School of Bus, IT & Management

CAPSTONE TERM 2

AIDI 2005-02

Winter 2020

AIDI_Project Objective

An Intelligent Customer Support and Feedback System with Sentimental Analysis

Due:

February 21, 2020 5:00pm

Submit to:

Marcos Bittencourt

Prepared by:

Ahmad Barakat #100495227

Pruthvi Rajendrabhai Ingole #100766711

Roshan Issac #100732626

Monika Slominski #100753368

Table of Contents

Introduction

In this competitive business environment, customers are more demanding, with more power and expectations than ever. Customer happiness is considered a key driver for any organization to grow its operations, and maximize its revenue at an optimum level. Therefore, many organizations are recognizing the importance of utilizing big data technology to better understand their customers and the factors that drive their satisfaction.

Incorporating artificial intelligence driven solutions enables companies to become more customer centric by analyzing unstructured data, highlighting key customer information and have it immediately accessible to decision makers.

The project will utilize Natural Language Processing (NLP) to cluster customers and classify if a customer is happy, neutral or angry, and be able to match a customer service representative to keep the customer pleased. In a nutshell this project is not utilizing AI to replace customer support teams, but using AI technology to make customer support more effective, and data driven.

Problem Statement

The objective of this project is to design and develop an intelligent customer support algorithm with sentimental analysis that will help a business serve their customers effectively while lowering support costs, and increasing customer loyalty. The below points define the key problems the project is trying to address throughout the project development life cycle:

- Gaining better insights and understand the factors that influence the customer opinions about a product and/or a service;
- Enhancing customer service processes by bringing customer feedback from multiple channels, into a single data lake to analyze;
- Eliminating the time required to analyze customer sentiment and provide decision makers with real time insight on customer analytics.

Rationale Statement

To remain competitive, companies are required to maintain a high customer engagement and satisfaction level, where all of their strategies and long-term investment need to be driven by data insight generated from the feedback they receive.

Analyzing customer feedback and generating useful insights can be very difficult and overwhelming. Therefore, this project will provide a tool powered by AI to better analyze customer feedback and better predict customer sentiment to help organizations maintain their customers.

The project will enable leading companies to exceed customer expectations by identifying the patterns and trends in their behaviors for which can be used to make an impact influencing future investments and strategic initiatives.

Business Requirements

No.	Requirement Type	Requirement	Comments
1	Business	A fully connected AI solution that can analyze customer sentiment and provide customer analytics to support decision making	
2	Business	The AI Solution should predict if a customer is happy, neutral or angry based on their feedback	
3	Technical	The AI Solution should display the customer sentiment and analytics on a dashboard that is user friendly	
4	Technical	The AI solution shall be able to receive data from multiple data source	
5	Technical	The AI solution should be able to perform sentiment analysis on voice and text datasets	Text is Mandatory where Voice is optional due to the availability of datasets
6	Technical	The AI solution should support real-time data processing (e.g twitter data)	
7	Technical	The AI Solution should be able to learn new data by itself	Optional

Table 1 Business Requirements

Key Metrics

1. The AI solution should be able to correctly predict customer satisfaction;
2. The AI Solution should provide end users with access to quantitative and qualitative insights;
3. The AI Solution should visibly display all the customer analytics on a dashboard.

Project Methodology

The below chart illustrates at high level the project methodology that will be utilized to deliver the project successfully:

Project Plan

TASK DESCRIPTION	ASSIGNED TO
Phase 1 Problem Definition	
Analyze and define a problem	All
Identify an ideal outcome	All
Define success metrics	Pruthvi
Define an ideal output	Roshan
Formulate your problem as an ML problem	All
Phase 2 Data Collection	
Design your data for the model	All
Define data requirements	Ahmad/Monika
Identify data source	Ahmad
Collect data and search for datasets	Ahmad
Examine and analyze datasets	Ahmad
Select datasets	All
Write and document phase 2	Ahmad/Monika
Phase 3 Data Preparation	
Assesses the condition of the data	Monika
Rectify and handle missing data	Roshan
Transform data	Ahmad
Normalize data	Ahmad
Clean and encode data	Pruthvi
Extract data features	Roshan
Transform data features	Ahmad
Select data features	All
Write and document phase 2	Ahmad/Monika
Phase 4 Data Segregation	
Select and set an evaluation protocol	Pruthvi
Split datasets into subsets	Pruthvi
Shuffle the records within each subsets	Pruthvi
Write and document phase 2	Pruthvi
Phase 5 Model Development	
Analyze and define algorithm to be used	All
Determine algorithm requirements	All
Develop and architect model	All
Write and document phase 2	Ahmad/Monika
Phase 6 Model Training and Evaluation	
Train Model using the train data subset	All
Compare prediction against true values	All
Refine model(s)/architecture	All
Test and assess the performance of the model	All
Select the best model to be deployed	All
Validate how it performs against new data	All
Write and document phase 2	Ahmad/Monika

Data Acquisition

The project will utilize data sets that are publicly available for research and application purposes on many websites such as companies' websites, Kaggle, Statista, Google dataset etc. Thus, in order to select a good quality dataset that fulfils our requirements and meets the objective of the project, a simple selection process was developed to assess project team during the data acquisition process.

Dataset Selection

In this section, we present the selection process that we used to select our datasets that will be ingested by our '*An Intelligent Customer Support and Feedback System With Sentimental Analysis'*. As illustrated in Figure 1 Selection Process, the framework illustrates the criteria that we used to narrow down and select the dataset that we will use for this project.

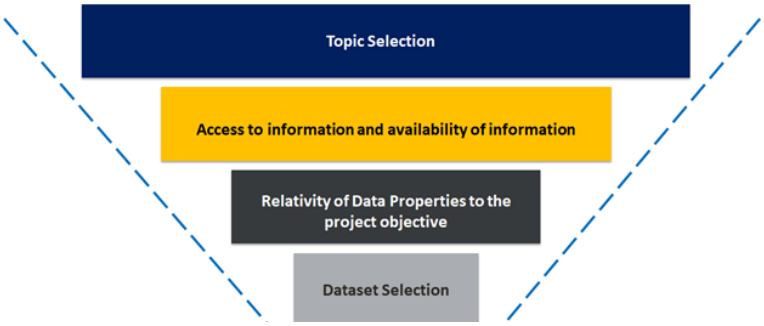

Figure 1 Selection Process

Data Requirements

Based on the scope and the defined issues mentioned in the problem statement, the data projected to be necessary for this project will be as follows:

1. **Data Type:** text and audio data
2. **Dataset Size**: The text dataset is required to contain a large volume of data rows – ideally over 10,000 rows and voice data
3. **Features for Text Dataset:** The text data set requires to contain
    a. Customer reviews and rating
    b. Product name, and brand
    c. tags and labels associated with available text observations
4. **Features for Audio Dataset:** The voice data set requires to contain
    a. Voice for all genders
    b. Audio dataset contain different emotions

Data Sources

This project will utilize a variety of sources and methods to acquire data. The below sources are under consideration to be further explored and analyzed:
Data source 1: Customer review and rating will be scraped from https://weedmaps.com/
Data source 2: Twitter will be used for real-time data
Data source 3: Audio dataset will be obtained from Ryerson Audio-Visual Database

Explanatory Data Analysis

The explanatory data analysis was done separately one for the text dataset and another one for the audio dataset. The EDA files can be viewed on the following links:

- **Text:** https://github.com/ahmadbarakt/cap_stone_2/blob/master/WM%20Data%20Profiling.ipynb
- **Audio:** https://github.com/ahmadbarakt/cap_stone_2/blob/master/Audio_Data_Profiling.ipynb

Data Assumptions

The below points illustrate the assumptions of the projects data:

- Real Time Data (Twitter)
    - It is assumed that using key words such as: “Cannabis,” “Reviews,” etc., will enable the project team to extract sufficient information about the customer opinion on social media
    - Its assumed all the gathered data is related to customer expression on the product but may contain unrelated information.
- Customer Review Data (WM)
    - It is assumed that the rating given by the customer and the content of the review is of same sentiment.
    - It is assumed that any rating of 0 means there was no rating given by the customer at all.
- Audio Data (Ryerson University)
    - It is assumed that all the span of audio clips is of the same length.
    - It is also assumed that the audio training and testing dataset have minimum and equal noise.
- By preprocessing the data it will not limit the models ability to accuracy predict sentiment.
- It is assumed that data will be processed and modelled using an open source of dialogue data
- Its assumed that across all of the data sets may contain inappropriate content.

Data Limitations and Constraints

- EDA shows most customer reviews display more positive than negative ratings and thus skewing the data to the right.
- Some of the data contains missing or null values.
- Data collected from such diverse sources may have different attributes and structures which can impact the quality and accuracy of the model
- Data extracted from social media may contain some level of social bias.
- Real time constraints which restrict the timings of events such that they occur on time.
- Challenge getting a pre-labeled training data related to different algorithms that were evaluated.
- The data obtained from different sources was in different format, adding difficulty to load that into the pandas data frame and to pre-process it.

Solution Overview

This project leverages Machine Learning (ML)/ Deep Learning (DL) algorithms and Natural Language Processing (NLP) to enable a computer to understand spoken or written customers language, and generate insights to improve experiences and revenue. The primary objectives of the solution are as follows:

- Develop a technical solution powered by AI that can analyze customer feedback and provide actionable insights
- Consolidate all customer reviews and feedback from various sources into a single data pipeline solution;
- Introduce a new information management tool that utilizes machine learning models to identify consumer behaviors patterns, segment customers and predict their satisfaction,
- Introduce a new dashboard to visually display customer analytics

Data Flow Pipeline

Figure 2 shows a ML pipeline applied to a real-time business problem where features and predictions are time sensitive**.** The data flow pipeline consists of three main stages, Data Ingestion, Data Preparation, and Data Segregation

Figure 2 Data Flow Pipeline

Model Architecture

The conceptual solution overview in Figure 3 gives visibility of the main users and the high level functional and technical components which make up the solution.

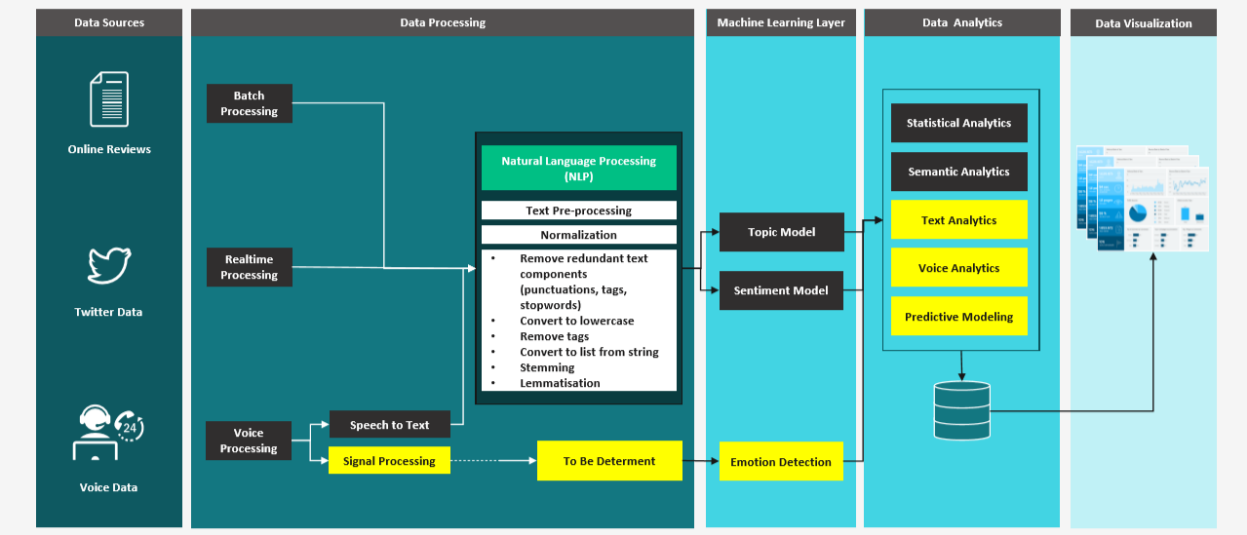

Figure 3 Model Architecture

Technology Requirements

Table 2 lists the software components that are required to develop this solution.

No	Software	Version	Comment
1	Python	3.8.1	Python is an interpreted, high-level, general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant whitespace.
2	Flask	1.1.1	Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions.
3	Docker	1.8.0	Docker is a set of platforms as a service product that use OS-level virtualization to deliver software in packages called containers. Containers are isolated from one another and bundle their own software, libraries and configuration files; they can communicate with each other through well-defined channels.
4	SQL Server	15.0.4013.40 (CU2)	SQL Server is a relational database management system developed by Microsoft. As a database server, it is a software product with the primary function of storing and retrieving data as requested by other software applications—which may run either on the same computer or on another computer across a network.
5	GitHub	2.17.4	GitHub is a Git repository hosting service, but it adds many of its own features. While Git is a command line tool, GitHub provides a Web-based graphical interface. It also provides access control and several collaboration features, such as a wikis and basic task management tools for every project.
6	Dash	1.9.0	Dash is Python framework for building web applications. It built on top of Flask, Plotly. js, React and React Js. It enables you to build dashboards using pure Python. Dash is open source, and its apps run on the web browser.

Table 2 Software Requirements

Solution Development

This section lists all the ML/DL algorithms that will be used to develop the solution, and describes the evaluation process that will be performed during model's training and testing to validate the operation and performance of this solution.

Algorithms

The following ML/DL models will be tested and used to develop the solution:

1. **Naïve Bayes Classifier Algorithm** - This algorithm is simple and easy to implement. Since this is a classification problem Naïve Bayes has the ability to categorize into multiple classes.
2. **Support Vector Machines Algorithm** - This algorithm is exceptional in high dimensional spaces and is memory efficient.
3. **Decision Tree Classifier** - Compared to other algorithms decision trees requires less effort for data preparation during pre-procession and works well with data that has missing values.
4. **Random Forest Classifier Algorithm** - Provides one of the best performances and can compete with most other supervised learning algorithms. This algorithm also produces efficient estimates of test error.
5. **K Nearest Neighbour Algorithm** - Complex concepts can be learned by local approximation using simple procedures.
6. **Latent Dirichlet Allocation (LDA) -** is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.
7. **Convolutional Neural Network (CNN)** – This algorithm is a class of deep neural networks, most commonly applied to analyzing visual imagery.

Model Validation

This section lists the evaluation methods that will be performed during model development to validate the operation and performance of this solution. The datasets will be split into training and testing dataset. The training dataset will be used to train the model, and then the test dataset will be used to evaluate the trained model using new datasets. The reason for doing so is to understand what would happen when the model is faced with data it has not seen before.

After model validation, one or more if the following evaluation metrics will be used to evaluate model's performance:

1. Confusion matrix
2. Accuracy
3. Precision
4. Recall
5. F1 score