of questions can be answered by utilizing this, Is the customer happy with the product, Will the customer buy or recommend the product to others, Was the customer happy with the service provided by Amazon, Will the customer order with Amazon again and so on. Similarly Time series forecast can be utilized for questions such as, How many orders of a particular product should be obtained from the producer for a particular month, Should Amazon continue to buy from a particular producer given his bad/good reviews, Should Amazon increase or decrease the amount spent on advertising a particular product and so on.

5 Target Question and Business Value

To classify the sentiment of a review into good, bad or moderate and identify fake review based on the predicted scores.
To predict the trend of good and bad reviews over a period and decide marketing strategies using Time-series forecast on the amazon fine dining data set.
The above results can also be utilized to order supplies of a particular product in the future.

6 Evaluation

Technology and Business perspective:

The accuracy of classifying into good/mod/bad is used to evaluate the bag of words method.
Akaike's Information Criterion of different models were used to determine the best one before forecasting.

7 Literature Survey

7.1 Sentiment Analysis using product review data

This paper aims to tackle the problem of Sentiment polarity categorization on both sentence-level and review-level categorization by implementing negation phrase identification and sentiment score computation. Three classification models are built to be evaluated and compared on a data set containing reviews from between February and April, 2014 on 4 major categories
To improve the accuracy, the subjective content of the reviews are extracted which contains the sentiment sentences (ie. That has atleast one positive or negative word). As the English words of pronouns or nouns contain no sentiment, they are filtered out.

Negation phrase identification
Most of the words such as "no", "nothing", "don't" give a negative sentiment but a sentence with such words might be termed as positive due to failure to detect negation of verb/adjective. To prevent this, presence of such words classifies the review as bad.

Ground Truth tables
The aim of this method is to classify a sentence and a review as good or bad by counting the number of good and bad words in a sentence implementing a bag-of-word model.

Feature Vector Formation
The training data is converted to vectors containing the sentiment scores as features. A feature vector is formed based on sentence/review, with the problem of curse of dimensionality and unequal number

of dimensions for each vector. To avoid this, binary strings are hashed to represent word tokens and phrase tokens.

Results
The performance of each classification is based on its F1 score. Precision and Recall were estimated using 10-fold cross validation.
In sentence level categorization, NaïveBayes outperforms SVM in a manually labeled dataset while Svm outperforms NaiveBayes on a machine labeled sentences
In review level categorization, SVM and NaiveBayes perform almost identically, and are superior than Random Forest.
However, sentiment level categorization is better than review level categorization. Though both produced a good F1 score, our analysis might not work in implicit sentiments

7.2 Time Series Analysis of Household Electric Consumption with ARIMA and ARMA Models

A comparative study is done to find a suitable forecasting method(Autoregressive Integrated Moving Average- ARIMA and Autoregressive Moving Average- ARMA) and also a forecasting period(daily, weekly, monthly, yearly). The dataset that is utilized for this analysis consists of individual household electric power consumption information from December 2006 to November 2010. Data is initially preprocessed to fill missing data entries and the date format is also changed. After preprocessing, a new time series object is constructed using the traditional method and this object is further decomposed into the required time series format – either weekly, monthly, yearly or daily. Models ARIMA and ARMA are built on this new decomposed dataset for every time frame – weekly, monthly, yearly and daily, to obtain the forecast plot and AIC values. Using the forecast plot and AIC values it is concluded that ARIMA model best represents monthly and quarterly while on the other hand, ARMA is best for representing daily and weekly data models.

7.3 Improving forecasting by estimating time series structural components across multiple frequencies

In this paper, a novel algorithm is proposed to eliminate the importance of model selection and also to increase the accuracy. Temporal aggregation is utilized to construct multiple time series. This novel algorithm is divided into 3 steps:

- The aggregation step deals with sampling the time series at given frequency and aggregating by considering the consecutive group of values of the original time series in sets of definite length.
- In the forecasting step, models are built on the aggregated data thereby combining forecasted time series in place of the traditional forecast combinations.
- ETS is utilized for forecasting as it covers level, trend and seasonal components of the fitted model directly and also Akaike's Information Criterion is used for selecting between the different types of ETS models. Forecasts for any horizon can be obtained by utilizing the state vector and the type of fitted ETS.

8 Method

8.1 Sentiment Analysis

To predict a new review as a good, bad or mod review and also to detect if it is a fake review or not.

8.1.1 Pre-processing

As the data set was large, a lot of pre-processing techniques were implemented to condense the dataset to only the necessary data. 3 sets of words were formed from the summary text- attribute of the data set (goorwords, badwords, mod words) as follows:
The helpfulness numerator and denominator attribute of the dataset gives us information about how helpful the review is, thus reviews with a helpfulness ratio less than 0.5 were ignored.
The data set was divided into train and test set, to check the accuracy after prediction.

Each word of the summary text attribute from the data set was extracted with a set of delimiters such as " ","." etc. and classified as good words if the score for that row was greater than 3, mod if the score was equal to 3 and bad if the score was less than 3.
The train set was used to form the 3 sets of words-good,bad or mod that can be compared for prediction.
The extracted words were converted to lower case to maintain uniformity and duplicates were also removed.
As there is a possibility that the same word would occur in good, bad and mod words a union of intersection of all the two combinations(good-bad,bad-mod,mod-good) were removed from the previously existing good, bad and mod words.
As most of the words of length less than 3 gives no sentiment, they were removed. Any identical character occuring more than 2 times continuously in a word was replaced with 3 such characters(eg. "cooooool" was replaced with "coool") to improve the model.
Nouns and Pronouns in English language do not contribute to any sentiment and hence were removed from the above obtained dataset.

8.1.2 Prediction

Based on the 3 sets of words, the words of a new review would be matched with each of the set to obtain a count of the number of good, bad, mod words in each review. As 2 negative words in a review would mean the review is good(2 negative words cancel out each other), a modulo was applied to the count before classifying the sentence.

8.1.3 Accuracy

The test data was used to determine the accuracy of the predicted words. Based on the reviews, each test data tuple was predicted as good, bad or mod as explained in section 8.2. The score's of the test data were compared with the predicted values (4,5 - good; 3-mod; 1,2-bad) to check how many such tuples were classified correctly.

8.1.4 Fake Reviews

Fake reviews are those that contains text that is not relevant to the score(a review of "very good" with a score of 1). To identify such fake reviews, the predicted value of the review and the scores are compared. Failure to match indicates a fake review.

8.2 Time Series Forecast

A time series model was built on the Amazon fine dining data set after suitable pre-processing techniques were applied and the dataset was condensed into the required form.

8.2.1 Preprocessing

To begin with, The Unix format Time attribute was converted to normal time in the EST zone using POSIXct(). All rows other than Score and date were removed and the new dataset was stored in $preprocessed\_reviews$.csv. Furthermore, to create a timeseries object the dataset was converted into a sequential one by finding the mean of scores across days in each month. Mean of scores for each month from 1999 to 2012 was obtained and stored. For months were there were no entries, the score was set to NaN. By utilizing these means and introducing NAs, the dataset was condensed into 168 rows and 2 columns.

8.2.2 Time-Series

A timeseries object as shown in Figure 2 was created with the above obtained rows and frequency as 12. Thus the time series consisted of 12 columns (January – December) and 14 rows (1999-2012).Before creating models, the stationarity of the time series was tested using acf, pacf plots and also the kpss test. These tests proved that the object is stationery. After the stationarity test, different models were built on time series object with appropriate parameters. These models were compared using their AIC values. The lower the AIC values the better the model is. After the model was built, Forecast() function was used to forecast the next 10 scores for the time period (Nov 2012 to August

2013). The scores that were obtained from each model and the trend of the forecast plot of different models were compared as well to obtain the best result.

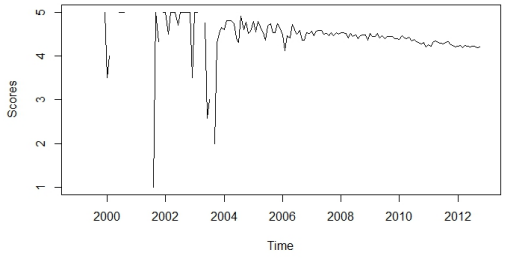

Figure 2: Time Series

9 Results

The sentiment analysis returned an accuracy of 82.2% after prediction on the initial test set.

Time-series

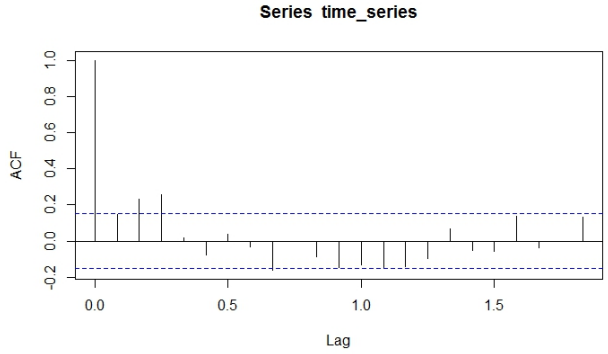

Figure 3: ACF

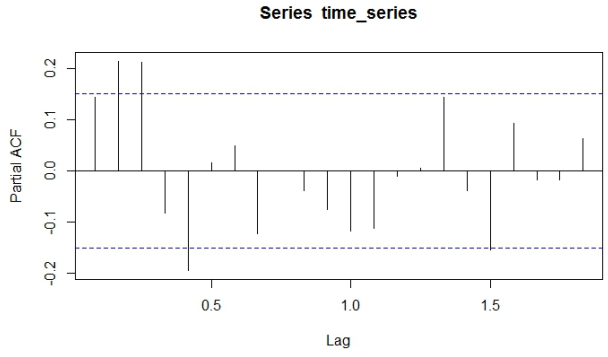

Figure 4: PACF

The ACF and PACF plots as shown in figure 3 and Figure 4, there is no gradual drop in the graph. Hence we can conclude that the data is stationary .
The Akaike's Information Criterion values of different models implemented are shown in table below.

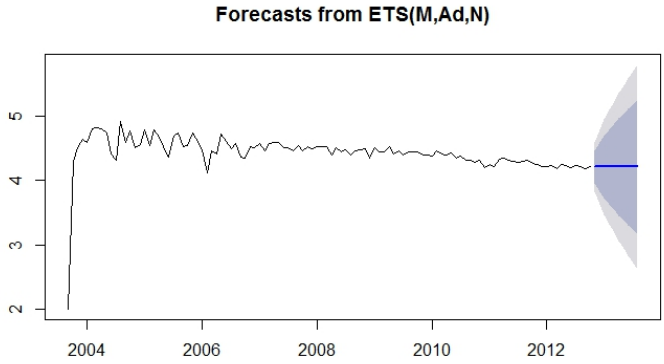

Figure 5: ETS model

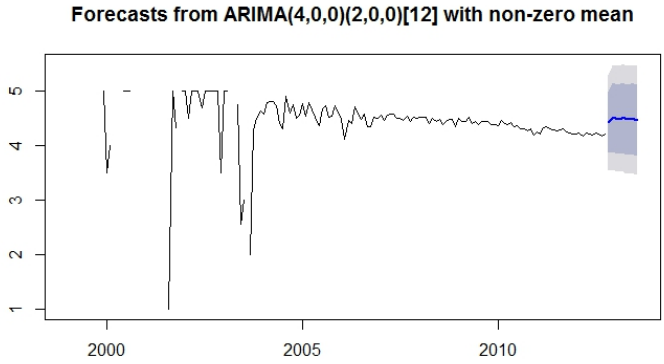

Figure 6: Auto.arima model

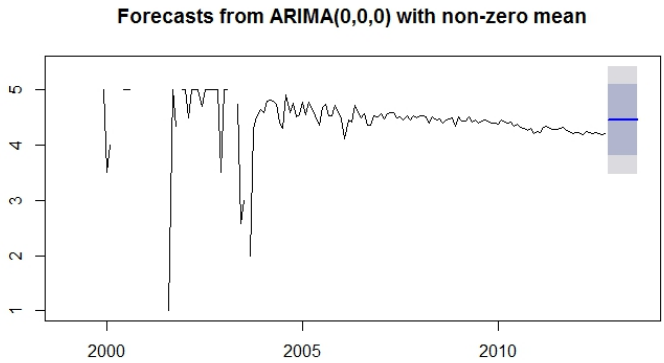

Figure 7: Holt model

From Table 1 we can conclude that ETS model has the lowest AIC value and is the best method to forcast data for the given dataset.

The plots of each model in Time Series Forecast is as shown in Figure 5 - Figure 9

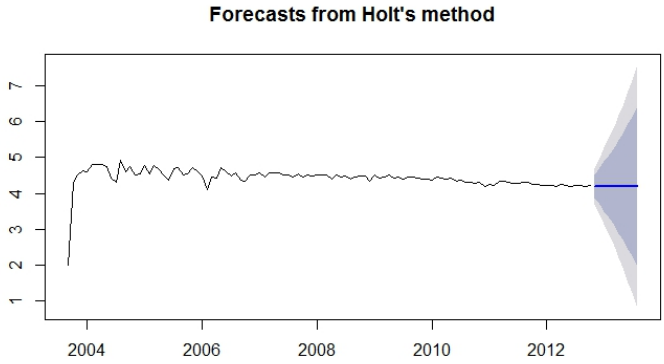

Figure 8: SES model

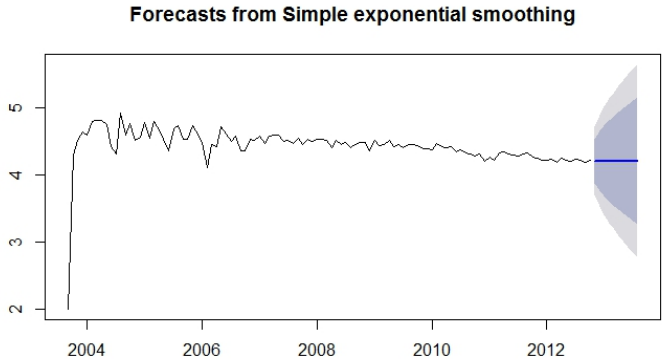

Figure 9: SNaive model

Methods	AIC	Forecast Score
ETS	178.87	4.208
AUTO.ARIMA	182.04	$4.432 - 4.514$
ARIMA	207.79	4.454
HOLT	226.1606	$4.18 - 4.20$
SES	222.5346	4.208
SNAIVE	–	$4.18 - 4.23$

Table 1: Model Comparison

10 Conclusion

Therefore, with an accuracy of 82.2%, the producer can identify fake review and aim to improve products of bad reviews to increase sales. Amazon can prevent such fake reviews from appearing on their sites that might have a negative effect on the sales of the product. The predicted value of score for the future was found to be 4.2 from the best model - ETS. This predicted score can be utilized by amazon to build their marketing strategy and focus on areas that needs improvement. The forecast plots can also be utilized to find the months in which the score was found to be low.

11 References

1. **Sentiment Analysis using product review data** : Journal of Big Data20152:5 DOI: 10.1186/s40537-015-0015-2© Fang and Zhan; licensee Springer. 2015

2. **Time Series Analysis of Household Electric Consumption with ARIMA and ARMA Models** : Journal of Proceedings of the International MultiConference of Engineers and Computer Scientists 2013 Vol I, IMECS 2013, March 13 - 15, 2013, Hong Kong

3. **Improving forecasting by estimating time series structural components across multiple frequencies** : International Journal of Forecasting, Volume 30, Issue 2, April–June 2014, Pages 291–302