Rishabh Shrestha
rshrest3@go.olemiss.edu

Data Mining News and Social media for Fertile Ground Project

Revised Design Specification

03-29-2021

Sponsor:
Dr. Dawn Wilkins
Department of Computer Science, UM
dwilkins@cs.olemiss.edu

Project Overview:

In 2017, Jackson, Mississippi was named the most obese city in the United States. It ranked in the top percentile in lacking access to healthy food, low fruit and vegetable consumption, diabetes, high blood pressure, and physically inactive adults. It also ranked at the top in food insecurity despite Mississippi being a leader in agriculture. The Fertile Ground campaign was launched in the Jackson, Mississippi area to provide food-related awareness like food access and food security among the public through the use of creative arts and public art exhibition events.

This project involves data mining news and social media for the Fertile Ground Project. This project helps to determine whether there has been a change in the frequency and nature/content of public discussions related to food access and food security in the Jackson area from 2018 to 2021. My goal is to collect the data from relevant sources and analyze the data to understand the situation and awareness among people regarding food and understanding the effect of the Fertile Ground project.

So, this project aims to understand how well-received the program is. The visualizations and the analysis of sentiment of public discussions on the internet platform such as social media and news media help to deliver valuable insight regarding the change in the frequency and nature of public discussions related to food access and food security in the Jackson area between 2018 and 2021.

Implementation Strategy:

The first task is to collect the data from 2018 to 2021 from the sources such as news media and social media such as Twitter through the use of APIs and web-scraping tools. The posts that refer

to terms like 'fertile ground', 'food access', 'food insecurity', 'food security', 'food health', 'food deserts', 'food swamps' and hashtags such as '#fertilegroundjxn', '#healthyfood' will be collected from Twitter by setting the geolocation as the latitude and longitude of the Jackson, MS and the radius of 50 kilometers. First of all, I will need an access to Twitter's developer accounts and use the API key from the developer's account to access the posts from Twitter. I will use Python over R and other programming languages for this project because of the web scraping tools that it contains and the wider range of access to Machine Learning tools and libraries. I will use the tweepy and sncrape library of Python to access the data from Twitter. Tweepy allows you to nicely collect the data from a specific user account after you set up the authentication using your Twitter developer's credentials. This will help me collect the data from the official Fertile Ground Project's account and the accounts that are closely related to it. However, tweepy allows you to collect data only from the past 7 days, when performing a query for a certain date or search term, due to a recent change in the Twitter API. So, I looked for another option to collect the Twitter data and snscrape seems to work for my purpose as it allows you to collect the historical posts related to specific terms and filter them by geo-location and the radius covered.

My initial plan was to also collect the data from Facebook and Instagram. However, Facebook and Instagram do not have access to API or scraping libraries to collect the data from their platforms due to privacy issues raised lately. I used python's Selenium Chrome Driver that allows you to control Google Chrome and create an automated bot that would log into Instagram and then go into each post after the hashtag search but I could only scrape pictures but not captions and comments which are the subjects of my interest for analysis. Instagram has made it harder to scrape and detects the bot quick enough to not let me collect the data.

For the news media sources, After researching numerous online tools and APIs that would be the best to collect the data that I needed, I landed on Google's news API since it lets me access the news articles for a custom date period and search term. The other APIs that I found had issues such as needing a paid subscription, limited access for the date period.

After the collection of data, the project will involve analyzing the data and determining whether there has been a change in the frequency of discussion or the nature of conversations relating to the terms above. The comma-separated value(.CSV) files mainly showing the date, source & text for each post or article will be constructed. To visualize and analyze the data, python's libraries such as matplotlib, seaborn plot, and sci-kit will be used. This is because python has a wider range of access to Machine Learning tools and tools for sentiment analysis and I want to be consistent throughout the project in the use of programming language for analysis.

After enough data is collected, another probable part of the project is to use Machine Learning techniques to evaluate the sentiment of the posts and a Machine Learning model to predict the sentiment of the posts. To do plain sentiment analysis on the posts and articles, I will be using Vader(Valence Aware Dictionary for Sentiment Reasoning). It is a model used for text sentiment analysis that is sensitive to both polarity(positive/negative) and intensity of emotion. It is available in python's Natural-Language Toolkit(NLTK) package and can be applied directly to unlabeled data. It works very well with short texts from social media containing emoticons and slang. It gives the sentiment scores such as positive score, negative score, neutral score, and the compound score of the text. Since, to use the Machine Learning(ML) technique for the prediction, I need to have a large number of data that are labeled, but the internet did not have as much data related to our content so my sponsor, Dr. Wilkins, and I agreed upon using an online

labeled dataset of tweets to perform Machine Learning to predict the sentiment. The data are labeled as positive, negative or neutral. This will show what we could have done if we had enough data scraped. And for Machine Learning, I will build models using python's sci-kit learn library and choose the one that works the best. The data could be imbalanced among different labels so that will be taken into consideration and confusion matrix will be generated to understand the accuracy of the model. Grid Search cross-validation technique will be used to tune the hyperparameters and determine the best combination of the parameters for the mode.

User Requirements:

The goal of this project is to deliver valuable insight regarding the change in the frequency and nature of public discussions related to food access and food security in the Jackson area between 2018 and 2021.

The Minimum Viable Product (MVP) of this project includes the spreadsheet files of the dataset after scraping and cleaning. Several visualizations like time-series graphs, visualization of sentiment scores from Vader by date, and other statistical visualizations to understand the frequency of the particular terms and nature of the content. An ML model to analyze the sentiments from the data to understand and predict the sentiment of the posts if it's positive, negative, or neutral.

For the user requirements, the user will need to have access to Twitter API and Google news API to access the data. I will set up the code to scrape the data but if the Center for Research Evaluation(CERE) at Ole Miss, who are the associates of Dr. Wilkins, my sponsor, at the Institute for Data Science and for whom this project is being done, needs to keep scraping data

for an extended period even after this project, then they will be able to manually enter the date, account information and other filters used in the project. Once the program is started to run, the user will be prompted for input such as search term, date, geolocation code, and then returned with the spreadsheet files of the dataset scraped from the web. The user will be able to see the visualizations from the data including the time-series graph of the frequency of the related terms that were scraped. The user will also be able to get the sentiment score such as scores for positive, negative, neutral sentiments for the posts and articles scraped. The user will also be able to predict the sentiment based on the labels from the Machine Learning model built.

User Interfaces/Interactions:

User story-As a research associate, I want to be able to scrape online tweets and articles based on my search terms for a certain date and understand the public sentiments of the related topic so that I can know how well-received a campaign is among the public.

With this project, the user can run the python script to collect the tweets from the console. The user will be asked to enter the search term, geo-coordinates, radius, start date, max number of tweets allowed as shown below.

After the user enters the information, the user can see the output on the console and the result will also be downloaded and stored into a .csv file. The user will get information like tweet date, tweet ID, tweet username, tweet text, reply count, retweet count, like count, quote count as the fields of the data frame.

```
PS C:\Users\rishabhstha\Desktop\8th\Csci 487\datascrape> python demo.py
Enter the term you want to search for in Twitter:food
Enter the coordinates of location(Example for Jackson= 32.2998,-90.1848):32.2998,-90.1848
Enter radius in km(example-50km):50km
Enter start date to search the data from(e.g. 2018-03-01):2018-03-01
Enter max number of tweets allowed:10
                   Datetime             Tweet Id  ... Like Count Quote Count
0  2021-03-29 14:14:52+00:00  1376538349473734656  ...          0           0
1  2021-03-29 13:45:50+00:00  1376531044799774720  ...          2           0
2  2021-03-29 13:44:10+00:00  1376530627508506624  ...          2           0
3  2021-03-29 13:15:07+00:00  1376523313904812033  ...          0           0
4  2021-03-29 12:58:40+00:00  1376519173346619393  ...          0           0
5  2021-03-29 12:25:15+00:00  1376510764144529417  ...          1           0
6  2021-03-29 05:49:40+00:00  1376411212339159044  ...          0           0
7  2021-03-29 04:48:47+00:00  1376395891926515713  ...         11           0
8  2021-03-29 03:57:14+00:00  1376382920986857473  ...          0           0
9  2021-03-29 02:26:37+00:00  1376360115633274885  ...          0           0
```

Similarly, the user can scrape the tweets from a specific Twitter account by giving the username of the account and the max number of tweets to collect. The user will be returned with information like tweet date, tweet ID, tweet text, tweet location as shown below.

```
PS C:\Users\rishabhstha\Desktop\8th\Csci 487\datascrape> python demo2.py
Enter the username of the twitter account:_fertileground_
Enter the maximum number of tweets you want to access:30
           Tweet Date            Tweet ID                                         Tweet text Tweet Location
0  2021-03-18 18:21:57  1372614266386329600  RT @NFUDC: When restaurants closed during the ...    Jackson, MS
1  2021-03-18 02:28:27  1372374308786466824  RT @cadwego: “Even as someone who is fairly cr...    Jackson, MS
2  2021-03-16 15:59:11  1371853563475988482  RT @BloombergDotOrg: We believe in the power o...    Jackson, MS
3  2021-03-11 23:04:26  1370148640539430919  Thank you for supporting our local community i...    Jackson, MS
4  2021-03-11 23:02:15  1370148088678060037  RT @MikeBloomberg: Many would like to forget t...    Jackson, MS
5  2021-03-11 22:10:59  1370135188559253508  RT @Inc: This morning, @Google is announcing t...    Jackson, MS
6  2021-03-09 22:18:20  1369412262671360000  Planning public space events during the COVID-...    Jackson, MS
7  2021-02-27 19:38:33  1365748174791581702  A haiku for your Saturday inspiration.  \n \nR...    Jackson, MS
8  2021-02-23 20:06:44  1364305716064960512  RT @foodprintorg: How can we ensure the domina...    Jackson, MS
9  2021-02-23 17:58:05  1364273338890223616  Grocery tax news, see below. MS House proposes...    Jackson, MS
10 2021-02-23 16:18:38  1364248311394414594  RT @grist: A beginner’s guide to talking to yo...    Jackson, MS
11 2021-02-23 15:53:39  1364242023017029637  RT @CivilEats: New donations to Black-led food...    Jackson, MS
12 2021-02-23 15:34:26  1364237187756269573  Invest in spring with a local statement tee, s...    Jackson, MS
13 2021-02-22 16:22:11  1363886815082610689  Spring is on the horizon, gear up for it with ...    Jackson, MS
14 2021-01-26 18:37:26  1354136382495711235  RT @FoodSolutionsNE: The 21-Day Racial Equity ...    Jackson, MS
```

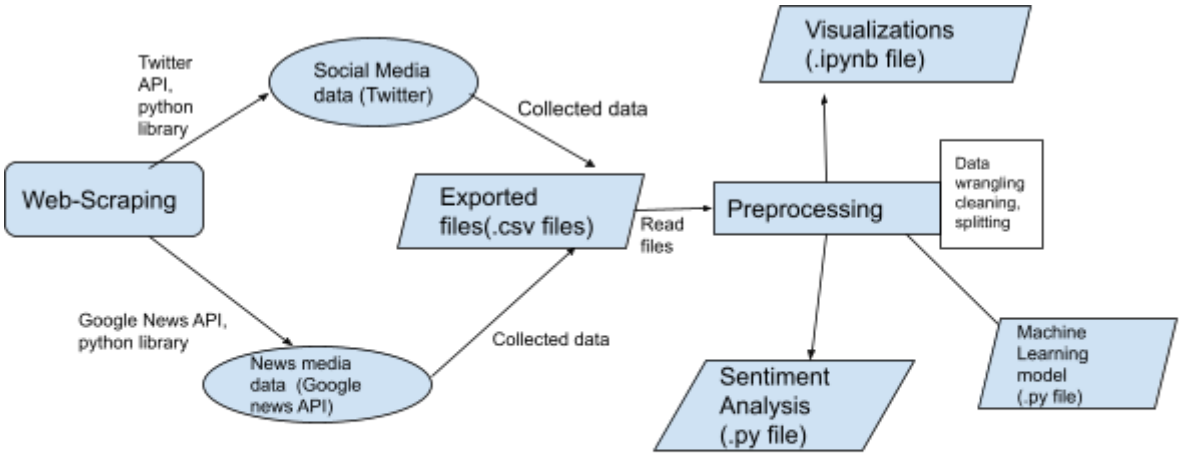
The user will be prompted for similar input for the data collection from news media sources where the user will be asked to enter the search term and date and the user will receive the date of the article, media source, article title, article content, article summary.

After the collection of data, the user can run the .ipynb file that takes the scraped data and shows the visualizations. Similarly, the user can run a python script for the sentiment scores of the text contents such as tweet contents, articles. And, the user can run another python script to perform Machine Learning and build a model to train and test the labeled data and predict the sentiment such as positive or negative.

Architecture:

This project will be run and hosted locally on a personal computer or exported to an executable script to run on any computer with Python and the required libraries. The data will be collected using Twitter API and python's tweepy and snscrape library to a pandas data frame and then exported and stored in a separate file. Similarly, for news media sources, the data will be collected using Google's news API and python's GoogleNews module. The user can input the search term, start date, geo-coordinates, radius, max count number to scrape. The data collected will be exported to CSV files. A python script(.ipynb file) to show the statistical visualizations from the collected data. Another python file will wrangle and clean the data and run the sentiment analysis on the collected text data and give the positive, negative and neutral sentiment scores along with the compound score. Again, another python script will read files with labeled tweets and then preprocess data such as cleaning and wrangling the data. After that, it will split the data into train/test data and then build a Machine Learning model trained using the labeled

data and then predict the sentiment label on test data. The chart below shows the workflow of the project architecture.



Development and Deployment Environment:

The development environment such as Jupyter Notebook and Python's libraries are used for the development of this project. Visual Studio Code is used to locally host the Jupyter Notebook from the personal computer. I am running Jupyter Notebook on Windows 10 OS on my laptop with Intel Core i5-7200U CPU @ 2.50 GH and 8 GB RAM. The Jupyter notebook provides a web-based interactive environment that combines code, text, images, plots into a single document. This makes it easier to retrieve the data for analysis and visualization and perform Machine Learning. The pandas data frame is nicely presented by the Jupyter Notebook which is easy to read and understand. The notebook can help perform reproducible interactive computing with ease. A block of code can be computed separately in the separate cells in the Jupyter Notebook which is faster and elegant. This helps in version control and to push the code

regularly. The testing of a part of the code is also easier and does not interfere with the rest of the code if done properly in separate cells in the Jupyter Notebook.

The code is pushed into the Github repository which gives the version control. The code is also uploaded to the Box folder shared by IDS(Institute for Data Science) which is my sponsor's, Dr. Wilkins's affiliations. The version control system will help me keep track of my code as it progresses and help me retrieve, in case I lose my data.

Since, this project was very different from anything I have done before, especially the scraping of data, so the research is done online about the tools to scrape the data. So, I have referenced posts from towardsdatascience.com and medium.com to help me with the collection of data such as using tweepy, snscrape library, and google news API. I referenced some online articles for text preprocessing and analysis. I also looked at some online tutorials for text mining and classification to help me better understand creating prediction models.

Once the final version of the project is pushed into the git repository, the user can download the project from the online repository. To run the project on an individual's computer, Python needs to be installed along with other important libraries used. The user also needs to have the Jupyter Notebook installed and that way it will be easier to view the plots. Once, those requirements are met, the user can run the project and see the results. The project can be deployed as a web application using python web frameworks such as Django or Flask which might take a long time. Therefore, for now, I am going to upload the whole project as an online repository on GitHub.

Test and Integration Plan:

This project will be tested within the python environment. The scraped data will be checked for correctness and relevancy during the data collection process. Different APIs will be tested for the best possible sources of data. The exported data will be stored in CSV files and will be sent to CERE for checking. Talking about the whole project, the weekly meeting with the sponsor and scheduled meetings with CERE to assess the results will be done. Their feedback will be used to see if it meets their expectations and works correctly which will help to improve the system. The project will be evaluated whether a claim or the result of the project can be true and see if the produced material is rational. Talking about testing the code itself, different search terms and radius for geolocation will be tested to collect the data. The unit testing will be done for error detection, functionality, logic, and efficiency. Several APIs and web-scraping tools will be tried to collect the data. For the sentiment analysis and Machine Learning part of the project, the Chi-square analysis and the confusion matrix will be created respectively. In order to answer the question if the change in the sentiments of public discussions is possibly related to COVID-19, Chi-square test of independence for time (pre-COVID, COVID) and sentiments (positive, neutral, negative) will be performed. The confusion matrix and cross-validation will be created to test different Machine Learning(ML) models. The tuning of the hyperparameters of Machine Learning classifiers using the GridSearch Cross-Validation will be done to determine the best combination of parameters for the classifiers.

After all the functionalities and components are joined together the testing will be done to see whether or not it affects the relative components that run along with it as a group. The client(CERE) will be given the application to test and run and see it has all the functionalities that they need.

Project Timeline:

- 29th March 2021- Finish Revised Design Specification.
- 1st April- Meet the CERE to discuss the data collected and results and possible analysis.
- 3rd April 2021- Create visualizations and other graphs needed for the project.
- 5th April 2021- Create Initial Oral Presentation by screen recording an informal oral slide presentation explaining their project.
- 7th April 2021- Start working on the ML part of the project and research the best techniques and tools for it.
- 10th April 2021- Perform ML on the labeled data or use sentiment analysis building lexicon.
- 13th April- Test different prediction models and determine the best one
- 16th April 2021- Perform testing and finalize the project and create a user manual.
- 19th April 2021- Create final presentation PowerPoint slides and present them.
- 26th April 2021- Create the final report of the project and wrap up the project.

Bibliography:

[1] S. Dandir, "Extracting Data from Twitter using Python," *Medium*, 02-Dec-2020. [Online]. Available:https://towardsdatascience.com/extracting-data-from-twitter-using-python-5ab67bff553a. [Accessed: 29-Mar-2021].

[2] Beck, M. (2021, January 14). How to scrape tweets from twitter. Retrieved March 29, 2021, from https://towardsdatascience.com/how-to-scrape-tweets-from-twitter-59287e20f0f1

[3] Beck, Martin. "How to Scrape Tweets With Snscrape." *Medium*, Better Programming, 18 Feb. 2021, betterprogramming.pub/how-to-scrape-tweets-with-snscrape-90124ed006af.

[4] S. Tomlins, "Scraping Tweets by Location in Python using snscrape," *The Startup*, 23-Dec-2020.[Online].

Available:https://medium.com/swlh/how-to-scrape-tweets-by-location-in-python-using-snscrape-8c870fa6ec25. [Accessed: 29-Mar-2021].

[5] R. Python, “How to make an Instagram bot with python and InstaPy,” *Realpython.com*, 06-Apr-2020. [Online]. Available: https://realpython.com/instagram-bot-python-instapy/. [Accessed: 29-Mar-2021]

[6] “GoogleNews,” *Pypi.org*. [Online]. Available: https://pypi.org/project/GoogleNews/. [Accessed: 29-Mar-2021].

[7] M. Dhingra, “GoogleNews API—Live News from Google News using Python,” *Analytics Vidhya*, 17-Jun-2020. [Online]. Available:https://medium.com/analytics-vidhya/googlenews-api-live-news-from-google-news-using-python-b50272f0a8f0. [Accessed: 29-Mar-2021].

[8] “Python,” *Geeksforgeeks.org*, 23-Jan-2019. [Online]. Available:https://www.geeksforgeeks.org/python-sentiment-analysis-using-vader/. [Accessed: 29-Mar-2021].

[9] C. J. Hutto, *vaderSentiment*.b

[10] S. M. Mohammad, “Sentiment Analysis,” in *Emotion Measurement*, H. L. Meiselman, Ed. Elsevier, 2016, pp. 201–237.

[11] I. A. Khalid, “Cleaning text data with Python,” *Towards Data Science*, 11-Oct-2020. [Online]. Available:https://towardsdatascience.com/cleaning-text-data-with-python-b69b47b97b76. [Accessed: 29-Mar-2021].

[12] J. Lee, “Kaggle Twitter Sentiment Analysis: NLP & Text Analytics,” Towards Data Science, 18-May-2020. [Online]. Available: https://towardsdatascience.com/twitter-sentiment-analysis-nlp-text-analytics-b7b296d71fce. [Accessed: 29-Mar-2021].