SENTIMENT ANALYSIS OF AMAZON FINE FOOD REVIEWS

GROUP 9 PHASE 4 PROJECT

BUSINESS UNDERSTANDING

- The rapid growth of e-marketplaces like Amazon has led to an enormous amount of customer-generated data, such as product reviews, ratings, and feedback.
- This data contains important information regarding customer feedback, that can directly affect business decisions like product improvements, marketing campaigns, and customer reach programs.
- For food business firms, understanding customer sentiment regarding their products is critical to product quality, improvement and customer satisfaction.

BUSINESS IMPACT

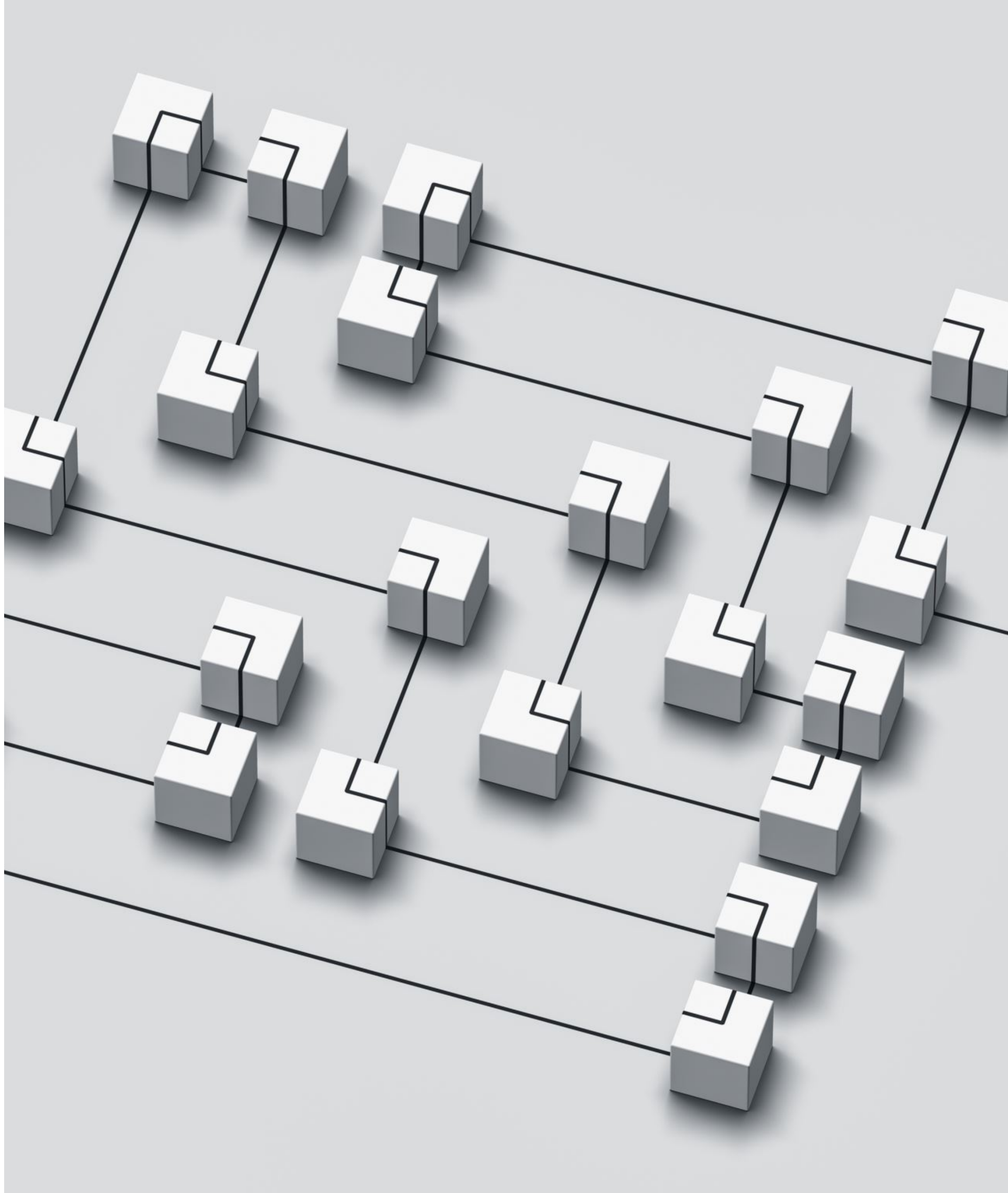
- The models used here-in can help in;
- Improved Product Quality: Finding negative comments informs restaurant owners of repeated problems in their products.
- Improved Customer Experience: Through customer sentiment analysis, organizations are able to customize services as per customer requirements.
- Data-Driven Decision Making: Sentiment analysis generates quantitative data for measuring customer satisfaction that enables more strategic decisions.

OBJECTIVES

- Create a robust sentiment analysis model using advanced Natural Language Processing (NLP) techniques for classifying customer reviews as positive or negative.
- Examine the performance of traditional ML (Logistic Regression, Random Forest, Naive Bayes and XGboost) and deep learning models (TextCNN, DistilBERT).
- Provide actionable recommendations to businesses based on examining patterns in customer feedback.

EXPECTED OUTCOMES

- A sentiment analysis model with high predictive accuracy and generalizability.
- Insights into customer sentiment trends that can inform business decisions.
- Recommendations for deploying the models based on performance.



LIMITATIONS

- Sentiment Classification: Automatically classifying reviews as "positive", "negative" or "neutral" from text content and matching ratings.
- Class Imbalance: The dataset exhibits a skewed distribution of ratings, with a majority of positive reviews (70% of ratings are 4-5 stars), which may lead to better model performance on positive reviews.
- High Dimensionality: Text data is inherently high-dimensional, and preprocessing and feature engineering must be effective processes transforming it into a state suitable for machine learning models.

DATA UNDERSTANDING

•The data set was sourced from kaggle https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews?resource=download

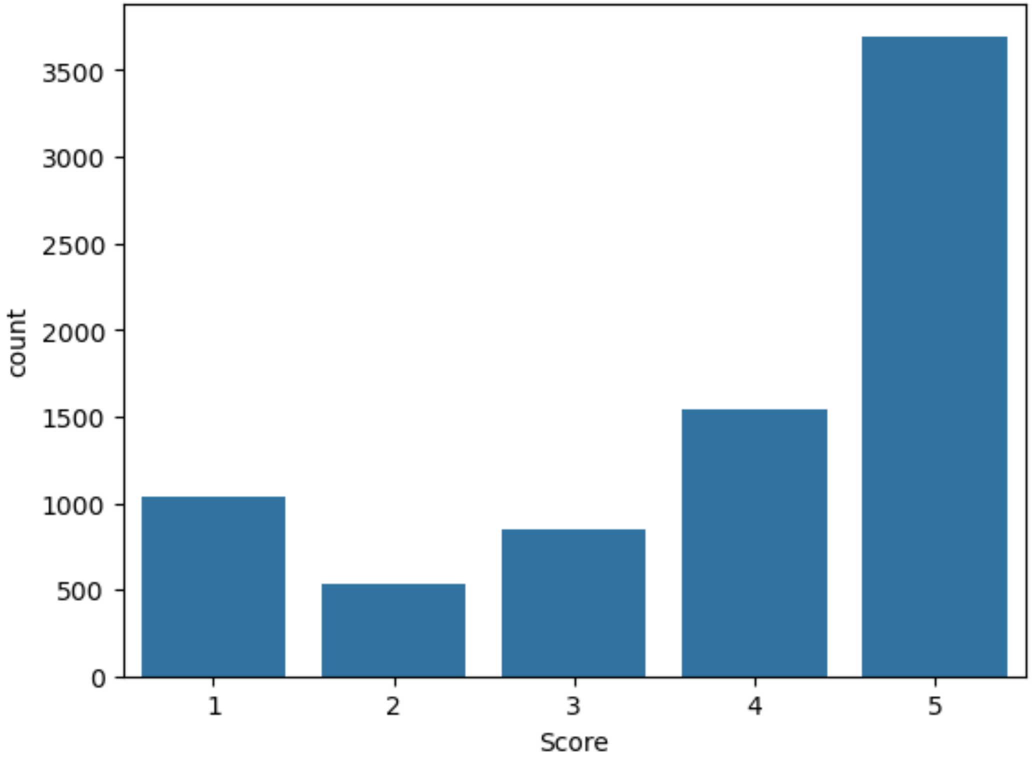
•Given the large size of the Amazon Food Reviews dataset, which contains over 500,000 entries, the group performed random sampling using excel(we generated random numbers in a new column and sorted them which sorted the whole data and picked the first 7,658 cells) to select a subset of 7,658 reviews for analysis in order to make the data more manageable and efficient to work with.

DISTRIBUTION OF SCORES

From the slide, we can tell that the food has received mostly good reviews (4stars and above), which would mean that the service is more than meeting expectations.
However, due to different tastes and preferences, and lapses in service delivery, we have few poor ratings.

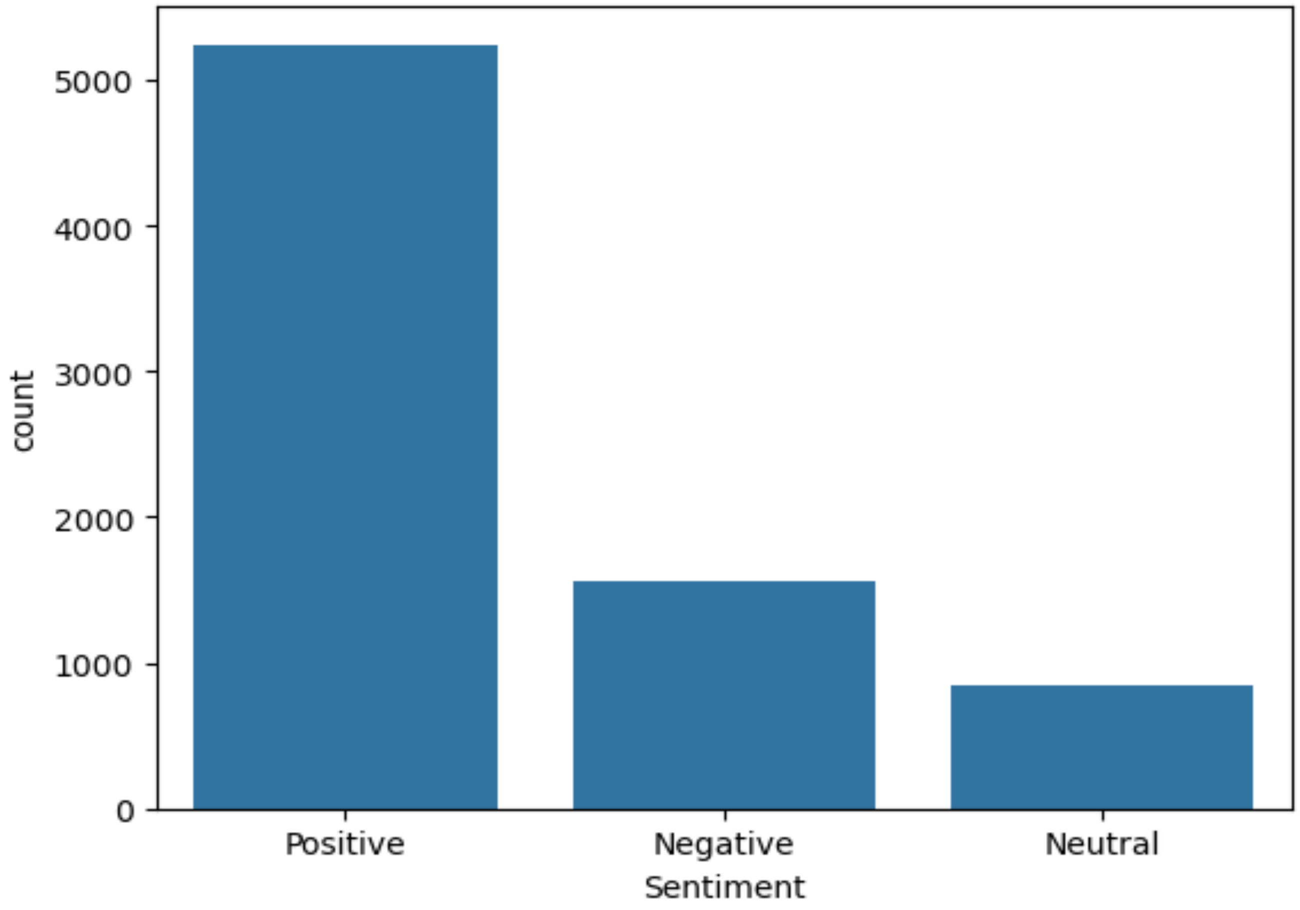
The models we are working with will lump the ratings into sentiments as follows;

1. Positive – 4,5
2. Neutral – 3
3. Negative – 1,2



DISTRIBUTION OF SENTIMENTS

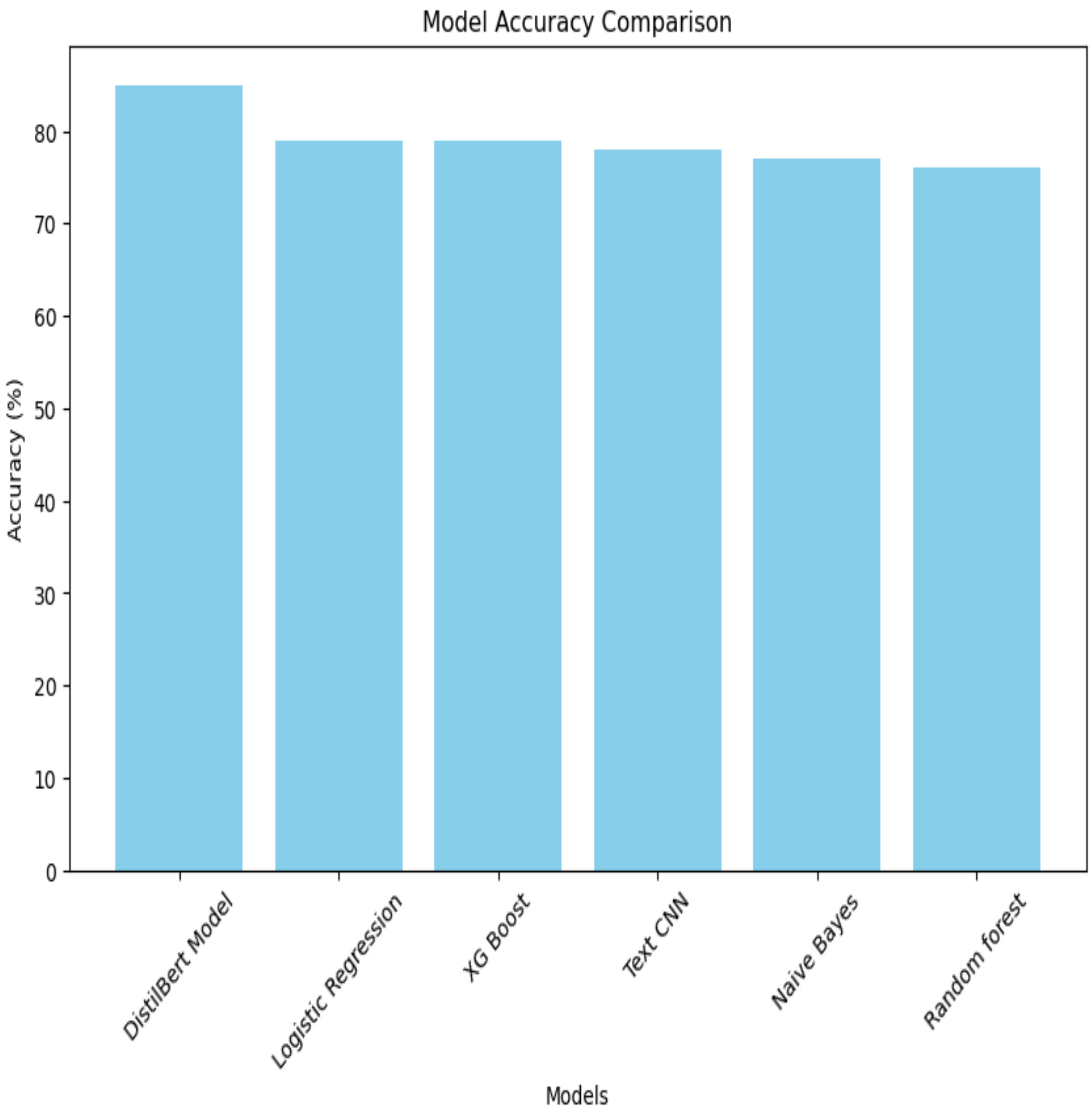
- As mention previously, we can now see the distribution of sentiments is more on the positive side. We can confidently say more than 70% of customers received the food well and truly enjoyed it and the service thereof.



MODEL EVALUATIO

We used the following models to analyse the data and h ranked them based on accuracy of each model, below is the ranking of each model:

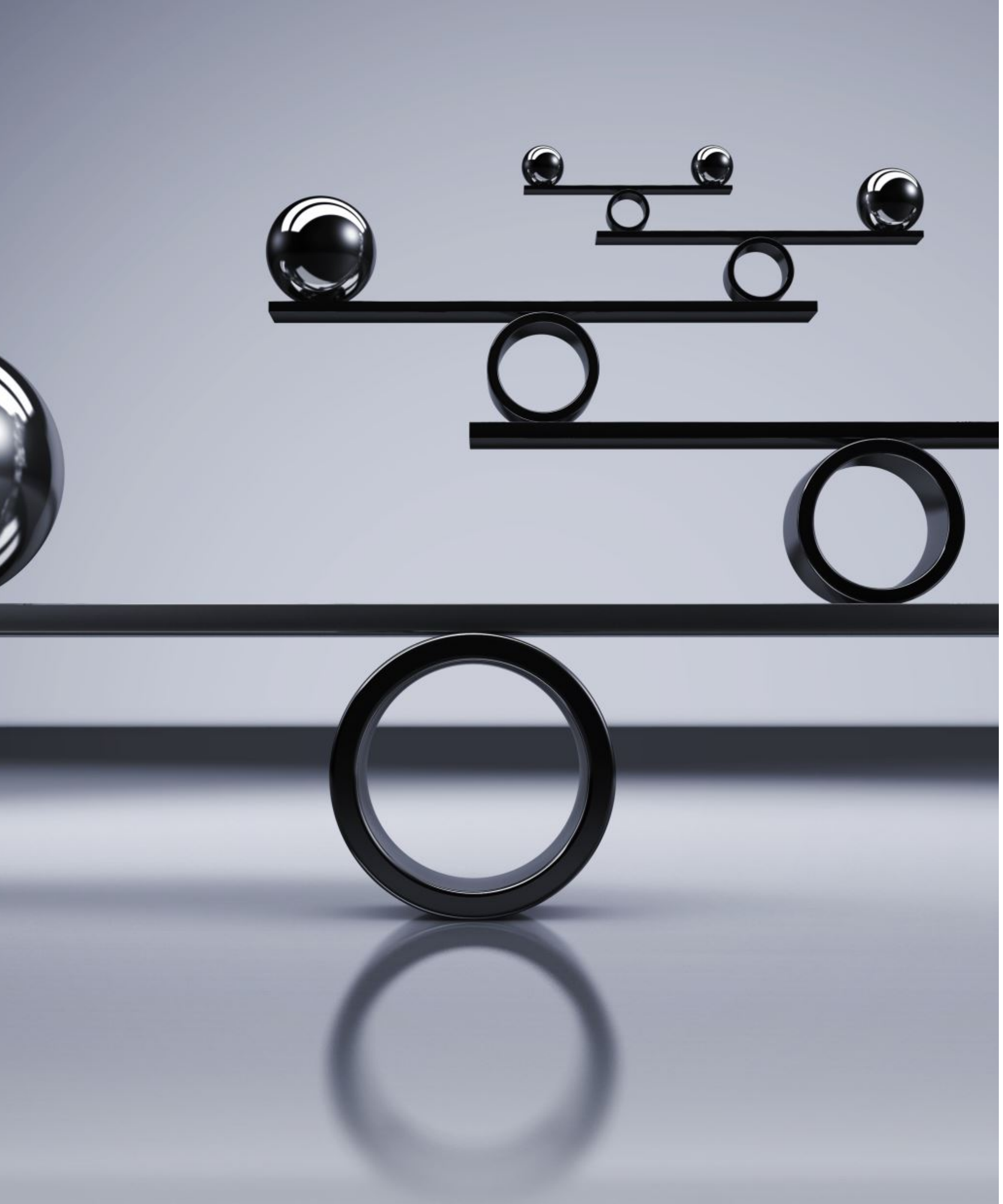
1. DistilBert Model - 85%
2. Logistic Regression- 79%
3. XG Boost - 79%
4. Text CNN - 78%
5. Naive Bayes - 77%
6. Random forest - 76%



CONCLUSION

MODEL PERFORMANCE

- DistilBERT achieved the highest accuracy at 87%, outperforming other models, while Logistic Regression offered a strong, resource-efficient alternative at 82%, and TextCNN showed potential but lagged behind despite tuning.

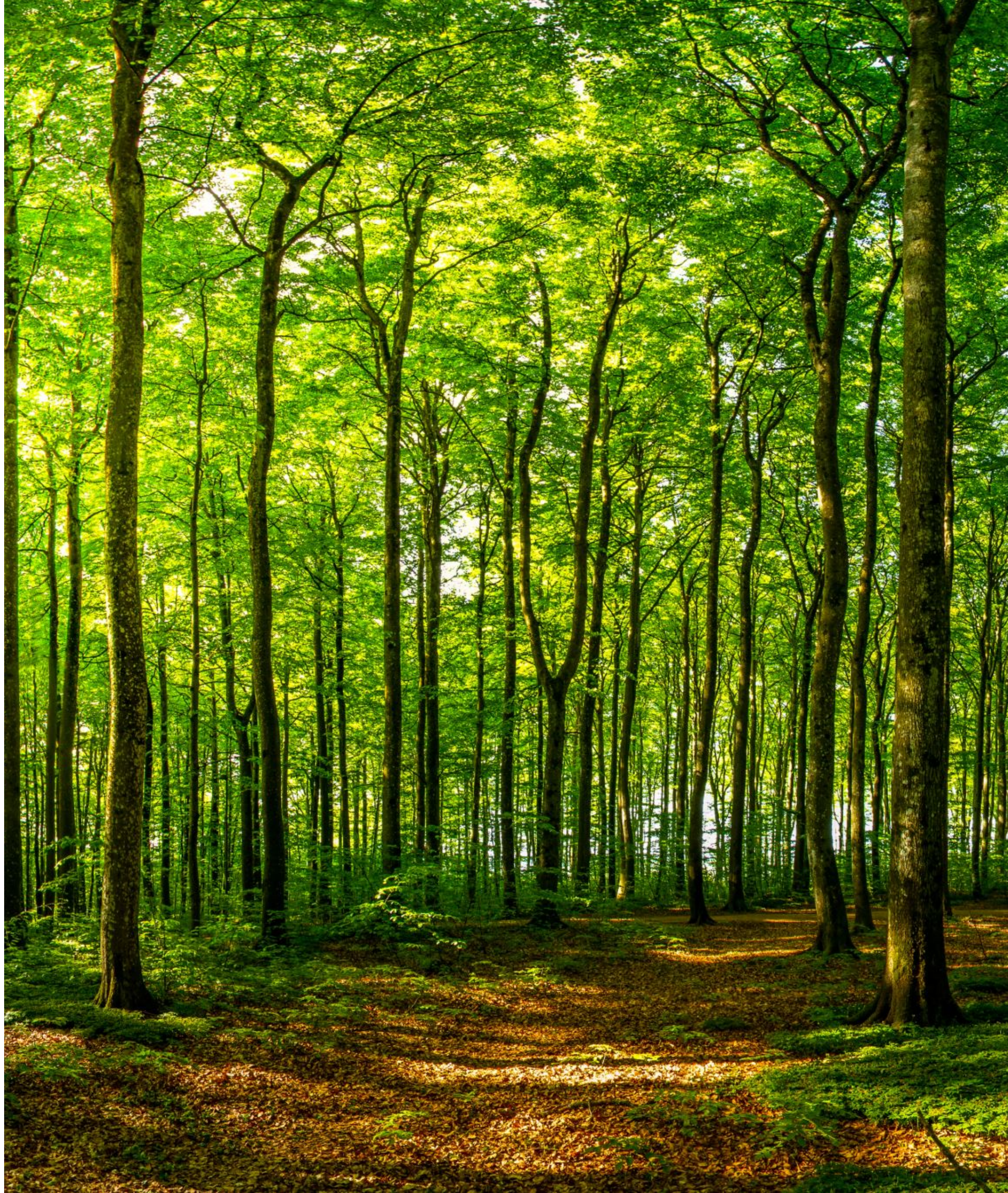


CLASS IMBALANCE

- The dataset exhibited a significant class imbalance, with the majority of reviews being positive. This imbalance negatively impacted the performance of all models, particularly for the Neutral and Negative classes.

FEATURE IMPORTANCE

- Logistic Regression and Random Forest models highlighted key features (words) that were most influential in determining sentiment. These insights can be valuable for understanding customer feedback.



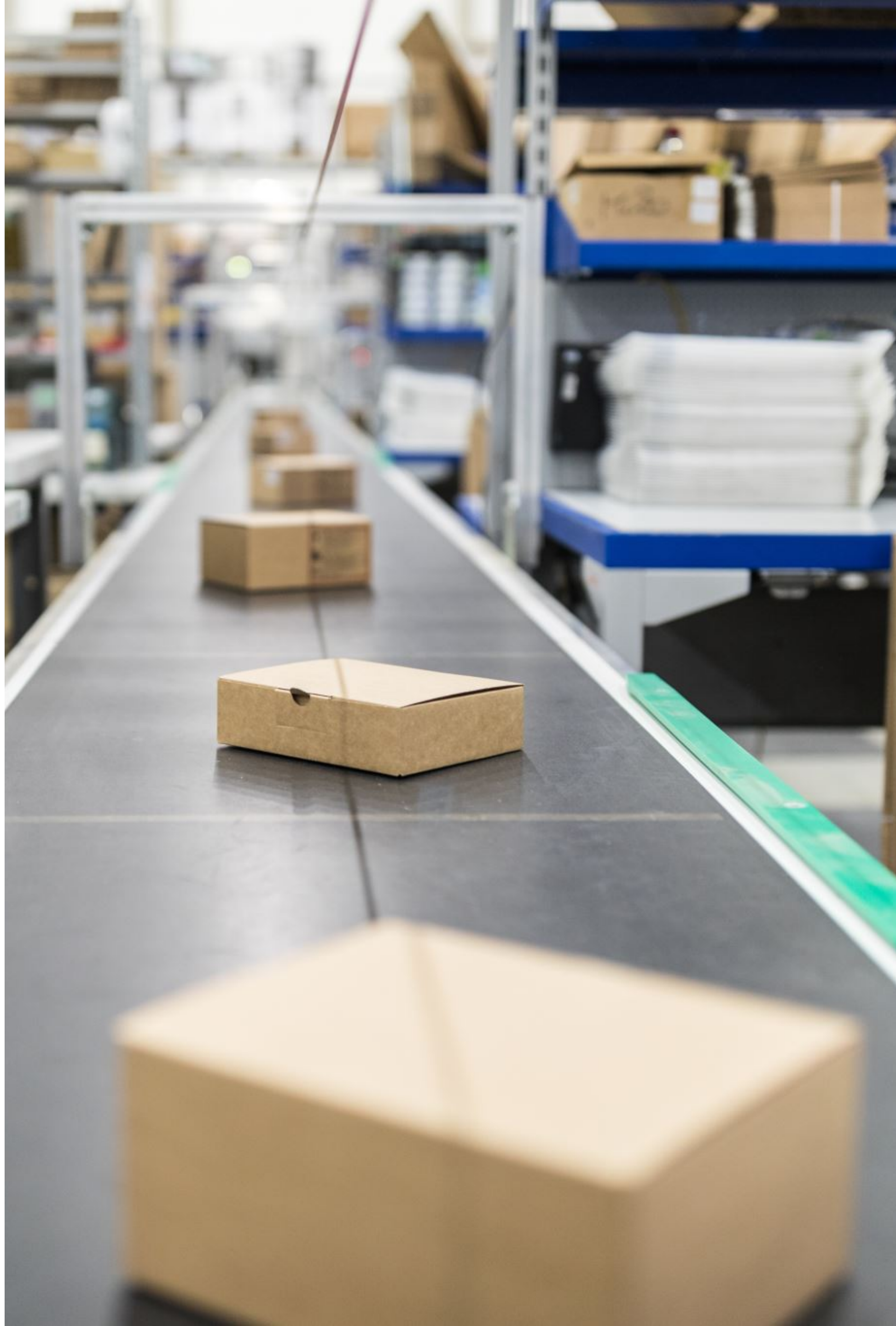
COMPUTATIONAL CONSTRAINTS

- Advanced models like DistilBERT and TextCNN require substantial computational resources for training and tuning. This limitation impacted the ability to fully optimize these models.

RECOMMENDATIONS

MODEL DEPLOYMENT

- For organizations with sufficient computational resources, the DistilBERT model is recommended for deployment due to its superior performance and ability to generalize well across classes.
- The tuned Logistic Regression model is a practical choice, offering a good balance between performance and computational efficiency.



REGULAR MONITORING

- Continuously monitor model performance post-deployment to ensure it remains effective as new data becomes available.
- Retrain the model periodically to adapt to changing customer sentiment trends.
- The TextCNN began to overfit and this is risky if the model is not regularly monitored

FUTURE WORK

- Allocate resources to tune advanced models like DistilBERT for potentially higher accuracy and better handling of class imbalance.
- By implementing these recommendations, Amazon can leverage sentiment analysis to make data-driven decisions, improve customer satisfaction, and maintain a competitive edge in the market

THANK YOU!