# Leveraging Sentiment Analyis via NLTK to Classify Electronics Reviews: Data Challenge 1

Jacob Farner
*Computer Science, COMP331*
*Occidental College*
Los Angeles, California
jfarner@oxy.edu

Leopold Ringmayr
*Computer Science, COMP331*
*Occidental College*
Los Angeles, California
lringmayr@oxy.edu

***Index Terms*—Electronics Reviews, Naive Bayes, NLTK, Sentiment Analysis, Text Classification.**

## I. Introduction

Retail purchases by individual consumers in the United States are quickly moving away from traditional brick-and-mortar store fronts toward online marketplaces such as Amazon.com, with e-commerce transactions accounting for 12.4% of all retail sales in the United States in 2020 up from 5.8% in 2012 [1]. Because consumers are unable to physically inspect products purchased online, decisions are often based on existing consumer reviews. So, customers have a vested interest in general sentiment associated with a product they are considering purchasing from those who already own the product, as well as a necessary trust in the legitimacy of the reviews they are reading.

In this paper, we use natural language processing methodology to classify product reviews gathered from Amazon.com into general categories of positive and negative opinions. This classification allows one to gauge overall sentiment toward a product better informing customers on the items they are considering purchasing. Leveraging this calculated sentiment against the star rating associated with a user review can help in flagging fake reviews which may have been generated by the seller to artificially increase their product's overall rating and subsequent sales, increasing consumer trust in the aggregate knowledge of online reviews.

## II. Related Work

Similar work in language analysis and classification has been used in various fields to gather information and gauge legitimacy of user posts, which we are able to compare our work against to confirm legitimacy of our own methodology.

In *Detecting Conditional Healthiness of Food Items From Natural Language Text*[2], Michael Wiegand and Dietrich Klakow use similar methodology to classify food items into categories describing varying levels of healthiness and suitability for individuals given their existing medical conditions. Given the immense variety of existing foods and medical conditions, there is no centralized database containing information outlining suitability for all combinations of food and afflictions. Wiegand and Klakow apply natural language processing models to text gathered from medical and food-related discussion forums in an attempt to automate the gathering of this information based on knowledge from the general population. The paper provides examples that they were able to extract information from regarding food suitability, such as:

*(a) Ginger is very good for your stomach - it helped me a lot against my heartburn.*

*(b) Iron deficiency can usually be prevented by consuming meat on a regular basis.*

In both of these examples the authors were not only able to determine suitability, but also extract information regarding how a certain food can alleviate an individual's symptoms. For this, forum posts were parsed into bags-of-words and compared against a specifically developed healthiness lexicon. We applied similar methodology in our categorization of sentiment of consumer reviews for products purchased off of Amazon.

Similar work in sentiment categorization can be seen in *Spatial and Temporal Sentiment Analysis of Twitter Data*[3] by Zhiwen Song and Jianhong (Cecilia) Xia, in which Tweets were analyzed to gauge sentiment polarity across specific time frames and geographic locations. For instance, the authors found that tweets from social science buildings on a college campus were generally more positive than those from science and engineering departments, with negativity peaking during exam periods. This was achieved through machine learning and lexicon-based methodologies including applied Support Vector Machines (SVM), Naive Bayes, and N-gram models. Our work relies largely on Naive Bayes, but it may be worth implementing other models in future iterations of this project or others.

## III. Methodology

The main components for this classification problem are the reading in and segmentation of the text data, feature engineering and data cleaning, and the classification portion on electronic reviews.

### A. Project Data and Data Preparation

For this work, the text files are read in locally; the data is stored locally within a PyCharm project. The project data consists of a total of 2000 plain text files of Amazon.com reviews for electronics. These reviews are labeled as either negative or positive, and are stored in respective directories. There are 1000 negative and 1000 positive reviews. Originally, the data is derived from a pseudo-XML format of the Amazon.com data.

For building the classifier using Python's NLTK, the reviews are read in and each file is assigned the respective classification label, stored in a list of all reviews. This list is shuffled for the separation into training and test data.

### *B. Data Cleaning and Feature Selection*

The text is tokenized using NLTK's word tokenizer function. The 2000 most fequent tokens are filtered out for further use. To improve model accuracy, certain token types are eliminated from the list of occurring tokens in the Amazon.com reviews. These tokens include stopwords, punctuation, and non-alphabetic characters. It is notable that applying a lemmatizer to the program did not significantly affect the model performance. For each document in the training set/training fold of the used cross validation, the frequency distribution of all words in the document is assessed–and the mere presence of a word in a document.

### *C. Classification*

Building the model entailed the use of 10-fold cross-validation on the 2000-file dataset. The cross validation allows for a more holistic assessment of the model's performance. This model relies on the NLTK Naive Bayes classifier. In addition to model accuracy, other performance statistics were collected, including precision, recall, and F-score for each of the labels. In addition, the ten most informative features were shown for each fold of the cross validation to give insight into which tokens were most impactful on the classification.

After the ten folds, the average of each of the performance statistics was taken (i.e. average accuracy over ten folds, average precision for the positive label, etc.). The final output consists of total accuracy along with precision and recall and F-score for each of the two labels.

In a separate file (secondmodel.py) a logistic regression model was used on the same dataset a a means for comparison. This model used the same steps for the feature selection and the same mode of tokenization, however, the accuracy of this model was on a test set after a simple training-test split (i.e. no cross-validation).

## IV. Results and Analysis

The Naive Bayes Classifier had an average accuracy of 77 percent over the 10 folds–with some folds having a classification accuracy of above 80 percent. Over the ten folds, other averaged results were the precision on the positive class (where the label ”positive” is deemed a success), at 0.725, the recall on the positive class at 0.867 percent, and the F-Score on the positive class at 0.789 percent. For the negative class (looking at the class ”negative”, for example, for filtering out negative reviews)), model performance was also assessed with a precision on the negative class at 0.835 percent, recall at 0.67 percent, and an F-score of 0.74. Thus there is a far lower recall on for the negative class as opposed to the positive class. The precision for both classes are similar as are the F-scores. In addition, some of the most informative features that occurred in the separate cross validation folds were tokens such as ”faulty”, ”waste”, or ”refund”.

In the separate file with the logistic regression model on a simple training and test data split, the accuracy was 77 percent as well. Thus, this result is not significantly higher or lower than the Naive Bayes classifier.

## V. Threats to Validity

Our model classifies user reviews of products into binary categories of positive or negative sentiment. In reality, many reviews have mixed emotions which are reflected in a star value assigned by the consumer. This project would be more comprehensive if we expanded our work to take existing star ratings into account, or assigned an estimated star rating based on calculated sentiment across a spectrum rather than strictly positive or negative. Leveraging existing star values assigned by users against the content of its associated review could also allow us to flag fake reviews in the database, increasing validity and trust of each review in our dataset.

## VI. Conclusion

This project was able to classify existing Amazon.com product reviews into general sentiment categories with an accuracy of 77 percent. We used a variety of approaches including the Naive Bayes classifier and logistic regression, eventually settling on the Naive Bayes model, relying on cross-validation, as we believe it provided the most robust model given the dataset. For future work, it would be interesting to explore more nuanced sentiment analysis techniques to potentially improve the feature selection once our dataset. Furhtermore, it wouls be productive to include a comprehensive comparison between the logistic regression model and the Naive Bayes model.

This project also acted as a comprehensive introduction to natural language processing through the Natural Language Tool Kit (NLTK), and provided us with a foundation of knowledge to expand upon in future projects.

## Acknowledgment

Thank you to Professor Chen, who taught us NLTK fundamentals and the theoretical and practical underpinnings of natural language processing.

## References

[1] Clement, J. “United States: e-Commerce Share of Retail Sales 2021.” Statista, 23 July 2019, www.statista.com/statistics/379112/e-commerce-share-of-retail-sales-in-us/.

[2] Wiegand, Michael, and Dietrich Klakow. ”Detecting Conditional Healthiness of Food Items from Natural Language Text.” Language Resources and Evaluation 49, no. 4 (2015): 777-830. Accessed February 23, 2020. www.jstor.org/stable/24710076.

[3] ISong, Zhiwen, and Jianhong (Cecilia) Xia. ”Spatial and Temporal Sentiment Analysis of Twitter Data.” In European Handbook of Crowdsourced Geographic Information, edited by Capineri Cristina, Haklay Muki, Huang Haosheng, Antoniou Vyron, Kettunen Juhani, Ostermann Frank, and Purves Ross, 205-22. London: Ubiquity Press, 2016. Accessed February 23, 2020. www.jstor.org/stable/j.ctv3t5r09.20.