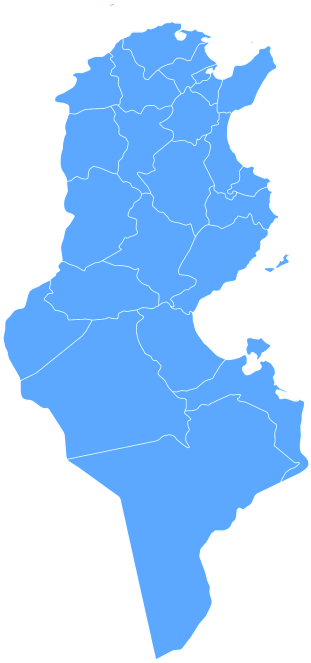

Identifying and Analyzing Twitter Data Related to Tunisia

CS 4624: Multimedia, Hypertext, and Information Access

Virginia Tech, Blacksburg, VA 24061

May 11, 2023

Instructor: Edward A. Fox
Clients: Andrea Kavanaugh, Steve Sheetz, Chreston Miller, Mohamed Farag
Team: Ryan Gniadek, Sraavya Gudavalli, Victoria Hardy, Steven Ruckert

Table of Contents

Table of Figures

1. ABSTRACT

Following the 2011 Tunisian Revolution, Tunisia is widely recognized as an Arab-Spring success story. With campaigns for civil resistance against corruption and civil oppression, the Tunisian Revolution consisted of mass demonstrations that ultimately inspired presidential elections and other democratic reforms across the nation, and a wave of similar protests across the Arab world. In 2021, amid ongoing demonstrations against government dysfunction and corruption, Tunsian President Kais Saied suspended the parliament, replaced the Prime Minister, and began drafting constitutional amendments which reversed nearly a decade of democratic reforms. As freedoms of speech and expression, the right to organize, and many local media outlets have been oppressed, Tunisians have taken to platforms like Twitter to speak truthfully. Thus, CS4624 Team 21 was focused on identifying and analyzing Twitter data relating to democracy, political reforms, and public sentiment in Tunisia since 2020.

Team 21 primarily worked with clients Drs. Kavanaugh, Sheetz, Miller, and Farag to analyze Tunisian Twitter data collected by a larger research team in collaboration with Virginia Tech's (VT) University Libraries. The team handled Twitter data previously collected at VT, as well as more recent data that extends the previous collection. After preprocessing all data to add sentiment scores and filter by English language, the team analyzed the cleaned collection of tweets for key terms and hashtags provided by their clients. The team determined counts for each keyword, extracted a list of URLs used in the tweets, and created visualizations of topic models to visualize monthly keyword and sentiment trends in the relevant timeline. Lastly, the team converted the cleaned tweet collection to consistent JSONL format determined via consultation with the client for eventual integration into the VT library repository. Ultimately, the team expects their project to revitalize research at VT related to Twitter data, inspire new publications about Tunisia based on Twitter data, and lead to a greater understanding of public sentiment about political reforms in Tunisia.

2. INTRODUCTION

2.1. Motivation

Tunisia, a small Arab country on the northern coast of Africa, gained widespread influence and attention following its 2011 political revolution. Resisting corruption, poverty, and civil oppression, the Tunisian Revolution consisted of mass demonstrations that ultimately led to a thorough democratization of the country. Events in Tunisia also ignited the Arab Spring, a wave of similar anti-government protests, uprisings, and armed rebellions across much of the Arab world. Over the last decade, Tunisia democratically elected presidents and prime ministers into office and reformed their constitution [1]. Following President Beji Caid Essebsi's death in 2019, Tunisia held an election which overwhelmingly voted President Kais Saied into office [2]; with this, Tunisia became the first country of the Arab Spring protests to undergo a peaceful transfer of power from one democratically elected government to another [3].

In early 2021, President Saied expressed his discontent with Tunisia's post-revolution parliamentary system, calling it the driver of Tunisia's political, social, and economic crises [4]. Four months later, amid demonstrations against government dysfunction and corruption, President Saied suspended the parliament, replaced the Prime Minister, and began drafting constitutional amendments which reversed nearly a decade of democratic reforms [4].

Both during the Arab Spring and now, Tunisians take to platforms like Twitter to voice discontent as freedoms of speech and expression, the right to organize, and many local media outlets have been oppressed [5]. Recently, Twitter has played an important role in providing Tunisians with the ability to organize networked protests in response to President Saied's actions and other political events [6]. Thus, CS4624 Team 21 focused on identifying and analyzing Twitter data relating to democracy, political reforms, and public sentiment in Tunisia between 2020 and the present date.

2.2. Problem

2.2.1. Objective

Team 21 aimed to use Twitter data to understand public sentiment regarding Tunisian President Saied's governmental reforms over the last three years. After collecting and filtering Twitter data, the team identified key words and phrases within tweets and connected terminology to the public opinion of Tunisians through sentiment analysis. Finally, the team worked to visualize their dataset to more clearly understand Twitter's relationship with political event timelines and civil involvement.

2.2.2. Clients

The team worked with Dr. Andrea Kavanaugh, Senior Research Scientist, and Associate Director of Virginia Tech's Center for Human-Computer Interaction, as well as Dr. Steve Sheetz, a Professor of Accounting and Information Systems with Virginia Tech's Pamplin College of

Business. Since 2007, Drs. Kavanaugh and Sheetz have undertaken many analyses related to Tunisia and other locations based on the analysis of Twitter data. Data was collected and transitioned into Virginia Tech's University Libraries by Xinyue Wang and Satvik Chekuri, and managed by University Libraries IT Director Bill Ingram. The team also collaborated with Chreston Miller, who specializes in data collection and analysis research with University Libraries. The team's final client is Mohamed Farag, who has been investigating and collecting Twitter data over the relevant two-year timeline.

2.2.3. Team

The team is composed of Ryan Gniadek, Sraavya Gudavalli, Victoria Hardy, and Steven Ruckert. Generally, Ryan handled the project's back end and data management, Sraavya worked on visualizations and documentation, Victoria the documentation, and Steven the back-end data management and visualizations. All members are senior students with Virginia Tech's Department of Computer Science.

2.3. Approach

Team 21 identified requirements used to divide work amongst members, with each member leading aspects that most closely relate to the roles outlined in the Team section. For greater accountability and communication, team members met twice a week during class to discuss project deliverables and progress. The team also conducted weekly stand-ups with their clients on Thursday afternoons. The team maintained active communication with their clients throughout the week to share ongoing developments or new information. The team also contacted and collaborated with contributors – experts on Tunisian politics, Twitter API, etc. – for help with data collection and learning new technologies to complete deliverables for their clients. Finally, the team checked in with Dr. Fox regarding questions about the project or deliverables.

3. REQUIREMENTS

To complete their Tunisia Twitter data project, the team identified four primary requirements capturing user goals and system abilities. Each requirement was broken down into units of tasks and subtasks; implementation-specific service information for each subtask and the overall project workflow can be found in Appendix A: Methodology.

3.1. Consolidate Twitter Data

The clients who will use the Tunisian Twitter Data project have provided a collection of Twitter data that needs to be filtered. With the goal of determining public sentiment on Tunisian politics within the last three years, Twitter data needs to be filtered for relevancy to Tunisian

politics and democratic reforms, and consolidated with a consistent format for use in eventual analysis and visualization. Consolidating Twitter data relating to Tunisian politics (Goal / Requirement 1) breaks down into the following subtasks:

a. Gather bulk Twitter data in target date range (2020-2023)
b. Convert data to standardized JSONL format
c. Perform an initial filter
d. Discard irrelevant / spam data
e. Format and compile collection of remaining tweets

3.2. Examine Twitter's role in reporting on Tunisian politics

Historically, Tunisians and individuals in other Arab Spring countries have taken to Twitter to speak out against governmental policies, report on current events, and exercise an otherwise suppressed freedom of speech. Clients are interested in examining the changing use of Twitter for reporting on Tunisian politics over the last three years. Subtasks for examining how Tunisians and other users report on Tunisian politics via Twitter (Goal / Requirement 2) are as follows:

a. Create a list of keywords (words and hashtags)
b. Filter cleaned Twitter data by keyword
c. Record keyword frequency
d. Record URLs of tweets containing keywords

3.3. Identify trends in public sentiment

After Twitter data is collected and filtered, the codebase must determine trends like tweet volume, keyword count, and sentiment. Like with the Arab Spring in 2011-2013, the client proposed that Twitter creates an insightful tool for sentiment analysis when freedom of speech and expression suppresses more traditional media like statewide news. Subtasks for identifying trends in public sentiment about Tunisian democratic reforms from 2020-2023 (Goal / Requirement 3) are as follows:

a. Find tweet sentiment using VADER (Valence Aware Dictionary for Sentiment Reasoning) [7]
b. Find appropriate time segments (ex., 1 month, 15 days, etc.)
c. Determine sentiment trends of tweets with any key terms over time
d. Determine sentiment trends for specific keywords over time
e. Record general sentiment for each key word or hashtag

3.4. Visualize trends in public sentiment and Twitter usage

Finally, the client requested data visualization on Tunisian Twitter data. After analysis on Twitter use and public sentiment is completed, the team graphics and charts to display Tunisian Twitter trends. Visualizations should be ethical, accurate, insightful, and supportive of the clients research goals. The team should visualize trends in public sentiment about Tunisian democratic reforms between 2020-2023 (Goal / Requirement 4) via the following subtasks:

a. Find an appropriate visualization tool or technique
b. Create bar graph visualizations for keyword uses
c. Create bar graphs on sentiment for each keyword over time
d. Research Tunisian events and align tweet data with political events

4. DESIGN

As stated in Requirements, the goal of the Tunisia Twitter data project was to collect, filter, classify, and visualize Tunisia-specific Twitter data from the last three years. This section describes the planned process for producing a relevant collection and depiction of tweet data. Overall, the team's design process focuses on collecting Twitter data from a relevant timeline, filtering out irrelevant data, and analyzing the remaining data to gain insights into public sentiment about Tunisian democratic reforms. The team design requires collaboration with clients for keywords and techniques to accomplish filtering and analysis, as well as to determine techniques for producing meaningful visualizations. Ultimately, Team 21 believes that this approach will provide valuable insights into the Tunisian political landscape and help their clients better understand how the Tunisian public feels about democratic reforms in their country.

4.1. Collection and Preprocessing

To begin the project, Team 21 was provided with a collection of Tunisian Twitter data by the project clients. The data contained information such as the tweet text, author, language, and timestamp. An initial evaluation of the collection involved determining data quality and relevance to project goals. The team worked closely with their clients to identify keywords and other metadata used to generate the collections, and discussed additional criteria to apply to the data to ensure its relevance to their analysis. The team finally identified any issues with the data that needed to be addressed, such as incomplete or incorrect information. Team members will then sort and classify tweets from the data collections provided by clients Mohamed Farag and Crestron Miller according to their filtering and classification heuristics, ensuring a comprehensive dataset that reflects client goals.

4.2. Filtering

After obtaining and pre-processing the collection, the team developed filtering and classification techniques. For their filter design, the team worked with their clients to identify the following keywords related to Tunisian political tweets: “tunisia”, “saied”, “25juillet”, “july25”, “referendum”, “tnelec2019”, “tnelection”, “tunisiavotes”, “tunisiadecides”, “tunisie”, “kaissaid”, “législatives”, and “constitutionélections.” The team’s design must then be capable of filtering by any or all of the 13 specified keywords.

The team’s design also involved sorting the tweets by overall sentiment. The team explored several possible methods of assigning quantitative values to sentiment, including Text Blob [8] and Vader Sentiment. Tweets with sentiment scores would then be filtered by key terms; ultimately, the team’s design would produce counts and output files of tweets containing each key term. Sub-collections of Twitter data would be used for data visualization, so that trends in key term frequency and sentiment could be researched by clients.

4.3. Data Visualization and Evaluation

The final component of the Twitter Data project design involved visualizing tweet data that was found to relate to Tunisian politics via key words and hashtags. The team investigated a variety of visualization techniques to gain insights into public sentiment about Tunisian democratic reforms from 2021-2023. For example, they considered timelines that show the frequency of tweets over time or word clouds to highlight the most common words used in tweets about Tunisian politics. Ultimately, the best solution would require working with their clients to determine the most effective visualization techniques and technologies.

The team then planned to analyze tweet usage trends by researching Tunisian political events and comparing event dates with usage spikes. Seeing the overall sentiment of tweets as well, the team planned to provide their clients with valuable insight on when Tunisians are discussing topics of interest and with what tone.

4.4. Timeline

The team divided the project into four main sprints to complete collection and preprocessing, filtering, and data visualization and analysis. Shown in Figure 1, the four focus areas align with the project design. The steps build upon each other, with Twitter data collection necessary for filtering, filtering necessary for analysis, and analysis necessary for visualization and evaluation. Thus, the timeline progressed from February 17, 2023 to April 14, 2023, when the team began working on final deliverables for the course and their clients.

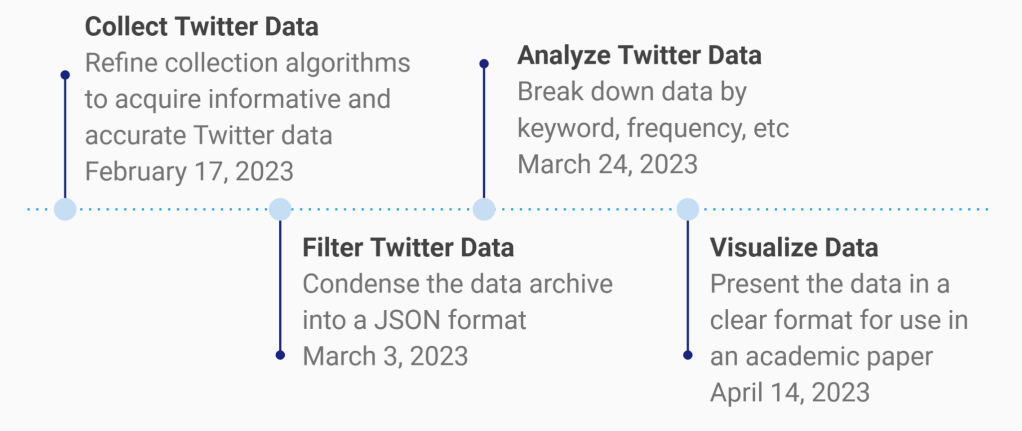

Figure 1: Project timeline.

5. IMPLEMENTATION

Considering the process for the tweet collection, filtering, and visualization in the Design section, Team 21's detailed technical implementation was divided into three phases: tweet collection and pre-processing, filtering, and visualization. Ultimately, this implementation produced file and graph outputs to be evaluated by the team and their clients for use in research about Tunisian political reforms.

5.1. Collection and Preprocessing

Team 21 was provided with an approx. 26 GB collection of Twitter data in JSONL format by clients Mohamed Farag and Crestron Miller. This format was established by project teams from this class in a previous semester. The team also generated smaller datasets, in the magnitude of several KB in the same format as the larger collection to be used for testing and frequent iterations. An initial evaluation of the dataset found the need for sentiment scores for each tweet. The team chose to utilize the Vader Sentiment Analysis Library to determine sentiment scores based on client advice. This decision was influenced by the Vader library being modeled on social media data, which allowed it to infer sentiment in short text with little context. Although tweets in languages like Arabic and French certainly contain meaningful information, the team pre-processed the Twitter collection to ensure that it contains only tweets in English (using tweet language tags), since the sentiment technique was trained on English datasets. The team also pre-processed the dataset to remove tweets that a sentiment score could not be produced for. Thus, pre-processing produced a data collection that would be of high quality and relevance to project goals, ultimately ensuring that the collection had complete information before filtering and analyzing tweets.

5.2. Filtering

Filtering began by encoding the 13 keywords provided by the clients, listed in Design, Section 4.1, into the filtering Python file. The filtering.py script loops through each tweet contained in the pre-processed dataset and checks for hits on any of the listed keywords. If a keyword is found, the associated keyword count is increased, and the tweet is added to a respective JSONL output file for the keyword(s) contained. For example, if a tweet in the pre-processed dataset containing keywords “25juillet” and “tunisiavotes”, the count for “25juillet” and “tunisiavotes” would each increment by one and the tweet would be added to two JSONL files - one of tweets containing “25juillet” and a second of tweets containing “tunisiavotes.” Filtering implementation is further discussed in the Developer Manual.

5.3. Data Visualization

Data visualizations were produced via Tableau, using output TSV files produced by the create_tsv.py and classify_tweets.py scripts. The create_tsv.py script converts tweet data filtered by keyword in JSONL format into TSV format; then, the classify_tweet.py script formats sentiment scores onto the TSV output file. Script implementation is described further in the Developer Manual, Section 8.4. Then, the team used Tableau to create visualizations for ten keywords. The team did not create visualizations for three keywords - “tunisie”, “législatives”, and “constitutionélections” - because each had counts of 0 or 1 since their contained tweet text was in French.

After initial Tableau creation, team members manually inspected and edited all visualizations so that the x-axis includes relevant timelines and thus the figure shows the highest-definition of tweet count data possible. In the most simple case, visualizations were given a 2019-2023 date range to match the by-year date range that the imported Twitter dataset included. Visualizations for keywords represented by Figures 2-5 were best shown by the default 2019-2023 date range, since high tweet volumes existed at every month in the range.

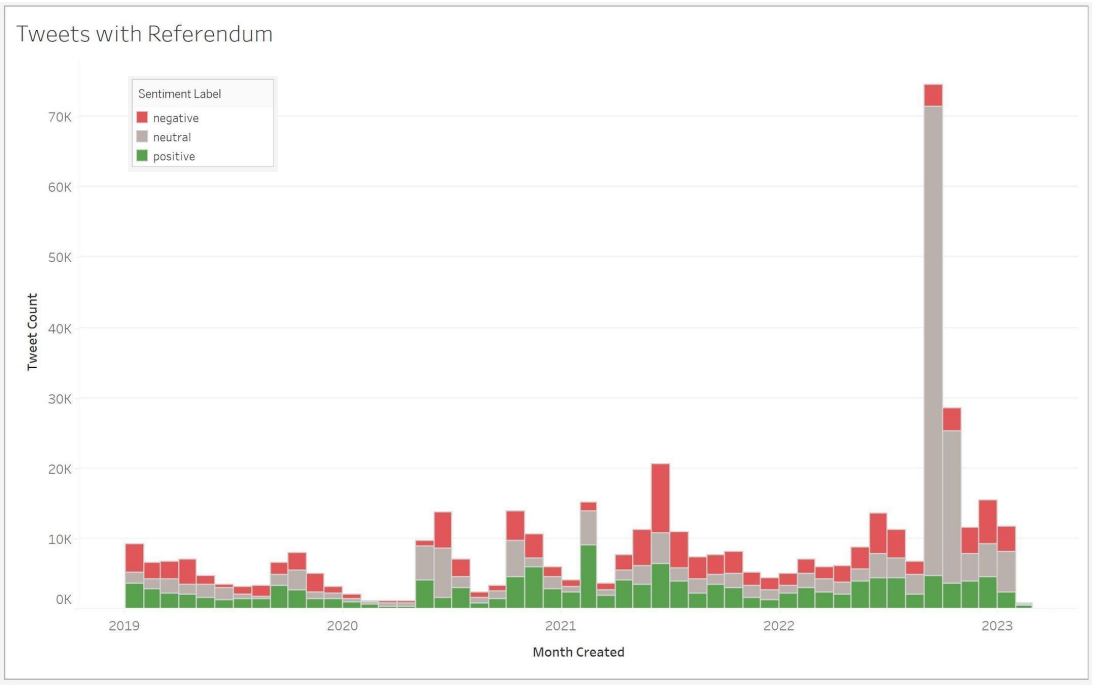

Figure 2: Tweet counts for the “referendum” keyword.

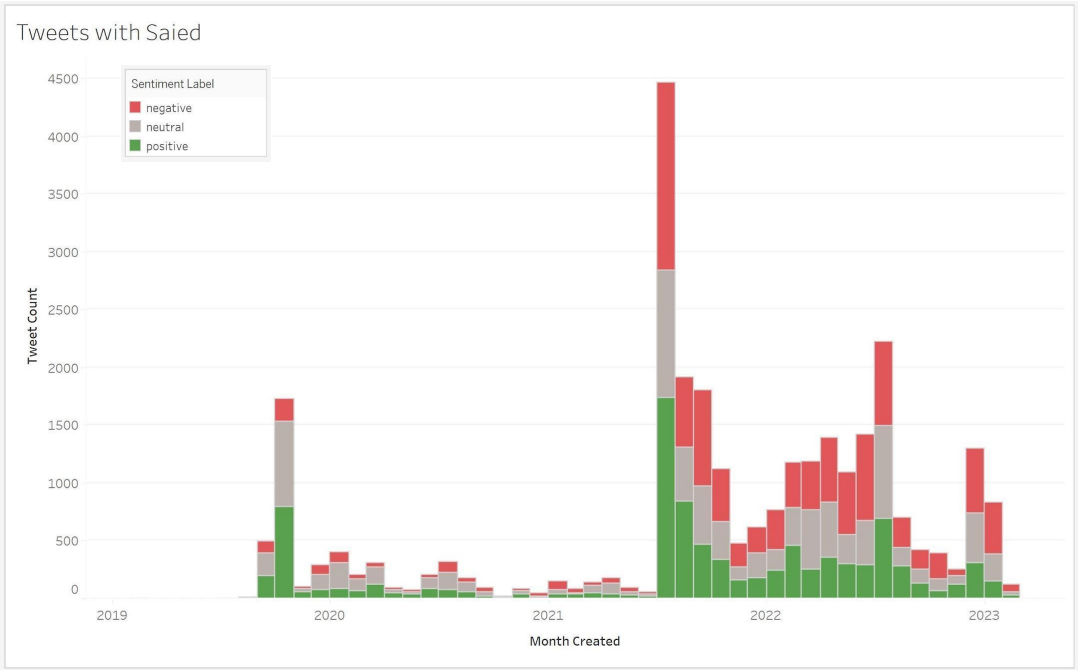

Figure 3: Tweet counts for the “saied” keyword.

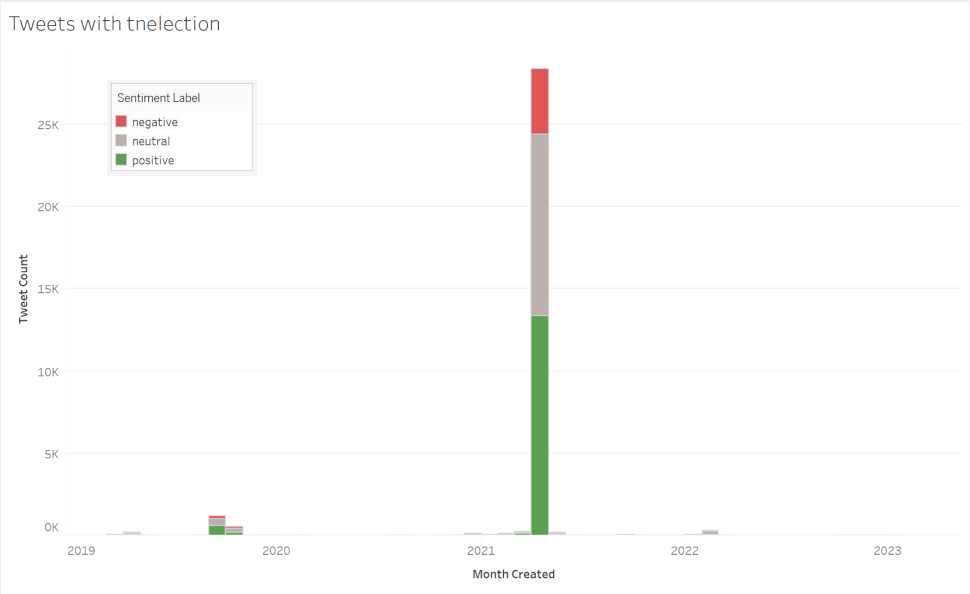

Figure 4: Tweet counts for the “tnelection” keyword.

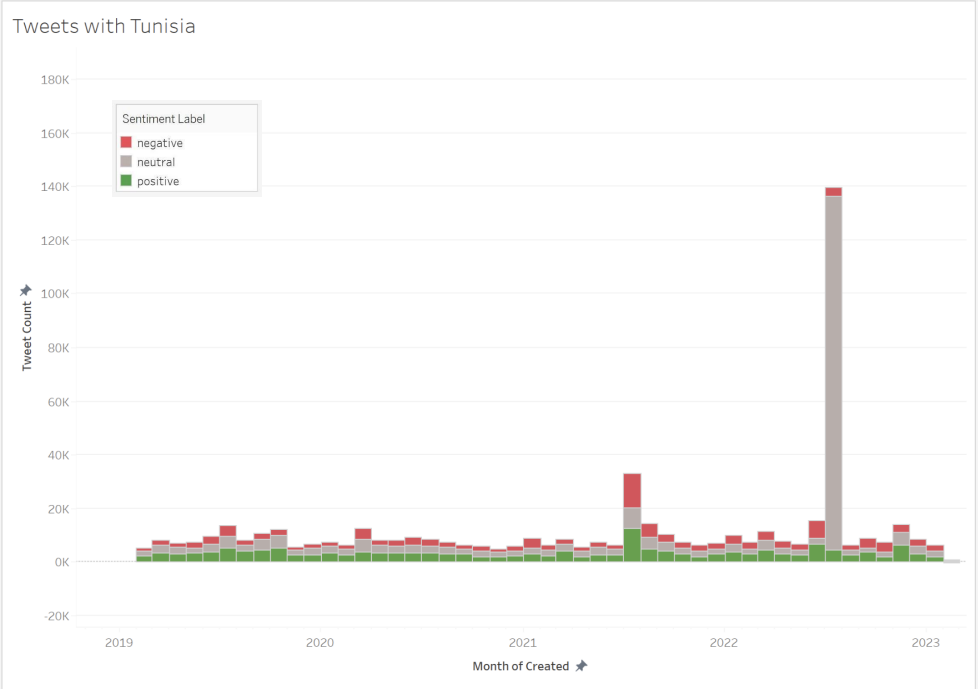

Figure 5: Tweet counts for the “tunisia” keyword.

Visualization start dates for Figures 6-11 were shifted according to the start date of tweet data, or the month where the tweet count is non-zero. Similarly, the last month shown in each visualization corresponded to the last month with a non-zero tweet count. Visualizations,

therefore, each show the rise and fall of keyword use over time, which will be analyzed further in the Evaluation section of the report.

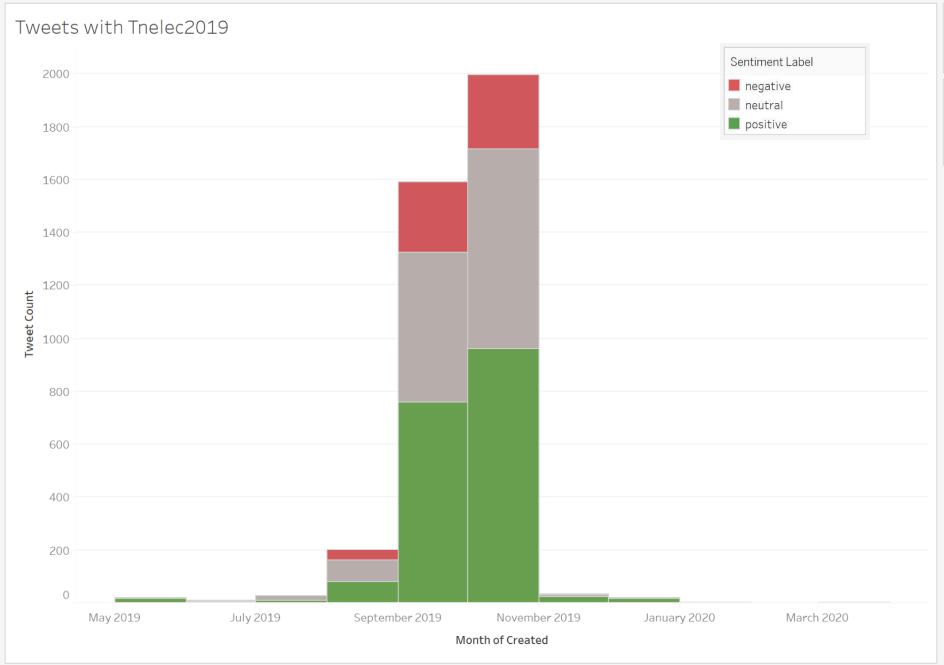

Figure 6: Tweet counts for the “tnelec2019” keyword.

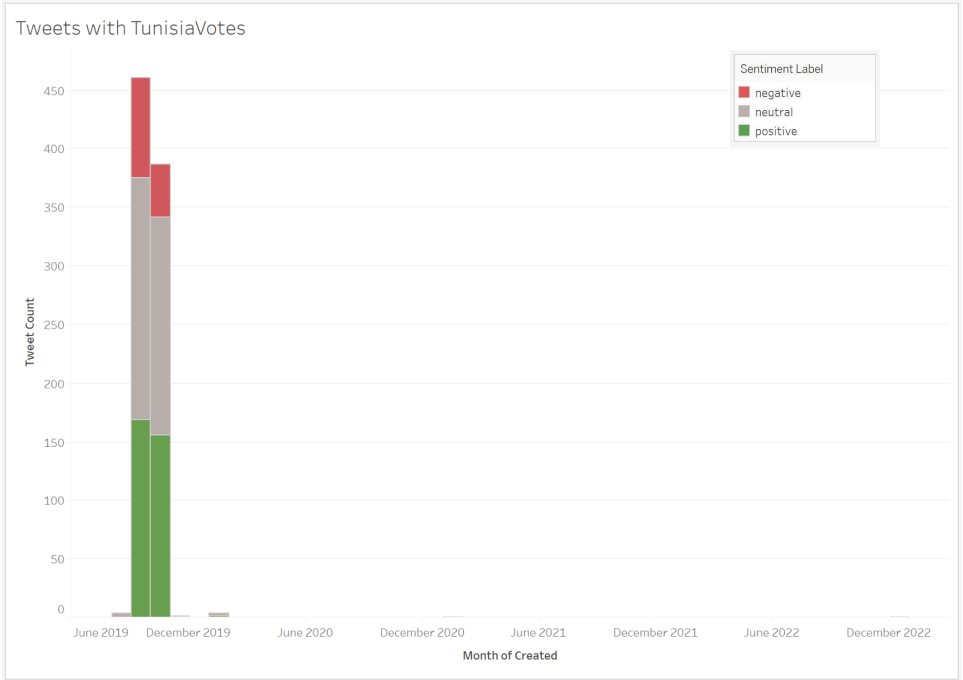

Figure 7: Tweet counts for the “tunisiavotes” keyword.

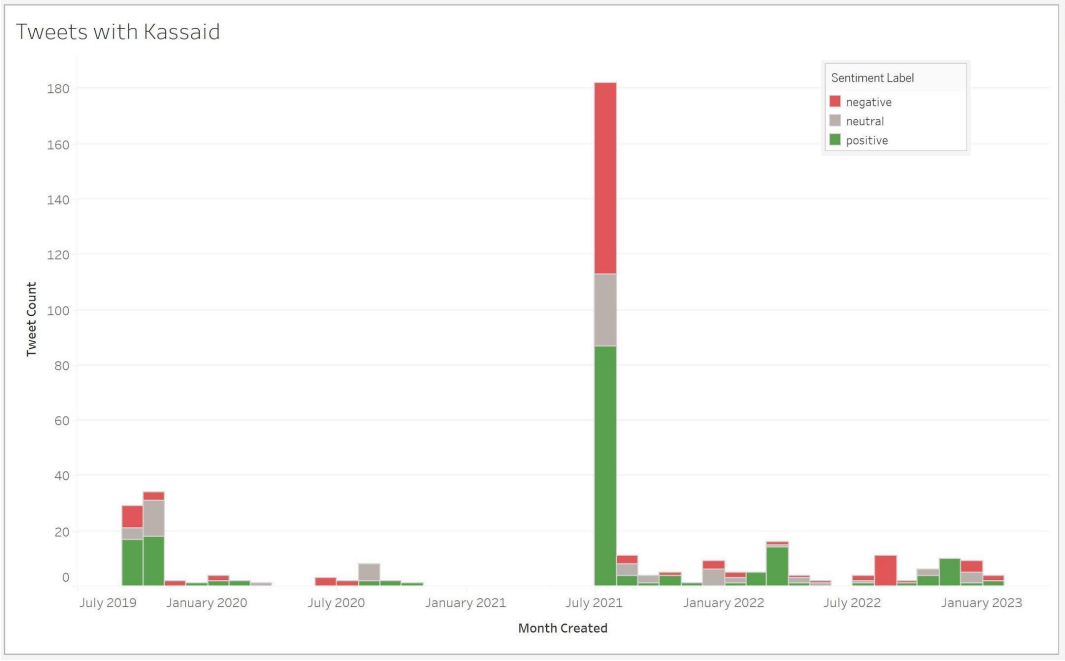

Figure 8: Tweet counts for the “kassaid” keyword.

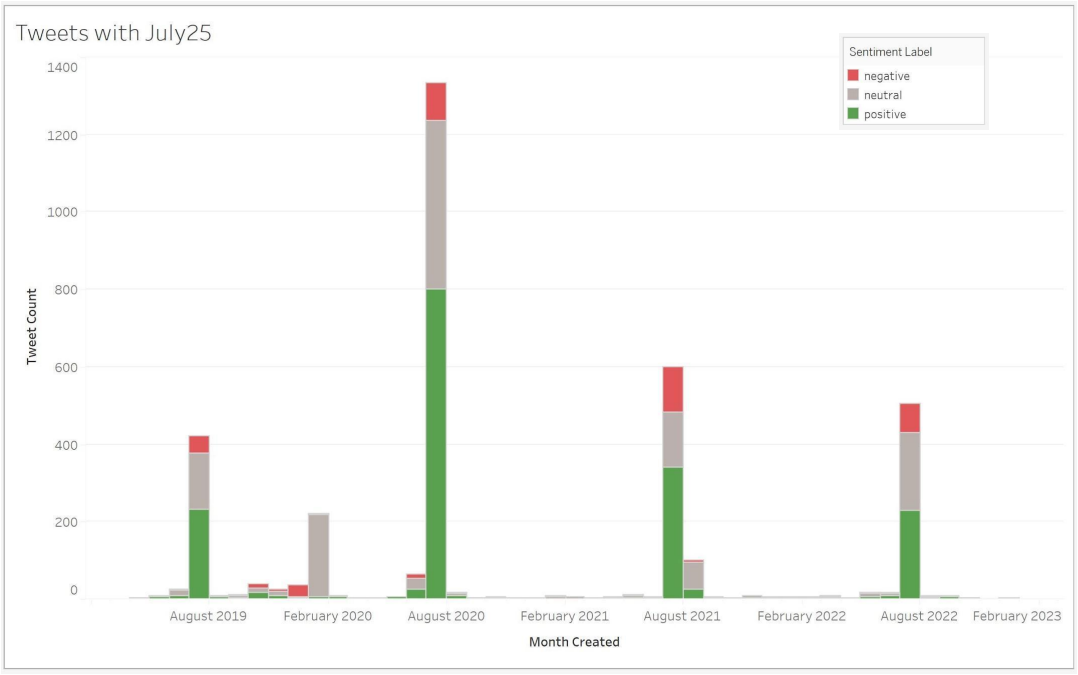

Figure 9: Tweet counts for the “july25” keyword.

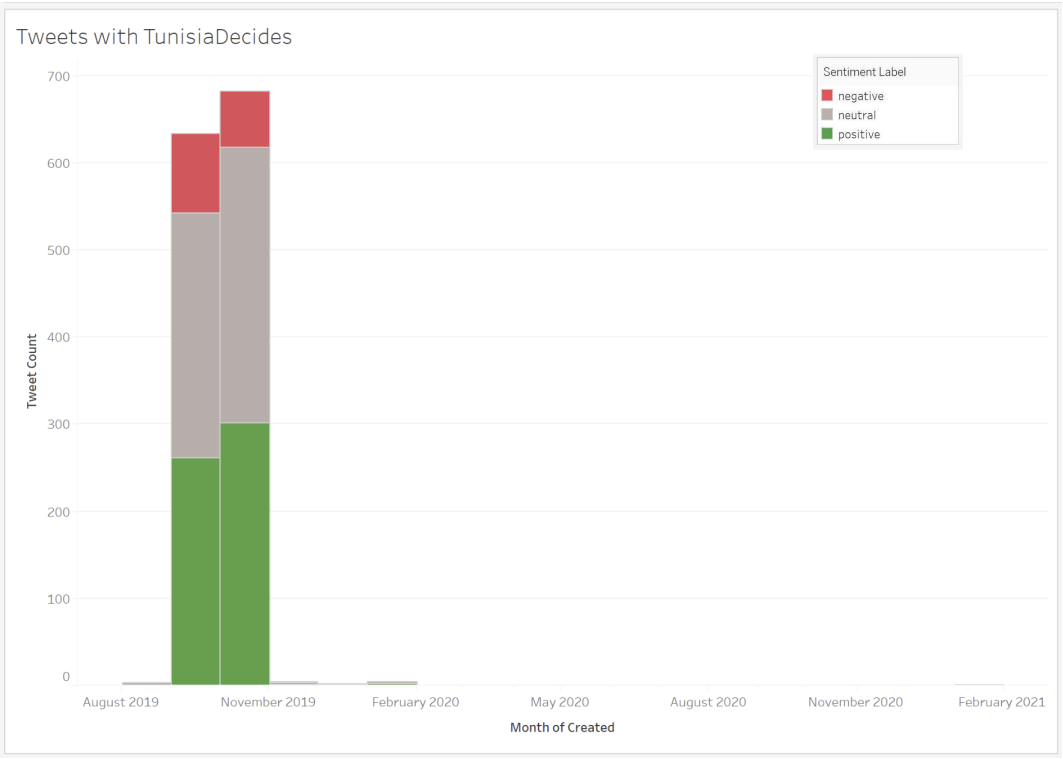

Figure 10: Tweet counts for the “tunisiadecides” keyword.

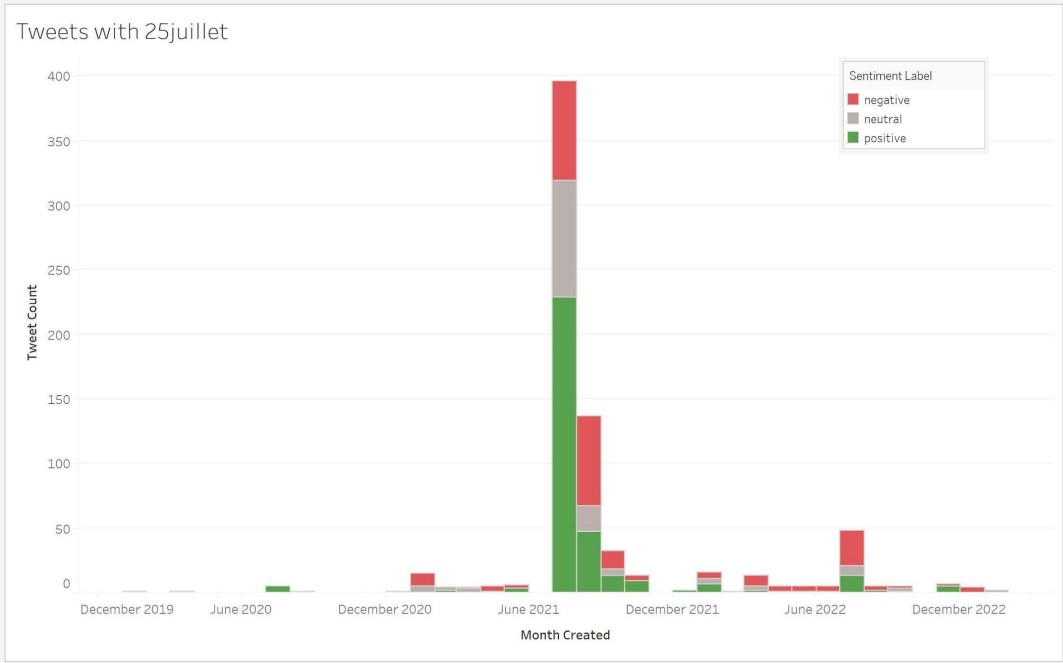

Figure 11: Tweet counts for the “25juillet” keyword.

6. EVALUATION

Team 21's visualizations summarize and demonstrate Tunisian Twitter data and how the political landscape has changed over time. In order to determine the effectiveness of keyword visualizations, the team conducted research on political events within the past three years around areas of keyword use spikes.

In the majority of visualizations, peaks in tweet volume occurred on July 25, 2021 and July 25, 2022. Peaks aligned with major political events: in 2021, President Saied declared a coup, and in 2022, the Tunisians voted on a referendum on a new Constitution.

Tweet trends also mirrored the meaning of the keyword. For example, the use of "25juillet" in tweets spikes on July 25, 2021, and "july25" use spiked every July between 2019 and 2023. Similarly, as shown in Figure 8, tweets contained "kaissaid" in the highest amount each July; tweet counts consistently ramped down during the months in aftermath.

Figure 4 shows a spike in the number of tweets containing "tnelection" in 2019 because President Kais Saied announced that he was assuming exceptional powers, following months of political deadlock and a severe economic crisis exacerbated by the Covid-19 pandemic [11].

The keyword "tunisiavotes" has the most tweets in 2019, as shown in Figure 7. This correlates with the presidential elections on September 15, parliamentary elections on October 6, and a presidential run-off on October 13 [10]. This timeline also is the same for the keywords "tunisiadecides" and "tnelec2019" which has spikes between September to November, when Tunisia voted to decide who their next president would be.

The final keyword "tunisia" has a lot of negative tweets in 2021 which are depicted in red. This is because President Kais Saied announced that he was assuming exceptional powers, following months of political deadlock and a severe economic crisis exacerbated by the Covid-19 pandemic [11]. The spike in 2022 aligns with July 25, 2022, when Tunisians went to the polls to vote on a referendum on a new constitution.

7. USER MANUAL

The user manual describes in detail what is shown to users – the project's clients – and how they should best interpret the output of the Tunisia Twitter project. In general, this user manual covers a high-level overview of the system including a discussion of the use environment, instructions on using the code for client research, and use cases.

7.1. Use Environments

The Tunisia Twitter project is designed to be used by the project clients for Virginia Tech research on public sentiment and opinion to recent democratic changes in Tunisia. The project handles tweet data in JSONL format supplied by the Data Collection team, and outputs a .tsv file containing the correctly formatted data of the keywords that were filtered.

7.2. Client Instructions

For clients who wish to obtain output counts and sentiment scores for each keyword, scripts should be run in the following order: preprocessing, filtering, and visualization formatting. In each step, the team used python3.6.8. If a different Python version is being used, the requirements file may require version changes for each package. Further details about the motivation for each script can be found in Section 8, Developer Manual.

7.2.1. Preprocessing

Preprocessing scripts should be run once a collection of Twitter data has been obtained. The preprocessing script, sentiment.py, takes the name of the input data collection file (JSONL) and the desired name for an output file. Usage via the terminal follows the convention:

```
python3 sentiment.py <input_filename> <output_filename>
```

7.2.2. Filtering

Filtering scripts can be run as often as desired, changing inputs as required for research goals. The script, filtering.py, filters by keywords according to:

```
python3 script.py <input_filename> <search terms .txt file> <output_directory>
```

Similarly, the group_by_date.py filtering script groups filtered data into one-month long time segments for use in visualizations, according to the usage:

```
python3 group_by_date.py <input_filename> <output_directory>
```

7.2.3. Visualization Formatting

Formatting data for visualization consists of creating a TSV file and then classifying tweets by sentiment. The usages for each function are as follows:

```
python3 create_tsv.py <input_filename> <output_filename>
```

```
python3 classify_tweets.py <input_filename> <output_filename>
```

7.3. Use Cases

The following use cases describe how to perform primary functions of the Tunisia Twitter projects and produce useful deliverables for research.

7.3.1. Determine Counts of Tunisia-Related Keywords on Twitter

After generating a filtered list of tweets from the above commands, a Linux terminal user can use the following command to get a count of how many tweets were generated by the filter:

```
wc -l <output .jsonl file>
```

The output will resemble:

```
<number of tweets> <output .jsonl file>
```

for example,

```
1120 kaissaid.jsonl
```

7.3.2. Determine Trends in Keyword Use on Twitter

To look for visual trends in the Keyword data, the team used Tableau. After generating a filtered .tsv table, this can be uploaded into Tableau in the Data Source window. Once the data has been uploaded, one can create a new Sheet from the Tableau dashboard. From here, the left-hand side of the screen shows valid tables to choose from.

To determine general trends in Keyword Use, simply drag the “Created At” table into the Columns bar at the top of the Sheet. Clicking the drop-down arrow allows one to change the date format. The next step is to select the “Tweet Id” table from the same left-hand side and place it in the Rows bar. Once that is completed, open the drop menu from the Tweet Id table, select the Measure button, and select count. This will provide a count of all of the Tweet Ids within a certain time frame. The graph used to represent this data can be changed from the “Marks” menu.

7.3.3. Determine Trends in Public Sentiment about Tunisian Events

To determine trends in tweet sentiment, simply drag the “Created At” table into the Columns bar at the top of the Sheet. Clicking the drop-down arrow allows one to change the date format. The next step is to select the “Sentiment Label” table from the same left-hand side and place it in the Rows bar. Once that is completed, open the drop-down menu from the Sentiment Label table, select the Measure button, and select count. This will provide a count of all of the tweets in a certain time frame. Change the graph to a bar chart by selecting the drop-down menu in the “Marks” section.

To break each bar into 3 distinct graphs, drag the “Sentiment Label” table from the left-hand side once more and place it in the Color box below the Marks table. Once that is in, there should be 3 separate fields representing positive, neutral, and negative sentiments. The colors of these values can be changed by selecting the color box and choosing the preferred color. Once that is completed, you will have a bar chart displaying the trends of public sentiment regarding a chosen topic.

8. DEVELOPER MANUAL

The developer manual serves as a source of information for any future teams that may continue work on this project, or on similar Twitter projects. Therefore, the manual explains the structure of the project, and contains directions and explanations on interacting with the codebase. Lastly, the developer manual is based on the project requirements defined in the Requirements section and Appendix A.1, and also references an updated project workflow diagram (Appendix A.2).

8.1. Prerequisites

This project is developed to work on an UNIX environment with access to python3.6+. The project relies on several external dependencies which are specified in a requirements.txt file and can be installed using the pip package manager.

8.2. Code Repository

The project code is hosted on GitLab at https://git.cs.vt.edu/tunisia/filtering. For development purposes, the team used the TML server owned by Dr. Fox’s lab, but any similar environment is sufficient for future development. Access to a shared server allowed the team to store certain preprocessed data between group members.

8.3. Version Control

Future contributors to the project should clone a git repository. The team used several development branches to work collaboratively, but current production code (at the time of publication) is tagged as “publication” on Git.

8.4. Data Processing Scripts

All code is in the repository root directory and instructions for setting up the environment are self-contained in a README file. There are several relevant Python scripts:

1. sentiment.py:

This script is for preprocessing of the dataset and accepts a JSONL collection as an input, as well as a JSONL file name to redirect output to. This script should be run once to filter the dataset to only include valid English language tweets and store sentiment values for those tweets and the new "preprocessed" collection should be used as the input for later scripts.

2. filtering.py:
   This script is for separating a JSONL collection of tweets into sub collections by topic. Topics are designated in a keyword file which is passed through as an argument and a sample keyword file is provided in the repository. Subtopic collections are stored in an output directory which is given as an argument.

3. group_by_date.py:
   This script works similarly to the topic filtering script except it separates the input collection into subcollections by month and year. No keyword file is needed.

4. create_tsv.py:
   This script accepts any preprocessed JSONL file as input (must already have the sentiment calculated) and converts it to the TSV format requested by client Crestron Miller. For certain visualization software used, TSV is a more convenient file format than the JSONL format established in previous work.

5. Classify_tweets.py:
   This script is also to complement certain visualizations, it takes in any previously TSV file as input and appends a column classifying tweets as "positive," "negative," or "neutral" based on a given threshold of 0.05 or greater being not neutral.

8.5. Visualizations

The Tunisia Twitter project team utilized Tableau to create visualizations. Future developers are recommended to continue using Tableau for consistency as well as for accurate data visualizations. To use Tableau, download the most-recent version of tableau and import the data desired for visualization. This TSV data serves as the data source; continue to create a sheet to create your preferred visualization.

9. CONCLUSION

9.1. Lessons Learned

Throughout the course of the project, the team learned how to stay organized and actively communicate with the clients and each other. The team consistently practiced professional skills by meeting with the clients every week and checking in during the week to ask questions or convey additional information. They also learned to use the agile methodology described in class to work efficiently and effectively towards their client's specifications.

Following completion of the Tunsia Twitter Data project, the team learned how to effectively utilize the Twitter API and data collection streamline, as well as about how to collect and sort the Twitter data within collections using tags. Team members also learned a new technology for data visualizations – Tableau – used to provide their clients with a better understanding of public sentiment and Twitter usage in Tunisia. Finally, the team gained a greater understanding of the relationship between modern politics and social media as global events unfold; directly accessing Twitter data can provide new context and insights on how society reacts to dramatic changes in politics and democracy.

9.2. Challenges

Overall, team 21 was challenged throughout the course of the project by working with unfamiliar concepts and data. The ending Twitter API access provided an additional challenge to the team, as their project scope changed from collecting current Twitter data to interacting with and analyzing tweets from 2019 to early 2023. Despite the threat of ending API access, team 21's clients were supportive to modify project goals and acquire a large dataset to work with before API access ended. Additionally, the team was not able to process and analyze tweets in Arabic and French, the most common languages in Tunisia, due to limitations of the sentiment analysis algorithm. Thus, the tweets handled by the team are likely meant for a global audience, and not the tweets commonly circulated by the average Tunisian. Despite the language-limitation, the team was still able to produce a meaningful codebase and analysis for their clients.

9.3. Future Work

Future collaborators on the Tunisia Twitter Data project should focus on analyzing tweets in various languages and producing topic models for each keyword. Though team 21 was limited by the VADER sentiment analysis algorithm, future teams could train and apply sentiment analysis algorithms on languages like Arabic and French, to obtain valuable sentiment scores for tweets in the languages of most Tunisians. Combining team 21's analysis of English tweets with analysis of Arabic and French tweets would provide valuable insights to clients for future research projects and publications.

10. ACKNOWLEDGEMENTS

Team 21 would like to thank their clients Dr. Andrea Kavanaugh (kavan@vt.edu), Dr. Steve Sheetz (sheetz@vt.edu), Dr. Chreston Miller (chmille3@vt.edu), and Dr. Mohamed Farag (mmagdy@vt.edu) for their expertise on Tunisian politics, Twitter data processing, and sentiment analysis, support through weekly meetings, and guidance with project data and deliverables.

The team also acknowledges their professor, Dr. Edward Fox (fox@vt.edu), and the CS 4624 teaching team for connecting them with the Tunisia Twitter Data project and supporting them in meeting project deliverables.

11. REFERENCES

[1] Abouaoun, Elie. Tunisia timeline: Since the Jasmine Revolution. *United States Institute of Peace* [online]. 6 November 2020. [Accessed 23 February 2023]. Available from: https://www.usip.org/tunisia-timeline-jasmine-revolution

[2] Zeidan, Adam. Kais Saied. *Encyclopædia Britannica* [online]. 19 February 2023. [Accessed 25 February 2023]. Available from: https://www.britannica.com/biography/Kais-Saied

[3] Tesch, Noah. Arab spring. *Encyclopædia Britannica* [online]. 14 February 2023. [Accessed 23 February 2023]. Available from: https://www.britannica.com/event/Arab-Spring

[4] Yekes, Sarah & Alhomoud, Maha. One year later, Tunisia's president has reversed nearly a decade of democratic gains. *Carnegie Endowment for International Peace* [online]. 22 July 2022. [Accessed 20 February 2023]. Available from: https://carnegieendowment.org/2022/07/22/one-year-later-tunisia-s-president-has-reversed-nearly-decade-of-democratic-gains-pub-87555

[5] Beaumont, Peter. The truth about Twitter, Facebook and the uprisings in the Arab World. *The Guardian* [online]. 25 February 2011. [Accessed 28 February 2023]. Available from: https://www.theguardian.com/world/2011/feb/25/twitter-facebook-uprisings-arab-libya

[6] Jones, Marc Owen. Tunisia crisis prompts surge in foreign social media manipulation. *Social Media News | Al Jazeera* [online]. 30 July 2021. [Accessed 28 February 2023]. Available from: https://www.aljazeera.com/news/2021/7/28/tunisia-crisis-prompts-surge-in-foreign-social-media-manipulation

[7] CJHUTTO. VADER Sentiment Analysis. *GitHub* [online]. 17 November 2014. [Accessed 1 March 2023]. Available from: https://github.com/cjhutto/vaderSentiment

[8] Loria, Steven. Simplified text processing. *TextBlob* [online]. 2020. [Accessed 1 March 2023]. Available from: https://textblob.readthedocs.io/en/dev/index.html

[9] Sun, Angel. What happened at the second round of Tunisian parliamentary elections. *The Boar* [online], 15 February 2023. [Accessed 1 March 2023]. Available from: https://theboar.org/2023/02/what-happened-at-the-second-round-of-tunisian-parliamentary-elections/

[10] Roth, Kenneth. World Report 2022: Rights Trends in Tunisia. *Human Rights Watch* [online], 13 January 2022. [Accessed 4 March 2023]. Available from: https://www.hrw.org/world-report/2022/country-chapters/tunisia.

[11] Yerkes, Sarah & Alhomoud, Maha. One Year Later, Tunisia's President Has Reversed Nearly a Decade of Democratic Gains. *Carnegie Endowment for International Peace* [online], 22 July 2022. [Accessed 4 March 2023]. Available from: https://carnegieendowment.org/2022/07/22/one-year-later-tunisia-s-president-has-reversed-nearly-decade-of-democratic-gains-pub-87555.

12. APPENDICES

Appendix A: Methodology

A.1 Requirements and Subtasks

The first goal shown in Figure 12 is to consolidate Twitter data on Tunisian politics from unfiltered tweet data from the Twitter API. To complete this goal, the first subtasks are to gather the data from 2020-2023 [1a] and convert it to JSONL format [1b]. The data is then filtered to obtain relevant data [1c] and the irrelevant data is removed [1d].

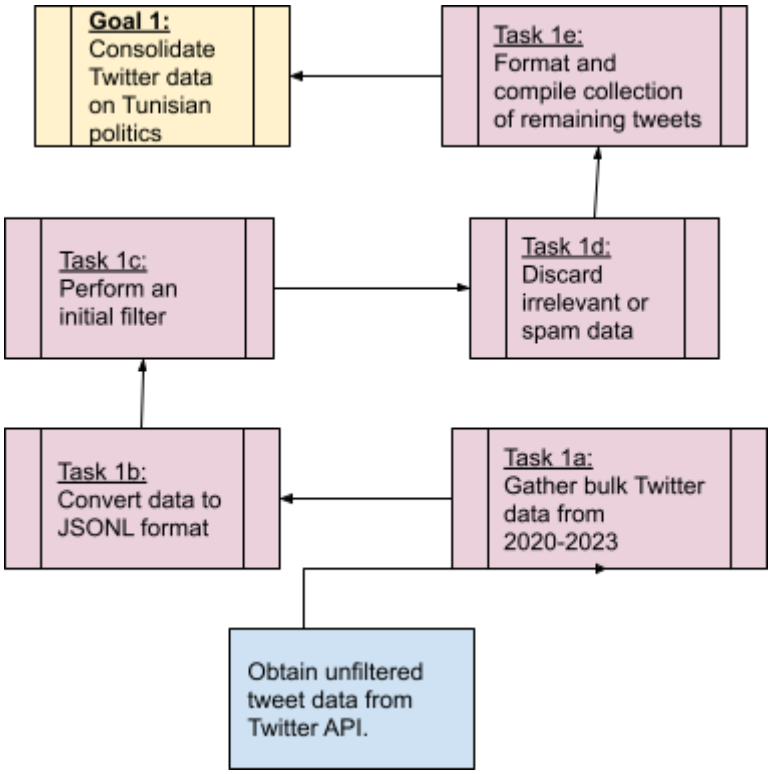

Figure 12: Workflow 1 including subtasks for achieving project goal 1.

As shown in Figure 13, the second goal is to examine how Tunisians and other users report on Tunisian politics using Twitter from imported, compiled and filtered tweets. To complete this goal, the first subtasks are to create a list of keywords [2a] and filter the tweets by those keywords [2b]. The next tasks are to record the keyword frequency [2c] then record urls of tweets containing keywords [2d].

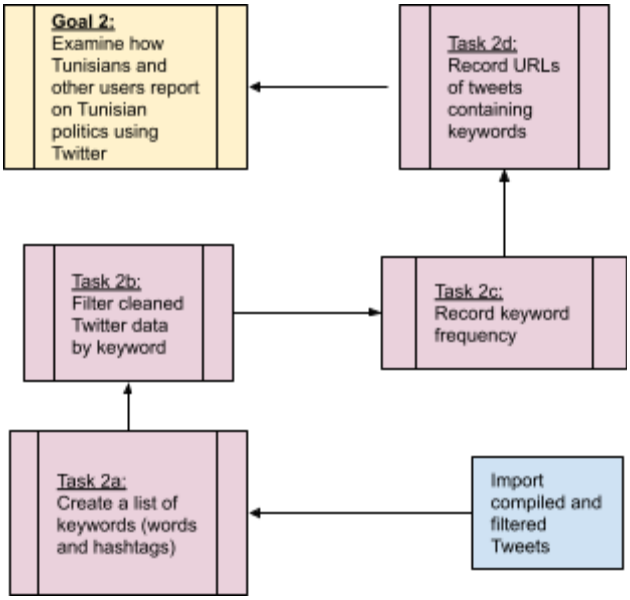

Figure 13: Workflow 2 including subtasks for achieving project goal 2.

The third task is to identify trends in sentiment about Tunisian democratic reforms from imported clean and formatted tweets. Shown in Figure 14, the first subtasks are to get the sentiment of the tweets using Vader [3a] and finding the appropriate time segments for parsing [3b]. The next subtasks involve determining sentiment trends of tweets with any key terms overtime [3c] and determining trends from that sentiment for specific keywords [3d]. The final subtask is recording general sentiment for each keyword of hashtag [3e].

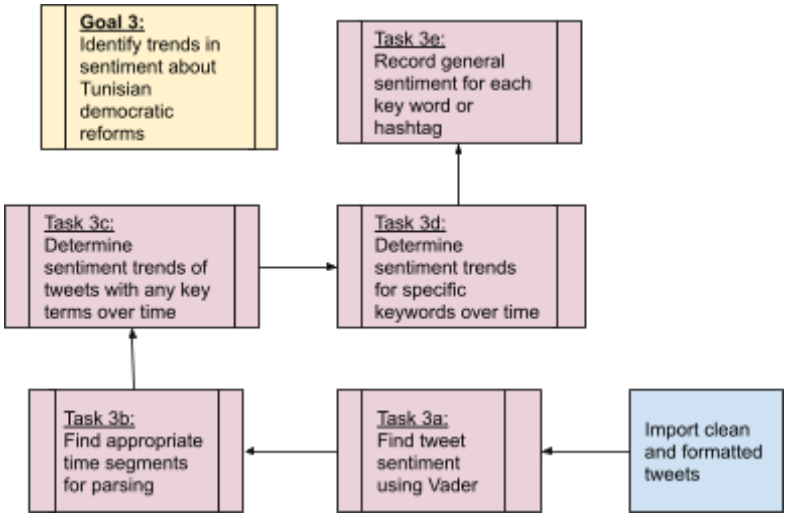

Figure 14: Workflow 3 including subtasks for achieving project goal 3.

The final goal, shown in Figure 15, is to visualize trends in public sentiment about Tunisian democratic reforms from imported clean and formatted tweets. The first subtask is to find an appropriate visualization tool [4a], which Tableau was used for. The next subtasks are to create bar graph visuals for keywords [4b] and then create bar graphs on sentiment for each keyword overtime [4c]. The final subtasks are to create a timeline graph of sentiment for tweets containing any keyword [4d], then to research Tunisan events and align tweet data with political events [4e].

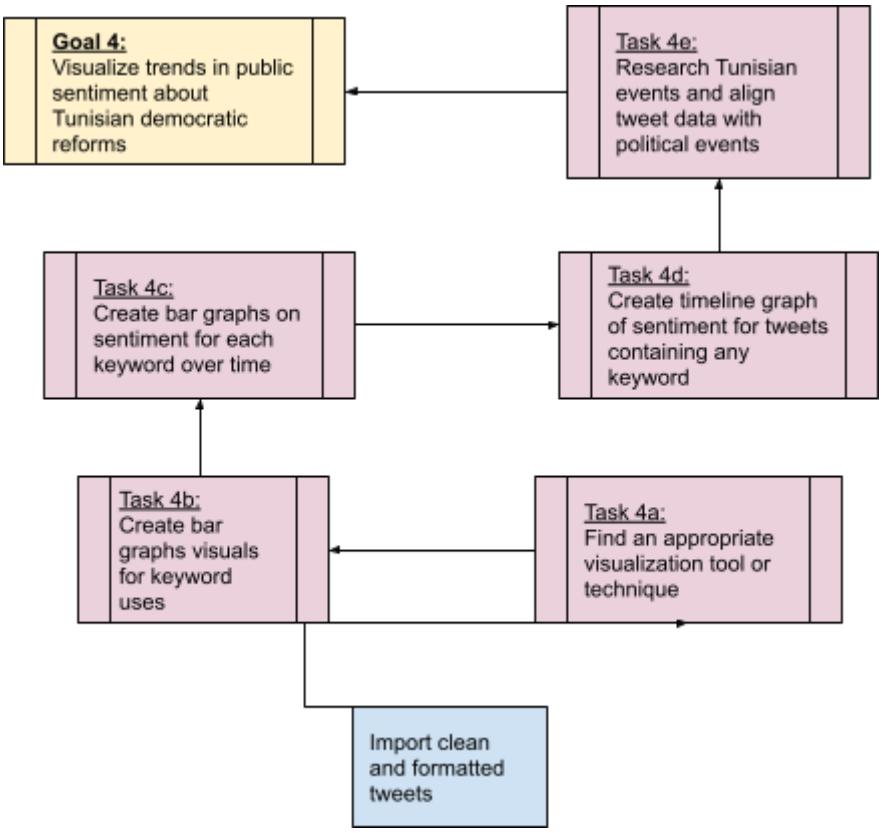

Figure 15: Workflow 4 including subtasks for achieving project goal 4.

A.2 Workflows covering each goal

Combining the four goals and workflows, Figure 16 outlines the workflow diagram for the Tunisia Twitter Data project driven by the needs of all user-clients.

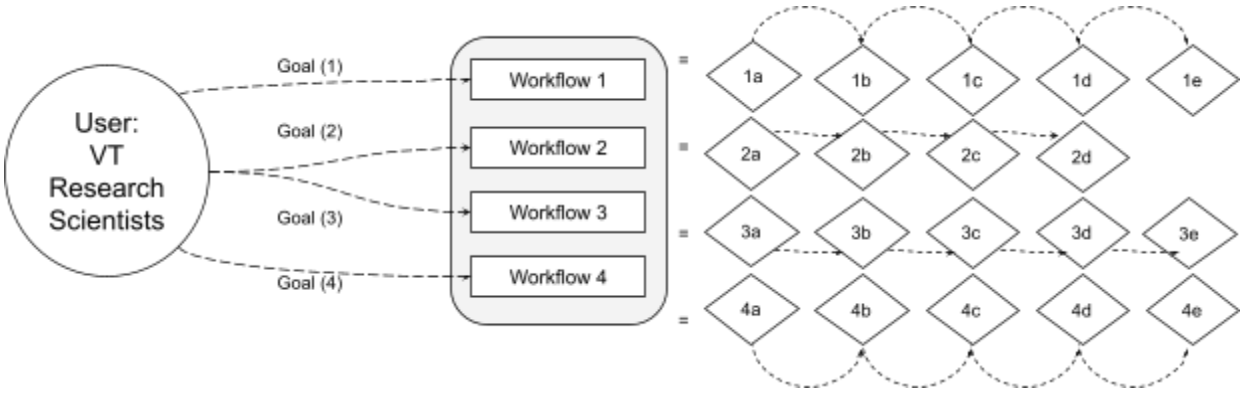

Figure 16: Workflow diagram for Tunisia Twitter project.