Report on

SENTIMENT ANALYSIS AND TEXT CLASSIFICATION

By

Aritra Ganguly

Data Visualisation

Infosys Springboard Internship 4.0

Email: [ganguly.aritra.03@gmail.com](mailto:ganguly.aritra.03@gmail.com)

Under the Mentorship of:

Mr. Sudheer Kumar Yudunuri

CONTENTS

|  |  |
| --- | --- |
| Topic | Page No. |
| Abstract | 3 |
| Introduction | 4 |
| Datasets | 5 |
| Data Preprocessing | 6 |
| Sentiment Analysis on Customer Reviews Data | 7 |
| Text Classification on Emotion Training Dataset | 8 |
| Text Classification Approach 1: Machine Learning | 9 |
| Text Classification Approach 2: LSTM based Neural Network | 10 |
| Business Use Cases | 11 |
| Conclusion | 12 |

ABSTRACT

This report explores the application of machine learning techniques in sentiment analysis and text classification, highlighting their significance in modern business contexts. Sentiment analysis, the process of identifying and categorizing opinions expressed in text, enables companies to gauge public sentiment, enhance customer service, and refine marketing strategies. Text classification, which involves categorizing text into predefined categories, aids in automating content moderation, spam detection, and document organization. The report delves into various machine learning algorithms such as Naive Bayes, Random Forest Classifiers, and deep learning models, emphasizing their effectiveness and practical implementation. Case studies and performance metrics illustrate the transformative impact of these technologies.

INTRODUCTION

In the realm of natural language data, **sentiment analysis** is a crucial tool for **categorizing texts** based on the emotions and opinions they convey. This technique, an **unsupervised** form of **artificial intelligence and machine learning,** empowers businesses to interpret and respond to customer sentiments effectively. By analyzing **customer reviews, social media posts**, and other textual data, companies can gain **valuable insights** **into public perception, enhance customer service, and refine their marketing strategies.**

**Text classification**, on the other hand, involves assigning labels to various types of text data, such as **customer reviews, descriptions, and social media posts**. This process enables organizations to structure and manipulate vast amounts of unstructured data, making it easier to generate meaningful insights. **Text classification** can be performed using both **supervised and unsupervised machine learning** techniques. In **supervised learning**, algorithms are trained on labeled data, learning to classify texts accurately based on predefined categories. In contrast**, unsupervised learning** does not rely on labeled data; instead, it identifies patterns and structures within the data to categorize texts.

Both **sentiment analysis and text classification** play a pivotal role in transforming raw textual data into actionable intelligence. These techniques allow businesses to automate content moderation, detect spam, organize documents, and monitor brand reputation, ultimately leading to more informed decision-making and enhanced operational efficiency.

DATASETS

In our project, we utilized two distinct datasets to achieve comprehensive analysis and insights:

* **Customer Review Dataset:**

This dataset was pivotal for preprocessing and detecting sentiments in customer reviews. By leveraging sentiment analysis techniques, we could accurately determine the emotional tone behind each review. This enabled us to understand customer satisfaction levels, identify potential issues, and improve overall service quality. The insights gained from this dataset are crucial for businesses to refine their strategies and enhance customer experiences.

* **Emotions Training Dataset:**

For our text classification project, which employed supervised machine learning, we utilized a natural language dataset that was pre-labeled with corresponding emotions. This labeled dataset was essential for training our machine learning models to recognize and classify different emotional tones in text accurately. By using a supervised approach, the models learned to associate specific textual patterns with predefined emotional categories, ensuring high accuracy in text classification tasks.

The combination of these datasets allowed us to tackle both sentiment analysis and text classification effectively. The Customer Review Dataset provided a real-world application of sentiment detection, while the Emotions Training Dataset facilitated precise text classification through supervised learning. Together, these datasets played a crucial role in transforming unstructured text data into structured, actionable insights, thereby supporting better decision-making and enhancing operational efficiency.

DATA PREPROCESSING

Natural Language Data Preprocessing is a crucial step in preparing textual data for analysis. To streamline this process, we can set up a pipeline to perform these preprocessing steps sequentially on the natural language dataset. The pipeline includes the following essential steps:

* **Lower Case:** Converting all text to lower case ensures uniformity and prevents mismatches due to case differences.
* **Removal of Links:** Extracting hyperlinks from the text to focus solely on the content and reduce noise.
* **Removal of New Lines (\n):** Eliminating newline characters to maintain consistent text formatting and facilitate easier processing.
* **Removal of Words Containing Numbers:** Discarding words that include numbers, as they often do not contribute to the sentiment or context.
* **Removal of Extra Spaces:** Removing additional spaces to standardize text spacing and enhance readability.
* **Removal of Special Characters:** Stripping special characters to focus on the meaningful words within the text.
* **Removal of Stop Words:** Eliminating common stop words (e.g., "the," "is," "in") that do not add significant meaning to the text.
* **Stemming:** Reducing words to their root form to ensure different variations of a word are treated as the same term.
* **Lemmatization:** Converting words to their base or dictionary form, which helps in understanding the actual meaning of the word.

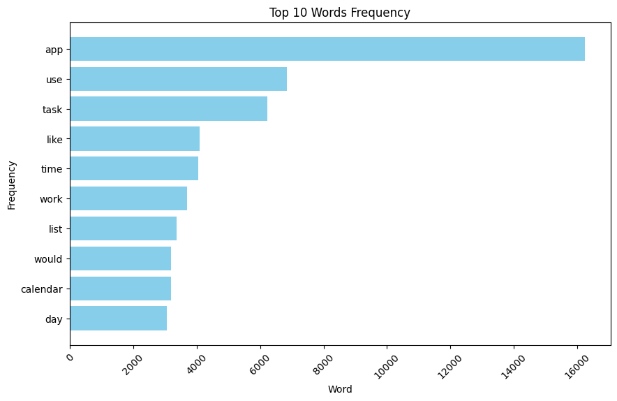
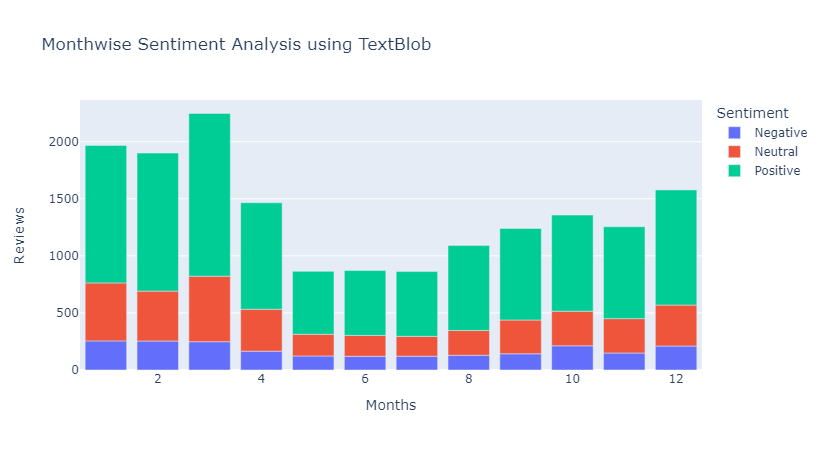
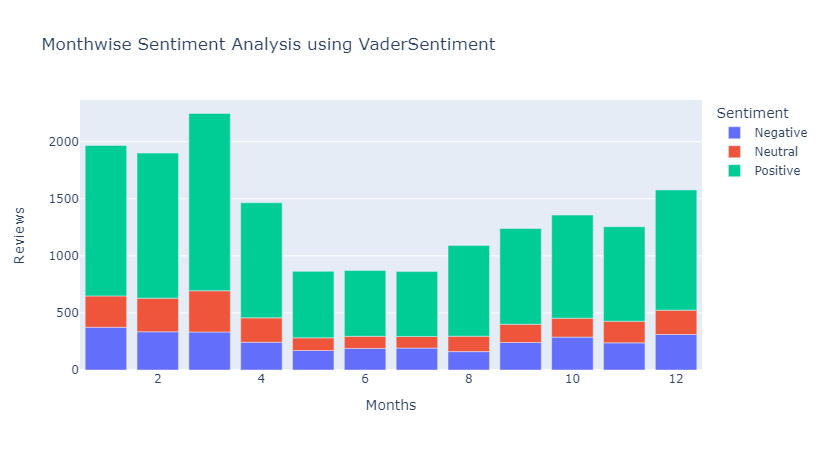
Implementing this pipeline ensures that the natural language data is clean, consistent, and ready for further analysis. By systematically preprocessing the text, we can enhance the performance of sentiment analysis and text classification models, leading to more accurate and meaningful insights. This structured approach to data preprocessing is fundamental for achieving high-quality results in natural language processing tasks.

SENTIMENT ANALYSIS ON CUSTOMER REVIEWS DATA

**Sentiment Analysis** is an essential **unsupervised machine learning technique** for **processing natural language data**. It categorizes text data into **three distinct sentiment categories: Positive, Negative, and Neutral**. This classification helps businesses understand **the emotional tone behind textual data**, such as customer reviews, social media posts, and feedback.

Several Python machine learning modules, such as **"TextBlob"** and **"VaderSentiment,"** facilitate sentiment analysis. These tools analyze the given text data and assign sentiment scores based on the keywords and context within the text. **TextBlob** provides a simple **API** for diving into **common natural language processing tasks**, while **VaderSentiment** is specifically attuned to social media texts and can accurately assess the sentiment of such content.

In our project, we applied sentiment analysis to the Customer Reviews Dataset, which consists of application reviews provided by customers. By leveraging **TextBlob and VaderSentiment**, we processed these reviews to determine their sentiment scores. Each review was then classified into one of the **three sentiment categories—Positive, Negative, or Neutral**—based on the assigned scores.

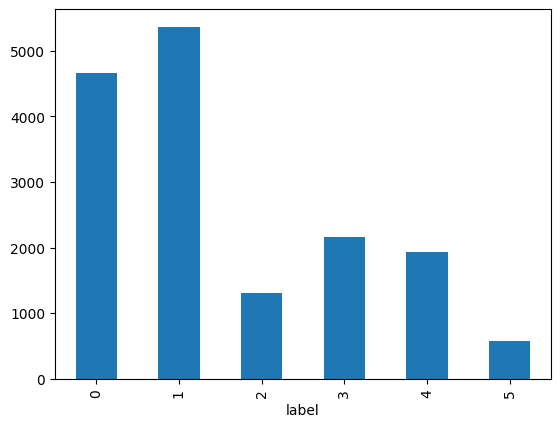
****This **sentiment analysis process** enabled us to gain valuable insights into customer opinions and experiences. By labeling the reviews with their corresponding sentiments, we could identify trends, highlight areas for improvement, and enhance customer satisfaction. The ability to systematically analyze and classify customer sentiments is crucial for making **data-driven decisions and refining business strategies.**

TEXT CLASSIFICATION ON EMOTION TRAINING DATASET

The Emotion Training Dataset comprises **pre-labeled text data**, where each entry is assigned an integer label corresponding to an emotion depicted by the text. The dataset features six distinct emotion classes: **Sadness, Joy, Love, Anger, Fear, and Surprise.** Notably, the dataset exhibits an **imbalanced distribution** of these classes, meaning some emotions are represented more frequently than others. This imbalance poses a challenge for training machine learning models, as it may affect the model's ability to **accurately recognize less frequent emotions**. Addressing this imbalance is crucial for developing robust and reliable emotion classification models.

Labels for Corresponding Emotions are:

* **0 : Sadness**
* **1 : Joy**
* **2 : Love**
* **3 : Anger**
* **4 : Fear**
* **5 : Surprise**



TEXT CLASSIFICATION APPROACH 1: MACHINE LEARNING

The Emotion Training Dataset comprises pre-labeled text data, with each entry encoded as an integer corresponding to an emotion conveyed by the text. This dataset includes **six distinct emotion classes: Sadness, Joy, Love, Anger, Fear, and Surprise.** However, it presents a challenge due to the imbalanced distribution of these classes, where some emotions are underrepresented compared to others.

To train machine learning models effectively, we first **vectorized** the Emotion Training Dataset, converting the textual data into **numerical representations suitable for model training**. This preprocessing step is crucial for enabling the algorithms to process and learn from the text data.

We then trained several **machine learning models** using the vectorized data and evaluated their performance based on accuracy scores. The models and their corresponding accuracy scores are as follows:

* **Gaussian Naive Bayes Model:** Achieved an accuracy of **35.46875%.** This model, while simple, struggled with the complexity and imbalance of the dataset.
* **Multinomial Naive Bayes Model:** Demonstrated a significant improvement with an accuracy of **76.90625%,** showing better handling of the categorical nature of the text data.
* **Random Forest Classifier:** Delivered an accuracy of **84.28125%,** indicating its robustness and ability to handle imbalanced datasets effectively through ensemble learning.
* **Extreme Gradient Boosting Classifier:** Achieved an accuracy of **83.96875%,** showcasing its strength in capturing complex patterns in the data despite the class imbalance.

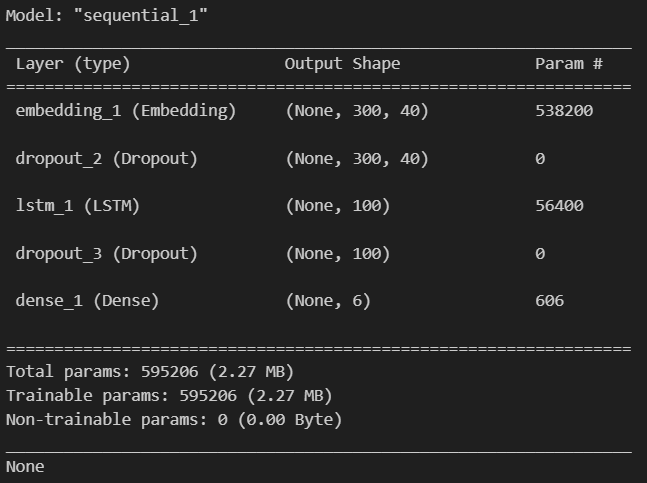
This **data modeling** approach highlights the importance of choosing appropriate models and preprocessing techniques to handle imbalanced datasets and achieve high accuracy in emotion classification tasks.

TEXT CLASSIFICATION APPROACH 2: LSTM BASED NEURAL NETWORK

We used the Emotion Training Dataset to train a deep learning model employing **Long Short Term Memory (LSTM)** networks. **LSTM** is a specialized type of **recurrent neural network (RNN)** designed to capture **long-term dependencies in sequential data**, making it particularly effective for **natural language processing** tasks. LSTM networks are **capable of remembering information for extended periods**, which is essential for **understanding the context and nuances** in text data.

The structure of our **LSTM-based neural network** was carefully designed to process the vectorized text data from the Emotion Training Dataset. This model was trained to recognize six different emotion classes: **Sadness, Joy, Love, Anger, Fear, and Surprise.** Despite the imbalanced distribution of these classes, the LSTM network demonstrated remarkable performance.

Our **LSTM model** achieved an impressive accuracy score of **89.53125%,** surpassing traditional machine learning models. This high accuracy underscores the effectiveness of LSTM networks in **handling complex, imbalanced datasets and capturing intricate patterns** in text. This approach highlights the potential of deep learning in advancing emotion classification and enhancing the understanding of natural language data.



BUSINESS USE CASES

There are numerous business use cases where implementing sentiment analysis and text classification approaches are essential. Through this data science project, we learned how to effectively work with text data in various real-world scenarios.

* One significant application is **Customer Service Optimization.** By analyzing customer inquiries, businesses can categorize and prioritize support tickets based on **sentiment and urgency**. This ensures that **urgent or negative sentiments are addressed promptly, enhancing customer satisfaction and retention.**
* Another critical use case is **Product Development**. By scrutinizing customer **reviews and feedback**, companies can **identify emerging trends and unmet needs.** This insight guides the development of new **product features and improvements,** aligning offerings more closely with **customer desires and market demand.**
* **Fraud Detection** is also a vital application of these techniques. By classifying text in emails and messages, businesses can **detect suspicious or fraudulent activities**. This proactive approach helps in **preventing potential scams and safeguarding company assets and customer information.**

Overall, implementing sentiment analysis and text classification can significantly benefit various business operations. These techniques enable organizations to transform unstructured text data into actionable insights, leading to more informed decision-making and enhanced operational efficiency. This project underscores the importance and versatility of data science in addressing diverse business challenges through advanced text analytics.

CONCLUSION

In conclusion, this report underscores the transformative potential of sentiment analysis and text classification techniques in extracting meaningful insights from natural language data. Through the exploration of various machine learning models and deep learning architectures, we have demonstrated their efficacy in addressing a wide range of business use cases.

From customer service optimization to product development and fraud detection, the applications of sentiment analysis and text classification are diverse and impactful. By leveraging these approaches, businesses can gain a deeper understanding of customer sentiments, identify emerging trends, and mitigate risks effectively.

Moreover, our analysis of different machine learning models, including Gaussian Naive Bayes, Multinomial Naive Bayes, Random Forest Classifier, Extreme Gradient Boosting Classifier, and LSTM networks, highlights the importance of choosing the right algorithm for the task at hand. Each model offers unique strengths and capabilities, and selecting the most suitable one is crucial for achieving optimal results.

As businesses continue to generate vast amounts of textual data, the importance of leveraging advanced text analytics techniques cannot be overstated. By harnessing the power of sentiment analysis and text classification, organizations can unlock valuable insights, drive informed decision-making, and stay ahead in today's competitive landscape. This report serves as a testament to the indispensable role of data science in unlocking the potential of textual data for business success.