# INSY669 Text Analytics Final Project Report

McDonald's is recognized globally as the leading and most successful fast-food chain. In an era where customer feedback is essential for enhancement and achievement, this project aimed to derive insights into customer priorities through text analysis, leveraging a Kaggle dataset of over 33,000 anonymized reviews from McDonald's outlets across the United States. The analysis aimed to diminish the disparity between customer expectations and McDonald's current offerings by unraveling customer preferences from both product-focused and store-specific perspectives to furnish McDonald's with strategic insights aimed at boosting customer satisfaction. This, in turn, could enhance loyalty, improve store performance, and ultimately generate increased business value.

Prior to analyzing insights, the dataset underwent preprocessing to ensure a solid foundation for analysis. A negligible fraction, less than 2% of the records, had missing values in latitude and longitude, all of which pertained to a single store in Hawaii. These records were thus excluded, considering their minimal impact on the overall quality of the analysis for other stores. Subsequent exploratory analysis of rating distribution revealed a significant polarization in customer satisfaction, with a majority of ratings concentrated at the scale's extremes (1 or 5). A compilation of reviews by individual stores through a bar chart revealed a wide range in the volume of comments, from over 1750 to fewer than 10, with the flagship store in Orlando recording the highest number of reviews. An analysis segmented by states indicated that stores in Washington and Annandale predominantly received ratings above 3.

After refining the dataset and uncovering distributions in customer feedback, the analysis progressed to the sentiment analysis pipeline. Multiple approaches were investigated to generate sentiment scores using combinations of different tools, with customer ratings serving as the benchmark for accuracy. To enable precise accuracy computations, numeric ratings were categorized into "Positive," "Negative," or "Neutral" based on their star value: ratings above 3 stars were classified as "Positive," below 3 stars as "Negative," and exactly 3 stars as "Neutral." The initial two methodologies employed the NLTK package for tokenization and lemmatization, and spaCy for tokenization and stopword elimination respectively. Subsequently, VADER was used to assess compound scores. Additional strategies incorporated various Part of Speech taggings, selecting either adjectives and adverbs from the corpus or solely adverbs. The accuracy rates for the first two methods were notably similar, both exceeding 67%. The technique focusing exclusively on adverbs reported an accuracy below 20%. While spaCy's preprocessing slightly edged out in accuracy, NLTK offered more depth in subsequent monogram and bigram analyses, thus emerging as the optimal preprocessing method.

The calculation of sentiment scores paved the way for store level analysis. By averaging the compound scores for each store, a ranking was established based on customer sentiment. Notably, the store in Miami emerged as the sole location with a negative sentiment according to VADER's categorization. The three stores having the highest average compound scores were located in New York, Washington, and Orlando correspondingly. Utilizing latitude and longitude data, a geographical distribution was mapped and revealed stores exhibiting both polarities of sentiment were predominantly located on the eastern coast. To discern the underlying factors contributing to sentiment disparities between the highest and lowest-performing stores, MDS was employed to visualize the cosine similarities among customer reviews. The configuration shows an oval shape with red dots–reviews from top stores–primarily positioned at the lower end, and golden dots–from lowest-performing stores–at the upper and outer regions. This visually highlights the differences in customer reviews. Interestingly, the proximity of some red

and golden dots within the central region suggested that certain reviews from top-performing stores bear resemblance to those from the lower-performing stores, indicating overlapping sentiments despite the stores' performance differences. LDA analysis of the top 3 stores revealed that customers frequently mentioned products not typically offered with McDonald's, such as “pizza” and “pasta” served only at the Orlando flagship store.

The analysis also entailed extracting and analyzing bigrams and trigrams associated with each store, which were visualized on an interactive map for enhanced insight. Specifically, the Miami store, identified by its comparatively low average sentiment score, revealed frequent mentions of phrases such as “long wait,” “wait line,” and “horrible experience.” This suggests that wait times significantly impact customer satisfaction. To address this, it is recommended that stakeholders consider augmenting the number of self-ordering machines or staff to improve service efficiency.

Topic modeling emerged as the second pipeline of the analysis, implemented to cluster all reviews into coherent groups. Various topic quantities, ranging from 5 to 7, were experimented with. The model with 5 topics yielded the most informative representative words. Consequently, the analysis proceeded with five distinct topics using LDA, extracting the top 20 words from each topic. The leading 20 terms within Topic 1 were primarily associated with food items and attributes, including "cheese," "sauce," "sandwich," "cold," and "large." Thus reviews within Topic 1 were selected for detailed product analysis.

The analysis began by identifying references to specific products using a predefined product dictionary and then analyzing the context of these references through the most commonly associated monograms for each product. However, this approach yielded limited insights, prompting a shift towards utilizing bigrams to more effectively extract and identify descriptive phrases linked to products. For instance, a frequently mentioned phrase alongside Big Mac was "ice cream," indicating a common customer habit of purchasing these two items together. Similarly, discussions about fries often included mentions of "cold fry" and "drive through," highlighting customer concerns regarding the quality of this product and its association with drive-through purchases. This methodological pivot enabled a more nuanced understanding of customer preferences and behavior related to specific products.

Text analysis informed business strategy by identifying customer service issues, guiding improvements in staff interactions and reducing wait times. It also influenced product development, leading to the refinement of offerings based on customer feedback. Location-specific insights shaped investment and marketing strategies, enhancing operational efficiency and customer retention. These strategies contributed to increased sales and profitability, with operational improvements reducing costs. Additionally, insights into store performance informed decisions on renovations, openings, and closures, optimizing market presence and business success.

Future steps could include expanding LDA topic exploration, particularly on environmental and service-related issues to improve store ambiance and service quality. Considering the value gleaned from bigrams, applying LDA to bigrams would also be beneficial. Broadening the database to include more reviews and social media feedback would offer richer insights and thus upgrading dynamic map visualizations will provide McDonald's leadership with clearer, real-time decision-making support.

In conclusion, customer feedback analysis is essential for McDonald’s. It facilitates decision-making, enabling McDonald's to adapt to stay attuned to consumer preferences and sustain a competitive edge.