

Social Media Sentiment prediction with Amazon reviews

CIS400 – Principle of Social Media Mining

Term Project Report

Prepared For:

Professor Edmund Yu

Prepared By:

Song Wei Kong – [sokong@syr.edu](mailto:sokong@syr.edu)

**Introduction**

This term project aims to explore and quantify the viability of sentiment prediction models across different online platforms. The two different platforms targeted are Twitter and Amazon. Twitter is an online microblog and social media networking platform where users can post and interact with other users using a 140-character limit message. Amazon is a leading American e-commerce company which host more than 34 million user-written reviews with ratings which provides a rich dataset for sentiment analysis. For this experiment, user review data from Amazon is used to build a sentiment prediction model on Twitter messages.

**Language & Tools**

The project is implemented in python with the following packages and libraries:

**Anaconda:** Anaconda is a freemium open source distribution of the Python and R programming languages for large-scale data processing, predictive analytics, and scientific computing. NLTK & scikit-learn, pandas and numpy included within the Anaconda package are used.

**Beautiful Soup**: A Python library for pulling data out of HTML and XML files

**Twitter API**: A Python library for interfacing to the Twitter REST and streaming APIs

**AFINN**: A wordlist-based sentiment analysis library in Python

**Wordcloud**: A python word cloud generator

**Preliminary data analysis**

The Amazon data set is retrieved from Kaggle’s Amazon Fine Reviews dataset provided by the Stanford Network Analysis Project which consists of 568,454 food reviews Amazon users left up to October 2012. Due to the large size of the dataset, the dataset was sliced and only 100,000 reviews were used to reduce the computing time and resource needed for the models.

Each amazon entry has the following fields:

|  |  |  |
| --- | --- | --- |
| **Field** | **Description** | **Sample data** |
| **Id** | Row id of entry | 1 |
| **ProductId** | Id of product | B001E4KFG0 |
| **UserId** | Id of user | A3SGXH7AUHU8GW |
| **ProfileName** | Profile name of user | delmartian |
| **HelpfulnessNumerator** | Number of users who found the review helpful | 1 |
| **HelpfulnessDenominator** | Number of users who indicated whether they found the review helpful | 1 |
| **Score** | Rating between 1-5 | 5 |
| **Summary** | Brief summary/headline of review | Good Quality Dog Food |

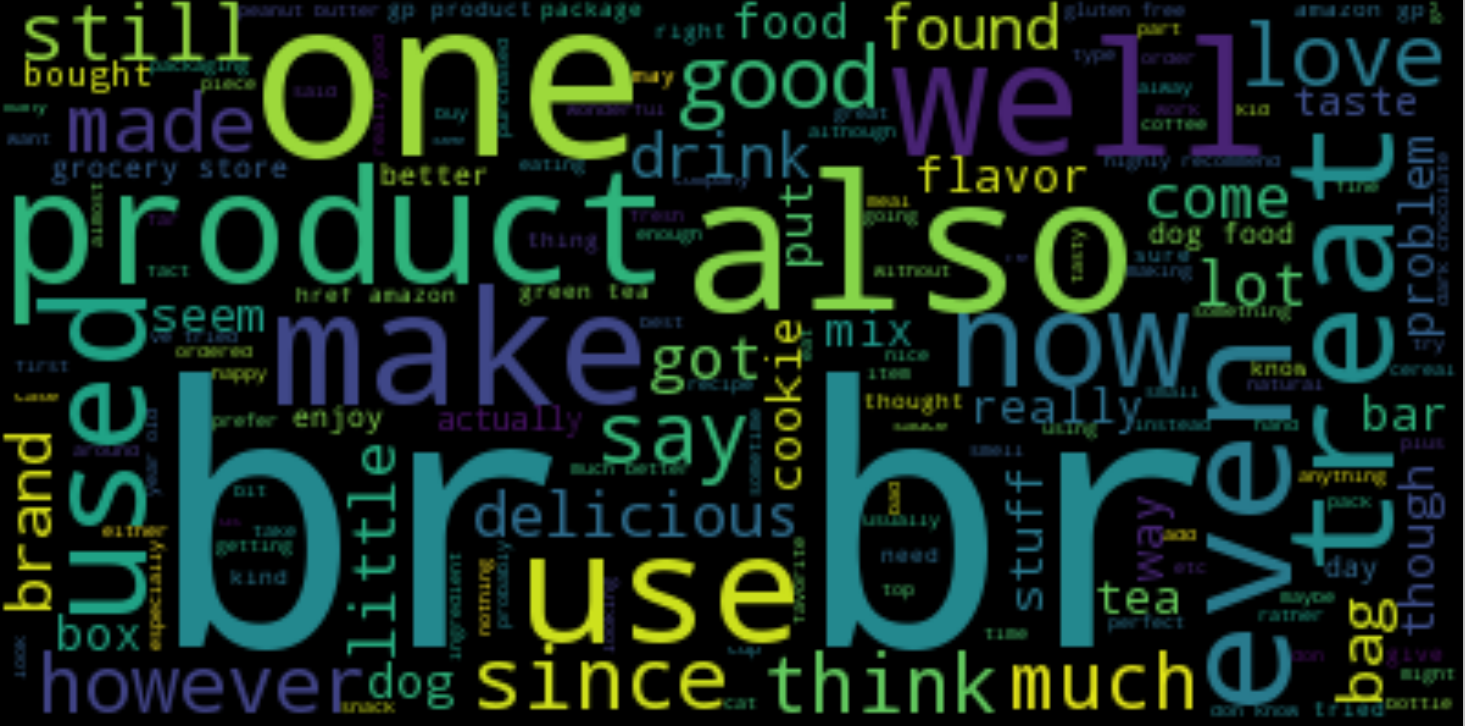
|  |  |  |
| --- | --- | --- |
| **Text** | Text of the review | I have bought several of the Vitality canned dog food products and have found them all to be of good quality. The product looks more like a stew than a processed meat and it smells better. My Labrador is finicky and she appreciates this product better than most. |
| **Sentiment** | A binarized feature created from score, with 1 = positive and 0 = negative. | 1 |

Only the score and text from the original dataset with the created feature of sentiment were used for this project.

Characteristics of the dataset:

|  |  |
| --- | --- |
| Total reviews | 100,000 |
| Total Words | 8,345,146 |
| Average words per review | 83.451 |
| No. of positive reviews (Score>3) | 77,055 (77%) |
| No. of negative reviews (Score<=3) | 22,945 (23%) |

Amazon reviews word cloud



It can be observed from the word cloud that html tags such as ‘br’ and stop words such as ‘also’ were common in the review which may not help in the prediction model.

Twitter dataset

The twitter test dataset was gathered with the python Twitter api using the search term “food”. The tweets were then labelled using a combination of the AFINN sentiment analysis library and hand-classification. It contains 51 entries with the following fields:

|  |  |  |
| --- | --- | --- |
| **Field** | **Description** | **Sample data** |
| **Text** | A tweet from twitter through the search “food” | Top 5 best feeling ever is comin home seeing the fridge/pantry packed w food |
| **Sentiment** | Prediction target. 1 = positive, 0 = negative. | 1 |

Characteristics of the twitter dataset:

|  |  |
| --- | --- |
| Total tweets | 51 |
| Total Words | 928 |
| Average words per tweet | 18.196 |
| No. of positive tweets | 29 (56.9%) |
| No. of negative tweets | 22 (43.1%) |

Twitter word cloud



A quick glance can see the frequency of the term “food” being the largest as the tweets are gathered with the search term “food”. It is interesting to note that there is less stop words as compared to Amazon review word cloud and the word choices appear to differ a lot.

Data cleaning & preprocessing

The datasets were cleaned in the following manner:

* **Removed punctuations**: punctuations do not appear to help in sentiment analysis and can interfere with tokenization of the letter. Therefore, all punctuations were removed
* **Removed HTML elements**: HTML tags such as <br> were included when crawling for the data which generally does not help with predictions
* **Removed links**: Tweets by users often include http links which provides context. However, the analysis of the linked content is outside of the scope of this project. Therefore, links were removed to focus on the message within the tweets
* **Removed stop words**: Stop words occurs frequently which doesn’t carry much meaning were removed using NLTK’s stop word list
* **Standardized all text to lower case**: python comparison is case sensitive, therefore all text were standardized to lower case for better comparison
* **Data binarization**: the amazon review score is converted into 1 for positive and 0 for negative and used as training and target labels to simplify the problem into a binary-class classification

Models

3 prediction models were created for sentiment prediction, namely the Naïve Bayes classifier, random forest classifier and the support vector machine (linear). Each model uses a 80:20 train test split on the dataset. The Naïve Bayes model used only a sample size of 20,000 reviews due to the large amount of memory required for training the model while the others used the full data set. The models trained were then used to predict the target label (sentiment) of the test results and tweets with accuracy being used as the main model evaluation method.

Result & Analysis

|  |  |
| --- | --- |
| Amazon sentiment prediction | |
| **Model** | **Accuracy** |
| Naive Bayes | 0.775 |
| Random Forest | 0.855 |
| Support Vector Machine (Linear) | 0.88 |

As we can see from the model results, SVM reported the highest accuracy out of the following closely by the random forest classifier. SVM with linear kernel was chosen as it reported the highest accuracy for this problem set as compared to other kernels ran.

|  |  |
| --- | --- |
| Twitter sentiment prediction | |
| **Model** | **Accuracy** |
| Naive Bayes | 0.569 |
| Random Forest | 0.62 |
| Support Vector Machine (Linear) | 0.58 |

The models ran for prediction on the Tweets reported significantly lower accuracy for all 3 models and the trained random forest classifier is the best performing model when predicting sentiment for tweets. The significant difference in results may be attributed to the following factors:

**Context of platform**: The context of the 2 different dataset may have contributed to the difference in prediction accuracy. The amazon reviews were written to share their thoughts on a targeted subject (the food product in question) while the tweets selected were based on the search term “food”. Therefore, the tweets may be a lot more general and wide in terms of context as compared to the Amazon training data, leading to very different lexicon as suggested by the vastly different word cloud. The models may provide a more accurate result for food reviews on other social media platform such as Instagram.

**Limitation of platform**: the 140-character limit by twitter may be a limitation for the models. The lexicon difference mentioned in regards to platform context may also be affected by the character limit of Twitter which may result in much different word choices as compared to writing for Amazon reviews. The short tweets with an average length of 18 words as compared to Amazon’s average of 83 words may also have contributed to the models’ low accuracy.

Future development

An ensemble of the different models may yield a better prediction result. Further feature pre-processing and engineering such as lemmatization as well as tuning of the models may also improve prediction performance.

Conclusion

According to Ogneva (2010), human analysis’s accuracy is estimated to be at about 79% of the time. Although the models did similar or better than that when analysing sentiments of text with similar context, the accuracy was significantly lower when the test data is from a different platform. The results appear to suggest some translatability between the two platforms across the three different models but fundamental differences between the context and limitation of the platforms may have contributed to the differences which are difficult to bridge.

Reference

**Ogneva, M. (2010, April 19). How Companies Can Use Sentiment Analysis to Improve Their Business. Retrieved May 07, 2017, from http://mashable.com/2010/04/19/sentiment-analysis/#Vv5zClkY25qa**