Infosys Springboard Internship 4.0

Sentiment Analysis and Text Classification

Name: Aritra Ganguly
Email: ganguly.aritra.03@gmail.com

Infosys Springboard Intern 4.0
Domain: Data Visualisation

Introduction

In case of Natural Language Data, **Sentiment Analysis** is a crucial tool to **categorize** texts. It is an unsupervised form of **Artificial Intelligence** and **Machine Learning** Techniques.

Text Classification involves assigning labels to text data, such as **Customer Reviews, Descriptions, Social Media Posts** etc. This enables organizations to **structure and manipulate** the data and generate **meaningful insights.** This process may be both **Supervised** or **Unsupervised** form of **Machine Learning**.

Business Use Cases 

- **Customer Service Optimization:** Analyzing customer inquiries to categorize and prioritize support tickets based on sentiment and urgency.
- **Product Development:** Identifying emerging trends and customer needs from reviews to guide new product features and improvements.
- **Fraud Detection**: Classifying text in emails and messages to identify suspicious or fraudulent activities and prevent potential scams.

About Datasets

Here, we have completed out work with two Datasets:

- **Customer Review Dataset:** This dataset is used to preprocess and detect the sentiments regarding the reviews given by the customers.

- **Emotions Training Dataset:** Since the Text Classification project was a form of Supervised Machine Learning, we used a Natural Language Dataset which was already Labelled with corresponding Emotions.

Data Preprocessing

Natural Language Data Preprocessing consists a few number of steps. We can set up a **pipeline** to perform these steps serially on the Natural Language Dataset. This pipeline includes the following steps:

- **Lower Case**
- **Removal of links**
- **Removal of next lines (\n)**
- **Removal of Words containing numbers**
- **Removal of Extra spaces**
- **Removal of Special characters**
- **Removal of stop words**
- **Stemming**
- **Lematization**

Sentiment Analysis

Sentiment Analysis is an **Unsupervised** form of **Machine Learning** for **Natural Language Data.** It classifies the given text data into **Three different categories - Positive, Negative** and **Neutral.**

There are some **Python Machine Learning Modules** for **Sentiment Analysis,** such as **“TextBlob”** and **“VaderSentiment”.** Based on the given text data, they assign a score based on the **Keywords** used in the texts. Using the scores we can analyze the sentiment of the data.

Sentiment Analysis on Customer Reviews Data

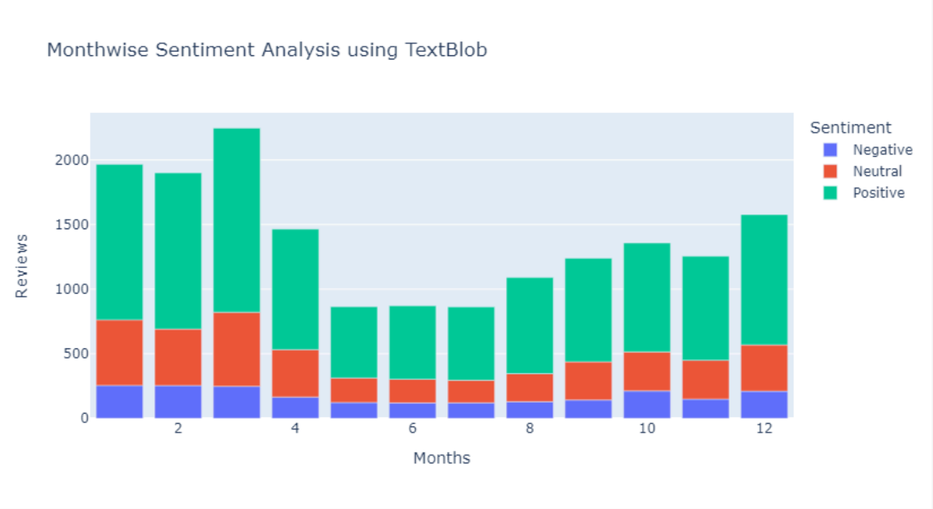
The **Customer Reviews Dataset** consists of application reviews given by the Customers. We check and **assign labels** about what **sentiment** they represent.

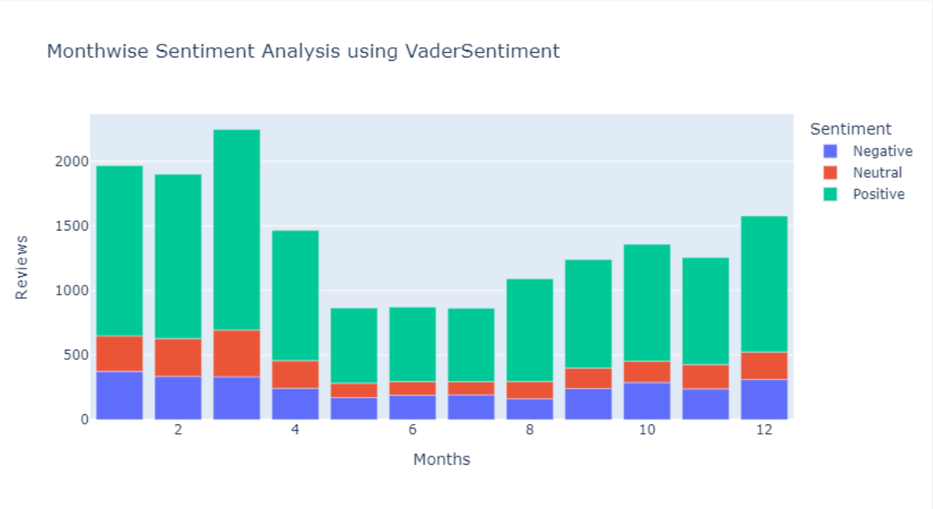
Here are some Insights of the dataset that we are using:

Sentiment Analysis on Customer Reviews Data

As mentioned earlier, **Sentiment Analysis** is a form of **Unsupervised Machine Learning.** We used some **Python Modules** like **TextBlob** and **VaderSentiment** for the **analysis** process.

The **corresponding labels** are assigned in the dataset itself. Here are some **visual representations** of the **Sentiment Analysis reults:**

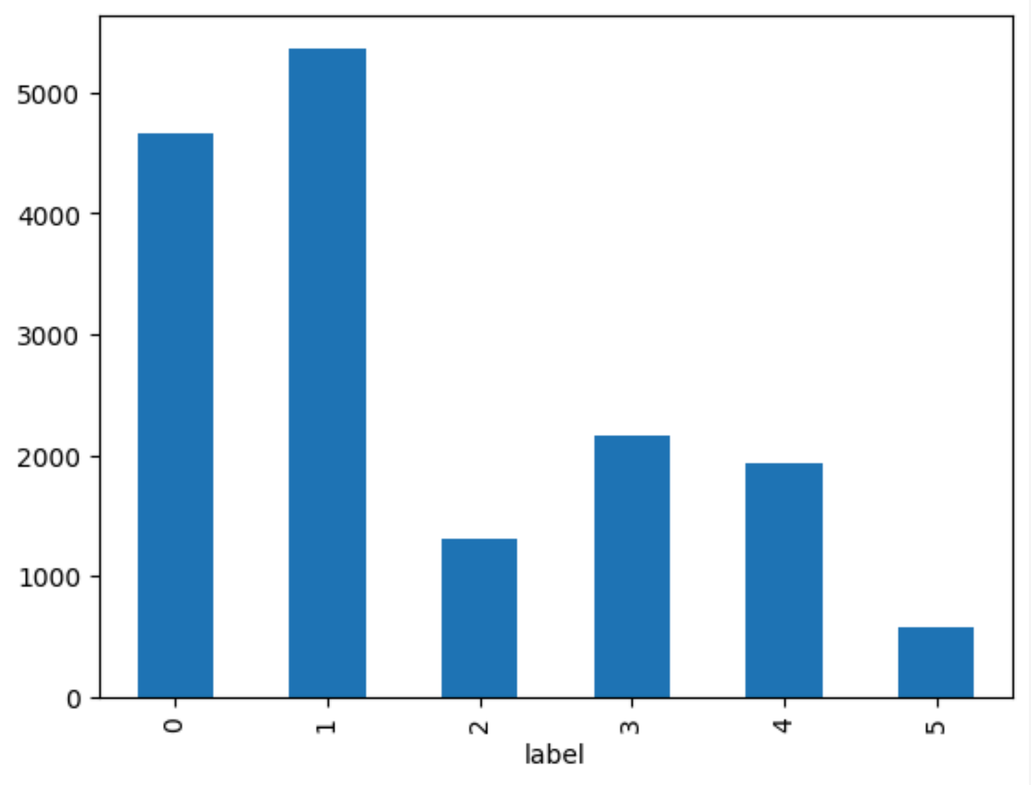




Text Classification on Emotion Training Data

The **Emotion Training Dataset** consists of **Pre-labelled text data.** This dataset is already **Label-encoded** i.e. each data is labelled with an Integer that corresponds to an **Emotion** portrayed by that text. There are **6 different emotion classes** shown in this dataset: **Sadness, Joy, Love, Anger, Fear, Surprise**

This dataset has **Imbalanced number of classes.**



Text Classification Approach 1: Machine Learning

We vectorized the **Emotion Training Dataset** to use it to train some **Machine Learning Models.** Here are the Models and their Corresponding **Accuracy Scores:**

- **Gaussian Naive Bayes Model: 35.46875 %**
- **Multinomial Naive Bayes Model: 76.90625 %**
- **Random Forest Classifier: 84.28125 %**
- **Extreme Gradient Boosting Classifier: 83.96875 %**

Text Classification Approach 2: LSTM Neural Network

We used the **Emotion Training Dataset** to train a **Deep Learning Model** that uses “**Long Short Term Memory**” in **Neural Networks.** The Structure of the **Neural Network** is shown here.

This Model got an **Accuracy Score** of **89.53125 %**

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_1 (Embedding)     (None, 300, 40)           538200

 dropout_2 (Dropout)         (None, 300, 40)           0

 lstm_1 (LSTM)               (None, 100)               56400

 dropout_3 (Dropout)         (None, 100)               0

 dense_1 (Dense)             (None, 6)                 606

=================================================================
Total params: 595206 (2.27 MB)
Trainable params: 595206 (2.27 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
None
```

Conclusion

Sentiment Analysis and **Text Classification** are **classifying the text data for further segmentation** to **organize** and **structure** them.

There are many **Business Use Cases** where implementing these approaches are essential. With this **Data Science Project**, we learned how to work with **text data** in these scenarios.

THANK YOU