**Feature-Based Sentiment Analysis**

**of ChatGPT**

Han (Hannah) Bao, Jeremy Montoya, Tse (Chester) Hsiao, Wanpeng Xu

Professor Yilu Zhou

 BYGB 7977-002: Text Analytics

May 5, 2022

1. **EXECUTIVE SUMMARY**

The introduction of Alexa and Siri gave the world an introduction to virtual assistants and a foundation of possibilities. Many new companies are now introducing their products belonging to the next generation of Artificial Intelligence (AI). Bard by Google and Microsoft Bing Chat are some of the older players still catching up to the level of the newest AI on the market: ChatGPT. ChatGPT is a Natural Language Processing (NLP) tool that was created to be used as bot to have conversation. The advanced uses that ChatGPT has makes some industries and professionals uneasy while others find it a new device they can use for everyday business practices.

News articles from online news sources and twitter posts will be the two main sources of new data from which insights will be derived. The data was cleaned separately and concatenated to gain an overall sentiment. The project goal is to build several crawlers to gather data from several different sources and gain insights as to which way the common public opinion may be leaning; to elaborate more on our goal there is a potential to gain insights as to where companies may optimize the use of ChatGPT.

**Business Goal Analysis**

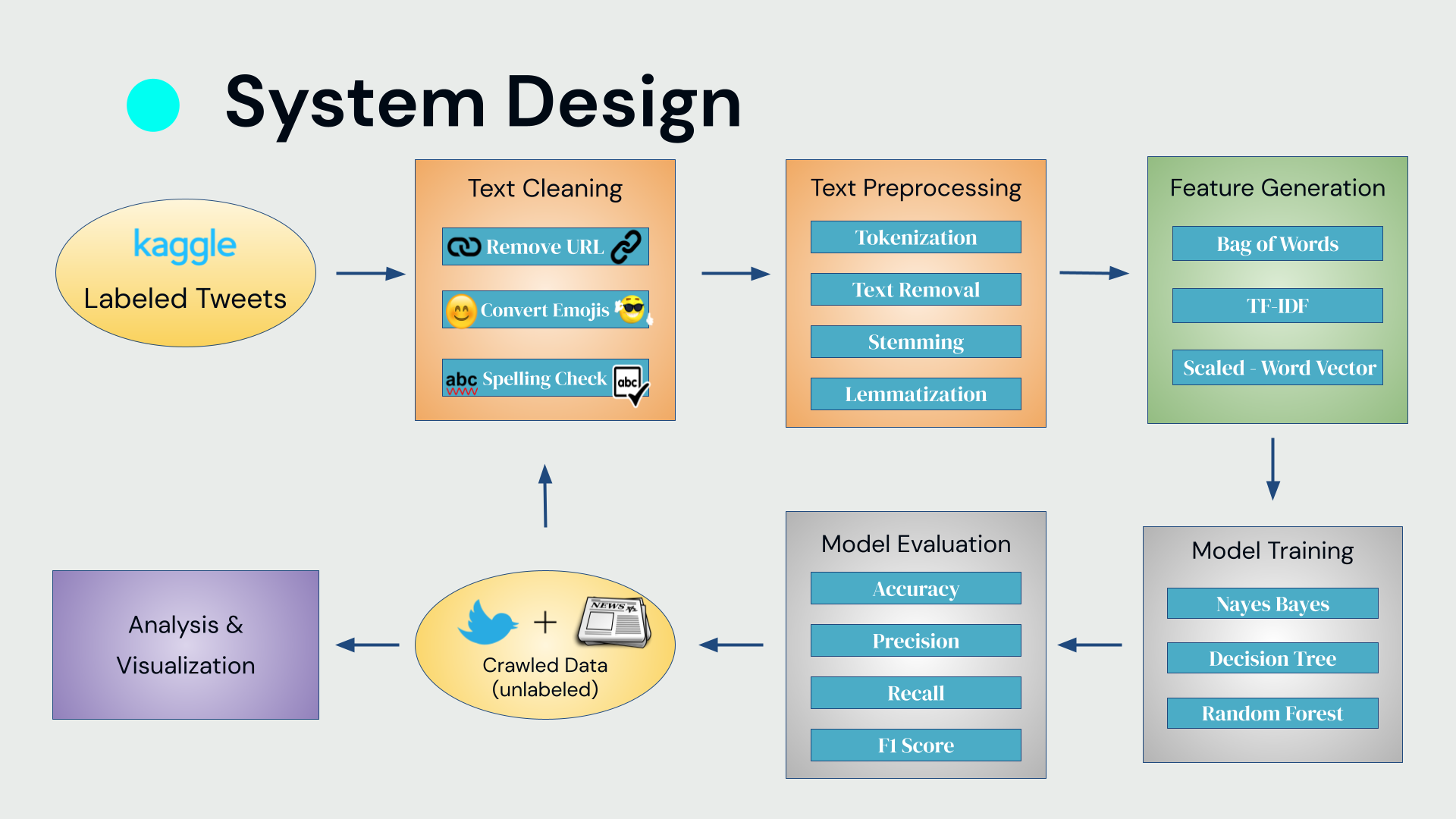
Currently, the use of ChatGPT is still being explored using unsupervised learning techniques and research is still being compiled. Many different industries and professions have varying degrees of worry or excitement for the new AI as some believe it can lead to students in education using the AI to gain an unfair advantage over others; while some individuals in the medical and finance industries are excited as it can be an additional layer, reducing risk when making decisions. ChatGPT at this point has had several issues as of late, these issues including discrimination against gender, race and minority. Other issues have been lack of emotional intelligence, lack of creativity, possibility of copyright infringement, privacy and misuse of the program.

Text analytics in this case has been great at gathering where many individuals' opinions are when it comes to the AI program and can continue to be used for this project to gather sentiment and also evaluate the product uses. As previously mentioned, there have been many different news articles and interviews with a wide range of industry professionals, many are optimistic and some who are resilient. The end goal of using text analytics in this project will be to increase AIQ of Chat GPT(AIQ: the use of AI to make a better world, business, project….) for individual goals and company goals and present other possible uses for ChatGPT in the future.

**Dataset Description**

The dataset comes from 2000 new articles taken between November 2022 and April 2023, 20K tweets that are related to Chat GPT, and finally a previously trained data set taken from Kaggle (https://www.kaggle.com/code/nkitgupta/aspect-based-sentiment-analysis).

**System Design & Implementation**



The first data processed was the tweets which after being gathered had the hashtags removed, words were split, and the stopword dictionary was created. Further a function for feature extraction was applied to the data. After the above was completed, the URL was removed and emoji were converted to a string character to remove noise. Further the data was then run through another function in which punctuation was removed and the stopwords dictionary was applied to remove additional noise. Many of the tweets also had incorrect spelling or shorthand, so additional preprocessing was needed to correct the spelling. Finally, words were lemmatized and all tweets were placed into a pandas dataframe with separate columns for the raw data and the cleaned data.

After all data was organized into a dataframe, the Sentiment Intensity Analyzer was applied to the clean data

**Evaluation**

|  | **Bag of Words** | **TFIDF** | **Scaled\_WV** |
| --- | --- | --- | --- |
| **Naive Bayes** | Accuracy: 0.8  Precision: 0.74  Recall: 0.92  F1 Score: 0.82 | Accuracy: 0.74  Precision: 0.67  Recall: 0.96  F1 Score: 0.79 | Accuracy: 0.61  Precision: 0.6  Recall: 0.7  F1 Score: 0.65 |
| **Decision Tree** | Accuracy: 0.87  Precision: 0.86  Recall: 0.87  F1 Score: 0.87 | Accuracy: 0.85  Precision: 0.85  Recall: 0.86  F1 Score: 0.86 | Accuracy: 0.55  Precision: 0.57  Recall: 0.49  F1 Score: 0.53 |
| **Random Forest** | Accuracy: 0.86  Precision: 0.84  Recall: 0.89  F1 Score: 0.87 | Accuracy: 0.84  Precision: 0.83  Recall: 0.86  F1 Score: 0.85 | Accuracy: 0.69  Precision: 0.78  Recall: 0.55  F1 Score: 0.65 |

**Conclusion and Future Direction**

**Reference**