

## Data Analytics Capstone Topic Approval Form

**Student Name:** Corey B. Holstege

**Student ID:** 009253881

**Capstone Project Name:** Sentiment Analysis to Indicate Customer Satisfaction

**Project Topic:** Classify Customer Text from Social Media via Neural Network to Calculate Customer Satisfaction

☒ **This project does not involve human subjects research and is exempt from WGU IRB review.**

**Research Question:** Can a neural network model be constructed on the dataset to accurately predict customer reviews as positive or negative, allowing these predictions to be used to calculate a customer satisfaction score?

### Hypothesis:

**Null hypothesis-**A neural network cannot be constructed from the dataset to accurately predict customer reviews as positive or negative.

**Alternate Hypothesis-** A neural network can be constructed from the dataset to accurately predict customer reviews as positive or negative with an accuracy greater than eighty percent (80%).

### Context:

The internet was invented in 1983 (Board of Regents). Social media took off in 2003 with the launch of Myspace(Wikimedia Foundation, 2023, *Timeline of social media*). According to Influence MarketingHub there are currently over 116 unique social media platforms with over 4.89 billion users. What does this mean for today's consumer company? There are more ways than ever for customers to interact with the businesses they spend their money with. With all these different platforms how are companies to know if their customers are satisfied or upset with them overall? Afterall, to keep customers coming back and making further purchases they need to keep their customers happy. A key metric companies track is the customer satisfaction score (CSAT). This metric is calculated by taking a count of all of reviews where the company was rated a 4 or 5 (satisfied or very satisfied) divided by the total number of reviews, multiplied by 100 to turn it into a percent. The benchmark for fast food restaurants for CSAT is 76% (SurveyMonkey).

How do companies take all these reviews, in text format, and turn it into a CSAT? This is where neural networks for sentiment analysis shine. Like all machine learning algorithms, neural networks require a training data set – a set of data (reviews) that are already classified as "satisfied" or "dissatisfied" to train the model. Fortunately, companies already have this – anyplace customers leave reviews with one-to-five-star rating. One to three stars can be considered dissatisfied, with four and five stars as satisfied. These reviews could be left on their own website, on products they sell, or on third-party websites such as Consumer Reports. Once a model is completed that can accurately predict an input (review) as "satisfied" or "dissatisfied", the companies can then take all of their reviews from all of the social media platforms they are on and enter then into the model to classified the review. After that, it's simply calculating the CSAT score.

### Data:

The dataset to be used for the analysis and to train the neural network is the publicly available McDonald's Store Reviews dataset on [Kaggle](https://www.kaggle.com/datasets/nelgiriyeithana/mcdonalds-store-reviews). The dataset was retrieved on October 24, 2023, and contained 33,396 rows. The dataset was created and is maintained by Nidula Elgiriyeithana and is updated annually, per the dataset's Metadata. The dataset contains anonymized reviews of McDonald's locations in the United States from scraped Google Reviews. The dataset does not detail exact methods used to scrap the reviews.

Per the metadata on the Kaggle page, the dataset was last updated "four months ago." For the purposes of this research, that will be interpreted as July 2023.

The dataset has the following columns:

Column Name	Description <sup>1</sup>	Type
-------------	--------------------------	------

<sup>1</sup> Descriptions are as described on the Kaggle dataset page: <https://www.kaggle.com/datasets/nelgiriyeithana/mcdonalds-store-reviews>

<b>reviewer_id</b>	Unique identifier for each reviewer (anonymized)	Integer, non-repeating
<b>store_name</b>	Name of the McDonald's store	Categorical: Nominal; String
<b>category</b>	Category or type of the store	Categorical: Nominal; String
<b>store_address</b>	Address of the store	Categorical: Nominal; String
<b>latitude</b>	Latitude coordinate of the store's location	Numeric: Continuous; Float
<b>longitude</b>	Longitude coordinate of the store's location	Numeric: Continuous; Float
<b>rating_count</b>	Number of ratings/reviews for the store	Numeric: Discrete; Integer
<b>review_time</b>	Timestamp of the review	Numeric: Interval; String
<b>review</b>	Textual content of the review	Categorical: Nominal; String
<b>rating</b>	Rating provided by the reviewer	Categorical: Ordinal; String

Limitations: The dataset contains several limitations that will need to be handled during the analysis. 1)The dataset contains 33,396 reviews from 40 unique locations from as recent as July 2023, to as far back as 2011<sup>2</sup>. According to ScrapeHero, McDonald's currently has 13,527 locations in the United States (ScrapeHero). The dataset contains a very small number of reviews from a few locations across a large time frame. 2) The dataset contains two columns that at first glance appear to be prime candidates for categorization during the analysis; however, will be dropped from the dataset. These are the "store\_name" column, which has a value of "McDonald's" for every row; and the "category" column, which has a value of "Fast food restaurant" for every row. 3) Other columns will require pre-processing to become tidy (Wickham, H.). For example, the "store\_address", "latitude", "longitude", and "rating\_count" columns are store metadata columns and as this information is the same for many rows should be store in a separate table and references back to the reviews. The "rating" columns contains values of "1 star", "2 stars", etc. when any machine learning algorithm will require these to be integers.

Delimitations: 1)The dataset will be delimited by removing the "store\_name" and "category" fields as each field contains only one value. 2)The project will set to obtain a neural network model that is reasonably accurate, as set forth in the hypothesis, not a model that is perfect.

#### **Data Gathering:** *Describe the data-gathering methodology you will use to collect data.*

The dataset used to train the neural network is publicly available the "McDonald's Store Reviews" dataset on [Kaggle](#). The dataset was created and is maintained by Nidula Elgiriye withana and is updated annually, per the dataset's Metadata. The dataset contains anonymized reviews of McDonald's locations in the United States from scraped Google Reviews. The dataset does not detail exact methods used to scrap the reviews.

A small sample of review's will be manually collected from other McDonald's Social Media sites as a validation set.

#### **Data Analytics Tools and Techniques:**

---

<sup>2</sup> From cursory examination of the dataset

Exploratory Data Analysis (EDA) will be performed to check the distribution of the data, to generate word clouds showing the most occurring words for both "satisfied" and "dissatisfied" reviews, and to show sentiment trends by location and over time.

Once EDA is completed, the data will be cleaned and prepared for modeling in a neural network. This will include standardizing the case of the text, remove punctuation and special characters, removing stop words, tokenizing the words, and padding the tokenized vectors to be equal length. Finally, the data will be split into a training and testing datasets with an 80/20 split.

A neural network will be generated and fit to the training data. This will then be used to complete a prediction on the test data set. Keras's accuracy metric (Team, K) will be used to determine the overall accuracy of the model's generated, and to select the most accurate model. For this project a model will be considered "accurate" to accept or reject the null hypothesis if the accuracy metric is greater than 80%. If an effective model is generated, it will then to be used to predict "satisfied" or "dissatisfied" sentiment on the small sample of reviews manually collected from McDonald's social media sites as a validation set.

Finally, the customer satisfaction metric will be calculated.

This project will use the following tools to complete the analysis and modeling:

- Jupyter Notebook
- Python programming language in the Anaconda Environment
- At a minimum the following Python libraries:
  - System
  - Pandas
  - Numpy
  - Missingno
  - SciKit-Learn (sklearn)
  - Tensorflow / Keras
  - Nltk
  - String
  - Matplotlib
  - Seaborn
  - Wordcloud

#### **Justification of Tools/Techniques:**

Python via the Anaconda Environment will be the primary tool used to complete the project. Python is a high-level, general-purpose programming emphasizing code readability with many libraries standard (Wikimedia Foundation. 2023. *Python (programming language)*). As such, this is an excellent option for data analytics projects.

Anaconda allows python virtual environments to be created and managed easier, allowing the data analyst to focus more on data analysis, and less on infrastructure setup and maintenance.

Finally, for presentation and sharing Jupyter Notebooks allows both code and code output (visualizations, etc.) to be shared in one concise manner for review and reproducibility and allows for iterative data analysis.

**Project Outcomes:** *List the key anticipated project outcomes and deliverables in less than 500 words.*

The project will seek to train a neural network to classify a text input (review) as either "satisfied" or "dissatisfied" with a greater than 80% accuracy. The trained model will be used to classify McDonald's reviews from other Social Media sources to calculate a Customer Satisfaction Score.

**Projected Project End Date:** December 15, 2023

#### **Sources:**

Board of Regents of the University System of Georgia. (n.d.). *A Brief History of the Internet*. Online Library Learning Center. Retrieved October 28, 2023, from:  
[https://www.usg.edu/galileo/skills/unit07/internet07\\_02.phtml#:~:text=January%201%2C%201983%20is%20considered,Protocol%20\(TCP%2FIP\)a](https://www.usg.edu/galileo/skills/unit07/internet07_02.phtml#:~:text=January%201%2C%201983%20is%20considered,Protocol%20(TCP%2FIP)a)

Santora, J. (2023, October 23). *116 social media sites you need to know in 2024*. Influencer Marketing Hub. Retrieved October 28, 2023, from: <https://influencermarketinghub.com/social-media-sites/>

ScrapeHero. *Number of McDonald's locations in the USA in 2023*. ScrapeHero. (2023, September 26). Retrieved October 28, 2023, from: <https://www.scrapehero.com/location-reports/McDonalds-USA/>

SurveyMonkey. *The Ultimate Guide to Customer Satisfaction Score*. SurveyMonkey. (n.d.). Retrieved October 28, 2023, from: <https://www.surveymonkey.com/resources/premium/customer-satisfaction-score-csat-guide/>

Team, K. (n.d.). *Keras Documentation: Accuracy metrics*. Retrieved October 28, 2023, from: [https://keras.io/api/metrics/accuracy\\_metrics/](https://keras.io/api/metrics/accuracy_metrics/)

Wickham, H. (n.d.). *Tidy Data*. Journal of Statistical Software. Retrieved October 28, 2023, from: <https://www.jstatsoft.org/article/view/v059i10>

Wikimedia Foundation. (2023a, October 14). *Timeline of social media*. Wikipedia. Retrieved October 28, 2023, from: [https://en.wikipedia.org/wiki/Timeline\\_of\\_social\\_media](https://en.wikipedia.org/wiki/Timeline_of_social_media)

Wikimedia Foundation. (2023b, October 18). *Python (programming language)*. Wikipedia. Retrieved October 28, 2023, from: [https://en.wikipedia.org/wiki/Python\\_\(programming\\_language\)](https://en.wikipedia.org/wiki/Python_(programming_language))

#### Course Instructor Signature/Date:

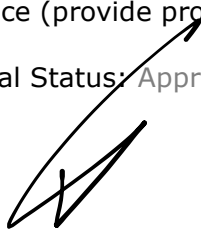
☒ The research is exempt from an IRB Review.

☐ An IRB approval is in place (provide proof in appendix B).

Course Instructor's Approval Status: Approved

Date: 10/29/2023

Reviewed by:



Comments: [Click here to enter text.](#)