Project Proposal for IAL620-01: Text Mining & Natural Language Processing

Bala Mallampati
Master's in informatics and Analytics
University of North Carolina, Greensboro, USA
b_mallampat@uncg.edu

Problem Statement:

After pandemic (Covid-19) started, every organization tracks their work from home employee activities digitally to predict and improve their business operations and employee's work time & increase their wellness time, retention rate.

In this project, we seek to answer the following data research questions:

A. Create chatbots and speech to text functionality to digitalize organization operations and better customer experiences as part of Data Generation process.

B. Analyze responses from Chatbots, Email subjects and extract useful information and convert to numbers as part of Text Analytics PCA (Principal Component Analysis)

C. Classify responses and emails based on subject text using Machine Learning Models (Decision Trees, SVM, Random Forest...).

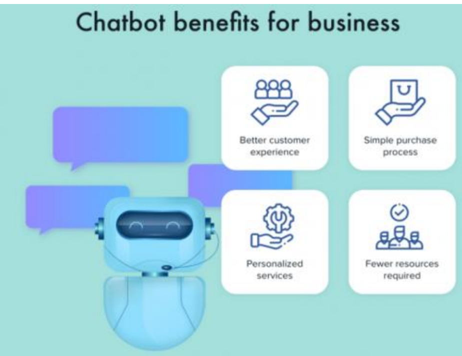
D. Integrate NLP (Natural Language Processing) to BI tools and present extracted insights to organization to improve business operations.

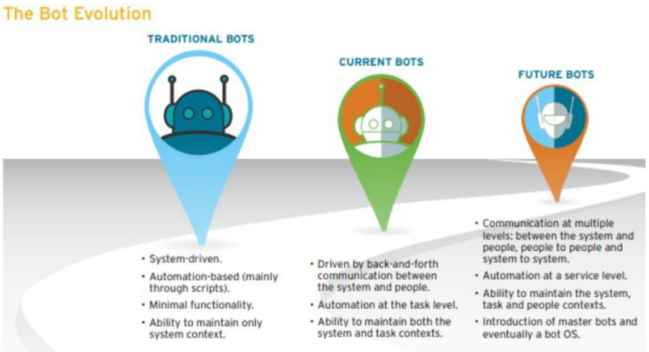
Introduction and Motivation:

Every organization wants to digitalize their operations and minimize human tasks, improve their employee satisfaction rating and retention rate. Every Text analytics and NLP(Natural Language Processing) needs accurate & digitalized data to predict employee's & customers satisfaction rate.

I love to play with data, now "The world's most valuable resource is no longer oil, but data" as per The Economist article (https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data).

Chatbots are most useful in day to day operations of Business, mainly Rule based, Self-Learning(Retrieval based , Generative) Chatbots.




The Bot Evolution
TRADITIONAL BOTS
CURRENT BOTS
FUTURE BOTS
• System-driven.
• Automation-based (mainly through scripts).
• Minimal functionality.
• Ability to maintain only system context.
• Driven by back-and-forth communication between the system and people.
• Automation at the task level.
• Ability to maintain both the system and task contexts.
• Communication at multiple levels: between the system and people, people to people and system to system.
• Automation at a service level.
• Ability to maintain the system, task and people contexts.
• Introduction of master bots and eventually a bot OS.

Text data will be useful for employee's sentiment analysis and predict their satisfaction rate.Email's classification using ML models for Spam, Important emails, collaboration, Network emails, Communication habits.

In future, I want to extend this project for web scrap organization complaints from websites (google reviews, company portals, Face book Pages..) and internal employees chats analysis to determine employee satisfaction rate, retention rate.

Data Source and description:

Most of the data will be generate from Chatbots of this project and export data from UNCG mailbox for emails Text analytics with below columns.

Subject, Body,From: (Name),From: (Address),From: (Type),To: (Name),To: (Address),To: (Type),CC: (Name),CC: (Address),CC: (Type),BCC: (Name),BCC: (Address),BCC: (Type),Billing Information,Categories,Importance,Mileage,Sensitivity

Methodology:

Data processing is an important step for in the data analysis. Data science involves methods of analyzing massive amounts of data for the purposes of knowledge extraction. It evolved from statistics and traditional data management. Data comes in many shapes and forms, and many times we need to get it ready to be able to analyze it. The phrase "garbage-in and garbage-out" is particularly applicable to text mining to Train and Test Data.

Preliminary Plan:

References:

- Sklearn, chatterbot, chatterbot_corpus, Worldcloud, TextBlob Python library
- https://www.upgrad.com/blog/how-to-make-chatbot-in-python/
- https://towardsdatascience.com/3-super-simple-projects-to-learn-natural-language-processing-using-python-8ef74c757cd9