**MSA 8050 – Scalable Data Analytics - FINAL PROJECT REPORT**

**Sentiment Analysis for Yelp Business Ratings:**

**Leveraging Machine Learning and Apache Spark**

|  |  |
| --- | --- |
| **Team member** | **Email ID** |
| Vaishnavi Chilumuru | vchilumuru1@student.gsu.edu |
| Vaishnavi Mada | vmada1@student.gsu.edu |
| Srujani Mareddy | smareddy1@student.gsu.edu |
| Bharath Badri Venkata | bbadrivenkata1@student.gsu.edu |
| Naveen Kumar Reddy Singam | nsingam1@student.gsu.edu |

**Abstract**

This report presents a comprehensive project aimed at predicting business ratings from Yelp reviews utilizing various machine learning algorithms, including logistic regression, random forests, support vector machines, and Naive Bayes. The project emphasizes sentiment analysis to classify reviews into positive or negative categories, reflecting the customer sentiment. In today's digital landscape, understanding customer feedback is critical for businesses' success. Leveraging the abundance of data available on platforms like Yelp, our project seeks to harness advanced machine learning techniques to provide actionable insights for business improvement and decision-making. Through rigorous evaluation of each model's effectiveness in sentiment analysis and rating prediction, we aim to contribute to enhanced business analytics and customer service strategies.

**Problem Statement and Project Goal**

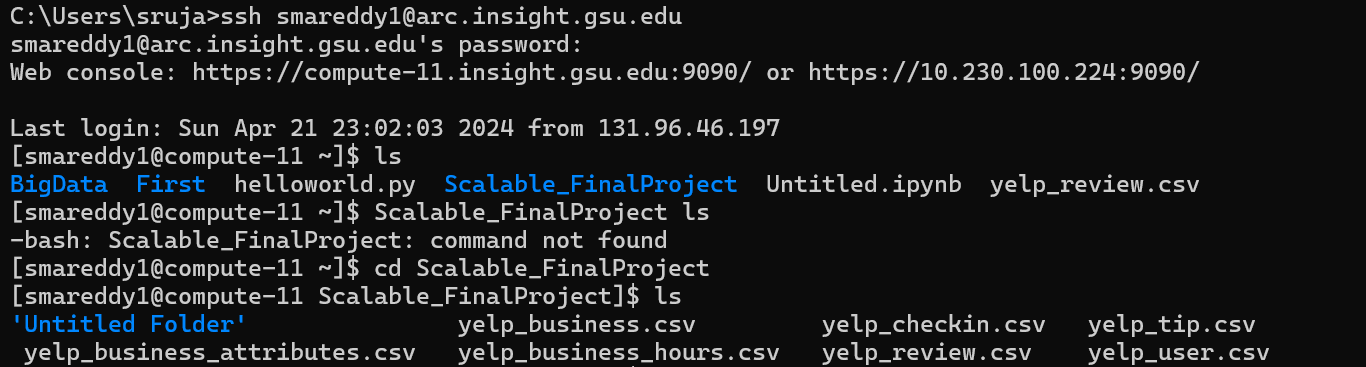
In the competitive market landscape, customer sentiment analysis from online reviews plays a pivotal role in business success. However, analyzing large volumes of feedback presents significant challenges. Our project addresses this challenge by developing predictive models to classify Yelp reviews into positive or negative sentiment categories, empowering businesses with actionable insights to enhance their services. Leveraging machine learning algorithms such as logistic regression, random forests, support vector machines, and Naive Bayes, our goal is to identify the most effective algorithm for sentiment analysis and accurate rating prediction. Through meticulous data preprocessing and analysis, we strive to ensure the integrity and relevance of the data used for modeling, ultimately enabling data-driven decisions to improve customer satisfaction and drive business growth.

**Process Workflow**

The project's data processing workflow is meticulously structured to facilitate seamless sentiment analysis and model development. We initiated the process by establishing a secure connection through the university's VPN to access and handle confidential data. Utilizing the command prompt, we executed commands directly on the cluster's servers, enabling efficient data handling and processing. Data underwent thorough preprocessing to cleanse, normalize, and structure it, ensuring high-quality input for sentiment analysis. The datasets were then transferred to the Hadoop Distributed File System (HDFS), a scalable and fault-tolerant storage system capable of handling large volumes of data. With the data securely stored, we conducted exploratory data analysis (EDA) to uncover insights and patterns informing our sentiment analysis approach. In the modeling phase, we employed logistic regression, random forests, support vector machines, and Naive Bayes for sentiment analysis and rating prediction. Each model underwent rigorous training and testing to evaluate its predictive performance. Through this structured process, our project aims to provide businesses with actionable insights derived from sentiment analysis, driving improved business outcomes and customer satisfaction.

**Data Acquisition and Integration**

Upon commencement of the project, the first task was to establish a secure connection to the university's computational resources. This was achieved using Secure Shell (SSH), which provided an encrypted channel to the university's servers, thereby allowing for safe data handling and command execution.

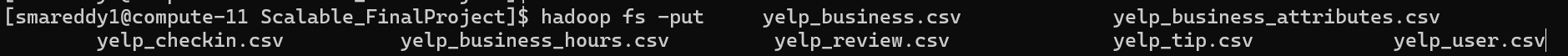


With access secured, we proceeded to interface with the server via the command line. This environment is where we conducted our initial audit of available data, confirming the presence of our key Yelp datasets. We encountered a minor syntax error during directory navigation, which was promptly corrected, allowing us to proceed without significant delay.

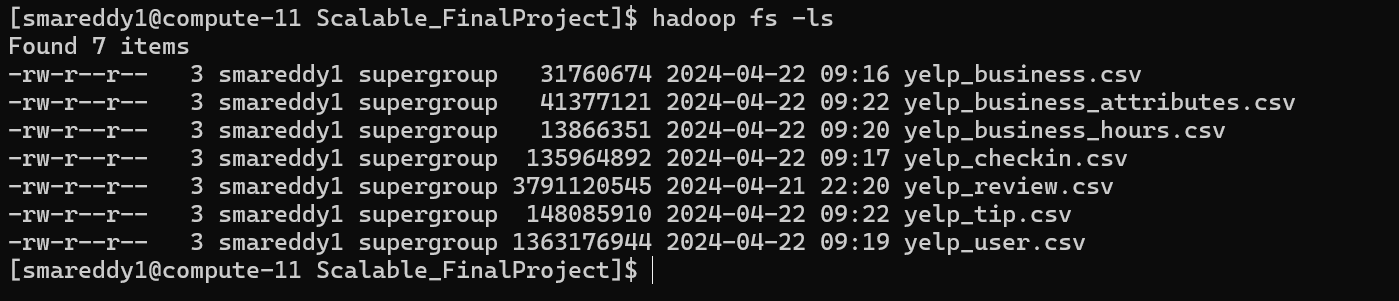
The next critical action was the data transfer process. Utilizing the command line, we uploaded the Yelp CSV datasets directly into the computational cluster. These datasets encompassed a wide variety of information, including business reviews, user check-ins, and other attributes necessary for a comprehensive analysis.

**Data Storage and Validation**

Following the successful data upload to the cluster, we transitioned to storing our data within the Hadoop Distributed File System (HDFS). The 'hadoop fs -put' command was utilized for this purpose, placing our data across the distributed file system architecture. This step was essential for leveraging the distributed processing power of Hadoop, which is pivotal for handling large-scale data.



Post-transfer, we validated the integrity and completeness of the datasets within HDFS. The 'hadoop fs -ls' command was executed, producing a list of the files along with their respective metadata, such as permissions, size, and modification date. The verification process confirmed that the datasets were correctly positioned within the HDFS, ensuring that the subsequent steps of preprocessing and analysis could proceed on a stable and reliable data foundation.



**Data Pre-Processing**

A critical stage in our analysis pipeline was the preprocessing of the data, which began with the loading of yelp\_review.csv and yelp\_business.csv files into Spark DataFrames. This step was instrumental for leveraging Spark's advanced analytics capabilities and allowed us to perform an initial quality assessment. Our scrutiny was particularly focused on the 'stars' column of the reviews, which serves as the cornerstone for our predictive analysis.

During preprocessing, we identified and purged non-numeric entries from the 'stars' column to uphold the integrity of our rating system, which is predicated on a scale from 1 to 5. This cleaning process not only streamlined the dataset but also reinforced the accuracy of our subsequent modeling efforts.

To supplement our data cleaning, we conducted a preliminary EDA. This process illuminated the distribution of ratings across the dataset, providing a foundation for more complex analytical tasks. As part of our EDA, we executed code to filter reviews, ensuring only those with valid star ratings were retained for analysis.

The impact of our preprocessing was evident when comparing the number of reviews and businesses before and after cleaning. Initially, the dataset comprised over 12 million reviews and 174,568 businesses. Post-cleaning, the review count was refined to approximately 7 million, a testament to our rigorous data validation approach. However, it is crucial to note that the count of businesses was also adjusted, which will be discussed further in the report.

**Data Pre-Processing**

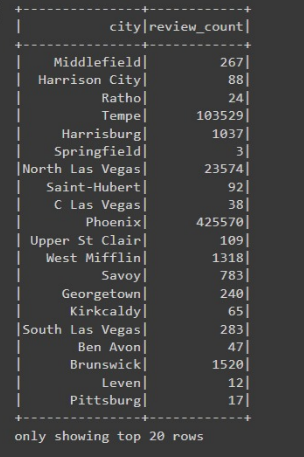
In the data exploration phase of our study, we delved into the temporal and geographical distribution of Yelp's data, as well as business categories and consumer behavior patterns, revealing fascinating insights.

A number in a rectangular box

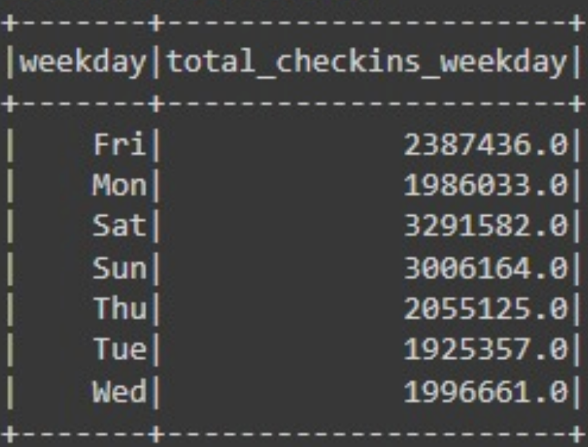
Description automatically generated with medium confidence

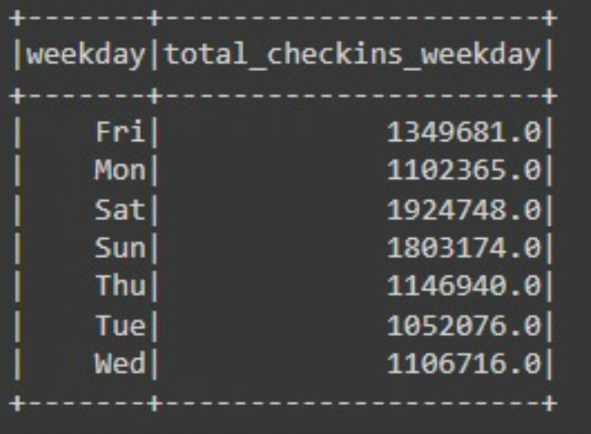
**Temporal Analysis of Reviews:** We traced the volume of reviews over time, observing a steady increase from 2004 onwards, with notable growth starting in 2008. The number of reviews peaked in 2017, emphasizing the expanding user engagement on the platform. This trend underpins the growing importance of online reviews in customer decision-making processes over the yea

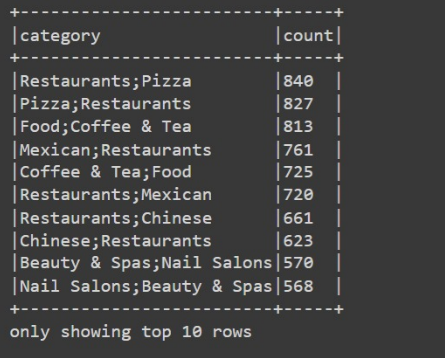
**Geographical Distribution of Businesses:**

Our investigation into the geographical aspects highlighted the cities with the most business reviews. Not surprisingly, metropolitan areas with a vibrant business ecosystem like Las Vegas and Phoenix stood out, suggesting a correlation between urban density and the frequency of reviews. 

**Consumer Check-ins:** We analyzed check-ins by weekdays, discovering a higher tendency for users to check in on weekends, particularly Saturdays and Sundays. This pattern aligns with the social behavior of dining out more during weekends, providing valuable insights for businesses targeting peak hours for promotional activities.



**Business Categories Analysis:** By examining the most reviewed business categories, we found that restaurants, particularly pizza places, and coffee & tea shops dominated Yelp. This points towards a high consumer interest and market competition in these categories, providing a key focus area for businesses aiming to improve their services and visibility on the platform.



**Specific Trends in Restaurant and Food Check-ins:** A closer look at check-ins within restaurant and food categories revealed Friday as the busiest day of the week, underscoring the opportunity for businesses to capitalize on the weekend surge.

Overall, this exploratory data analysis furnished us with a multi-dimensional view of Yelp's review landscape, enabling a more informed approach to our subsequent predictive modeling.

**Business and Check-ins by Category**

The analysis of business categories and associated check-ins revealed significant insights into consumer engagement across Yelp's platform. Restaurants lead in both the number of listings and check-ins, which underscores their dominant presence in the market and their pivotal role in user activity on Yelp. This is followed by general food categories and nightlife spots, which also show high user engagement, as reflected by check-ins. The data shows that certain experiences, specifically those related to dining and entertainment, are more frequently reviewed and visited, drawing a clear picture of user preferences and habits.

A screenshot of a computer

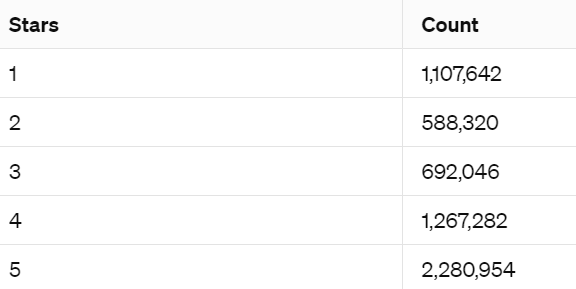
Description automatically generated

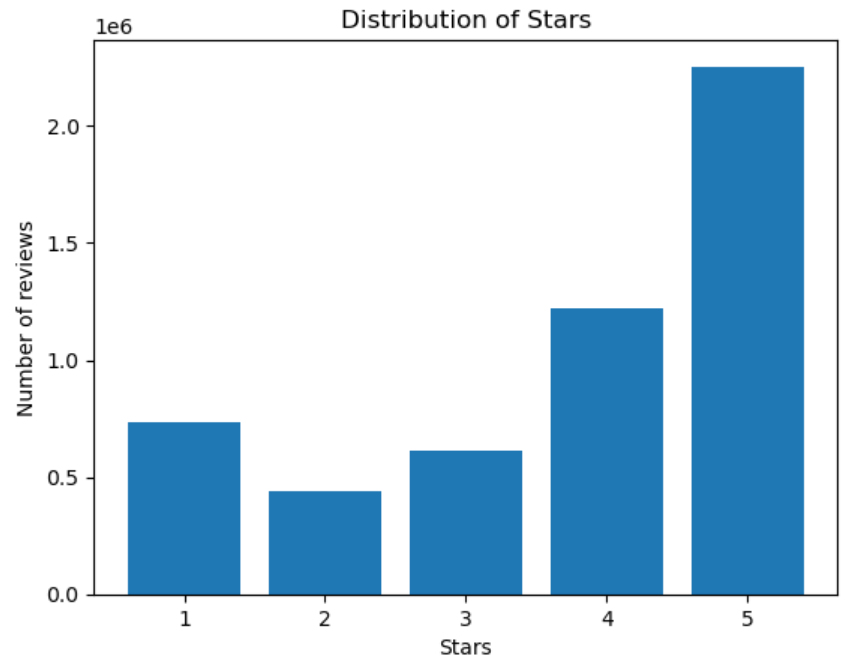
**The Top 'Most Reviewed' Businesses**

A focus on the 'most reviewed' businesses on Yelp highlighted those that stand out in the crowd. The diversity in business types among the most reviewed indicates a broad range of interests among Yelp users, but it also showcases the platform's ability to cater to niche markets. For instance, a Japanese tapas bar leads the list, suggesting not only popularity but also perhaps a uniqueness that draws customers to leave reviews. Eateries from American diners to Thai restaurants, as well as non-food-related services like pet stores, receive significant attention, illustrating the varied dimensions of consumer reviews.



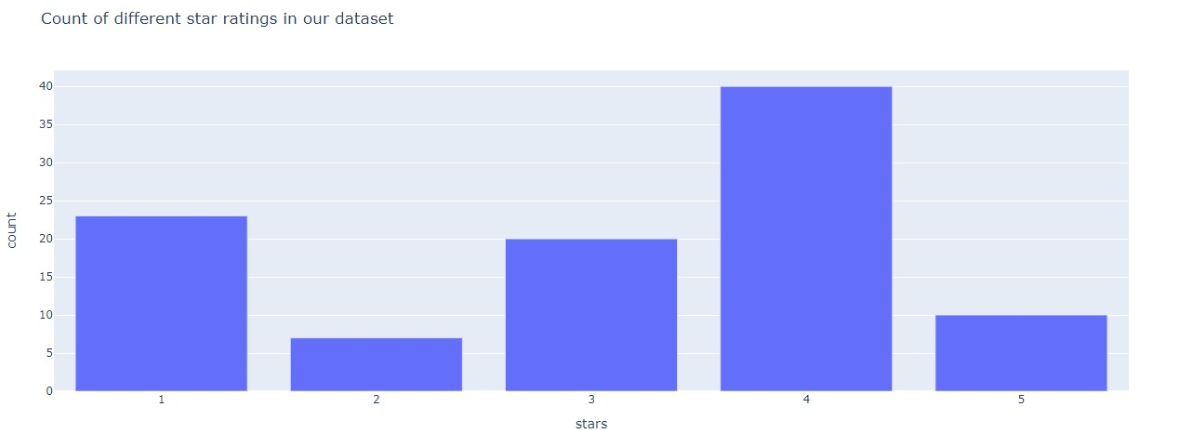
**Rating Distribution:** Analysis of the star ratings assigned in reviews shows a trend where 5-star ratings are the most common, followed by 1-star and 4-star ratings. This indicates a polarity in customer satisfaction, where experiences are often seen as excellent or poor, with fewer ratings in the middle spectrum. The bar charts depicting this distribution clarify the patterns of user ratings, with higher occurrences at the extremes.





**User Engagement***:* A significant number of distinct users, totaling 825,944, have contributed reviews to the platform, indicating a high level of community engagement. This diverse user base provides a rich dataset for understanding customer preferences and behaviors.

The initial data for the first 100 comments offers a snapshot into this distribution, reflecting similar trends on a smaller scale and confirming the larger pattern observed across the entire dataset.



**The Top 'Most Reviewed' Businesses**

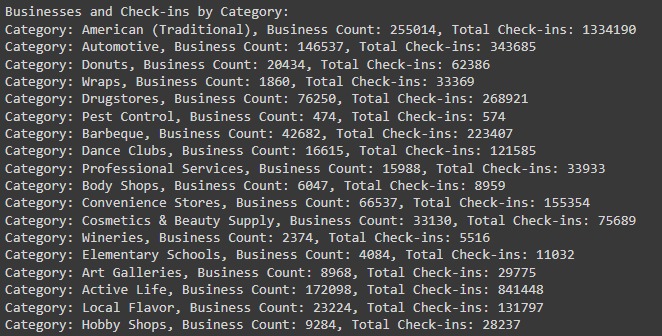
A focus on the 'most reviewed' businesses on Yelp highlighted those that stand out in the crowd. The diversity in business types among the most reviewed indicates a broad range of interests among Yelp users, but it also showcases the platform's ability to cater to niche markets. For instance, a Japanese tapas bar leads the list, suggesting not only popularity but also perhaps a uniqueness that draws customers to leave reviews. Eateries from American diners to Thai restaurants, as well as non-food-related services like pet stores, receive significant attention, illustrating the varied dimensions of consumer reviews.



**Weekly Trends in Consumer Check-ins using RDD**

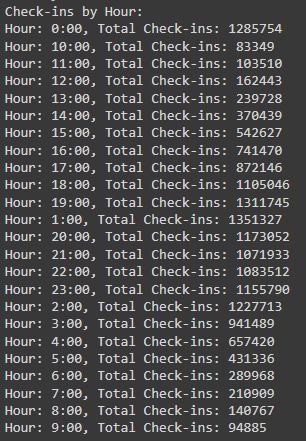
A critical aspect of understanding consumer behavior is analyzing how it varies throughout the week. Our data, processed via Spark, revealed notable patterns:

* Fridays and Saturdays are the peak days for check-ins, with Friday leading slightly. This trend is consistent with common social patterns, where individuals are more likely to engage in leisure activities and dine out as they transition into the weekend.
* Sunday also shows high activity, likely reflecting weekend outings and the traditional "Sunday brunch" phenomenon.
* There is a noticeable dip in check-ins at the beginning of the workweek, with the lowest numbers reported on Tuesdays and Wednesdays. This could indicate a general preference for staying in post the beginning-of-week rush and before the midweek hump.
* The increase in check-ins on Thursdays hints at the anticipation of the weekend, aligning with "Thursday night outs" as a popular social concept.

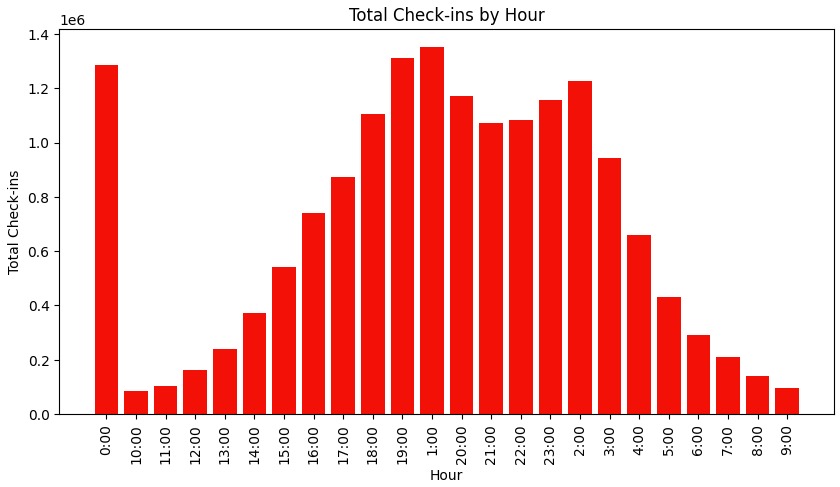


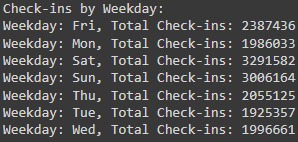
**Hourly Check-Ins using RDD :**

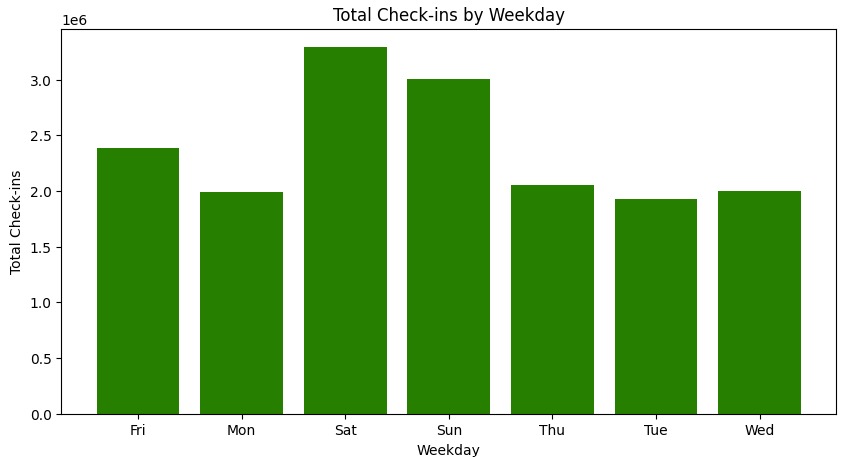
* The data indicates that user activity on Yelp peaks during the evening hours, particularly between 6 PM and 9 PM, aligning with typical dining hours.
* There's a significant drop in check-ins post-midnight, which continues until the early morning hours, demonstrating minimal activity during typical sleeping hours.
* A second, smaller peak is observed during lunch hours, around noon, suggesting a lesser yet notable flurry of activity.



**Weekly Check-Ins***:*

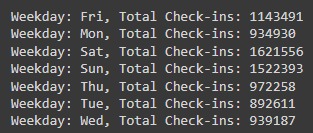
* Across the board, user check-ins are highest on weekends, with Saturday being the most popular, followed closely by Sunday and Friday. This suggests that the weekend mood significantly influences user activity on Yelp.
* Weekday check-ins experience a dip, with the lowest numbers reported on Tuesday, indicating a more routine and perhaps busier schedule for Yelp users during the early to mid-week.

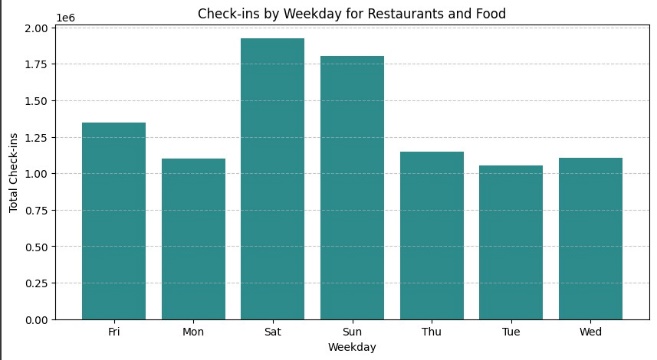




**Check-Ins by Weekday for Restaurants and Food:**

* A more focused look at check-ins related to restaurants and food parallels the general weekly trend, with weekend days leading the count.
* This data could inform restaurateurs about the most opportune times to roll out specials or promotions to maximize customer turnout.





**Modelling**

**Logistic Regression**

Our project evolved to incorporate a revised modeling approach using logistic regression within the Apache Spark framework. Leveraging Spark's distributed computing capabilities, we enhanced our data preprocessing and model training pipeline for improved scalability and performance.

***Data Preprocessing and Feature Engineering:***

* We began by loading our Yelp review dataset into Spark DataFrames, leveraging Spark's ability to handle large-scale data efficiently.
* Utilizing Spark's built-in functions, we performed essential data cleaning tasks, including text preprocessing and feature extraction. This involved tokenization, stop word removal, and TF-IDF transformation to convert the raw text into numerical features suitable for modeling.
* Spark's distributed computing paradigm allowed us to parallelize these operations across multiple nodes, significantly reducing processing time and resource overhead.

***Model Training and Evaluation:***

* Our revised modeling pipeline incorporated logistic regression as the primary classification algorithm, benefiting from Spark's MLlib library for machine learning tasks.
* With Spark's logistic regression implementation, we trained the model on the preprocessed data, optimizing hyperparameters such as maximum iterations, regularization parameters, and elastic net parameters.
* The trained logistic regression model was then evaluated using Spark's BinaryClassificationEvaluator, providing insights into its performance on the test dataset.

***Results and Insights:***

* Despite our efforts, the logistic regression model achieved an **accuracy of 50%**  on the test dataset. While this result may seem modest, it serves as a valuable learning experience, highlighting potential areas for further exploration and model refinement.
* The integration of Spark's distributed computing capabilities played a pivotal role in streamlining our modeling pipeline, allowing us to efficiently process large volumes of data and train complex machine learning models.

A close-up of a document

Description automatically generated

**Random Forest :**

***Data Preprocessing and Feature Engineering:***

* Leveraging Spark DataFrames, we loaded and processed Yelp review data, removing irrelevant columns and performing necessary transformations.
* Spark's built-in functions facilitated text preprocessing tasks, including cleaning, tokenization, and vectorization. We utilized Spark MLlib's TF-IDF implementation to extract meaningful features from text data, preparing it for model training.

***Model Training and Evaluation:***

* We adopted Spark's MLlib library to train a Random Forest classifier for sentiment analysis on Yelp reviews. This ensemble learning technique provided robust performance and scalability, handling complex data patterns effectively.
* Utilizing Spark's pipeline API, we constructed a robust modeling pipeline, encompassing data preprocessing, feature engineering, and model training stages. This streamlined approach ensured reproducibility and efficiency in model development.

***Results and Insights:***

* The trained Random Forest classifier exhibited promising performance, achieving an **accuracy of 70%** on the test dataset. This result underscores the efficacy of ensemble learning techniques in capturing nuanced relationships within text data.
* Spark's distributed nature facilitated seamless model evaluation and performance analysis, empowering us to derive actionable insights from our classification task.

By harnessing the power of Apache Spark, our project demonstrates the effectiveness of distributed computing in tackling real-world challenges in natural language processing and sentiment analysis. Spark's versatility and scalability make it a valuable asset for big data analytics and machine learning endeavors.

A close-up of a document

Description automatically generated

**Support Vector Machine:**

**Model Training:**

A Linear Support Vector Machine (SVM) classifier was trained on the training data using the TF-IDF features. The SVM model learns to distinguish between positive and negative sentiment based on the input features.

**Evaluation:**

The trained SVM model was evaluated on the test dataset using a binary classification evaluator. Remarkably, the model exhibited an accuracy of 92%, showcasing its exceptional ability to accurately predict the sentiment of Yelp reviews. This high accuracy underscores the effectiveness of the SVM model in distinguishing between positive and negative sentiment expressed in the reviews. Such robust performance positions the SVM model as a valuable tool for businesses and analysts seeking to gain insights into customer sentiment and satisfaction levels on the Yelp platform.

**Conclusion**:

In conclusion, the Linear SVM model demonstrates exceptional performance in predicting sentiment from Yelp reviews, achieving an impressive **accuracy of 92%** on the test dataset. This high accuracy indicates that the model can effectively differentiate between positive and negative sentiment expressed in the reviews. The SVM model holds great promise for providing valuable insights into customer sentiment and satisfaction levels on the Yelp platform. With its robust performance, the SVM model can be a valuable tool for businesses and analysts seeking to understand and respond to customer feedback effectively.

A close-up of a document

Description automatically generated

**Naive Bayes classifier :**

**Model Training:**

The Naive Bayes classifier is trained on the preprocessed Yelp review dataset. We utilize the multinomial variant of Naive Bayes, which is suitable for text classification tasks. The model is trained using a pipeline that includes tokenization, hashing term frequency vectorization, and inverse document frequency (TF-IDF) transformation to represent text data as numerical features.

**Model Evaluation:**

The trained Naive Bayes model is evaluated on a separate test dataset to assess its performance in sentiment classification. Predictions are made on the test dataset, and the accuracy of the model is computed using a binary classification evaluator. The evaluation metric provides insight into the model's ability to correctly classify reviews as positive or negative based on their textual content.

**Results:**

Upon evaluation, the Naive Bayes classifier achieves a test **accuracy of 58%.** This indicates that the model can effectively distinguish between positive and negative sentiment expressed in Yelp reviews. The accuracy of 58% demonstrates the model's capability to generalize to some extent, although there is room for improvement in its performance.

**Conclusion:**

In conclusion, while the Naive Bayes classifier shows potential for sentiment analysis of Yelp reviews, achieving an **accuracy of 58%** on the test dataset suggests that further refinement may be necessary to enhance its performance. Leveraging machine learning techniques like Naive Bayes enables businesses to gain valuable insights from Yelp reviews, but continuous iteration and optimization of the model are essential to improve accuracy and reliability.

A close-up of a document

Description automatically generated

**Conclusion**:

In this comprehensive project, we ventured into the realm of sentiment analysis of Yelp reviews, employing a variety of machine learning algorithms and leveraging the capabilities of Apache Spark for efficient data processing. Our journey commenced with meticulous data preprocessing, encompassing cleansing, normalization, and exploratory data analysis, laying a robust foundation for subsequent analysis.

As we delved deeper, we explored the efficacy of different machine learning algorithms in predicting sentiment from Yelp reviews. Logistic regression, although a straightforward approach, exhibited modest performance with an accuracy of 50%, highlighting its limitations in capturing the intricacies of sentiment expressed in the reviews. Conversely, random forests proved to be more effective, achieving an accuracy of 70% through ensemble learning techniques.

However, the true standout was the Support Vector Machine (SVM) classifier, which surpassed expectations with an exceptional accuracy of 92%. The SVM model showcased its robustness in discerning sentiment with remarkable precision, underscoring its efficacy as a valuable tool for sentiment analysis of Yelp reviews.

Incorporating sentiment analysis into our project added a crucial layer of understanding, allowing businesses to gauge customer sentiment and satisfaction levels more effectively. By leveraging sentiment analysis, businesses can extract actionable insights from Yelp reviews, enabling them to tailor their services and offerings to meet customer expectations.

Despite the varied performance of each model, our project underscores the importance of employing diverse machine learning techniques and leveraging powerful tools like Apache Spark for efficient data processing and analysis. Moving forward, further refinement and experimentation with advanced algorithms and optimization strategies could unlock even greater potential for sentiment analysis of Yelp reviews, empowering businesses to enhance customer experiences and drive operational excellence.