**Sentiment Analysis for Amazon Mobile Reviews**

**A.  Research Question**

My research question for my project today is “Could Text mining technique make accurate prediction for sentiment analysis?”

“Sentiment analysis is the interpretation and classification of emotions (positive, negative, and neutral) within text data using text analysis techniques. Sentiment analysis tools allow businesses to identify customer sentiment toward products, brands, or services in online feedback.” (MonkeyLearn, n.d.) In modern data world, there is about 80% of the data is unstructured or unorganized. With tremendous amount of text data such as emails, chat, social media conversations, surveys, articles, reviews, and many more had been created every day but hard to analyze them. Sentiment analysis help the businesses classifying and tagging the unstructured data. (MonkeyLearn, n.d.)

**Hypothesis**:

Null Hypothesis (H0)**:** Amazon mobile reviews dataset indicates no significant accurate predictions for Sentiment Analysis.

Alternative Hypothesis (H1): Amazon mobile reviews dataset indicates significant accurate predictions for Sentiment Analysis.

When we make the predictions of sentiment analysis, Amazon mobiles reviews, and rating data set could be used for my project. Those traditional machine learning models such as logistics regression and NaÏves Bayes could be used for large scale sentiment analysis since they scale well. (Komban, Narayan, Orsi 2019) Sentiment analysis has been widely used in the modern world. Many businesses look forward to predicting how likely the consumer will purchase their products. The data analyst could collect the data from an e-commerce site such as eBay and Amazon, social media sites such as Twitter, Facebook, etc. to make the prediction. With my project, I will use logistic regression and NaÏves Bayes methods with cross-validation to predict the outcome of each mobile phones’ ratings based on the previous Amazon mobile’s consumer reviews.

**B.  Data Collection**

The dataset is public information from Amazon.com. The dataset is non-restricted information that could help me with my sentiment analysis project. It contains over 400,000 product reviews that are large enough for my capstone project study. In the modern world, sentiment analysis can play a vital role in any industry. Classifying tweets, Facebook comments or even product reviews from e-commerce site using an automated system that can save a lot of time and money. Most importantly, it’s more accurate. (Nasrin Sucky, 2019)

This Amazon mobile reviews dataset contains 413840 observations and 6 attributes such as “Product Name”, “Brand Name”, “Price”, “Rating”, “Reviews”, and “Review Votes”. This dataset was achieved at data.world. It was extracted by PromptCloud to use for the public. (PromptCloud n.d.) The dataset was extracted in December of 2016 from Amazon.com with over 400 thousand reviews of unlocked mobile phones that sold. The file is located at <https://data.world/promptcloud/amazon-mobile-phone-reviews>

**Attribute Information:**

|  |  |
| --- | --- |
| *Field* | *Type* |
| *Product Name* | *Categorical* |
| *Brand Name* | *Categorical* |
| *Price* | *Continuous* |
| *Rating* | *Integer* |
| *Reviews* | *Text* |
| *Review Votes* | *Continuous* |

**Advantages:**

The advantage of sentiment analysis of data is to use robust techniques to quickly analysis of language, word usage and writer interpretation. We could also do digital editing, mapping and also visualization. (Danjustis, 2015) The business could benefit from it with information extraction, information retrieval, categorization, clustering, and summarization. The sentiment analysis technique could easily extract information from different resources. Such as web pages, social media, reviews, surveys and many others. As an example with my project, I intend to use the Amazon mobile reviews (Public reviews data) to determine how likely will potential consumer buy the mobile products with the previous reviews information we extracted from the website.

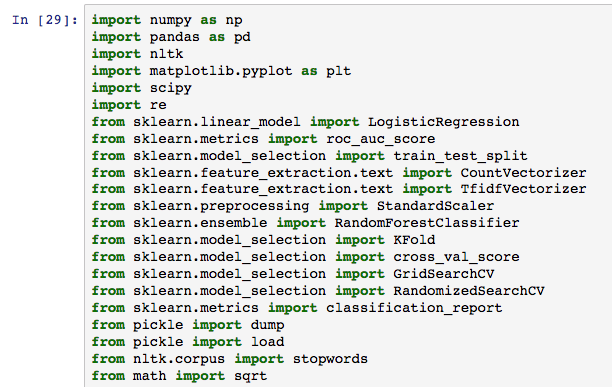
**Disadvantages:**

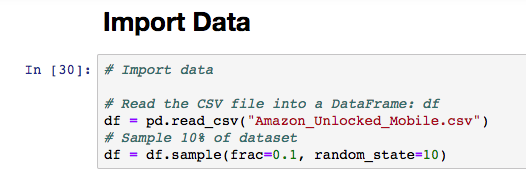
The downside with the sentiment analysis of data is that we will need to be aware if the information we extracted or obtained from the web against the privacy policy. We also need to beware that data accuracy and security concerns. We will need to ensure that all the information collected is for the ethical purposes and will not be misused for unethical used. (Zentut, n.d.)

**C.  Data Extraction and Preparation**

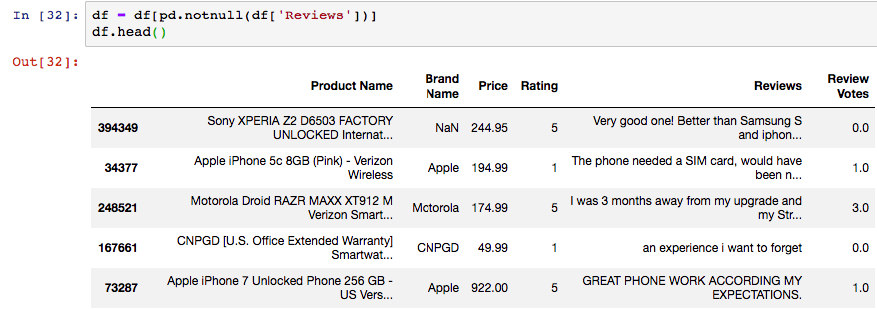
This Amazon mobile reviews dataset contains 413840 observations and 6 attributes such as “Product Name”, “Brand Name”, “Price”, “Rating”, “Reviews”, and “Review Votes”. I use the following steps for data cleaning: I remove stopwords, tokenize, and capitalize on the dataset. I also remove unwanted characters and punctuation. I drop all the rows with missing values. I drop rows with the rating with 3, which indicate neural review. I also conduct data transformation for target variable. I assume the rating with greater than 3 are rated as positive (1), and the remaining are rated as negative (0). I feature text to convert text into vectors as an independent variable using countvectorizer, tfidfvectorizer, and n-gram.

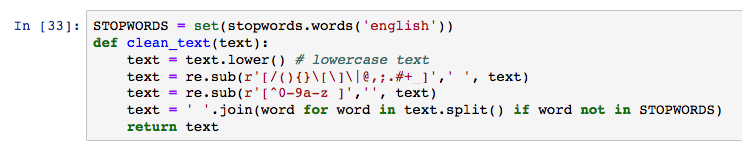
**Import libraries**

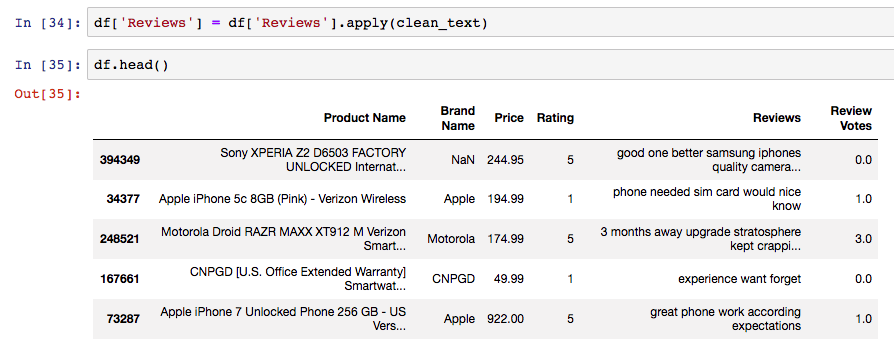


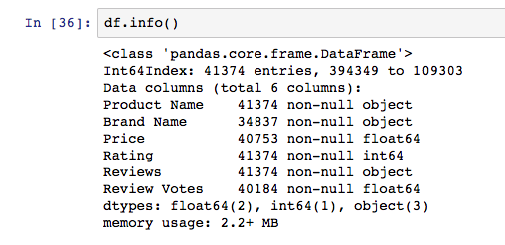


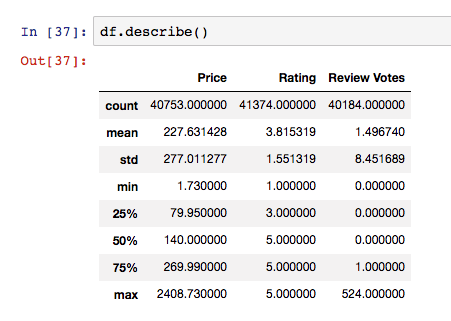


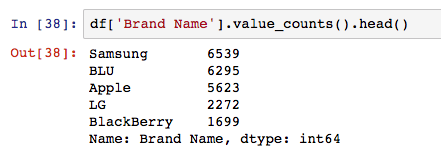




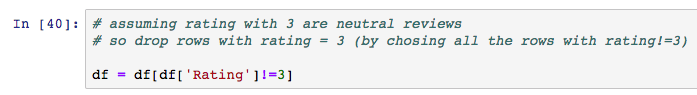


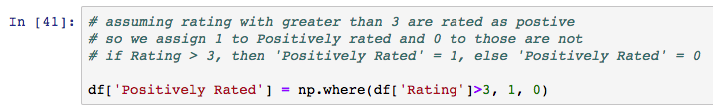


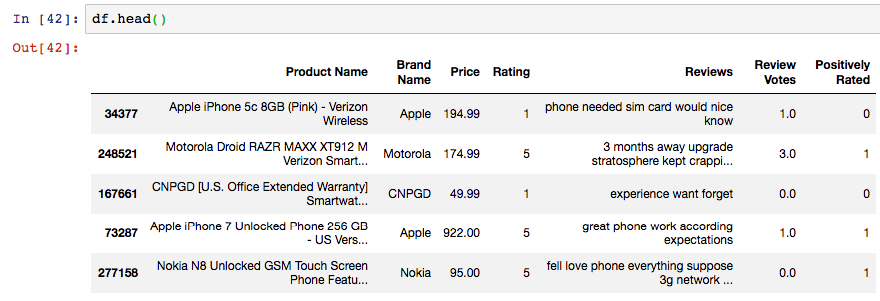


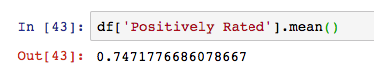




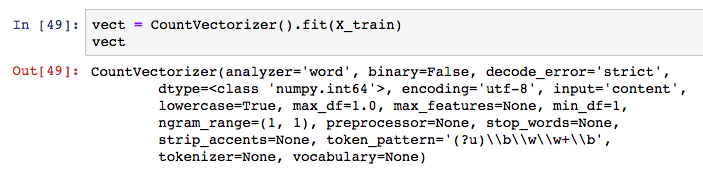


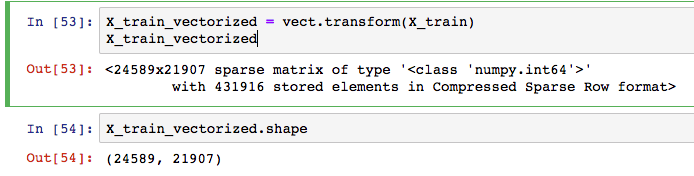




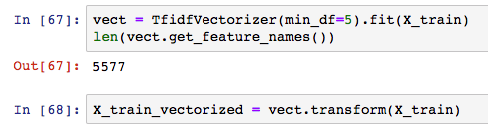




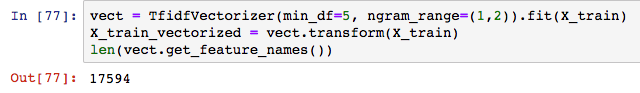












**D.  Analysis**

To investigate the sentiment analysis of Amazon mobiles’ dataset, logistics regression, and Naïve Bayes model are implemented in Python (scikit-learn). Machine Learning and text mining pipelines are conducted to develop, tune, and optimize the model. Rating with positive and negative values is treated as a dependent variable. I feature text to convert text into vectors as an independent variable using countvectorizer, tfidfvectorizer, and n-gram. For logistic regression, cross-validation is applied and two important hyperparameters such as Hypermeters-C and Penalty are tuned. For Naïve Bayes, cross-validation is applied, and alpha is tuned. These two conventional machine algorithms Naïve Bayes and logistic regression are very popular and an efficient algorithm to conduct text mining with very high accuracy. Text mining pipelines include import data, data cleaning, feature creation, train and tune model, and test model to make the prediction. The dataset contains 413840 observations, and I sample 10% of the dataset. I will use 80% as train and validation dataset, and 20% as test dataset.

Python is an open-source and general-purpose programming language, which is the most popular programming language for machine learning and artificial intelligence. Python is cross-platform compatibility and it’s an objected-oriented tool. Python is also a preferable programming language and it’s much faster than “R” programming language. Scikit-learn is a great library ecosystem used for handling Machine Learning algorithms like classification and logistic regression. (Alferd, 2019)

In CountVectorizer section, logistic regression and Naïve Bayes algorithms used. Cross-Validation is applied, and Hyper-Parameter is tuned with random search.

For logistic regression classifier, 5-folds cross-validation is applied and two Hypermeters-C and Penalty are tuned. C describes the inverse of regularization strength. Like in support vector machines, smaller values specify stronger regularization. C is set to 15 numbers spaced evenly on a log scale, the starting point is 1e-05, the stop point is 100000000. Penalty is used to specify the norm used in the penalization. Penalty is set to “l1”, “l2”. After cross validation, the best parameters are selected. The best model with highest accuracy is that C is set to 0.4393970560760795 and Penalty is set to l2.

For Naïve Bayes classifier, 5-folds cross-validation is applied and one hypermeter alpha is tuned. alpha is set to 10 numbers spaced evenly between 0 and 1. After cross validation, the best parameters are selected. The best model with highest accuracy is that alpha is set to 0.4.

In TfidefVectorizer section, logistic regression and Naïve Bayes algorithms used. Cross-Validation is applied, and Hyper-Parameter is tuned with random search.

For logistic regression classifier, 5-folds cross-validation is applied and two Hypermeters-C and Penalty are tuned. C describes the inverse of regularization strength. Like in support vector machines, smaller values specify stronger regularization. C is set to 15 numbers spaced evenly on a log scale, the starting point is 1e-05, the stop point is 100000000. Penalty is used to specify the norm used in the penalization. Penalty is set to “l1”, “l2”. After cross validation, the best parameters are selected. The best model with highest accuracy is that C is set to 0.05179474679231213 and Penalty is set to l2.

For Naïve Bayes classifier, 5-folds cross-validation is applied and one hypermeter alpha is tuned. alpha is set to 10 numbers spaced evenly between 0 and 1. After cross validation, the best parameters are selected. The best model with highest accuracy is that alpha is set to 0.4.

In TfidefVectorizer with n-gram section, logistic regression and Naïve Bayes algorithms used. Cross-Validation is applied, and Hyper-Parameter is tuned with random search.

For logistic regression classifier, 5-folds cross-validation is applied and two Hypermeters-C and Penalty are tuned. C describes the inverse of regularization strength. Like in support vector machines, smaller values specify stronger regularization. C is set to 15 numbers spaced evenly on a log scale, the starting point is 1e-05, the stop point is 100000000. Penalty is used to specify the norm used in the penalization. Penalty is set to “l1”, “l2”. After cross validation, the best parameters are selected. The best model with highest accuracy is that C is set to 268.2695795279727 and penalty is set to l2.

For Naïve Bayes classifier, 5-folds cross-validation is applied and one hypermeter alpha is tuned. alpha is set to 10 numbers spaced evenly between 0 and 1. After cross validation, the best parameters are selected. The best model with highest accuracy is that alpha is set to 0.1.

AUC is selected to evaluate performance cross-validation. The results of cross-validation are shown below.

|  |  |  |
| --- | --- | --- |
|  | **Cross-Validation Results** | |
| **Logistic Regression** | **NaÏve Bayes** |
| **CountVectorizer** | **0.959506** | **0.946520** |
| **Tfidfvectorizer** | **0.956385** | **0.964037** |
| **Tfidfvectorizer with n-gram** | **0.965729** | **0.970141** |

As we can see the AUC of Tfidfvectorizer with n-gram method with Naïve Bayes is the highest, and logistic regression method is second highest.

**Advantage:**

The advantage of these two conventional machine-learning algorithms is that they could be widely used for large scale sentiment analysis since they scale well. (Komban, Narayan, Orsi, 2019) They are efficient algorithms to build the model to make prediction of sentiment analysis.

**Disadvantage:**

The disadvantage of the study is that we only use logistic regression and Naïve Bayes models and more machines techniques could be used on the project such as SVM, XGBoost, etc. for further study to conduct model selection.

**E.  Data Summary and Implications**

Logistic regression and Naïve Bayes is implemented to build the model to make prediction and feature is created with countvectorizer, tfidfvectorizer, and tfidfvectorizer with n-gram. Accuracy, AUC, Precision, Recall, and F1 Score are selected to evaluate the performance of test set, which they can help to determine how good the model generalize to the unseen data.

The results of the performance of the test set of these methods are shown in the table below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Metrics** | **CountVectorizer** | | **Tf–idfVectorizer** | | **TfidfVectorizer with n-gram** | |
| Logistic Regression | NaÏve Bayes | Logistic Regression | NaÏve Bayes | Logistic Regression | NaÏve Bayes |
| **Accuracy** | **0.96** | **0.95** | **0.96** | **0.96** | **0.97** | **0.97** |
| **AUC** | **0.88** | **0.88** | **0.72** | **0.85** | **0.90** | **0.88** |
| **Precision** | **0.92** | **0.91** | **0.87** | **0.91** | **0.92** | **0.92** |
| **Recall** | **0.92** | **0.92** | **0.86** | **0.91** | **0.93** | **0.92** |
| **F1 Score** | **0.92** | **0.91** | **0.84** | **0.90** | **0.92** | **0.92** |

Six models by using machine learning methods such as logistic regression, Naïve Bayes, and three feature creation methods such as countvectorizer, tfidfvectorizer, and tfidfvectorizer with n-gram are developed. The highest accuracy model is TfidfVectorizer with n-gram methods with logistic regression and Naïve Bayes, which are 0.97. The highest AUC model is TfidfVectorizer with n-gram methods with logistic regression, which is 0.90. The highest precision model is TfidfVectorizer with n-gram methods with logistic regression and Naïve Bayes, which are 0.92. The highest recall model is TfidfVectorizer with n-gram methods with logistic regression, which is 0.93. The highest F1 score model is TfidfVectorizer with n-gram methods with logistic regression and Naïve Bayes, which are 0.92.

Based on our analysis, logistic regression model with TfidfVectorizer with n-gram methods is selected as the best model since it has the highest accuracy, AUC, precision, recall, F1 score, and it perform very well for sentiment analysis on a large-scale of dataset to make prediction. In conclusion, Amazon mobile reviews dataset indicates significant accurate predictions for Sentiment Analysis.

One direction of future study is to use many other machine learning algorithms, such as Gradient boosting method, Xgboost and SVM (Support Vector Machine) and deep learning algorithms such as MLP (Multiple-layer Perceptron) and RNN (Recurrent Neural Network) to improve the performance of model. Also trying to tune more hyper-parameters in a wide range can be used to find the best combination of hyper-parameters. The other direction of future study is to use the word embedding to create the feature. The “Word embedding” is a type of word representation that allows words with similar meaning to have a similar representation. (Brownlee, 2017) This is a very effective method to generate the model to unseen text.

**Reference:**

Alferd S. (November 21, 2019) Advantages of Using Python for Machine Learning. Retreived from <https://medium.com/@stenalferd/advantages-of-using-python-for-machine-learning-9568a2a62524>

Brownlee Jason (October 11, 2017) What are word embeedings for text? Retrieved from <https://machinelearningmastery.com/what-are-word-embeddings/#:~:text=A%20word%20embedding%20is%20a,challenging%20natural%20language%20processing%20problems>.

Danjustis (July 11, 2015) What are the advantages and disadvantages of text mining historical texts? *Digging up the past* Retrieved from <https://dannajustis.wordpress.com/2015/07/11/what-are-the-advantages-and-disadvantages-of-text-mining-historical-texts-what-did-you-learn-about-your-soldiers-life-from-ngram-viewers/>

Komban S, Narayan R. Orsi G. (August 22, 2019) Deep Learning Models for Sentiment Analysis Retrived from <https://underthehood.meltwater.com/blog/2019/08/22/deep-learning-models-for-sentiment-analysis/#:~:text=Traditional%20machine%20learning%20methods%20such,analysis%20because%20they%20scale%20well>.

MonkeyLearn (n.d.) Sentiment Analysis the baics explaination retrieved from <https://monkeylearn.com/sentiment-analysis/>

Nasrin Sucky, Rashida (Jun 2019) Sentiment Analysis in Python with Amazon Product Review Date Retreived from <https://towardsdatascience.com/sentiment-analysis-in-python-with-amazon-product-review-data-b4e75b5d8e78>

PromptCloud (n.d.) Amazon Mobile Phone Reviews Retrieved from <https://data.world/promptcloud/amazon-mobile-phone-reviews>

Zentut (n.d.) Advantages and Disadvantages of Data Mining Retretried from <https://www.zentut.com/data-mining/advantages-and-disadvantages-of-data-mining/>