Sentiment Predictions and Analysis on Amazon Reviews of Cell Phones and Accessories

Mohil Tanti  
Master of Data Science  
Swinburne University of Technology  
Melbourne, Victoria  
mohiltanti7@gmail.com

Table of Contents

[Introduction 2](#_Toc55849632)

[Methodology 3](#_Toc55849633)

[A. Data Source 3](#_Toc55849634)

[B. Data Preprocessing 3](#_Toc55849635)

[I. Data Loading and Cleaning 3](#_Toc55849636)

[II. Data Preprocessing 3](#_Toc55849637)

[C. Sentiment Prediction and Analysis 3](#_Toc55849638)

[1. Automate Approach 3](#_Toc55849639)

[2. Rule-based Approach 4](#_Toc55849640)

[Findings 5](#_Toc55849641)

[1. Number of Reviews based on Sentiments. 5](#_Toc55849642)

[2. Number of words used in reviews. 6](#_Toc55849643)

[3. Products Likeliness. 6](#_Toc55849644)

[4. Most used words for different Sentiments. 7](#_Toc55849645)

[Conclusion 8](#_Toc55849646)

[GITHUB Repository 8](#_Toc55849647)

[References 8](#_Toc55849648)

# Introduction

From research, it was derived that from the overall world’s data almost 80% of it is unstructured. Tons of data are created on daily basis from various sources such as emails, social media, surveys, articles, documents, etc. It is quite difficult to analyze, sort, understand this amount of data, also is time-consuming and very expensive. Here sentiment analysis helps businesses make sense of this lots of unstructured text by analyzing, understanding, processing, and tagging it in a much smarter way and automatically.

Sentiment analyses are text qualitative mining, recognizing and collecting subjective data from databases, and helping an organization recognizes the social sentiment of its company, product, or service as online conversations are tracked. The capability of algorithms has increased sufficiently because of the recent advances in deep learning. Creative AI techniques can be used to execute in-depth research on social media streams as they are usually restricted to just basic sentiment analysis and count-based metrics.

Various types of Sentiment Analysis can be conducted on the data based on the requirement; the types are:

1. Standard sentiment analysis
2. Fine-grained Sentiment Analysis
3. Emotion detection
4. Aspect-based sentiment analysis
5. Intent detection

Sentiment analysis utilizes various NLP (Natural Language Processing) techniques and algorithms, they are:

1. **Automatic** –the system usesmachine learning algorithms to learn and classify the data.
2. **Rule-based** – the system uses human developed rules to identify subjectivity, polarity, or subject.
3. **Hybrid** – the system uses a combination of rule-based and automatic approach

# Methodology

## Data Source

Dataset required for the study was chosen from the blog of Prof. Julian McAuley, an Associate Professor at the University of California, San Diego. The data was sufficient in size to predict and analyze the sentiments to a quite significant accuracy. The dataset description is provided in below Table 1:

*Table 1: Dataset Description*

|  |  |
| --- | --- |
| **Title** | Amazon Product Data (Cell Phones and Accessories) |
| **Description** | This dataset contains reviews from Amazon spanning May 1996 - July 2014 for Cell Phones and Accessories. |
| **Attributes** | reviewerID, asin, reviewerName, helpful, reviewText, overall, summary, unixReviewTime, reviewTime |
| **Dataset Size** | 194439 |

## Data Preprocessing

For the sentiment analysis, some processing of data was required to successfully implement NLP techniques on the review dataset. The preprocessing steps were:

1. Data Loading and Cleaning

Firstly, the JSON file of the dataset was loaded, and visualizing of the data was performed, also empty rows were removed from the dataset to avoid any casualties.

1. Loading and visualizing the data.
2. Removing the unnecessary rows.
3. Data Preprocessing

Secondly, after cleaning the data, using Natural Language Toolkit(nltk) the reviews were processed to feed it to Machine Learning Classifier. NLTK provides a collection of libraries and programs for the symbolic and statistical NLP for English written text in Python.

1. Changing Reviews to lowercase for uniformity.
2. Tokenizing the review strings.
3. Removing StopWords
4. Removing Punctuations
5. Stemming the words

## Sentiment Prediction and Analysis

To perform sentiment predictions on reviews, a hybrid approach was utilized as a methodology. The hybrid approach is explained below:

### **Automate Approach**

Machine Learning algorithm was used to predict the sentiments automatically. Random Forest Classifier was used to predict the individual words from a particular review from the list of stemmed reviews list prepared in the pre-processing step. As the labels were not assigned to reviews, the Random Forest Classifier was trained with the words for different sentiments such as positive, negative, and spam. Then the output of the classifier would be a list of predicted sentiments for reviews and then mean/average is taken to classify the whole review in the next step.

*Table 1: Labels and Their Encoding*

|  |  |
| --- | --- |
| **Label** | **Encoded** |
| Negative | 0 |
| Positive | 1 |
| Spam | 2 |

Example:

**Review:** They look good and stick good! I just don't like the rounded shape because I was always bumping it and Siri kept popping up and it was irritating. I just won't buy a product like this again

**Stemmed Review:** ['look', 'good', 'stick', 'good', "n't", 'like', 'round', 'shape', 'alway', 'bump', 'siri', 'kept', 'pop', 'irrit', 'wo', "n't", 'buy', 'product', 'like']

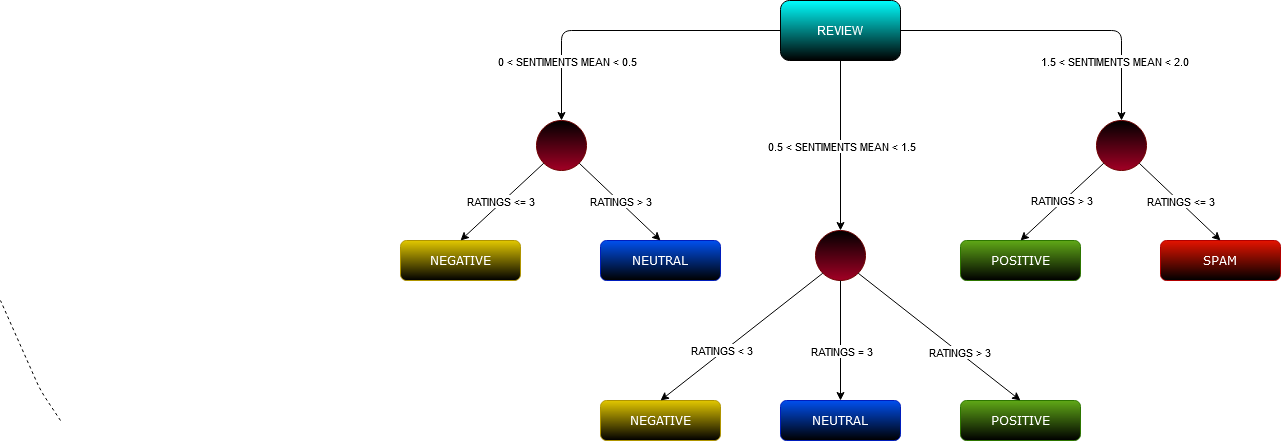
**Random Forest Classified List:** [ 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 0 2 0 1]

The model was not accurate in classifying the words, but around 70% accuracy was achieved.

### **Rule-based Approach**

Human-crafted rules are used in the rule-based approach, in that the mean of every review was taken and based on mean scores, along with ratings the reviews were classified in 4 sentiments. The mean of classified sentiments for individual reviews is taken and is matched with certain rules and based on the mean and ratings the review is classified under 4 sentiments: “POSITIVE”, “NEUTRAL”, “NEGATIVE”, and “SPAM”.

The below tree shows the classification rules.

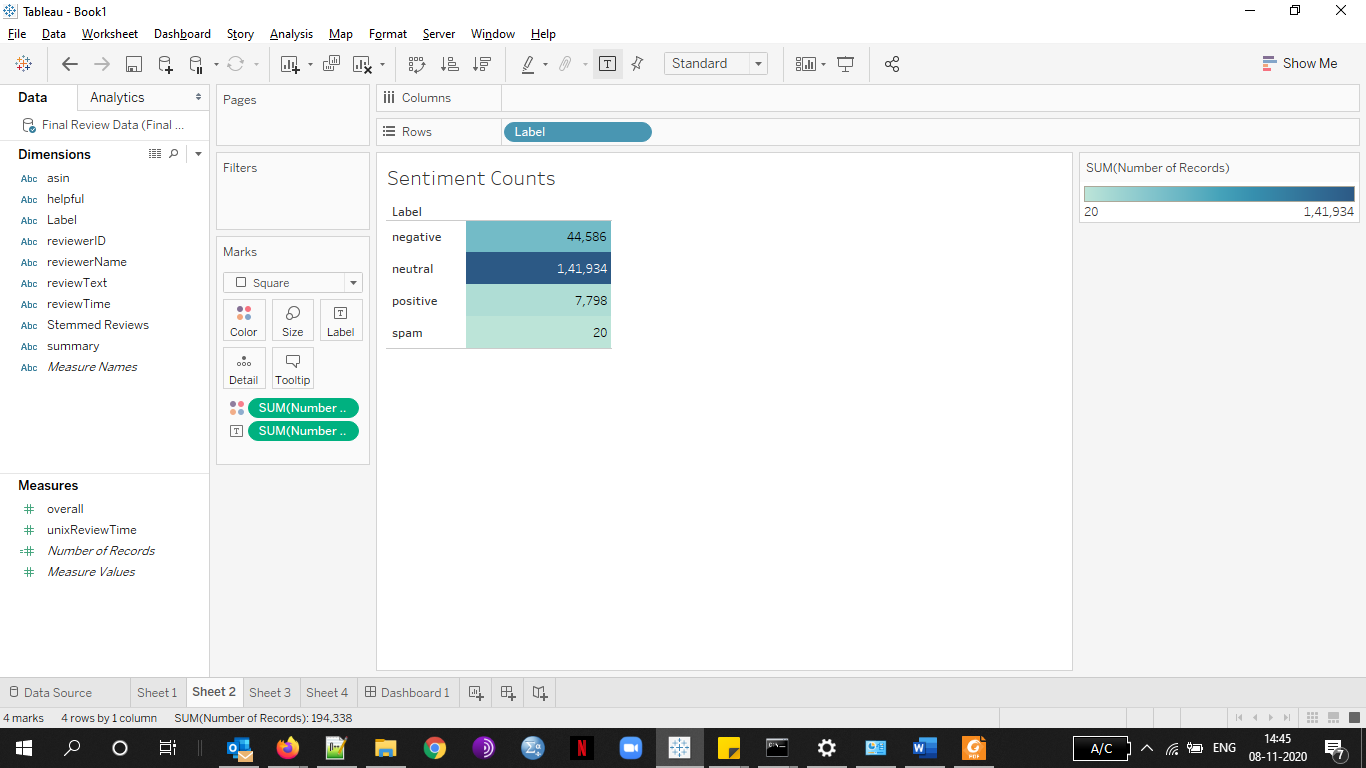
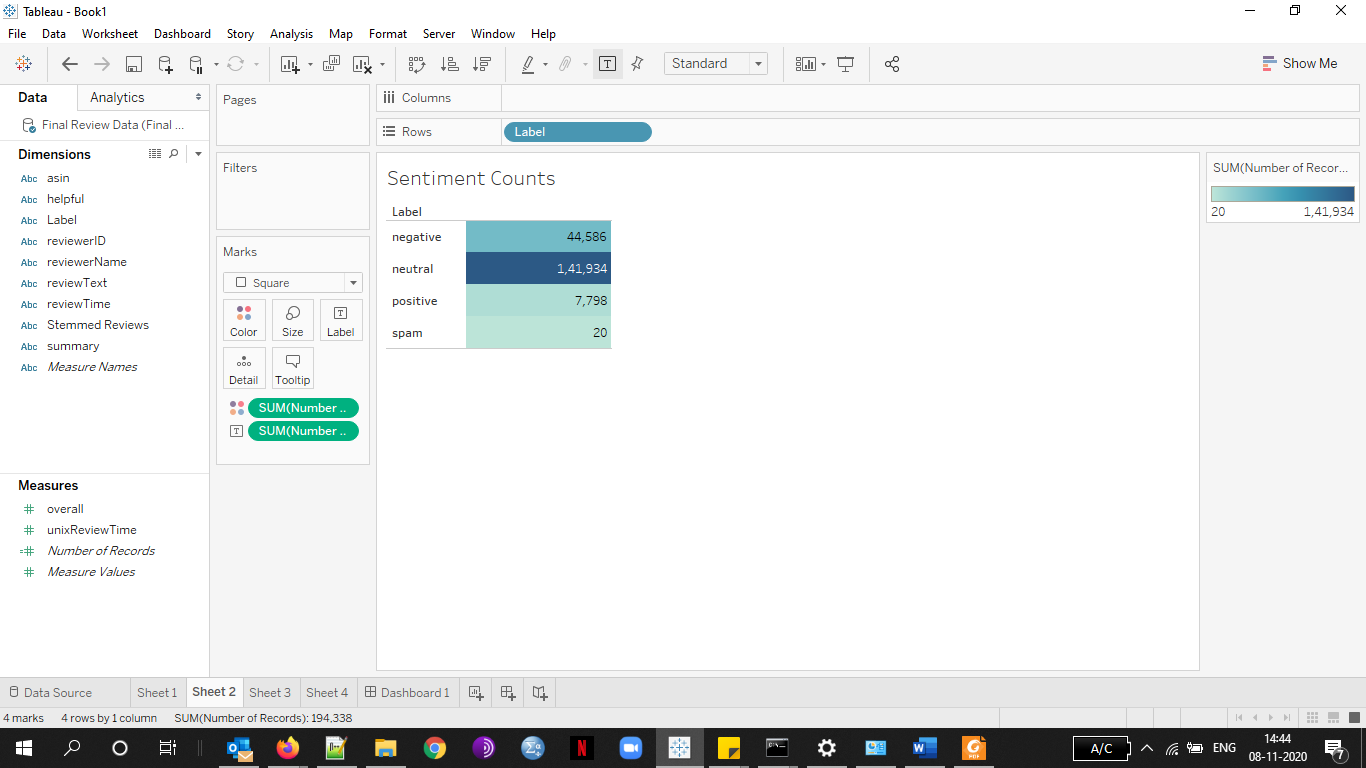


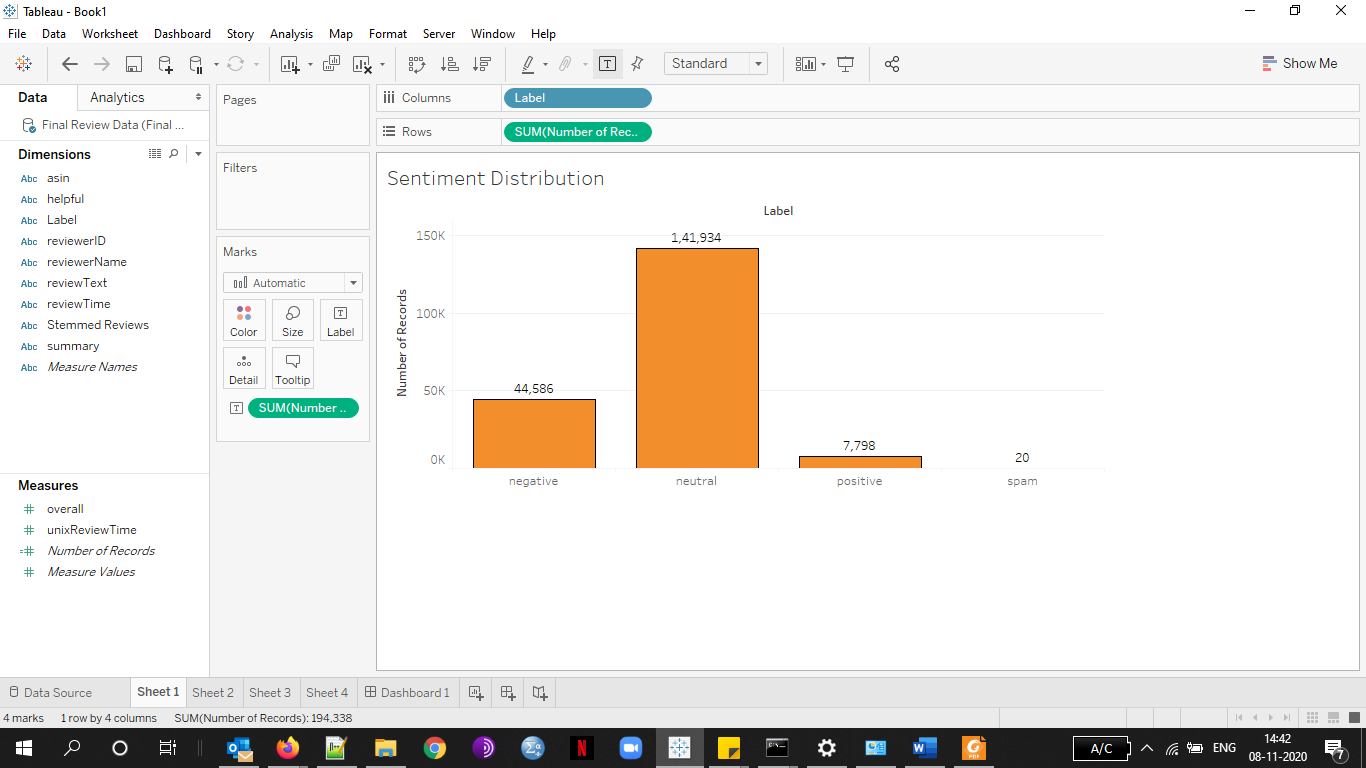
# Findings

After predictions of sentiments for the reviews, analysis of the sentiments and product likeliness was performed. Visualizations were derived from the data using Tableau Software. The major findings from the analysis were:

### **Number of Reviews based on Sentiments.**

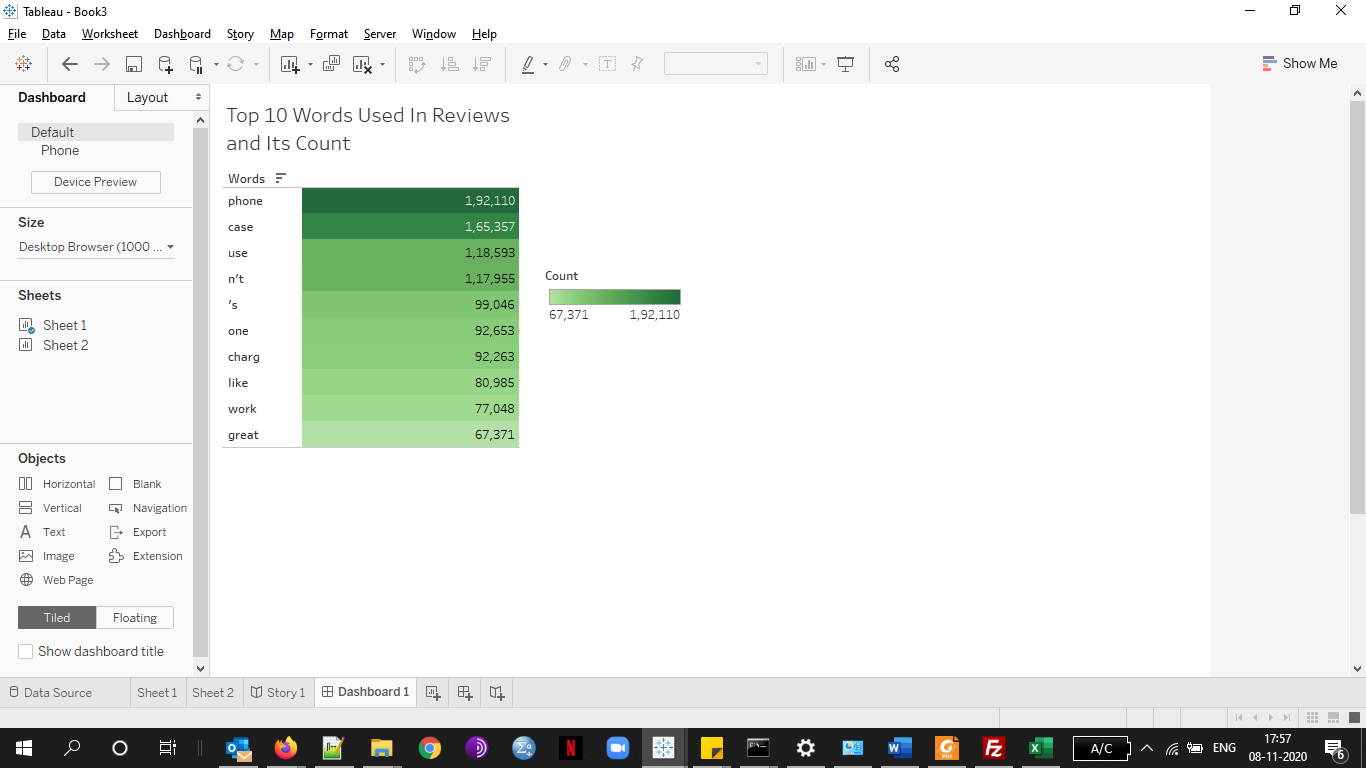
The distribution of reviews is described in the below figure. It shows that the maximum number of reviews was Neutral i.e. 1,41,934, followed by 44,586 Negative reviews, also 7,798 and 20 number of Positive and Spam reviews respectively.





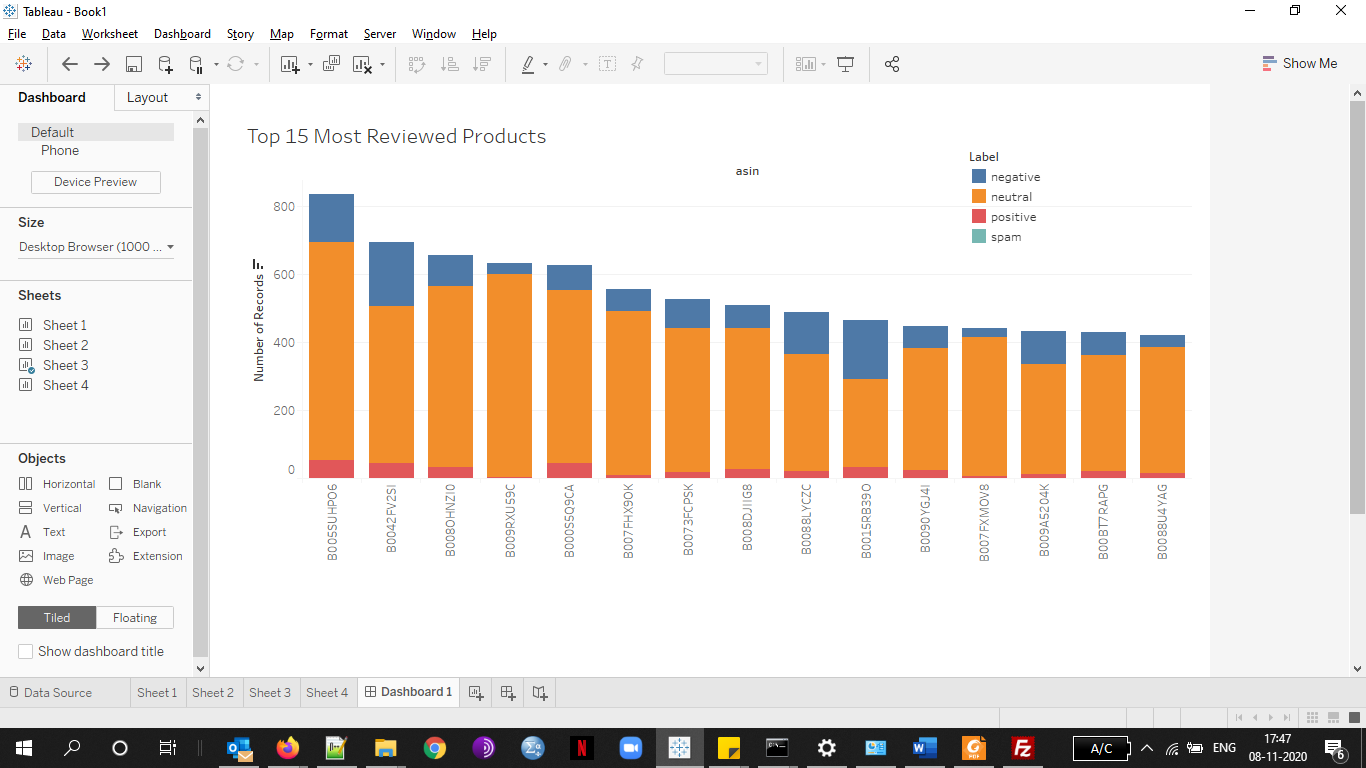
### **Number of words used in reviews.**

The total distinct words in all reviews were calculated and sentiments were predicted on them. The graph shows the composition of the words. Also, the counts and sentiments for the top 10 used words are displayed below in the table. As the dataset is of cell phones and accessories the maximum used words were “phone” and “case” i.e. more than 1,50,000 occurrences.



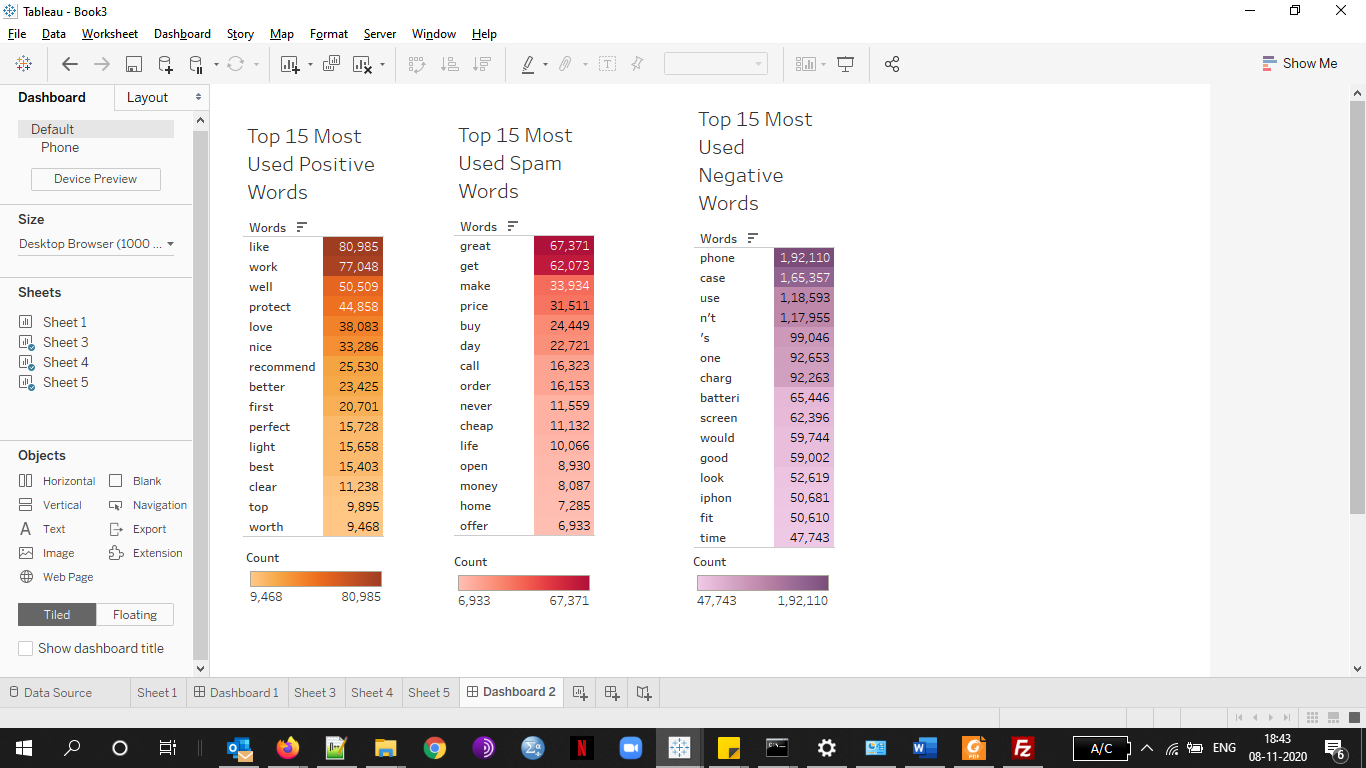
### **Products Likeliness.**

Getting insights on whether products and detecting whether the product is good or bad, can be done efficiently with the help of sentiment analysis. The below insights show the top 15 products that were highly reviewed and observing whether the product was loved or not by the customers. The stacked bar chart depicted the products having product Id “B005SUHPO6”, “B009RXU59C”, “B000S5Q9CA”, and “B007FXMOV” seems to be promising to customers with minimal negative reviews and maximum positive and neutral reviews. In contrast, the products having product Id “B0042FV2SI”, and “B001SRB39O” can be classified as a bad product as having maximum negative reviews and minimum positive reviews.



### **Most used words for different Sentiments.**

Different types of words are used to express different sentiments, below are the insights with the top 15 words with their count and to which sentiment they belong. The top 15 most used positive words ranged from 9468 to 80985 where “like”,”work”, and ”well” were the top 3 words respectively. The top 15 most words used spam words ranged from 6933 to 67371 in which “great”, “get” and “make” were the most 3 used words. While considering negative words the “phone”, “case” and “use” were the top 3 words used in the range of 47743 to 192110 words in the top 15 list.



# Conclusion

To conclude, through this project, sentiment analysis has proven to be a quick and efficient way of analyzing and getting fruitful outcomes which could be beneficial for the business to improve their products according to the reviews provided by the users. The analysis such as which product is best based on its ratings and reviews, different types of reviews and their counts, different words used in reviews to identify the most occurred problem, and much more could give in-depth analysis for the products that a particular business supplies. Although using more better and much accurate classifying algorithm could give much more accurate results.

There were some key takeaways from the project, that were working with Big Data, the process of Sentiment Analysis, implementing Natural Language Processing (NLP) techniques, and Machine Learning algorithms to classify text and Visualization of data using Tableau.

# GITHUB Repository

<https://github.com/MohilTanti/Sentiment-Predictions-and-Analysis-on-Amazon-Reviews-of-Cell-Phones-and-Accessories->

# References

McAuley, J 2020, *Amazon review data*, viewed 9 November, 2020, <http://jmcauley.ucsd.edu/data/amazon/>.

Anon 2020, *5 Real-World Sentiment Analysis Examples*, viewed 9 November, 2020, <https://monkeylearn.com/blog/sentiment-analysis-examples/>.

Sharma, P 2020, *6 NLP Datasets Beginners should use for their NLP Projects | MLK - Machine Learning Knowledge*, viewed 9 November, 2020, <https://machinelearningknowledge.ai/nlp-datasets-beginners-should-use-for-nlp-projects/>.

Anon 2020, *Sentiment analysis*, viewed 9 November, 2020, <https://en.wikipedia.org/wiki/Sentiment\_analysis>.

Anon 2020, *Sentiment Analysis: A Definitive Guide*, viewed 9 November, 2020, <https://monkeylearn.com/sentiment-analysis/>.

Hana, L 2020, *A step by step Guide on Sentiment Analysis with RNN and LSTM*, viewed 9 November, 2020, <https://medium.com/@lamiae.hana/a-step-by-step-guide-on-sentiment-analysis-with-rnn-and-lstm-3a293817e314>.

Yiu, T 2019, *Understanding Random Forest*, viewed 9 November, 2020, <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>.

Anon 2020, *Natural Language Toolkit — NLTK 3.5 documentation*, viewed 9 November, 2020, <https://www.nltk.org/>.