# **Project Report for IAL620-01: Text Mining & Natural Language Processing**

**Chatbot, SpeechToText, UNCG Emails, NYCTimes data for sentiment analysis**

Bala Mallampati Friday, December 10, 2021

Master’s in informatics and Analytics

University of North Carolina, Greensboro, USA

[b\_mallampat@uncg.edu](mailto:b_mallampat@uncg.edu)

**Abstract:**

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, I seek to answer the following data research questions:

1. Create chatbots and speech to text functionality to digitalize organization operations and better customer experiences as part of Data Generation process.
2. Web scrape NYCTimes data & Extract UNCG emails data to analyze and convert to numbers as part of Text Analytics PCA (Principal Component Analysis)
3. Run exploratory analysis on UNCG Email data and show comparisons using Business Intelligence tools and also do sentiment analysis on subjects.
4. Clean and Visualize NYCTimes Top Words, Bigrams, Correlate words and run sentiment analysis, Topic Modeling to get top topics in news.
5. Integrate NLP (Natural Language Processing) to BI tools (Power BI, Tableau) 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.

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 will 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:**

I have used Chatbot input text with fixed labels and for exploratory analysis & sentiment analysis from below sources.

Data extracted from below NYC Times news website. I had extracted latest top articles using html\_nodes in R.

<https://www.nytimes.com/>

|  |  |  |  |
| --- | --- | --- | --- |
| **NYC Times** |  |  |  |
| **Attribute** | **Description** | **Type** | **Data Sample** |
| Rownum | Serial Number | Int64 | 1,2,3.. |
| URL | Uniform Resource Locator | String | <https://www.nytimes.com/2021/12/10/us/politics/texas-abortion-supreme-court.html> |
| Title | NYC Website Title | String | Supreme Court Allows Challenge to Texas Abortion Law but Leaves It in Effect - The New York Times |
| Date | Date, News posted | Date | 12/10/2021 |
| Text | NYCTimes News description | String | Advertisement Supported by The law, which bans most abortions after about six weeks of pregnancy, was drafted to evade review in federal court and has been in effect since September. Send any friend a story As a subscriber, you have 10 gift articles to give each month..... |

<https://mail.google.com/mail/u/0/#inbox>

|  |  |  |  |
| --- | --- | --- | --- |
| **UNCG Email Outlook:** |  |  |  |
| **Attribute** | **Description** | **Type** | **Data Sample** |
| Received | Email Received Date | Date | 10/28/2021 11:39:00 AM,9/27/2021 12:25:00 PM |
| Created | Email created date | Date | Thu 10/28/2021 10:23 PM, Thu 9/23/2021 1:38 PM |
| From | email Sender name | String | UNCG Athletics, Genevieve Smith |
| To | Email receiver Name | String | [inan-students-l@uncg.edu, Bala Mallampati](mailto:inan-students-l@uncg.edu,%20Bala%20Mallampati) |
| Subject | email subject | String | Re: Capstone Presentation TONIGHT at 5:00 PM - Zoom |
| Size | Email Size in KB, MB | String | 2 MB,4 KB |
| Conversation | If same topic discusses then its Conversation | String | Capstone Presentation TONIGHT at 5:00 PM - Zoom |
| Categories | Email caregories | String | [bridges2-users] Bridges-2 (Including VMs and Filesystems) |
| Sensitivity | email Organization sensitive flag | String | Normal, High, Low |
| Contacts | Same as receiver email id | String | [b\_mallampat@uncg.edu](mailto:b_mallampat@uncg.edu) |
| Cc | CC emailer names | String | Bala Mallampati; Richa Kurkure |
| Sent | Email Sent from sender time | Date | 12/9/2021 6:01:00 PM, 12/9/2021 10:15:00 AM |
| Follow Up Flag | Followup flag enabled time | Date | 10/28/2021 11:10:00 AM,9/27/2021 12:10:00 PM |

**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.

In this project I have used the following languages, frameworks, tools, libraries, packages from web scrape data to do Sentiment analysis and extract from UNCG email.

*Technologies & Libraries*: R(ndjson, tidyverse, tm, lubridate, fliptime, stringr,ggmap, maps, quantenda, readtxt, textplots, widyr, topicmodels, ldatuning, lubridate, rvest), Python(numpy, pandas, matplotlib, seaborn, sklearn, google.colab), R Selenium

*Tools*: Rstudio, Colab for python, Power BI, Tableau, Outlook 365.

Chatbot creation using Python Chatterbot, Chatterbot\_corpus, ChatterBotCorpusTrainer

Speech Recognition using Google Colab and Google cloud platform developer authentication.

Power BI with Text Analytics – It needs Power BI Premium License, So imported Wordcloud Visualization.

Tableau Prep Builder BI tool integrate with Text Analytics using Python Script – In progress

**ChatBot Creation:**

I have created simple chatbot using python files and below libraries.

A picture containing diagram

Description automatically generatedChart, funnel chart

Description automatically generated

from chatterbot import ChatBot

from chatterbot.trainers import ListTrainer

my\_bot = ChatBot(name='PyBot', read\_only=True,

                 logic\_adapters=['chatterbot.logic.MathematicalEvaluation',                                 'chatterbot.logic.BestMatch'])

***Output:-***

**Graphical user interface, text, application, email

Description automatically generated**

**A picture containing text

Description automatically generated**

***Speech Recognition:***

Google Colab speech recognition also created using Google Authentication Json File.

<https://console.cloud.google.com/iam-admin/serviceaccounts/details/105018423178735813720/keys?project=balamallampatiuncg>

!pip3 install google-cloud-speech

!curl -LO https://github.com/mozilla/DeepSpeech/releases/download/v0.6.0/audio-0.6.0.tar.gz

!tar -xvzf audio-0.6.0.tar.gz

!ls -l ./audio/

**Graphical user interface, text, application

Description automatically generated**

It accesses your local computer Input devices (audio and video)

**Data Analysis:**

UNCG email data Exploratory analysis and comparisons

**Chart, bar chart

Description automatically generated**

**Chart, bubble chart

Description automatically generated**

**Chart

Description automatically generated**

**Chart, bubble chart

Description automatically generated**

**A picture containing histogram

Description automatically generated**

**Chart, bar chart

Description automatically generated**

**Graphical user interface, Word

Description automatically generated**

**Chart

Description automatically generated**

**NYC Times Exploratory Analysis:**

**Chart

Description automatically generated**

***Word Embeddings:***

**Graphical user interface, text, application, email

Description automatically generated**

**Correlation:**

**Chart, scatter chart

Description automatically generated**

Chart, bar chart

Description automatically generated

***Bigrams Correlation >0.7:***

**Map

Description automatically generated**

***Sentiment Analysis NYC Times:-***

**Chart, funnel chart

Description automatically generated**

**Chart, bar chart

Description automatically generated**

**Chart, bar chart

Description automatically generated**

**Topic Modeling:**

**top 2 topics and their top 10 terms**

**A picture containing chart

Description automatically generated**

**words with greatest beta spread between 2 topics**

**Chart

Description automatically generated**

***Number of Topics in NYC Times using different models:***

**Chart, histogram

Description automatically generated**

***Power BI tool integration with Text Analytics:***

Power BI needs Premium license to do Text analytics and Azure Machine learning. I have only Pro license.

**Graphical user interface, application, Word

Description automatically generated**

So Imported WordCloud visualization provided by Microsoft Into PowerBI and run below visualizations.

**Graphical user interface, application

Description automatically generated  
Graphical user interface, text, application

Description automatically generated**

*UNCG Email Subject WordCloud Using Power BI:*

**Text, whiteboard

Description automatically generated**

*NYC Times World Cloud using Power BI:*

**Text

Description automatically generated with medium confidence**

***Tableau with Text Analytics using Python*:** We can do Text Analytics using Tableau Prep Builder and then feed output file to desktop

<https://towardsdatascience.com/a-guide-to-integrating-text-analytics-into-tableau-94d1a1331117>

**Graphical user interface, application

Description automatically generated**

From Python script we can run Sentiment Analysis and identify named entities and assign back to dataframe to export data. We can refer project files BI\_Tools/tabpy\_server\_script.py script

# Importing the requisite libraries for this function

from textblob import TextBlob

import en\_core\_web\_sm

nlp = en\_core\_web\_sm.load()

import pandas as pd

**Conclusion:**

Github URL for code: <https://github.com/BalaMallampatiGIT/IAL621_TwitterProject>

I learnt very new subject social media network analysis using NodeXL & how we can utilize twitters tweets & connections contents data for organization (UNCG) growth prediction.

Massmine made my life very easy for twitter data extraction with simple commands, thanks to Dr Aaron.

UNCG tweets have more positivity and most of them related to games (Baseball, Soccer) & ssu(science, asu), uncg.

Though my models ultimately did not result in an effective way to classify and predict UNCG tweets data, the improvements I made to our ML model over time did significantly reduce the MSE loss. Given more time and resources to do further research and feature engineering, I believe I could retrain this model to better success.

I believe that I did not have enough of the contributing features that drive classification of UNCG Twitter dataset & not enough data. Additional data points year over year Tweet growths, seasonal trends of tweets, other modern bigdata techniques may have improved these model results.

**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>
* <https://www.nytimes.com/>
* <https://docs.gspread.org/en/latest/oauth2.html>
* <https://towardsdatascience.com/a-guide-to-integrating-text-analytics-into-tableau-94d1a1331117>
* <https://www.red-gate.com/simple-talk/databases/sql-server/bi-sql-server/text-mining-and-sentiment-analysis-power-bi-visualizations/>